Deep Learning for Hydrological Modelling: From Benchmarking to Concept Formation

Thomas Lees
Christ Church College
University of Oxford

supervised by

Prof. Simon Dadson, University of Oxford
Dr. Steven Reece, University of Oxford

A thesis submitted for the degree Doctor of Philosophy

Oxford, March 2022

Example is the school of mankind,
and they will learn at no other.
Edmund Burke

Abstract

Hydrological modelling seeks to address the question: what happens to water once it falls on the land surface? Water can flow into river systems, it can pass through soils into the subsurface, it can be absorbed by the biosphere, or it can be released back into the atmosphere as evaporation. The ultimate purpose of hydrological modelling is twofold, to improve our predictions about the system of interest, and to understand how the system works. In recent decades, advances in science and technology have been made by using techniques from the field of Deep Learning, whereby flexible models are calibrated on large datasets to deduce relationships and make predictions. These techniques have begun to be applied across the environmental sciences.

In this thesis I will explore a particular model architecture for deriving relationships between inputs and outputs from data, to provide accurate simulations of hydrological systems as well as to improve our understanding of the hydrological processes themselves. The Long Short-Term Memory (LSTM) is a neural network architecture from the field of Deep Learning which has shown promise for time-series modelling. This model architecture was chosen for its correspondence with our perceptual model of hydrology, whereby we consider the hydrological system to be characterised by a description of its state, and processes that govern the transfer of energy and materials from that state. This input-state-output architecture is similar in many ways to traditional process-based and conceptual models. However, unlike these models the LSTM is capable of searching a much wider range of possible functions that map inputs to outputs, capable of learning any process that can be deduced from the data, as opposed to being limited by the encoding in the traditional models.

The chapters that make up this thesis first demonstrate that the LSTM is an appropriate architecture for rainfall-runoff modelling on the island of Great Britain. I trained the model using meteorological and catchment averaged attributes as input, and river discharge as outputs over a large sample of catchments. In comparison with often used conceptual model architectures, I show that the LSTM demonstrates state-of-the-art performance and justify further interrogation of what the model has learned. In a follow up study, I explore what the LSTM has learned about the hydrological system by taking the trained

model weights and interpreting them with reference to intermediate stores of water that relate the meteorological inputs to the outputs of discharge. Despite the complexity of translating rainfall to discharge, hydrology is not limited to rainfall-runoff modelling. The final chapter in this thesis turns to the problem of forecasting a satellite-derived vegetation health metric which is used operationally as a proxy for drought conditions. Like rainfall-runoff modelling, the system being simulated is driven by the complex interaction of meteorological and land surface attributes, however, the target variable is now a store of water (vegetation) rather than a flux (discharge).

This dissertation provides the hydrological community with three important outcomes. Firstly, the LSTM model results are provided as a benchmark for future work looking to develop a national rainfall runoff model for Great Britain. Secondly, this dissertation demonstrates a method used elsewhere in machine learning research that allows a scientist to diagnose what the LSTM has learned about the hydrological system. Finally, this dissertation demonstrates the utility of the LSTM in a drought monitoring context, forecasting a satellite derived vegetation health metric with the potential to improve the ability of national agencies to respond to drought events.

Ultimately, this dissertation offers a demonstration of the power of Deep Learning models in hydrology, and calls on the community to interrogate these tools further to not only advance our predictive goals, but also our scientific ones.

Contents

1 Introduction

This dissertation explores a particular neural network architecture with useful properties for modelling hydrological systems, the Long Short Term Memory Network (LSTM). I argue that the LSTM offers a flexible data driven approach with state-of-the-art accuracy for rainfall-runoff modelling and vegetation health forecasting. Just as important, however, is the opportunity the LSTM offers to leverage large sample and multi-source datasets to derive interpretable concepts and patterns about the modelled system. Throughout this DPhil, I have sought to inspect the "black box" of machine learning, and in this dissertation I contribute to our knowledge about the hydrological systems in which an LSTM-based approach adds predictive and scientific value.

1.1 Background and Motivation

Hydrological modelling describes the process of simulating the movement of water through the atmospheric and land-surface system, including the biosphere. This involves encoding theories about the Earth system into computer code and developing a set of instructions that process inputs to make predictions of an output of interest, for example river discharge.

As an applied scientific discipline, hydrology is focused on tackling societal needs for managing water resources and mitigating water related hazards, motivated by both episteme and techne [Nearing *et al.*, 2020a; Parry, 2003]. Curiosity and a desire to understand the underlying system drive improvements in our understanding (episteme), but ultimately, these improvements in understanding aim to increase predictability and control (techne) [Beven, 2011a]. Indeed, water managers need robust and reliable predictions in order to enable sustainable water resource management and the development of effective drought and flood mitigation [Hrachowitz *et al.*, 2013]. Modelling is ultimately required due to a lack of access to the phenomena of interest. The model provides a means of extrapolating information in space, time, or to unobserved variables [Beven, 2011b; Oreskes *et al.*, 1994]. Therefore, models are used for practical reasons (techne) and for gaining knowledge about the system of interest (episteme), which is considered the domain of scientific discovery and testing hypotheses. In this dissertation, I start with the practical problem of benchmarking model performances for rainfall-runoff modelling, before moving on to evaluate the potential of these models to discover hydrological concepts. Finally, I blend both motivations when exploring the

potential of these models in an alternative hydrological context, drought forecasting.

In hydrological modelling, three modelling paradigms exist: physical, conceptual and data-driven, categorized by the degree of process realism encoded in the model [Dadson *et al.*, 2019]. Physically-based models explicitly describe the physical processes that govern the surface and subsurface fluxes of matter and mass. Representing these fluxes through partial differential equations, physically based models require detailed descriptions of catchment characteristics [Abbott *et al.*, 1986a; Clark *et al.*, 2011; Freeze & Harlan, 1969]. Conceptual modelling approaches use empirical equations to simplify the processes, reduce the number of required parameters and increase the speed at which these models can make predictions [Beven, 2011b; Beven & Kirkby, 1979].

An alternative approach has become known as a data-driven approach [Reichstein *et al.*, 2019; Shen, 2018; Shen *et al.*, 2018] variously called an empirical [Dadson *et al.*, 2019] or statistical approach to hydrological modelling. This category of models loosely corresponds to those models that assume very little prior structure and seek to relate inputs to outputs. They can be more or less interpretable as mechanistic models (see Young [2003]), but ultimately, they start from data products (often precipitation, temperature and streamflow) and work backwards to infer the state and structure of the hydrological system, hence why this approach has been labelled "*doing hydrology backwards*" [Kirchner, 2009].

Deep learning (DL) is one such data driven approach. DL describes a subfield of machine learning that builds upon artificial neural networks, stacking multiple layers of these network structures into a directed graph that can learn increasingly abstract representations of data enabling complex function fitting [Huntingford *et al.*, 2019]. This is valuable because it enables scientists to automatically extract information from large, complex and highly non-linear systems given data about those systems [Schmidhuber, 2015]. Increasing availability of large data sources, access to increased computer processing power and advances in algorithms for model training and inference have coalesced, causing developments in DL with wide ranging impacts across science, technology and society. These advances offer exciting opportunities to improve our understanding and predictions of environmental systems, [Reichstein *et al.*, 2019].

One model architecture is of particular interest for hydrological applications, the Long Short Term Memory Network (LSTM) [Hochreiter, 1991; Schmidhuber, 2015]. The LSTM is an effective architecture for modelling timeseries. It is a recurrent neural network architecture with a mechanism for encoding the current state of the system of interest at any given time, the state vector, and a series of gates that control the information flow into and out of the state vector. That state vector defines the memory of

the system, and this corresponds neatly to our understanding of these processes as encoded in physical and conceptual models, since storage processes are important for the transformation of rainfall into river discharge, or water used by vegetation. The hydrological motivation of the LSTM is described in detail in Chapter 2.

Deep learning and LSTMs specifically have been applied to hydrological systems with notable success [Fang *et al.*, 2017; Gauch *et al.*, 2021a; Kratzert *et al.*, 2018, 2019e]. This dissertation builds upon and contributes to this endeavour in three ways. Firstly, by testing the LSTM and its variants in different contexts, both geographically and conceptually. The LSTM is tested on two systems, rainfall runoff modelling in Great Britain and drought forecasting in Kenya. Secondly, this dissertation develops and tests methods to learn from the LSTM based models. Finally, the LSTM is tested in a forecasting setting, where it has the potential to be used to improve the timeliness of responses to developing drought conditions.

Given the LSTM has been shown to be a powerful tool for modelling rainfall-runoff behaviour in the Continental USA [Kratzert *et al.*, 2018, 2019c], this thesis seeks to test how robust the LSTM is to different geographies and different hydrological systems. Therefore, the central research question of this thesis is as follows:
Is the LSTM capable of accurately modelling hydrological systems characterised by a hidden state with long-term time dependencies generally? If so, how can we interrogate an LSTM to better understand what has been learned about our system of interest?

1.2 Thesis Outline

This DPhil addresses several open questions regarding the application of deep learning methods to the hydrological sciences. What can we learn from large-sample benchmarking experiments comparing LSTM-based models against traditional hydrological models (Chapter 3)? How can we extract information about what the LSTM has learned (Chapter 4, Chapter 5)? Can LSTM-based models be used to deliver skillful forecasts of drought properties, measured by a satellite-derived index using EO data in an otherwise data sparse region (Chapter 5)?

Chapter 2 introduces the philosophical approach of data-driven hydrological modelling. This chapter also introduces some of the early work utilizing neural network models for rainfall-runoff modelling. The Long Short Term Memory (LSTM) architecture is introduced and the suitability of this architecture for hydrological modelling is discussed.

The field of machine learning interpretability is introduced and the chapter concludes with a discussion of how data driven models can be used to advance our scientific as well as our predictive goals.

Chapter 3 benchmarks the LSTM against a number of conceptual models for rainfall-runoff modelling in Great Britain (GB). This chapter evaluates LSTM performances over a large sample of catchments from across GB and using a variety of performance metrics. Uncertainty is estimated by training an ensemble of models all with different random seeds and hence initial conditions. Building on community benchmarking efforts in GB and elsewhere, this provides a reference for model performances that future studies can compare against to ensure that we continue to improve simulation accuracy. This chapter concludes with a call for further analysis of what the deep learning models have learned about the hydrological system.

Chapter 4 investigates what the trained LSTM from Chapter 3 has learned, seeking to determine whether the state vector of the LSTM corresponds with intermediate hydrological stores, snow water and soil moisture. We define a simple model, the "probe", to extract meaningful concepts from the LSTM internal states.

Chapter 5 applies the LSTM and its variants to a more data-scarce region, Kenya in East Africa. Moving away from rainfall-runoff modelling, this chapter uses meteorological data from a global reanalysis product to forecast a satellite-derived vegetation health index, used operationally in Kenya to define drought stress. The LSTM-based models are tested against alternative approaches and against previously published results. Furthermore, alternative methods for interpreting the models are explored. Firstly, diagnosing how the model groups together similar behaviours, and secondly, exploring the importance of different input features using gradient-based methods to explore the contribution of input features.

Chapter 6 provides a summary of the previous chapters and discusses avenues for further research, outlining possible next steps for the exploration of deep learning methods in the hydrological sciences.

2 Literature Review

This chapter provides an overview of the existing literature from which the research questions and corresponding objectives for this thesis are derived. The chapter covers key topics such as: (i) data driven modelling; (ii) deep learning applied to hydrological systems; (iii) interpretability and concept formation. This chapter begins with an introduction to recent advances in Deep Learning. The chapter proceeds with a discussion around data-driven models as they have been applied to environmental systems, focusing on applications to rainfall-runoff modelling and drought modelling. It then introduces and motivates the technologies used in this DPhil, the LSTM and EA LSTM, before discussing outstanding issues with this approach to hydrological modelling. Finally, this chapter will address some of the concerns with Deep Learning methods in hydrological sciences, arguing that given the accuracy of model simulations, it is beholden upon us as scientists to interrogate these models and determine what it is they have learned about the hydrological system.

2.1 Deep Learning

"*The ability to learn hierarchies of concepts, building up multiple layers of abstraction, seems to be fundamental to making sense of the world.*" [Nielsen, 2015]

Deep Learning (DL) is a subfield of machine learning that has developed rapidly since 2012 [Krizhevsky *et al.*, 2012]. DL refers to a variety of neural-network architectures that have achieved "unreasonable effectiveness" in a variety of complex, high-dimensional tasks that computers had previously struggled to model accurately [Sejnowski, 2020]. "Deep" refers to the fact that models are made up of multiple layers stacked upon each other, allowing the automatic extraction of increasingly complex concepts from input datasets, and largely negating the need for hand-crafting features for input to models [Bengio *et al.*, 2013]. The process of feature engineering describes the selection and transformation of raw data into a format that can be used for modelling. This is an extremely significant opportunity, since automated feature engineering allows the model to discover new concepts from the input datasets, freeing scientists to interpret results and identify what patterns have been discovered from the data [Shen, 2018].

The origins of DL are much older than the rapid growth since 2012, with fundamental advances in the development of DL, which broadly occurred in two waves, the 1950s and 1980s. The fundamental structure on which DL models are based is the artificial neural network, which describes the linear combination of input values passed through a non-linear activation function [Rosenblatt, 1957], itself an extension of the early McCulloch & Pitts [1943] neuron. After a cooling of interest, multi-layer perceptrons were explored in the 1980s and an efficient training algorithm using stochastic gradient descent was discovered, backpropagation [Rumelhart *et al.*, 1986]. The multi-layer perceptron is a nonlinear statistical model where each layer extracts a linear combination of the inputs, producing derived features, and then models the target variable as a nonlinear function of these features [Friedman *et al.*, 2001]. The multi-layer perceptron describes the stacking of these layers, useful for constructing hierarchical features at different levels of abstraction. Backpropagation is short for "backward propagation of errors", and is an algorithm for efficiently calculating the gradient of each weight in a multi-layer model which is then utilised by an optimisation algorithm (such as stochastic gradient descent) to update the model weights [Schmidhuber, 2015]. For a more complete history of deep learning see Sejnowski [2018].

The recent revival of interest in DL models is largely due to the success of these approaches in benchmark competitions for tasks such as computer vision (e.g. ImageNet, Deng *et al.* [2009]) and natural language processing [Sutskever *et al.*, 2014] and speech recognition (e.g. TIMIT [Garofolo *et al.*, 1993]). These advances have been made possible due to: (a) the advancing availability of computational power, especially the growth of graphical processing units for highly parallelised training of neural network weights, (b) the availability of large, openly available data sets [Schmidhuber, 2015] and (c) the availability of software for flexibly specifying model architectures and training deep learning models (Pytorch Paszke *et al.* [2017], TensorFlow Abadi *et al.* [2016]).

Two common architectures that are worth mentioning are convolutional neural networks (CNNs) which are commonly used for image recognition tasks, and Long Short Term Memory Networks (LSTMs), which are commonly used for tasks with recurrent inputs (e.g. time series and natural language processing). Given the ubiquity of time series modelling tasks in the environmental sciences, the LSTM warrants further exploration in the following dissertation.

Given the demonstrated effectiveness of DL in a variety of highly complex tasks, the application of these techniques to hydrological modelling [Shen, 2018; Shen *et al.*, 2018] and to (geo)scientific contexts more generally [Reichstein *et al.*, 2019] is an active area of research ripe with possibility.

2.2 Hydrological Modelling

"*Let me begin, as all science begins, with the data*" (Kirchner [2006], p.2)

Hydrological modelling describes the process of simulating the movement of water through the atmospheric and land-surface system. This involves encoding our theories about the earth system into computer code, developing a set of instructions that process inputs to make predictions of an output of interest. As outlined in the introduction, modelling is required because the modeller cannot observe everything they wish to know about the hydrological system [Oreskes *et al.*, 1994]. This ignorance can arise either in space (i.e. in ungauged catchments) in time (i.e. future conditions) or for unobserved variables (i.e. groundwater fluxes). The motivations for developing hydrological models can be either scientific (episteme) or technical (techne), but ultimately, I agree with Beven that "the ultimate aim of prediction using models must be to improve decision making about a hydrological problem" ([Beven, 2011b], p2). Given the need to extrapolate in time and space to improve decision making, hydrologists require models that offer useful predictions, defined both in terms of their accuracy and their ability to provide insights into the system being modelled. Various approaches have been developed to address the demand for hydrological models. These approaches can be broadly categorised as spatially-explicit, physically-based models; lumped conceptual models and data driven models [Dadson *et al.*, 2019]. There has been a lot of investment in developing physically-based and conceptual models in the hydrological discipline in recent decades [Dadson *et al.*, 2019]. These two families of models (conceptual and physically based) are sometimes referred to as traditional hydrological models, despite the long history of data driven models [Mulvaney, 1851].

Conceptual models are a specific kind of hydrological model [Kavetski *et al.*, 2006] which can also be referred to as bucket models [Sivapalan *et al.*, 2003a] or explicit soil moisture accounting models [Beven, 2011b]. The final term refers to the mechanism by which soil moisture is treated as a "leaky bucket" with fluxes of water and the current state of the system determined by how full the bucket is at time t. These models are often used in operational settings because they have low data requirements and are computationally efficient. A wide variety of model structures have been developed and tested including TOPMODEL [Beven & Kirkby, 1979], SACRAMENTO [Burnash *et al.*, 1973b], ARNO/VIC [Liang, 1994], and PRMS [Markstrom *et al.*]. These models differ in terms of their internal structure and how they parameterise the flux equations, determining how mass is stored and processed from input to output. Furthermore, they

can differ in terms of how strictly they encode conservation laws, and some have parameters describing the loss of water from a system (e.g. GR4J, Perrin *et al.* [2003]), potentially capturing either systematic errors in the input data, or else inter-catchment transfers of water i.e. through groundwater or anthropogenic means. They can be run in a lumped form (treating the catchment as a single entity) or in a semi-distributed form, grouping the catchment into hydrological response units which have similar hydrological behaviours. These models often have relatively few parameters, require only one or two input variables and they are less computationally expensive to calibrate and run [Knoben *et al.*, 2019]. Furthermore, they are often relatively accurate in the settings in which they are applied. This has led to them being widely used as operational models, and there is extensive literature comparing their performance in various scenarios [Adams & Pagano, 2016; Lindström *et al.*, 2010; Thielen *et al.*, 2009; Wesemann *et al.*, 2018].

Recent advances in computational resources, the availability of high resolution datasets and improved understanding of physical processes and data assimilation schemes have led to the development of increasingly sophisticated physically based models. Physically -based, spatially explicit models describe the surface and subsurface flow processes, linking them with boundary conditions through nonlinear partial differential equations. The seminal paper in this area, from which many of today's process-based models derive their structure, was written more than 50 years ago [Freeze & Harlan, 1969]. These equations are solved using numerical schemes which replace continuous differentials with finite differences by chunking time and space into individual units. There are a plethora of choices a modeller must make, in terms of what and how these equations are represented and solved, and this has led to a variety of hydrological models, including the Joint UK Land Environment Simulator (JULES) [Best *et al.*, 2011], CLASSIC-GB [Crooks *et al.*, 2014] and MIKE SHE [Abbott *et al.*, 1986a,b]. In agricultural systems, where we are interested in simulating the vegetation and crop dynamics, these process-based models are often referred to as mechanistic models [Jones *et al.*, 2017], and include models such as STICS [Brisson *et al.*, 2003] and WOFOST [Van Diepen *et al.*, 1989].

Both approaches, physically based and conceptual models, are human-interpretable in an important sense. For both families of models we can extract the values of hydrological stores, latent variables whose values may be of interest, such as soil moisture [Bouaziz *et al.*, 2021]. While there may be an issue in validating the models in this way, due to the incommensurability of measurements and modelled quantities (because the observed values are at a point scale, whereas the modelled values are at a grid- or

catchment- averaged scale) it can still be meaningful to extract this information. Furthermore, we can ensure that fundamental conservation laws are met, conservation of mass for conceptual models, and conservation of both mass and energy for physically based models. Fundamentally, we can interrogate physically based models since we have built them from first principles, describing our understanding of the hydrological system mathematically and representing that in the computer model.

2.2.1 Key issues in Hydrological Modelling

The promise of the physically-based approaches is that because they are derived from fundamental equations that describe the movement of fluids through porous media (Darcy's Law and Richard's Equation) we should expect them to generalise well, since fundamental laws of physics remain the same in unseen locations (ungauged basins) and unseen times (the future). As Kirchner [2006] has pointed out, however, this premise is based on the assumption that the physics derived from laboratory-scale experiments will "scale-up" to the model grid scale, where state variables are averaged and effective parameters capture the heterogeneity of the subsurface. Firstly, it is possible that this assumption is wrong, and that the scale of model application is important such that the governing equations do not apply at this scale. Given the variability of catchment characteristics and the resulting hydrographs, some hydrologists argue that the "uniqueness of place" means that scale-relevant theories do not exist [Beven, 2000].

A second problem relates to the effective parameters of hydrological models. There are many parameters for each grid box. While procedures exist to estimate parameters in a spatially consistent manner [Franchini & Pacciani, 1991], the reality is that physically-based hydrological models tend to utilise this complexity to overfit, dancing like "mathematical marionettes" to the tune of the calibration data [Kirchner, 2006](p3). The overfitting of these highly parameterised models potentially suggests that we lack sufficient information in current datasets (either due to the number of samples in of those datasets or the information content of the data itself) to calibrate these models effectively.

Discussions around the parameterisation of hydrological models leads neatly to the introduction of two key concerns for hydrological modelling, overparameterization and equifinality. Both concepts refer to the difficulty of fitting models that generalise well. Beven [2011a] summarises the problems occurring due to overparameterisation and equifinality as so: "*A good performance in fitting the learning set does not guarantee a good performance in prediction when the conditions go outside the range seen in the learning set*"

(p.101). Equifinality describes the phenomenon where different model structures and parameter values can lead to equivalent outcomes [Beven & Binley, 1992; Beven & Freer, 2001a]. This can also describe situations where different components of a model cancel each other out, compensating for error and structural weakness [Hrachowitz *et al.*, 2013]. Interestingly, an equivalent problem occurs in neural networks and is called "co-adaptation of weights" [Hinton *et al.*, 2012]. Closely related is the concept of overparameterisation. Overparameterisation refers to the problem of too many parameters leading to identifiability issues, whereby two different values of parameters give equivalently good model “fits”. One can imagine trying to fit a linear regression with only 1 data point; any line you draw through that one point will fit the data equally well. Ultimately, overparameterization describes complex models with too few data points (or too little information in the data) to effectively constrain the parameter values.

Another key issue in hydrological modelling is the nature, role of, and treatment of uncertainty in hydrological modelling. Uncertainty affects all model typologies and is an issue that the hydrological community has actively engaged with [Beven & Binley, 2014; Beven, 2011a; Nearing *et al.*, 2016, 2020b; Weijs & Ruddell, 2020].

One can broadly classify two classes of uncertainty, epistemic uncertainty (owing to a lack of knowledge) and aleatoric uncertainty (uncertainty due to inherent randomness) [Beven, 2016; Gong *et al.*, 2013]. Mathematically, one can think of uncertainties as being described by a probability distribution. Epistemic uncertainty describes not knowing what the correct distribution is. Aleatoric uncertainty describes not knowing what value a random variable drawn from that probability distribution will have. Uncertainty in hydrological modelling is largely epistemic, and arises from data uncertainty, parameter uncertainty and model structure uncertainty [Beven, 2011b].

Model structures differ in their degree of process realism, the level of discretisation and the stores and fluxes that are represented. Data driven models also differ in structure, although as we discuss in Section 2.3, the aim of data driven models is to impose as few structural assumptions as possible. These model structures can be considered as alternative hypotheses for the structure of a hydrological system. One source of uncertainty is, therefore, model structure uncertainty. An alternative source of uncertainty is the parameter values. The process of calibration involves the fitting of model parameters to ensure that model outputs match observed data, which themselves may be subject to uncertainty. Above we introduced the concept of equifinality, describing the idea that multiple parameter sets or model structures can give similarly plausible predictions. This is a key driver of uncertainty in hydrological systems.

Complex, open systems require computational modelling in order to derive gener-

alisable insights [Oreskes *et al.*, 1994], however, these models are necessarily built up from interacting hypotheses which makes it difficult to assess these hypotheses individually [Cartwright & McMullin, 1984]. One approach for addressing the difficulty of separating multiple hypotheses from a single model is the Framework for Understanding Structural Errors (FUSE) [Clark *et al.*, 2008]. As far as possible, this approach seeks to disentangle modular components of hydrological models, often conceptual model components, and this therefore allows us to test these individual components to evaluate our hypotheses [Clark *et al.*, 2015].

One approach that addresses parameter uncertainty is the Generalised Likelihood Uncertainty Estimation [Beven & Freer, 2001a]. This involves randomly sampling parameter sets and keeping "behavioural" simulations, those whose model performance is above a predefined threshold. Rather than having one single optimal solution, therefore, the modeller has a distribution of reasonable values and can propagate that uncertainty through to the simulation. Furthermore, one can use these parameter sets to determine the sensitivity of the model outputs to uncertainties in particular parameters, guiding the modeller towards the parameters that require extra attention in their calibration. This has been applied in a number of studies, including a large benchmarking study in Great Britain [Lane *et al.*, 2019].

A final source of uncertainty arises from uncertainties inherent in the data used to calibrate models. Our interpretation of observational data is sometimes termed our observation model, to make clear that even our interpretation of raw data requires certain assumptions. Discharge data can be biased by the ratings curves that are used to translate river stage into volumetric fluxes [McMillan *et al.*, 2018], and input variables, for example rainfall and evapotranspiration, are themselves estimated using models. This is especially the case for national scale datasets, where assumptions are required to determine how point based estimates (such as rain gauges), or satellite measured reflectances are converted into a consistent data product for the national scale. Estimating data uncertainties is important for knowing whether our conclusions about catchment processes are supported, or whether, given the uncertainties, other hypotheses are equally plausible.

2.2.2 Summary

In this subsection I have introduced some of the most pressing challenges in hydrological modelling, overparameterization, equifinality and uncertainty. I have deliberately outlined them prior to introducing the third family of models, data driven models. The

reason for doing so is to highlight that these criticisms have been levied at all instances of hydrological models, and that these issues are not unique to data driven models.

I give a very brief summary of the current information that has been outlined in this subsection.

1. There exists process-based and conceptual models, which differ in their degree of process realism and the spatial discretisation of the simulation (catchment - sub-catchment)
2. These models are derived from physical concepts, such as conservation laws (conceptual and process-based models) and partial differential equations describing the flow of fluids through porous media (process-based models).
3. These model families give us mechanisms for interpretability, allowing us to extract physically meaningful parameters and stores that we can use to describe and understand the modelled system.
4. There are issues with the application of physical equations derived at the laboratory scale to larger scales categorised with heterogeneities.
5. Experiments have shown that conceptual and process-based models struggle to generalise. Hypotheses for why process based models struggle focus on the large numbers of parameters in these models. These parameters (and even model structures) can differ, suggesting the catchment structure also differs, yet these different models give equivalent predictions.
6. The simplicity of the conceptual models often leads to them being preferred in operational settings, since they are faster to run, easier to calibrate and often provide sufficiently accurate simulations at a catchment scale.

2.3 Data Driven Approaches

An alternative set of approaches to hydrological modelling have become known as "data-driven" approaches [Reichstein *et al.*, 2019]. Data driven models correspond to those models that assume very little prior knowledge of hydrological processes and seek to statistically relate inputs to outputs. This approach is "unashamedly empirical" [Beven, 2011b] (p84) seeking to generate models not from physical processes, but from analysis of time series data [Solomatine *et al.*, 2009]. These approaches contrast with the "knowledge driven" models describing physical behaviour and outlined above, which impose

strong assumptions about the physical behaviour of the system prior to observing the data. The data driven approach is an inductive approach that seeks to learn the relationship (mapping) between a system's inputs and its outputs, allowing the modeller to predict the system's outputs from known inputs (in a different location or time).

Regression-based approaches have been explored by hydrological modellers since the middle of the 19th Century [Mulvaney, 1851] and an alternative data driven approach, unit hydrograph theory, was developed later in the 20th Century [Sherman, 1932]. This suggests that the data driven approach has a long history in hydrological modelling. Further development of data-driven approaches occurred first in the early 1990s under the banner of Hydroinformatics [Abbott *et al.*, 1991] which is itself an active area of hydrology with named research groups, conferences and journals [Abrahart *et al.*, 2008].

2.3.1 Artificial Neural Networks in Hydrology

Data driven models include a wide variety of approaches that include but are not limited to: linear regression, artificial neural networks [Dawson & Wilby, 1998], genetic programming based approaches [Chadalawada *et al.*, 2020a] and tree based models [Gauch *et al.*, 2021b; Solomatine & Dulal, 2003]. As we explored in Section 2.1, ANNs are an important building block of modern DL approaches, and we therefore focus on the historical development and testing of ANN based approaches for hydrological modelling. We first turn to ANN approaches in drought modelling, and we then turn to ANNs in rainfall-runoff modelling.

Drought is a complex hazard that involves the complex interaction of geology, ecology and hydrology. There are comparatively fewer physically based models used for drought modelling, although land surface models are used for modelling drought events [Marthews *et al.*, 2015; Uhe *et al.*, 2018]. Given this, data driven models generally, and ANN based approaches specifically, have been important for drought modelling. Anshuka *et al.* [2019] give an overview of data driven approaches for meteorological drought modelling, exploring papers that have sought to model the standardised precipitation index [McKee *et al.*, 1993] as the target variable of interest. Mishra & Desai [2006] explored an ANN to forecast the SPI for the Kansabati River Basin in India. They found that the ANN performed better than the ARIMA models that they benchmarked the ANN performance against. Alternative approaches have combined ANNs with wavelet modelling, with notable success [Belayneh *et al.*, 2014, 2016]. Indeed, in a review of approaches, Fung *et al.* [2020] note that these wavelet-ANN approaches produce better goodness-

of-fit metrics than standard ANN approaches. In their review of the application of ANNs to drought modelling, they found that the ANNs generally performed well in the studies where they were tested, however, they suffered from a lack of a physical interpretation, they were strongly limited by the availability of data and they tended to overfit (Fung et al 2020). This is a brief review of the ANN literature in drought modelling, however, it illustrates the fact that ANNs have been applied for drought modelling in numerous geographical contexts, for various drought indicators (see Anshuka *et al.* [2019], Mishra & Singh [2011], Fung *et al.* [2020] for more information).

Various studies have explored ANN based models for rainfall-runoff modelling [Dawson & Wilby, 2001]. These approaches can be broadly classified as lumped, deterministic data-driven models [Wilby *et al.*, 2003]. The first studies in the early 1990s applied feedforward, fully connected neural networks to rainfall runoff modelling [Daniell, 1991; Halff *et al.*, 1993]. Since then the application of ANNs to rainfall runoff modelling has expanded (see reviews in Dawson & Wilby [2001] and Abrahart *et al.* [2012]). Abrahart *et al.* [2012] describe a number of themes in which ANNs have been developed since the early 2000s. The majority of research focused on the implementation of these methods as non-linear regression models, and explored a large number of applications of various training approaches, model structures and geographical applications. However, their conclusion was somewhat disheartening: "*The number of papers per annum may be increasing but the scientific progress is very sluggish. In common with many other areas of NN-based hydrology, the field appears to operate as something of an anarchic 'free-for-all', which lacks a common will and strategy for taking on the higher-risk research opportunities which can offer the greatest scientific rewards for the discipline*" (Abrahart *et al.* [2012] p501). They argue for further focus on interpretability, incorporating uncertainties and integrating findings with traditional approaches.

2.3.2 Concerns with Data Driven Approaches

"*Although hydroinformatics and data-driven modelling have been in use for more than two decades, it is struggling to find full acceptance within the hydrological community, which is dominated by large groups of traditional hydrologists because of inherent problems in these models (e.g., chances of overfitting, redundancy of input, lack of modelling rigor, lack of transparency in reproducing results, uncertainty issues, etc.).*" (Remesan & Mathew [2015], p21)

In addition to the issues of overparameterization outlined above (Section 2.2.1), to which we return later (Section 2.4), there are specific concerns with data driven models

that need to be outlined. The main concern points to the fact that data driven methods lack a physical basis. An associated, but distinct issue, is around model interpretability, something that is discussed more fully later in this Chapter (Sect. 2.4.4).

A pressing concern for data driven models is the lack of a physical basis in the model calibration and architecture. This is closely related to the issue of poor ability to generalise outlined above. There are three reasons for concern that can be discerned from the hydrological literature. The first is that the lack of a physical basis disregards physical theories developed over the past half century. The second is that data driven models are fully dependent on the information present in the training (calibration) period, and so if nonstationary behaviours can be expected, hydrologists should be concerned about the generalisation capabilities of their models [Slater *et al.*, 2020]. The third reason is that the latent parameters in the model are not directly related to the physical phenomena being modelled.

The lack of a physical basis for the models is problematic because, it is argued, data driven modelling disregards progress made in recent decades and fails to integrate domain knowledge. This idea is frequently cited [Hrachowitz *et al.*, 2013; Kirchner, 2006]. Indeed, while writing a history of the development of hydrological modelling Todini [2007] identified that data driven approaches "*derive [model structure] from the observations, thus disregarding de facto the results of at least 50 years of research efforts aimed at specifying the physical hydrological mechanisms that generate floods*" (p.471). In order to motivate their study into the internal dynamics of an ANN, [Wilby *et al.*, 2003] wrote "*there are concerns within the hydrological modelling community that research into neural solutions may be a scientific cul-de-sac … research (to date) has done little to build on existing knowledge, or to provide greater understanding of hydrological processes per se*" (p.164).

The second reason is that since a data driven model must learn physical laws for itself from the data, if there are statistical differences in the training (calibration) data compared with the test data, then it is likely that the data driven model will struggle to generalise. Given this dependence on the training set, how can we have any confidence that a model trained on data from one period/location will continue to be useful in different conditions? The ability to extrapolate our findings is one of the key motivations for modelling in the first place [Oreskes, 2003]. Many hydrologists agree with the sentiment that a model must "*work well for the right reasons*" [Kirchner, 2006], since correlations between variables that allow the model to work in the training conditions may break down in future conditions, especially, given non-stationary climate conditions. The promise of developing models from fundamental physical laws is that we can safely assume that the laws of physics will continue to operate in these conditions. However, when our

model is unmoored from these principles, we have less faith in its ability to continue to be useful.

The third reason is that the latent variables in a data driven model may no longer be physically interpretable. In physically-based and conceptual models, the unobserved parameters that are estimated after model fitting can be measured. In principle, this allows hydrologists use their models as hypotheses and go out into the field and make measurements of the latent parameter that has been fit, and thereby estimate whether the model has accurately learned to simulate the physical phenomena [Beven, 2011a]. In data driven models there is no guarantee that the parameters are physically meaningful, and this makes it harder to use these models as hypotheses.

This is something of a paradox. Many scientists who use and develop process-based models dislike the data driven approach because it is not derived from known physical laws. However, the fact remains that these data-driven approaches outperform process-based models. It remains an open question as to where, when and why data driven approaches outperform the physical models. This research seeks to begin to answer these questions, by focusing on what we can learn from the locations and times where the DL approaches outperform more traditional hydrological models, and seeking to interpret why the DL approaches are able to better simulate hydrologic processes.

Despite some pessimism about data driven modelling approaches, studies from the early 2000s began to explore methods for interpreting data driven models. We will first discuss one data driven approach that explicitly focuses on post-hoc interpretations of statistical models. We will then return to the approaches from DL.

One response to the criticism that data driven models are non-physical is that data driven models offer useful insights into the hydrological system because of their lack of physical theory, not despite it. One interesting approach that builds on this idea is Data Based Mechanistic (DBM) modelling [Young, 2003]. The DBM approach argues that the lack of a physical basis is an important strength of the data driven approach. The following steps outline the approach taken [Young, 2003](p2198).

1. Define the objectives of the model and consider the data and possible model structures that are appropriate. Young writes: "*the prior assumptions about the form and structure of this model are kept at a minimum in order to avoid the prejudicial imposition of untested perceptions about the nature and complexity of the model needed to meet the defined objectives.*" (p2197)

2. Use statistical inference to select appropriate model structures.

3. Fit parameter values for the chosen model.

4. Consider nonstationary or nonlinear aspects of the system behaviour, perhaps using time-varying parameters.

5. Fit non-linear model parameters if required in step 4.

6. Consider physical ("mechanistic") interpretations of the model, does it "*provide a description that has direct relevance to the physical reality of the system under study at the scale of interest*." (p2198)

This approach is interesting because it emphasises the need to rely on inductive inference, letting the data drive the choice of model structure. Furthermore, it provides an explicit reaction to process-based models which impose strong prior perceptions of model form, and have many parameters, making it hard to fit these parameters with the given data. In contrast, the data based mechanistic modelling approach tries to fit a minimally parametrized data-driven model that explains the data in a "*statistically efficient*" manner [Young & Chotai, 2001]. This prevents the danger of overconfidence in prior assumptions about the system being modelled. In fact, Young explicitly labels his method as "top-down" [Jothityangkoon *et al.*, 2001] to contrast this method with the "bottom-up" approach of determining the fundamental units and how they behave and then aggregating them to produce simulations at the level of interest (national or catchment scale). He writes: "*Conceptual and physics-based models tend to be the slaves of deterministic reductionist thinking*" [Young, 2002](p1435). While no modelling approach can be perfect, Young's comment communicates that writing physical laws from the laboratory scale into computer code and scaling them up to the whole system can cause difficulties. These difficulties arise from imposing structure on the data with too much certainty, and preventing a more complete search of the space of functions that translate meteorological inputs to hydrological outputs.

For process-based models parameters correspond to (potentially) observable quantities. This allows for easy interpretability, since after calibration, model parameter values can be used as hypotheses and tested by comparing these parameter values against observations. While there are difficulties with the mismatch between the scale of observations and parameter values [Beven, 2011b], the possibility of this hypothesis testing is a powerful argument in favour of process-based models. One approach motivated by the DBM framework is to first fit a flexible model with minimal assumptions about the nature of our system of interest. After fitting the model we attempt an interpretation of what the model has learned, seeking to use the model for knowledge extraction (see Chapter 4).

Examples of the post-hoc interpretation of ANNs exist in the literature. One such experiment can be found in Dawson & Wilby [2001]. They explored an ANN trained for two catchments, the River Ouse in England and the Kentucky River, USA. The ANN had good goodness-of-fit metrics, but what was innovative about this study was that they were explicitly able to relate individual neurons to different components of the hydrograph. They found that the different hidden units in their network modelled different components of the hydrograph, quickflow, delayed surface flow and subsurface flow. This was further corroborated in different contexts by Jain & Srinivasulu [2004] looking at the Kentucky River in the USA and See *et al.* [2009] looking at the River Ouse. These studies demonstrated how ANNs may have a mechanistic interpretation and that the information encoded in the model could be extracted in a human-interpretable way. Despite claims to the contrary, data driven models do often contain interpretable information about the system of interest.

2.3.3 Summary

In this subsection we have introduced data driven hydrological models, and outlined the historical application of data driven approaches generally, and ANNs specifically, to hydrological modelling, including examples from both streamflow simulation and drought forecasting. Early applications of algorithms from traditional ML approaches for streamflow simulation have been met with widespread, and justified, scepticism about neural network based approaches.

However, recent advances in DL approaches have demonstrated powerful learning algorithms that can perform well on problems in machine learning that were previously difficult (image recognition, natural language processing). DL is an approach that falls within the data-driven family of models, seeking to minimise prior assumptions about the form and structure of the model to avoid imposing uncertain preconceptions about the nature of the modelled system [Young, 2003]. It is to the application of the DL approaches in hydrology that this dissertation now turns. I will first outline the specific DL architecture that I explore in this thesis. I will then turn to the current state of the art in DL hydrological modelling.

A brief summary of the key points outlined in this section:

- Data driven modelling is the oldest approach to hydrological modelling and involves using a flexible model encoding very few prior physical assumptions and then fitting that model to data.

- ANNs are one data-driven approach that have been applied to hydrological modelling, both for rainfall-runoff tasks and for drought modelling.
- It is often assumed that the lack of a physical basis of data driven models makes them uninterpretable, since parameters do not refer directly to measurable quantities.
- It is also argued that data driven models will fail to generalise since they tend to overfit the peculiarities of the training dataset, and will therefore struggle to learn the physical relationships that govern the stores and fluxes of water in the hydrological system.

2.4 Deep Learning in Hydrological Modelling

ANNs were extensively tested in the 1990s and 2000s for rainfall runoff systems, and continue to be used today for hydrological models. However, an issue regarding the choice of a lag parameter arises when using ANNs. The sequential structure of the data is lost, and a lag parameter must be selected. The lag parameter defines the number of lagged timesteps used as input to the model. In order to address this problem, recurrent neural networks [Rumelhart *et al.*, 1986] were first tested for hydrological modelling by lin Hsu *et al.* [1997]. These models process input data at each timestep, thereby encoding the sequential nature of the input data. Recurrent Neural Networks (RNNs) were found to outperform artificial neural networks by Kumar *et al.* [2004] who compared the performance of these two architectures for monthly streamflows in the river Hemavathi, India. RNNs can be considered a form of deep learning, since when calibrated using backpropagation, each RNN cell (one per timestep) must be unrolled into a feedforward structure, stacking the RNN layers on top of one another. The "feedforward structure" refers to the fact that each timestep can be considered a new layer, and so one can imagine each timestep ingesting the outputs from the previous timestep.

Despite improvements by moving from an ANN to an RNN, problems with this architecture remain. The RNN cannot learn dependencies of more than 10 time steps due to “vanishing-”, or “exploding gradients” [Bengio *et al.*, 1994; Hochreiter & Schmidhuber, 1997]. The gradient calculation for early layer parameters includes a product of all trainable parameters in later layers. Training recurrent neural networks uses backpropagation through time, which “unrolls” the recurrent neural network so that it becomes a feedforward network of T copies. As the number of timesteps (copies) increases, the multiplication becomes unstable if values are smaller than one (vanishing gradients) or

greater than one (exploding gradients) [Nielsen, 2015]. Therefore, updates to early parameters are either too slow, or too fast [Pascanu *et al.*, 2013]. Therefore, as networks become deeper, the gradient gets smaller as the optimizer moves backward through the layers, meaning that the updates to weights and biases in early layers update increasingly slowly [Nielsen, 2015; Pascanu *et al.*, 2013].

In a hydrological system, there are processes that require information from prior to 10 time steps (days in most applications), such as catchment wetness, soil moisture and snow accumulation/ablation (stores of water that can persist from the winter until spring or summer). The current state of the art for sequential data where we need to incorporate effects occurring many timesteps ago, such as rainfall-runoff modelling, is the Long Short Term Memory Network (LSTM).

The section that follows provides an overview of the architecture of the LSTM, and how it can be used to model hydrological systems. It is important to note that the LSTM architecture is not a novel contribution of this thesis, indeed, the LSTM has been used in machine learning for three decades [Hochreiter, 1991]. The novelty of this thesis is in the application of these methods to the rainfall-runoff system in Great Britain (Chapter 3), to the modelling of drought dynamics in Kenya (Chapter 5) and to the systematic exploration of the LSTM internal states for a large sample of catchments (Chapter 4).

2.4.1 The Long Short Term Memory Network (LSTM)

The Long Short Term Memory Network (LSTM) was developed to solve the problem of long term dependencies [Gers *et al.*, 2000; Hochreiter, 1991]. Through architectural changes, the LSTM is able to learn long term dependencies, i.e. retaining information from many steps ago in the sequence. Remembering information for long periods of time is an essential feature of this architecture. The LSTM and the Entity Aware LSTM (EA LSTM) have demonstrated superior performance in modelling rainfall-runoff systems in CONUS [Kratzert *et al.*, 2018, 2019c,e].

LSTMs overcome the issues associated with learning long-term dependencies by maintaining two state vectors, a cell memory vector that captures slowly evolving processes ($\mathbf{C_t}$) and an output state vector, colloquially named the "hidden" vector ($\mathbf{h_t}$, Eq. 2.6). The $\mathbf{C_t}$ vector, accounts for longer-term dependencies, and a series of 'gates' control the information passing into and out of the memory vector. The $\mathbf{h_t}$ vector evolves more quickly depending on input information and the output of the memory vector (see Fig. 2.1). The gates include: the forget gate ($\mathbf{f_t}$), which controls the elements of the cell memory vector that are forgotten (i.e. how long water persists in the system,

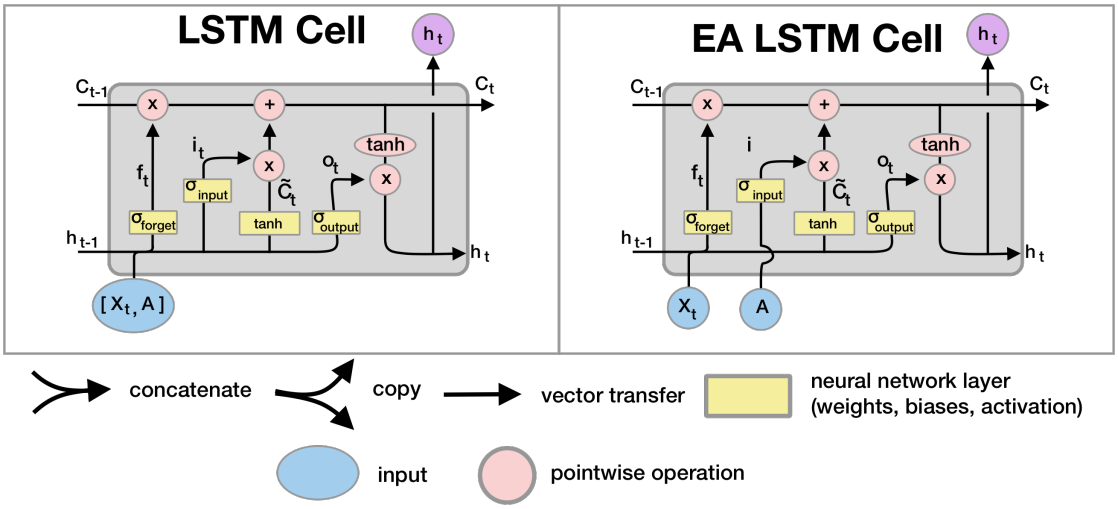


Figure 2.1 | Wiring diagram for the LSTM and Entity Aware (EA) LSTM recurrent cells, adapted from Olah [2016]. These cells are repeated for each input timestep in our sequence length. The key difference between the EA LSTM and the LSTM is the separation of the static data, $\mathbf{A_n}$), from the dynamic data $\mathbf{X_{t,n}}$. In the EA LSTM, the static data is the sole input to the input gate, producing an embedding, $\mathbf{i_t}$. In both LSTM models there is a cell state $\mathbf{C_t}$, that passes from cell to cell, capable of modelling longer-term dependencies. Note that the neural network layers correspond with the weights ($\mathbf{W}$), biases (b) and activation functions (σ, tanh). These operations correspond to the yellow layers in the diagram.

Eq. 2.1); the input gate ($\mathbf{i_t}$), which controls what information from the new input data at that timestep will be incorporated into the cell memory vector (i.e. what information is stored for future timesteps, Eq. 2.2); and finally the output gate ($\mathbf{o_t}$), which determines what information from the cell memory will be used to update the hidden state (i.e. what information will impact discharge at the current timestep, Eq. 2.3). These gates are neural network layers, made up of weights (W_{layer}), biases (b_{layer}) and activation functions. The activation functions allow the LSTM to model nonlinear processes. During training, we seek the values for these weights and biases that best describe observed discharge. The information that passes through the input gate to the cell state ($\mathbf{C_t}$ - see Eq. 2.5) is itself processed through a neural network layer, producing a series of candidate values that may be used to update the cell state (Eq. 2.4). Finally, information from the cell state is passed through the output gate ($\mathbf{o_t}$) to produce the hidden output ($\mathbf{h_t}$) at that time-step (Eq. 2.6). Note that for the LSTM we have explicitly defined the inputs as the concatenation of the dynamic meteorological data and the static catchment attributes, $[\mathbf{X_{t,n}}, \mathbf{A_n}]$.

$$\mathbf{f_t} = \sigma\left(\mathbf{W_f}\left[\mathbf{X_{t,n}}, \mathbf{A_n}, \mathbf{h}_{t-1}\right] + \mathbf{b_f}\right) \tag{2.1}$$

$$\mathbf{i_t} = \sigma\left(\mathbf{W_i}\left[\mathbf{X_{t,n}}, \mathbf{A_n}, \mathbf{h}_{t-1}\right] + \mathbf{b_i}\right) \tag{2.2}$$

$$\mathbf{o_t} = \sigma\left(\mathbf{W_o}\left[\mathbf{X_{t,n}}, \mathbf{A_n}, \mathbf{h}_{t-1}\right] + \mathbf{b_o}\right) \tag{2.3}$$

$$\tilde{\mathbf{C}_t} = \tanh\left(\mathbf{W_C}\left[\mathbf{X_{t,n}}, \mathbf{A_n}, \mathbf{h}_{t-1}\right] + \mathbf{b_C}\right) \tag{2.4}$$

$$\mathbf{C_t} = \mathbf{f_t} * \mathbf{C}_{t-1} + i * \tilde{C}_t \tag{2.5}$$

$$\mathbf{h_t} = \mathbf{o_t} * \tanh(\mathbf{C_t}) \tag{2.6}$$

We can think about this very similarly to a state space hydrological model, as expressed below (Eq. 2.7, 2.8, 2.10, 2.11, Kratzert *et al.* [2019b]). What is clear is that we have some state of the system ($\mathbf{C_t}$, $\mathbf{h_t}$) that depends on the previous state ($\mathbf{C_{t-1}}$, $\mathbf{h_{t-1}}$), the description of the system (captured by the weights and biases, $\mathbf{W}$, b) and the inputs at that timestep ($\mathbf{X_t}$).

$$\mathbf{C}_t = f\left(\mathbf{x}_t, \mathbf{C}_{t-1}, \theta_i\right) \tag{2.7}$$

$$\mathbf{y}_t = f\left(\mathbf{x}_t, \mathbf{C}_t, \theta_j\right) \tag{2.8}$$

$$\tag{2.9}$$

$$\mathbf{C}_t, \mathbf{h}_t = f_{\text{LSTM}}\left(\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{C}_{t-1}, \theta_k\right) \tag{2.10}$$

$$y_t = f_{\text{Dense}}\left(\mathbf{h}_t, \theta_l\right) \tag{2.11}$$

This LSTM structure is particularly well suited to hydrological modelling for the following reasons:

1. The LSTM can be thought of as representing a state-space model, with an input-state-output structure that is familiar to hydrologists and can describe many hydrological models.
2. The LSTM is capable of learning long term dependencies that are important for discharge.
3. The LSTM has a recurrent structure allowing for the processing of inputs at each timestep.

4. LSTMs are universal function approximators [Schäfer & Zimmermann, 2006], and therefore, can search a very large space of potential functions for those that are compatible with the data to minimise the loss function.

5. In contrast with traditional models, the LSTM is tasked with learning relationships between inputs and outputs purely from the data in the training (calibration) period.

As we have outlined above, ANN methods have been criticised on the basis of being unphysical and uninterpretable. These criticisms could equally be lobbied against the LSTM-based approach outlined here. While we have good reasons to choose the architecture in terms of the similarities between the network architecture and our understanding of the nature of the hydrological system, one of the outstanding questions for LSTM based approaches is how do we interpret what they have learned, and how can we further impose physical realism onto the network structure.

2.4.2 The Entity Aware LSTM (EA LSTM)

One recent development of LSTM architectures sought to directly address the desire for interpretability and physical realism.

The EA LSTM was developed specifically for rainfall-runoff modelling to aid interpretability, allowing the scientist to extract the learned similarity between catchment behaviours, stored in the internal LSTM embedding. The key difference between the EA LSTM and the LSTM is that the input gate (**i**) is no longer conditional upon the dynamic (time-varying) data (Eq. 2.13). Instead, the static (time-invariant) catchment attributes ($\mathbf{A_n}$) exclusively influence the input gate (Eq. 2.2 is replaced with Eq. 2.13), and all other gates are solely influenced by the dynamic input data (Eq. 2.12, 2.14, 2.15).

$$\mathbf{f_t} = \sigma\left(\mathbf{W}_\mathrm{f}\left[X_t, \mathbf{h}_{t-1}\right] + \mathbf{b}_\mathrm{f}\right) \tag{2.12}$$

$$\mathbf{i} = \sigma\left(\mathbf{W}_\mathrm{i} A + \mathbf{b}_\mathrm{i}\right) \tag{2.13}$$

$$\mathbf{o_t} = \sigma\left(\mathbf{W}_\mathrm{o}\left[X_t, \mathbf{h}_{t-1}\right] + \mathbf{b}_\mathrm{o}\right) \tag{2.14}$$

$$\tilde{\mathbf{C}_\mathbf{t}} = \tanh\left(\mathbf{W}_\mathrm{C}\left[X_t, \mathbf{h}_{t-1}\right] + \mathbf{b}_\mathrm{C}\right) \tag{2.15}$$

The EA LSTM is described as “entity-aware” because it explicitly learns how to use $\mathbf{A_n}$ to distinguish between similar dynamic inputs ($\mathbf{X_{t,n}}$) for different catchments ("entities"). For the EA LSTM, **i** is determined solely by the catchment attributes (Eq. 2.13).

Therefore, each catchment has one unique vector of size *hs* (hidden-size, a chosen hyperparameter) dimensional which controls what information should persist in future timesteps ($\mathbf{i} \in \mathbb{R}^{hs}$). In contrast, the LSTM learns to modify the input gate $\mathbf{i_t}$ based upon the meteorological forcing data ($\mathbf{X_{t,n}}$) *and* the catchment attributes ($\mathbf{A_n}$). The output of the input gate ($\mathbf{i_t}$ or $\mathbf{i}$) is a vector of values between 0 and 1, which is learned from data. This vector, also known as an "embedding", translates our catchment attributes into a *hs*-dimensional space that represents catchments in a manner optimised to differentiate between catchment rainfall-runoff behaviours. Kratzert *et al.* [2019e] demonstrated how this embedding represents what the model has learned about our catchments.

For the sake of clarity, it is important to note that both models receive the same information. The LSTM still receives the static catchment attributes ($\mathbf{A_n}$). However, rather than affecting only the input gate, the static data can influence all gates, since they are appended to a vector of dynamic inputs ($[\mathbf{X_{t,n}}, \mathbf{A_n}]$) and so the same information is given to the LSTM at each timestep. The static attributes are used by the LSTM in the same way as the dynamic data. This offers extra flexibility for the LSTM compared with the EA LSTM, since the LSTM is able to modify the input gate based on information from time-varying data, whereas the EA LSTM is not. We are using the static nature of the data as a constraint on the EA LSTM to reflect the nature of the input data (separated into static and dynamic inputs - see Fig. 2.1). Both models have a final layer, a fully connected linear layer, which transforms the $\mathbf{h_t}$ vector into a single discharge prediction, $\hat{\mathbf{y}}_{\mathbf{t,n}}$.

2.4.3 The Current State of the Art

Approaches from DL have found their way into hydrological modelling, and recent advances have fundamentally challenged some of the prevailing wisdom, directly addressing concerns outlined in Section 2.2.1 and Section 2.3.2. Research into DL approaches to hydrological modelling has rapidly developed over the course of this DPhil and this dissertation directly contributes to advancing this area of research.

LSTMs have been tested and shown to be successful for a variety of tasks in hydrological modelling. We focus here on the application of LSTMs to rainfall runoff modelling, but other deep learning approaches have been tested elsewhere, including for precipitation nowcasting [Ravuri *et al.*, 2021], extracting precipitation estimates from satellite imagery [Tao *et al.*, 2018], predicting soil moisture [Fang *et al.*, 2017; Lee *et al.*, 2019], groundwater modelling [Afan *et al.*, 2021] amongst others.

The LSTM was originally applied to streamflow simulation in a large sample of US catchments in the Kratzert *et al.* [2018] paper which benchmarked LSTM-based approac-

hes against the Sacramento hydrological model. The paper ran three experiments, first training an LSTM on each catchment separately. Second, trained a single regional model on all catchments. Finally, Kratzert *et al.* [2018] simulated the effectiveness of predictions in ungauged basins (PUB) or short data records by training on a large sample and fine tuning the weights of the network to specific catchments. They demonstrated three things:

1. LSTMs are competitive with traditionally used conceptual models
2. It is possible to examine their internal states for hydrological intuition
3. It was possible to transfer information from basin to basin, thus directly addressing the regional modelling problem.

Thus, they directly addressed the criticism that DL models are uninterpretable, that data driven models won't generalise well and offered evidence that DL models might solve the regional modelling problem. The regional modelling problem describes the outstanding issue in hydrological modelling whereby models trained on discharge in one catchment struggle to generalise to other catchments with different hydrological, geological or meteorological conditions, and thus struggle to apply generally to an entire region.

Later work further developed the LSTM and demonstrated two key findings. Firstly, regional LSTMs outperform conceptual models trained on individual catchments (the conditions in which traditional hydrological models work best) [Kratzert *et al.*, 2019e]. Secondly, by using static catchment attributes as inputs to the model, the DL methods were able to effectively generalise over space, learning the interaction between geological and climatic attributes and the meteorological drivers of catchment discharge processes. This research contributed more evidence that DL models address the regional modelling problem. This demonstrated that in the continental US (CONUS), LSTM based models were the most accurate models (defined by a variety of metrics) across a large sample of catchments.

Secondly, LSTMs trained with regional attributes are able to produce PUB to a comparable accuracy with in-sample catchment predictions [Kratzert *et al.*, 2019d]. The performance in ungauged basins is particularly interesting, since, this task, training or calibrating on one catchment, and then making predictions on another, was identified as one of the key tasks for hydrological modellers, outlined as one of the 23 unsolved problems in hydrology [Blöschl *et al.*, 2019]. PUB has been identified as a key challenge in hydrological modelling [Hrachowitz *et al.*, 2013; Sivapalan *et al.*, 2003b].

Both findings directly address the first two concerns outlined above; namely, that data-driven models are unable to generalise and that data-driven models are overparameterised (see Sect. 2.2). Indeed, it is often said that models have too many parameters, which is one reason for poor extrapolation performance. Evidence from this study suggested that the flexibility of the deep learning models in terms of the numbers of parameters was not a hindrance, but precisely the reason why these models were able to generalise. Unlike traditional hydrological models that are required to search within a very small possible space of functions that map from inputs to outputs, the deep learning LSTM based models can approximate a much larger space of possible functions, and the lack of structure is perhaps the fundamental cause of their ability to better generalise than alternative approaches tested thus far. Nearing *et al.* [2021b], offer three consequences from this finding:

1. Scale-relevant theories exist in available datasets, but as a community we have thus far been unable to encode them in a way that traditional model structures can use.

2. Neither uniqueness of place, nor lack of data hold back our ability to find these theories, since the LSTMs effectively generalise.

3. Beliefs about why conceptual or process based models fail to extrapolate to ungauged basins (overparameterization, lack of regularisation), does not hurt performance for data driven models, even when tested on hold-out data in space (PUB) or time (our validation set).

As we outlined in Section 2.1, uncertainty is a key component of actionable hydrological predictions. Klotz *et al.* [2020] showed that the LSTM architecture can be modified to produce distributional predictions rather than scalar values. They tested a number of different approaches including using dropout to approximate ensembles of different neural network structures, Monte Carlo Dropout [Gal & Ghahramani, 2015], and directly producing parameters of a distribution from the LSTM (mixture density networks). Their study outlined that DL networks can produce uncertainty estimates that chime with hydrological intuition, and that provide accurate predictions.

One of the promises of the DL approach is that information can be extracted from any dataset that is made available, given sufficient training data [Nearing *et al.*, 2021b]. This has been demonstrated in a number of studies. Frame *et al.* [2021b] were able to post-process the US National Water Model, improving the discharge predictions by

using all of the NWM outputs as well as the raw meteorological observations (temperature, precipitation) as input to the LSTM. Gauch *et al.* [2021a] showed how an LSTM might be used to produce predictions at multiple resolutions, combining data from an hourly resolution with data from a daily resolution. Kratzert *et al.* [2021] showed how DL models can improve their estimates by using multiple precipitation products as input, since the different products are associated with different errors, and information can be extracted from the model that describes which product is most informative in what locations. Thus they demonstrated how to combine information in spatially and temporally dynamic ways.

Another key criticism of data driven methods is that due to an overreliance on the training dataset, we need to be skeptical of the ability of these models to perform in future conditions. A key motivation for this concern is that the climate is nonstationary [Slater *et al.*, 2020] and it is assumed that extremes will continue to get more extreme in the future [Arias *et al.*, 2021]. Frame *et al.* [2021a] test the LSTM by training it on periods with lower magnitude extremes than the test periods. This simulates the idea that the future will have bigger flood peaks than the past. They found that the LSTM was the most accurate model at simulating these out of sample extremes. Interestingly, they also found that the original LSTM performed better for these extremes than an LSTM with an imposed mass-conservation law [Hoedt *et al.*, 2021] which was surprising, since it was assumed that in these extremes, a model forced to use the excess precipitation leading to the flood peaks would outperform a model that is supposedly free to remove water from the system. Interestingly, a follow up study [Frame, 2022] has shown that the LSTM itself has lower conservation biases than physically based models that obey conservation laws.

To summarise the findings from the current state of the art for DL based modelling:

- DL models show state of the art accuracy for streamflow modelling, especially when generalising to a large sample of catchments (regional modelling problem) [Kratzert *et al.*, 2019e] and predicting in ungauged basins (PUB) [Kratzert *et al.*, 2019d].
- Methods exist to interpret DL models [Kratzert *et al.*, 2018, 2019a] and the hydrological community should focus significant attention into interpreting what these models have learned and how they generalise effectively.
- Over parameterization is not a cause for lack of generalisation capabilities in DL models, since these models have many more parameters than traditional modelling approaches [Kratzert *et al.*, 2019e; Nearing *et al.*, 2021b].

- Methods exist to directly produce uncertainty estimates from DL models [Klotz *et al.*, 2020].
- DL allows hydrologists to combine multiple sources of information in an optimal way, performing data assimilation in a straightforward way [Fang *et al.*, 2017; Kratzert *et al.*, 2021; Nearing *et al.*, 2021a].

We now turn to an overview of approaches to DL interpretability, which will give us tools to explore how DL methods can advance our scientific goals as well as our predictive ones.

2.4.4 Deep Learning Interpretability

"*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.*" [Nearing *et al.*, 2021b](p5)

"*the potential of models to teach us more about the system can and should be thoroughly exploited by using models as learning tools, as stressed by Beven (2007)*" [Hrachowitz *et al.*, 2013](p1221)

"*Improving predictive accuracy is important but insufficient. ... Interpretability has been identified as a potential weakness of deep neural networks, and achieving it is a current focus in deep learning*" [Reichstein *et al.*, 2019](p199)

We outlined in Section 2.3.2 the criticism that data driven models are not useful, since they disregard progress in hydrological theory and they do little to advance our understanding of hydrological systems. We saw two examples for shallow learning where interpretability was used to better understand what the models had learned, the DBM approach [Young, 2003] and the ANN interpretation of Wilby *et al.* [2003]. In the section that follows we will outline the issue with interpreting DL models. We will then highlight some of the methods for interpreting DL models that are currently being explored in hydrology and in the field of Machine Learning Interpretability [Molnar, 2020; Samek *et al.*, 2019].

As we have explored, DL offers state-of-the-art accuracy for modelling in a variety of scientific and technical domains. These advances have reached hydrology, and there is a growing body of evidence suggesting that these models outperform traditional hy-

drological models. This dissertation forms a part of this body of evidence (Chapter 3). However, given the large numbers of parameters and lack of a physical basis, there is a meaningful sense in which DL models are difficult to interpret, since model weights cannot be individually interpreted, prima facie, as weights in a linear regression model perhaps can be. As an aside, we should note that given the complexity of modern Earth system models, it is also difficult to trace results back to the assumptions that drive them, also limiting their interpretability [Reichstein *et al.*, 2019]. However, given the remarkable performance of these models on large-sample studies, there is an opportunity to use these DL models as knowledge discovery tools. We should utilise existing techniques and develop new techniques to interrogate what these models have learned and translate the learned patterns into something that is human interpretable [Nearing *et al.*, 2021b].

Many approaches now exist for extracting information from trained DL models. For this dissertation we explore two approaches that seem particularly useful for scientific discovery and hydrological modelling. These approaches can be broadly separated into estimating the importance of inputs and interpreting internal weights. However, other approaches exist including approximating DL models with simpler, more interpretable models [Ribeiro *et al.*, 2016b] and generating counterfactual examples to test what would cause a model to change its predictions [Wachter *et al.*, 2017].

Feature importance describes the sensitivity of the model to given inputs, which is important to understand what information the model is using, and how we should expect the model to behave in future conditions. This has been called the "attribution problem" [Sundararajan & Najmi, 2020]. Furthermore, we can use this information to select informative features and discard uninformative features. There are numerous methods for estimating the importance of input features, however, we restrict our analysis to Integrated Gradients [Sundararajan *et al.*, 2017], although others exist (Permutation Importance, Fisher *et al.* [2019]; Parr *et al.*, Shapley Additive Explanations, (SHAP), [Lundberg & Lee, 2017]; LIME, Ribeiro *et al.* [2016a]).

DL networks are differentiable and the software used to train DL networks includes an automatic method for calculating gradients (required for the backpropagation algorithm). Given this, one can imagine assigning feature importances by calculating the gradient of the output with respect to the input. This would tell a user how much the output varies for each unit change in the inputs. In image recognition, we want to answer which aspects of the input image the model finds noteworthy. Unfortunately, using the raw gradient calculations fails to capture what the model deems important, since the gradient saturates at the raw input, where the gradient of the output with respect

to the inputs (X) may be small even if X is important. This is known as the saturated gradients problem [Sundararajan *et al.*, 2017]. Sundararajan *et al.* [2017] proposed a method they call "integrated gradients" where they use a scaling parameter, α, to represent pixel intensities, changing the brightness of the image such that you start with an all-black image, and increase α to 1 until you have your original input. Along that path from an all-black image to the full-brightness image, you calculate the gradients, and then sum them, which gives a more representative estimate of the importance of individual pixels.

Integrated gradients have been used in hydrology by Frame *et al.* [2021b], who post-processed the US National Water Model with an LSTM. This study was mentioned previously (see Section 2.4.3), however, this section is interested in the approach Frame *et al.* [2021b] took to determine how much information the LSTM was extracting from the process-based model predictions compared with the raw input data when using the LSTM to "correct" the model errors. Firstly, they found that the LSTM trained on the raw data produced more accurate simulations than when trained on the physically based model outputs. Secondly, they used the integrated gradient method to attribute the contribution of different National Water Model outputs, finding that the raw meteorological outputs contributed most of the information, and while the LSTM "listened" to the National Water Model by using those features, there was no extra information encoded in the National Water Model states or fluxes [Frame *et al.*, 2021b]. In order to apply the technique to hydrological time series, rather than defining an all black image (with pixel values of 0), they instead use the mean values of meteorological inputs as the baseline.

Other studies by the same authors have used the method in a variety of ways, including estimating the contribution of different rainfall products [Kratzert *et al.*, 2021] and determining the inputs that most influence different cell state values [Kratzert *et al.*, 2019a]. This thesis builds on this work and applies this idea in Chapter 5, albeit with a slightly different approach that uses fundamentally the same technique.

Interpreting internal weights of the network has been explored in hydrology, specifically interpreting what the static input gate of the EA LSTM (Section 2.4.2) has learned about catchment similarity. Kratzert *et al.* [2019e] visualised the static gate activations (valued [0, 1]) in a 2-dimensional matrix, which represented similarity and difference in terms of how catchments respond to meteorological forcings. Given the ability of this model to produce accurate simulations when tested on a large sample of catchments, this matrix contains information about how catchments differ. As Nearing *et al.* [2021b] point out, the job of the hydrologist is therefore to extract this information in

a meaningful way. Kratzert *et al.* [2019e] used cluster analysis and dimensionality reduction on this matrix to determine a) what catchments were deemed most similar and b) what catchment attributes were used to determine that similarity. This offers exciting opportunities for using DL methods for knowledge extraction, and suggests that DL techniques can be used to answer our scientific goals as well as our predictive ones.

2.4.5 Overparameterization and Overfitting

"*Overparameterisation of neural net models relative to the information in the learning set is an issue with this type of model (as in any empirical model). The danger of over parameterisation is that, in general, it will lead to greater uncertainty in prediction or extrapolation, particularly in prediction or extrapolation beyond the range of the learning or calibration set. A good performance in fitting the learning set does not guarantee a good performance in prediction when the conditions go outside the range seen in the learning set (e.g. Cameron et al., 2002; Gaume and Gosset, 2003; Han et al. 2007).*" [Beven, 2011a](p101)

This chapter has already addressed the key outline of Beven's argument in Section 2.2.1. Here, we return to this argument and address the particular concerns for data driven approaches. A frequent concern about DL approaches is that they are prone to overfitting due to the number of trainable parameters, which can be tens of thousands.

Overparameterization is one hypothesised cause for poor performance in conditions that differ from the training set. It is argued that too many parameters cause poor generalisation of hydrological models, including data driven, conceptual and physically based models. Too many parameters begin to fit noise in the training data, which does not extrapolate to new conditions (e.g. the future, ungauged basins). Furthermore, poor generalisation is also potentially a result of different plausible parameter values that can provide equivalent simulations (equifinality - Beven & Freer [2001b]; Beven [2000]). If multiple model "fits" give equally good outputs, then how can the model choose between them? Hrachowitz *et al.* [2013] argue that the "*high number of parameters together with the limited number of constraints . . . resulted in high degrees of freedom, i.e., poorly conditioned parameter estimation problems, so that models 'behaved more like mathematical marionettes'*" (p. 1203). While Hrachowitz et al and Kirchner, from whom the dancing imagery was cited [Kirchner, 2006], were explicitly referring to highly parameterised physically based models, identifiability issues are well known in statistics (where it is often related to the "bias-variance tradeoff" [Friedman *et al.*, 2001]) and systems engineering [Young, 2003]. This criticism equally applies to data driven and DL-based modelling.

There are two responses to this concern. The first outlines the procedures to minimise this danger, a number of which are used in this dissertation. The second highlights that the reasons for the improved generalisation capability of the DL models is an open area of research in the machine learning community.

We first consider procedures to minimise the danger of overfitting. The most valuable approach is to explicitly separate a training dataset, a validation dataset and a test dataset. This three way split allows one to fit the model parameters on the training dataset, test the generalisation capabilities on the validation set, and then finally, publish the results on the test dataset. This does not necessarily prevent overfitting, but the performance on the validation set is a useful guide to judge ones model, and can be used to adjust model hyperparameters (those parameters chosen by the modeller and not chosen through backpropagation of errors) and prevent model overfitting. More advanced approaches exist such as leave one out cross validation [Vehtari *et al.*, 2017], or k-fold cross validation [Kohavi *et al.*, 1995]. However, for modelling time-series, to prevent model leakage (i.e. instances of your test data appearing in your training data), it is often safest to simply split your time series into three. For example, in this dissertation we split data by decades (Chapter 3, 4) or by years (Chapter 5).

Another approach is called early stopping [Yao *et al.*, 2007]. DL models are trained for a number of epochs, where an epoch refers to a single pass of the training dataset, such that every sample has been used to update model parameter values. At each epoch, the model performance can be evaluated against the validation data. A criterion is set for the change in model performance on the validation, such that if the performance of the model does not improve for a number of consecutive epochs, training will be stopped and the final weights taken from the best performing model to that point. This is another reason why splitting a training dataset into three components, train, validation, test, can be useful. Early stopping prevents overfitting by stopping model training before the model parameters start to fit the noise in the training dataset.

A third approach involves the inclusion of a penalty term in the loss function, such that the model is encouraged to minimize the size of parameter values. This is commonly used as an extension of linear regression, and is variably called ridge regression (imposing an L2 norm penalty on the squared weight size) and the lasso (imposing an L1 norm penalty on the absolute weight size) [Friedman *et al.*, 2001]. These penalties can be combined (the ElasticNet is one such combination of both Lasso and Ridge regression) and can be applied to DL architectures. This is often called regularisation. A related idea is to use information based criterions to select optimal models, where there is a penalty on the model complexity (often defined by the number of parameters). The

Akike Information Criterion and Bayesian Information Criterion are two commonly used metrics in hydrology [Weijs & Ruddell, 2020].

One commonly used approach which has been developed specifically for DL, is called dropout [Hinton *et al.*, 2012]. Dropout forces randomly selected weights to zero ("dropping" them). This has two benefits: (1) dropout creates reduced-order subnetworks that are then effectively ensembled in the final prediction [Gal & Ghahramani, 2015]; and (2) dropout encourages the network to use a simpler and more robust representation of processes. Section 2.2.1 mentioned that equifinality can describe situations where different parts of a model cancel out, compensating for errors in other parts of the model [Hrachowitz *et al.*, 2013]. This is called coadaptation of weights, and was the motivation for developing dropout [Hinton *et al.*, 2012].

One of the most interesting questions that currently remains unanswered is how the DL approaches overcome issues of over-parameterisation and overfitting associated with traditional statistical methods. The traditional viewpoint holds that too few parameters causes a model to underfit the given data (high bias) and too many parameters cause a model to overfit the data (high variance), and that there is a trade-off as complexity of our model increases [Friedman *et al.*, 2001; Neal, 2019]. However, recent research suggests that approaches from machine learning (DL, random forests and gradient boosted approaches) show instead a "double-descent curve" where after a critical point, more parameters lead to further decline in test error [Belkin *et al.*, 2019].

While there is certainly a danger of overfitting with a highly parameterised model, empirically, DL has been demonstrated to generalise well in a variety of situations [Schmidhuber, 2015], and further fundamental research is required to consider what conditions prevent the model overfitting. What is important for this dissertation is that a) approaches exist to mitigate the dangers of overparameterization b) the danger of overfitting may be exaggerated.

Certainly, recent research in hydrology suggests that DL approaches are better suited to generalising to unseen conditions than less parameterised alternatives (outlined in Section 2.3) [Kratzert *et al.*, 2019d]. Nearing *et al.* [2021b] outline that the benchmarking studies outlined above have presented evidence that despite the increased number of parameters, this enormous flexibility does not hinder generalisation. Instead, "*it is this lack of regularization that allows them to learn general and transferable hydrological relationships*" [Nearing *et al.*, 2021b](p3).

2.4.6 Summary

This section has explored the application of the LSTM, a particular DL approach, to rainfall runoff modelling. The LSTM has a favourable architecture for modelling hydrological systems with long-term dependencies in sequential data and can be viewed as a state-space model that implicitly models a latent state.

Recent studies that have explored LSTM based modelling have demonstrated that these DL approaches:

- perform well in unseen conditions (in time, space and in terms of simulating extremes).
- offer the ability to deal with uncertainty in their predictions.
- have approaches to deal with overparameterization.
- are interpretable and we can apply and develop methods to examine what they've learned.

We have explored how recent studies have addressed some of the key criticisms of the data driven approach, and demonstrated why the LSTM is worthy of further study. In the following we motivate the work presented in this dissertation.

How robust is the finding that LSTM models solve the regional modelling problem? All previous studies were performed on data from CONUS. As one of the most densely gauged and monitored locations across the world, Great Britain represents an important test-bed for hydrological models [Clark & Khatami, 2021a]. Furthermore, inspired by the benchmarking results published by [Lane *et al.*, 2019], this chapter uses the intercomparison of four conceptual rainfall runoff models with the LSTM performances across a large sample of GB catchments to diagnose missing processes in the conceptual models.

Kratzert *et al.* [2019a] explored the information captured in the cell state, manually extracting individual cell states that corresponded to snow or soil moisture signals. These initial findings suggested that it would be possible to interpret hydrological models as being physically realistic. However, further work remains to be done to address what the models have learned about hydrology. How is it that the LSTM is capable of making better predictions than traditional hydrological models? Furthermore, can we automatically identify cell states that correspond to our intermediate timeseries of interest, and how might we account for the fact that LSTMs can distribute information across

several cells, rather than being forced to store information in one cell. These questions are explored in Chapter 4.

An open question remains: LSTMs work well for catchment systems, and the availability of large sample datasets offer a great test-bed for these approaches. However, there are other hydrological systems where data is available but less structured (remote sensing data), and uncertainties about the physical processes are greater yet a similar input-state-output approach chimes with a hydrologists perceptual model of the system. One such system is how the biosphere responds to meteorological conditions. Chapter 5 explores how LSTMs perform in an altogether more challenging environment, but one with potentially impactful consequences by using LSTMs for forecasting vegetation health in Kenya.

2.5 Hydrology beyond streamflow

Hydrological modelling is often concerned with modelling streamflow. One particularly common task for hydrological modellers is to create rainfall-runoff models, which map meteorological inputs to a discharge estimate. In this dissertation we explore both rainfall-runoff modelling (Chapter 3, 4) and an applied forecasting task exploring remotely sensed drought metrics (Chapter 5).

As well as being motivated by the desire to improve our understanding of hydrological systems (episteme), hydrologists are also tasked with helping decision makers better manage water resources and respond to water related hazards (techne). One such hazard is flooding, and rainfall-runoff modelling is an important component of flood modelling, although other components are required to model the area that is inundated by water [Teng *et al.*, 2017]. Another water-related hazard occurs at the opposite end of the hydrograph, specified by a lack of water. Drought can be simply defined as a "deficit of water relative to normal conditions' [Sheffield & Wood, 2012]. Droughts are extremely damaging natural hazards, and since 2000 there have been damaging droughts on every continent [Baudoin *et al.*, 2017; García-Herrera *et al.*, 2010; Muller, 2018; Nicholson, 2014; Spinoni *et al.*, 2015; Swain *et al.*, 2014; Zeng *et al.*, 2008] and the most recent IPCC report expects that drought frequency and magnitude are expected to increase globally [Arias *et al.*, 2021].

Drought is a difficult hazard to specify in time and space, with many hydrometeorological variables that can be used to monitor and define drought [Van Loon, 2015]. Broadly, we can think about four categories [Sheffield & Wood, 2012]:

- Meteorological drought, referring to a lack of precipitation.
- Agricultural drought refers to a deficit of soil moisture [Seneviratne, 2012] and/or anomalously poor vegetation health [Agutu *et al.*, 2017].
- Hydrological drought related to negative anomalies in river discharge [Cammalleri *et al.*, 2020] and/or groundwater levels [Peters *et al.*, 2003].

In developing world contexts where the impacts from drought hazards are largest, remotely sensed vegetation indices are often used to define agricultural drought [Klisch & Atzberger, 2016]. In Chapter 5 we explore whether the LSTM can be applied for drought forecast, using the Vegetation Condition Index (VCI) as our target variable of interest. Vegetation health forecasting is an interesting problem for hydrological modellers for two reasons. Firstly, vegetation is an important hydrological store. Secondly, drought conditions and negative vegetation health anomalies have the potential to be extremely damaging, especially for pastoralist communities in Kenya. Therefore, this task offers the potential to contribute to the mitigation of a hazard, the reduction of drought impacts, and ultimately, minimise the loss of life.

There is an important research gap to identify the characteristics of hydrological systems that DL approaches can be used for, to explore their potential in a forecast setting and to assess the potential for using DL in an operational context where accurate model predictions have the potential to save lives. In this dissertation we explore drought modelling in Kenya using remote sensing imagery. This addresses three open questions in the application of DL to hydrological systems:

- Does the LSTM generalise to other hydrological contexts, especially in contexts where spatial outputs are required?
- How does the LSTM cope with alternative data sources such as remote sensing data, that may be associated with different uncertainties?
- Can the LSTM be used in an operational forecast setting?

In Chapters 3 and 4 this dissertation contributes to the growing body of literature applying the LSTM to rainfall runoff modelling. However, there are various other hydrological contexts. Given that the LSTM can be considered a state-space model, and in Chapter 4 we explore whether these states correspond to known hydrological stores, it remains an open question whether we can model a hydrological store, such as vegetation health, explicitly.

Ultimately, the same epistemological preoccupations apply, and issues of overparameterization, model interpretability and questions about generalisation capacity of our models remain unanswered. Therefore, Chapter 5 explores the LSTM applied to vegetation health modelling as an alternative hydrological system.

2.5.1 Other Environmental Applications of Data Driven Methods

Environmental science is a broad field with a wealth of modelling problems. While the application of DL to modelling catchment systems and drought conditions is the focus of this thesis, there are many other applications of DL. While a full treatment of all the applications of DL in environmental science is beyond the scope of this thesis (see Reichstein *et al.* [2019]), we turn briefly to the related problem of modelling *global* water and energy fluxes at the land surface.

DL can be used to produce global, gridded estimates of environmental variables from sparse observations. This approach has been successfully applied for various metrics including runoff. The Global RUNoff dataset Ghiggi *et al.* [2019] is a great example of this approach. They built upon the methods of Gudmundsson & Seneviratne [2015] and used an ensemble of tree-based regression models, the random forest, to produce monthly runoff values at 0.5° resolution for the period from 1902 to 2014 [Ghiggi *et al.*, 2019]. Runoff here is calculated as the river discharge at a gauged point (cubic meters per month) divided by the area of the upstream catchment (km2) to give the monthly runoff over that area (mm per month). They used reanalysis data from the Global Soil Wetness Project Phase 3 [Kim, 2017] to provide predictors, and then the model was trained to predict monthly runoff from the reanalysis precipitation and temperature for the previous 6 months. This project was extended to account for the uncertainties in the driving data, and Ghiggi *et al.* [2021] trained random forest models on 21 different atmospheric datasets (including reanalysis and interpolated station data) to build an ensemble with 525 members. Ultimately both studies found that the random forest (data-driven) based models outperformed the traditional global hydrological models (process-based) in their ability to reconstruct out-of-sample runoff estimates for large catchments.

A related research project sought to produce gridded estimates of gross primary productivity and terrestrial ecosystem respiration [Jung *et al.*, 2020]. Using the FLUXNET dataset of eddy-covariance observations from flux-towers around the globe, Jung *et al.* [2020] were able to demonstrate that producing global estimates of ecosystem exchange parameters by incorporating their relationships with remote sensing data and meteoro-

logical products was possible. They tested a multitude of data-driven methods including tree methods (random forests, model tree ensembles), kernel methods (support vector machines, kernel ridge regression, gaussian processes), ANNs and regression splines [Tramontana *et al.*, 2016]. The modelling task is similar to the research undertaken by Jung *et al.* [2009], although the diversity of models was smaller in Jung *et al.* [2009] and they only tested tree-based models.

Finally, a similar methodology was applied for producing global estimates of soil moisture data (SoMo.ml) using the EA LSTM [Orth *et al.*, 2021]. Orth *et al.* [2021] created a target dataset by combining point-based measurements of soil moisture (International Soil Moisture Network, ISMN, and the National Center for Monitoring and Early Warning of Natural Disasters of Brazil, CEMADEN) with the gridded reanalysis soil moisture values (ERA5) from ECMWF (similar to the data used in Chapter 4 and Chapter 5). ERA5 data was also used for the meteorological forcing data (the regressors) used as input to the EA LSTM. They found that the SoMo.ml product produced more accurate soil moisture estimates than other gridded products for three soil depths, such as GLEAM [Martens *et al.*, 2017], ESA-CCI Soil Moisture [Gruber *et al.*, 2019] and the ERA5 products.

2.6 Conclusion

This chapter presented key concepts in hydrological modelling, outlining a data-driven approach to hydrological modelling and situating the use of deep learning methods within the wider context of modelling approaches. The LSTM and EA LSTM were introduced. The LSTM is a deep-learning architecture developed for modelling time-series and its variant, the Entity Aware LSTM, captures some prior knowledge of the nature of static input data. There are strong hydrological motivations for using the LSTM, which can be considered as comparable to commonly used conceptual models and statistical dynamical models. We have discussed the pitfalls of using LSTMs, and identified interpretability as a key component in the successful implementation of deep learning methods.

In this thesis I suggest that the LSTM is ideally suited to hydrological modelling and seek to fulfil the following research objectives:

- To determine the efficacy of the LSTM in different hydrological contexts (Chapter 3, 5).
- To identify what a trained LSTM has learned about hydrological systems (Chapter 3, 4, 5)

- To explore the potential of the LSTM in an operational context (Chapter 5).

The body of research presented in this thesis builds upon existing literature to test the validity and effectiveness of using the LSTM in a variety of hydrological contexts, seeking to understand what the LSTM has learned about the system under investigation and to demonstrate that the LSTM not only advances our predictive goals, but also our scientific goals.

3 Benchmarking LSTMs for rainfall-runoff modelling

Contributions This chapter is largely based on the following publication*

T Lees, M Buechel, B Anderson, L Slater, S Reece, G Coxon and SJ Dadson, 2021. *Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models*, **Hydrology and Earth System Sciences**, accepted. Preprint 10.5194/hess-25-5517-2021

Abstract. Long Short-Term Memory models (LSTMs) are recurrent neural networks from the field of Deep Learning (DL) which have shown promise for time-series modelling, especially in conditions when data are abundant. Previous studies have demonstrated the applicability of LSTM based models for rainfall-runoff modelling, however, LSTMs have not been tested on catchments in Great Britain (GB). Moreover, opportunities exist to use spatial and seasonal patterns in model performances to improve our understanding of hydrological processes, and to examine the advantages and disadvantages of LSTM-based models for hydrological simulation. By training two LSTM architectures across a large sample of 669 catchments in GB, we demonstrate that the LSTM and the Entity Aware LSTM (EA LSTM) simulate discharge with median NSE scores of 0.88 and 0.86 respectively. We find that the LSTM based models outperform a suite of benchmark conceptual models, suggesting an opportunity to use additional data to refine conceptual models. In summary, the LSTM based models show the largest performance improvements in the North East of Scotland and in South East England. The South East of England remained difficult to model however, in part due to the inability of the LSTMs configured in this study to learn groundwater processes, human abstractions and complex percolation properties from the hydro-meteorological variables typically employed for hydrological modelling.

*with the following author contributions. Conceptualisation: TL, LS, SD, SR. Data curation: GC, TL, MB, SD. Formal Analysis: TL. Methodology: TL, SR. Visualisation: TL, MB, BA, GC. Writing – original draft: TL. Writing – review & editing: TL, MB, BA, LS, SR, GC, SD.

3.1 Introduction

Rainfall-runoff models have evolved over many decades, reflecting a diversity of applications and purposes. These models range from physically based, spatially explicit models such as SHETRAN [Birkinshaw *et al.*, 2010], CLASSIC [Crooks *et al.*, 2014] and PARFLOW [Maxwell *et al.*, 2009] to lumped conceptual models, such as TOPMODEL [Beven & Kirkby, 1979] and VIC [Liang, 1994]. Additionally, data-driven models have also been used for modelling rainfall-runoff processes [Elshorbagy *et al.*, 2010; Gauch *et al.*, 2021c; Le *et al.*, 2019; Nourani *et al.*, 2014; Reichstein *et al.*, 2019; Wilby *et al.*, 2003]. The diversity of modelling approaches reflects the diversity of user objectives, uncertainty in terms of how to best represent the stores and fluxes of water and energy and the trade-offs in terms of data requirements, degree of realism and computational costs [Beven, 2011a].

Data-driven models range from simple regression models to large neural networks with thousands of parameters. These methods draw on empirical relationships between inputs and outputs to form a representation of how the hydrological system operates more generally [Beven, 2011a]. Other approaches from the class of data-driven models, such as statistical modelling and machine learning include genetic programming [Chadalawada *et al.*, 2020b; Herath *et al.*, 2020], random forests [Booker & Woods, 2014] and support vector regression models [Elshorbagy *et al.*, 2010]. Alternative empirical approaches also exist, including data-based mechanistic (DBM) modelling [Young, 1998, 2003]. DBM approaches suggest that rather than imposing model structures from the outset, hydrologists should in the first instance allow the data to suggest an appropriate model structure. Then, the modeller should see if there is a mechanistic interpretation of the learned model structure [Young & Beven, 1994]. Our modelling approach uses Deep Learning (DL) techniques, which have produced accurate predictions on a wide variety of tasks, including rainfall-runoff modelling [Huntingford *et al.*, 2019], and represent a fruitful area of further exploration for hydrologists and Earth scientists [Reichstein *et al.*, 2019]. For a more complete picture on the uses of DL techniques in hydrology, an interested reader is referred to Beven [2020]; Kratzert *et al.* [2018]; Nearing *et al.* [2020a]; Shen [2018].

DL methods have been used in hydrology and meteorology for decades [Daniell, 1991; Dawson & Wilby, 1998; Halff *et al.*, 1993; Peel & McMahon, 2020; Wilby *et al.*, 2003]. However, one architecture explicitly designed for time-series simulation, the Long Short-Term Memory network (LSTM) [Hochreiter, 1991; Hochreiter *et al.*, 2001], has recently demonstrated credible performance for modelling hydrological signatures across the Continental United States (CONUS) [Duan *et al.*, 2020; Fang *et al.*, 2018, 2020; Feng *et al.*,

2020a; Gauch *et al.*, 2021c; Kratzert *et al.*, 2018, 2019e]. More recent work has begun not only to explore the accuracy of forecasts, but also to use LSTMs to: (i) provide estimates of uncertainty [Klotz *et al.*, 2020]; (ii) explore the ability of the LSTM to integrate prior physical knowledge into DL model architectures [Hoedt *et al.*, 2021; Jiang *et al.*, 2020]; and (iii) use LSTMs to produce predictions at multiple timescales from a single model [Gauch *et al.*, 2021a].

By contrast with the physically-based, spatially-explicit hydrological models, lumped conceptual models have relatively few parameters and simulate the stores and fluxes of water on a catchment scale, for example, using a single store to represent the catchment-wide upper-soil water storage [Beven, 2011a]. Lumped conceptual models have lower data and computational requirements when compared to the spatially-explicit, physically-based, models, which is one reason they are often used for operational purposes [Clark *et al.*, 2008]. There exist many lumped conceptual models, differing in their internal structures, the equations that govern fluxes of water and energy, and the processes that are included [Knoben *et al.*, 2019]. As an evidence-guided discipline, performance benchmarks provide hydrologists with an objective means for selecting between different models, instead of model selection by lineage or affiliation [Addor & Melsen, 2019]. Furthermore, when applied over a large sample of catchments, differences in model performance can be instructive with regards to the hydrological conditions that are well simulated by one model compared with others [Gupta *et al.*, 2014]. Increasingly, "we need large-scale evaluations of model capability to identify which processes are important and which model structure(s) are most appropriate" [Lane *et al.*, 2019, p4012].

This paper seeks to address three research gaps. First, there exists no large-sample performance benchmark of LSTMs in a GB context. This is important because scientists and practitioners are interested in using LSTMs as hydrological models for hazard impact assessment, hazard early warning and rainfall-runoff modelling [Shen, 2018]. Therefore, a rigorous assessment of LSTM performances is necessary to determine whether such a model choice is appropriate in the GB context. Furthermore, given that the data archives are rich in GB, there exists a very good opportunity to learn more about the capabilities and limitations of LSTM-based methods [Clark & Khatami, 2021b]. Second, there exists only one other comparison of the EA LSTM performance against the LSTM [Kratzert *et al.*, 2019e]. Finally, there exist no studies that explore the relationship of performance differences (between conceptual and deep learning models) with the hydrological conditions in which those differences occur. The aim of studying the relationship between performance differences and hydrological conditions is to determine how best to improve our conceptual models. What information might be present

in the underlying data that can help identify processes that are currently missing from our conceptual models?

The research questions that this study seeks to address are determined by the research gaps identified above.

1. How well do LSTM-based models (including the EA LSTM) simulate discharge in Great Britain?

2. How do LSTM-based model performances compare with the conceptual models used as a benchmark?

3. Can we extract information from the spatial and temporal patterns in diagnostic measures? e.g. What is the relationship between LSTM performance and catchment attributes?

To address these questions, we have trained an ensemble of 8 LSTMs and 8 EA LSTMs on 669 catchments in Great Britain. We compare the results of the LSTM models with four deterministic lumped-conceptual models from a previous benchmarking study [Lane *et al.*, 2019]. This paper provides an evaluation of LSTM model ability across a large sample of GB catchments. We explore the association between catchment characteristics and the differences in model performances and present a data-driven benchmark that reflects the null-hypothesis of what information is present in a large sample dataset [Nearing *et al.*, 2020a]. Future modelling efforts may seek to assess whether hydrological theories encoded in conceptual and process-based models may contain more information than the benchmarks provided here [Nearing *et al.*, 2021b].

We believe that the research addresses the following needs of the hydrological community: (i) practitioners wishing to know whether the LSTM is a justifiable model choice in the GB context, (ii) scientists and practitioners interested in understanding under what hydrological conditions (e.g. catchment attributes) the LSTM performance differs from conceptual models, and (iii) as a reference for future GB-wide modelling studies.

3.2 Methods

3.2.1 Data - CAMELS GB

All data employed in this analysis originate from the CAMELS-GB data [Coxon *et al.*, 2020a]. CAMELS-GB is a recently-released, large-sample, long-term, daily data set that offers the potential for GB-wide modelling studies. CAMELS-GB collates hydrologically

relevant data for 671 GB catchments between the years of 1970 and 2015. The data set includes daily time series for meteorology (dynamic data - $\mathbf{X_{t,n}}$); and discharge (target data - $\mathbf{y_{t,n}}$). Also included are catchment attributes (static data, $\mathbf{A_{t,n}}$) such as topography, climate, hydrologic signatures, soil and land cover, hydrogeology, and human influence. These features are, in reality, not static over time. However, for the purposes of this study we treat these features as time-invariant. Further information on the variables we used as input to our model can be found in Table 3.1. The reader is directed to Coxon *et al.* [2020b] for details of the source of the data, how the data were processed and a discussion of data limitations.

The data set contains novel inputs compared with previous CAMELS (US, Chile, Brazil) data sets [Addor *et al.*, 2017; Alvarez-Garreton *et al.*, 2018; Chagas *et al.*, 2020], such as human attributes, calculated potential evapotranspiration (pet) and uncertainty estimates. We do not use all of these features here. The static attributes we use to train the LSTM models are listed in Table 3.1. These static attributes were chosen to reproduce the experimental framework of Kratzert *et al.* [2019e], however, the differences reflect the fact that the CAMELS-US and CAMELS-GB have slightly different attributes. These include both catchment properties and climate properties, describing the conditions relevant for rainfall-runoff modelling in different catchments.

Table 3.1 | Catchment attributes from the CAMELS-GB data set [Coxon *et al.*, 2020b] used to train the LSTM based models, the static features included in *A*.

Static Variables	Static Variable Description	Median	Range
area	catchment area (km^2)	152	[2, 9931]
elev_mean	mean elevation (m a.s.l)	163	[25, 682]
dpsbar	slope of the catchment mean drainage path (mkm^{-1})	79	[12, 488]
sand_perc	percent sand (%)	43	[19, 82]
silt_perc	percent silt (%)	30	[9, 43]
clay_perc	percent clay (%)	24	[7, 51]
porosity_hypres	soil porosity calculated using the hypres pedotransfer function (-)	47	[34, 81]
conductivity_hypres	hydraulic conductivity calculated using the hypres pedotransfer function (cmh^{-1})	1	[0.5, 3]
soil_depth_pelletier	depth to bedrock (m)	1	[0.5, 42]
frac_snow	fraction of precipitation falling as snow (for days colder than 0)	0.02	[0.00, 0.17]
dwood_perc	percent of catchment that is deciduous woodland (%)	6	[0, 37]
ewood_perc	percent of catchment that is evergreen woodland (%)	2	[0, 93]
crop_perc	percent of catchment that is cropland (%)	13	[0.00, 91]
urban_perc	percent of catchment that is urban area (%)	3	[0.00, 81]
reservoir_cap	catchment reservoir capacity (ML)	0	[0, 8 x 10^7]
p_mean	mean daily precipitation ($mm\,day^{-1}$)	2.57	[1.54, 9.61]
pet_mean	mean daily PET ($mm\,day^{-1}$)	1.38	[1.03, 1.51]
p_seasonality	seasonality and timing of precipitation (estimated using sine curves)	-0.14	[-0.42, 0.14]
high_prec_freq	frequency of high-precipitation days ($\geq$ 5x mean daily precipitation)	15.69	[7.58, 20.73]
low_prec_freq	frequency of dry days (< 1 $mm\,day^{-1}$)	214.23	[1.63, 259.23]
high_prec_dur	average duration of high-precipitation events ($\geq$ 5x mean daily precipitation)	1.14	[1.05, 1.25]
low_prec_dur	average duration of dry periods (number of consecutive days < 1 $mm\,day^{-1}$)	3.70	[2.64, 4.67]

3.2.2 An Overview of the LSTM and EALSTM

In this paper, we test two neural network architectures used in other hydrological studies [Kratzert *et al.*, 2019e; Shen *et al.*, 2018]. The first is the LSTM, which has been used in a variety of time-series modelling applications. The second model is the EA LSTM, which conditions the discharge response to meteorological forcings on time-invariant properties of river catchments, such as soil and topographic attributes, treating these time-invariant properties separately. For a summary of notation used throughout the paper please refer to Table 3.2:

Table 3.2 | Table describing the notation used throughout the paper.

Symbol	Description	Notes
$\mathbf{y_{t,n}}$	Our target variable, specific discharge at time t, catchment n	mm day^{-1}
$\hat{\mathbf{y}}_{\mathbf{t,n}}$	Simulated specific discharge at time t, catchment n, predicted by the model $\mathrm{M}_{\boldsymbol{\theta}}$	mm day^{-1}
n	Gauge ID	-
$\mathrm{p}_{t,n}$	Precipitation	mm day^{-1}
$\mathrm{pet}_{t,n}$	Potential evapotranspiration	mm day^{-1}
$\mathrm{T}_{t,n}$	Temperature	°C
$\mathbf{A_n}$	Catchment attributes (static data)	
$\mathbf{X_{t,n}}$	Hydro-meteorological data (dynamic data)	$\left[\mathrm{p}_{t,n}, \mathrm{pet}_{t,n}, \mathrm{t}_{t,n}\right]$
hs	Hidden size	$\mathrm{hs} = 64$
$\mathbf{W}_{\text{layer}}$	The matrix of learnable weights	-
b_{layer}	The vector of learnable biases	-
$\boldsymbol{\theta}$	Learned model parameters, representing all $\mathbf{W}_{\text{layer}}$ and $\mathbf{b}_{\text{layer}}$	-
$\mathrm{M}_{\boldsymbol{\theta}}$	The model (LSTM or EA LSTM) with parameters $\boldsymbol{\theta}$	-
$\mathbf{C_t}$	The cell state of the LSTM models.	$\mathbb{R}^{\mathrm{hs}}$
$\tilde{\mathbf{C}}_\mathbf{t}$	The candidate cell state values	$\tilde{\mathbf{C}}_\mathbf{t} \in \mathbb{R} \mid -1 < x < 1\}$
$\mathbf{h_t}$	The hidden state of the LSTM models.	$\mathbb{R}^{\mathrm{hs}}$
$\mathbf{f_t}$	The forget gate of the LSTM models	$\{\mathbf{f_t} \in \mathbb{R} \mid 0 < x < 1\}$
$\mathbf{i_t}$	The input gate of the LSTM models	$\{\mathbf{i_t} \in \mathbb{R} \mid 0 < x < 1\}$
$\mathbf{o_t}$	The output gate of the LSTM models	$\{\mathbf{o_t} \in \mathbb{R} \mid 0 < x < 1\}$
ℓ	The Loss Function used to train the model (Nash Sutcliffe Efficiency)	-

What follows is a brief introduction to the LSTM model architectures. For a more complete description of these models please refer to the Literature Review Section 2.4.1 and 2.4.2 and Kratzert *et al.* [2018, 2019e].

The LSTM has a strong inductive bias towards retaining information over long sequences [Bengio *et al.*, 1994; Hochreiter, 1991]. This means that the LSTM architecture is designed to retain information that is important over both long and short term time horizons. LSTMs do this by maintaining two state vectors, a cell memory vector that captures slowly evolving processes ($\mathbf{C_t}$) and a more quickly evolving state vector, colloquially named the "hidden" vector ($\mathbf{h_t}$). Information flow is controlled by a series of 'gates' which are neural network layers that determine what information is removed from $\mathbf{f_t}$ (forget gate), what information is stored in $\mathbf{i_t}$ (input gate) and what information

is passed to the $\mathbf{o_t}$ (output gate) respectively.

The Entity Aware Long Short-Term Memory (EA LSTM) modifies the forget gate, so that the output of the forget gate is a function of only the static catchment attributes ($\mathbf{A_n}$) rather than both the catchment attributes and the dynamic data ($[\mathbf{X_{t,n}}, \mathbf{A_n}]$). The EA LSTM was developed specifically for rainfall-runoff modelling [Kratzert *et al.*, 2019e]. For the sake of clarity, it is important to note that both models receive the same information. The LSTM still receives the static catchment attributes. However, rather than affecting only the input gate, the static data can influence all gates, since they are appended to a vector of dynamic inputs ($[\mathbf{X_{t,n}}, \mathbf{A_n}]$) and so the same information is given to the LSTM at each timestep. The static attributes are used by the LSTM in the same way as the dynamic data. This offers extra flexibility for the LSTM compared with the EA LSTM, since the LSTM is able to modify the input gate based on information from time-varying data, whereas the EA LSTM is not. We are using the static nature of the data as a constraint on the EA LSTM to reflect the nature of the input data (separated into static and dynamic inputs).

3.2.3 Model Training

We used the "neuralhydrology" codebase, written in Python 3.6 [Van Rossum *et al.*, 2007], to train and evaluate the models, found here: `github.com/neuralhydrology/neuralhydrology/`. The configuration files used to run the models can be found using the links at the end of this article. The predictions and error metrics for the fitted models can be found online at Zenodo (zenodo.org/record/4555820, last accessed: 19 July 2021).

The goal of rainfall-runoff modelling is to predict time-varying specific discharge, $\mathbf{y_n} = (\mathbf{y_{1,n}}, ..., \mathbf{y_{T,n}}) \in \mathbb{R}^T$ (mm day^{-1}) for time $t = \{1, ..., T\}$ at measuring gauge n of N, given hydro-meteorological forcing data, $\mathbf{X_n} = (\mathbf{X_{1,n}}, ..., \mathbf{X_{T,n}})$, and catchment attributes ($\mathbf{A_n}$ - Table 3.1) within the catchment area upstream of the gauge. In the present case for GB, $N = 669$. Although the underlying CAMELS-GB data has 671 station gauges, we trained on data from only 669 stations because two basins have missing data in the static attributes; stations 18011 and 26006 have missing mean elevation (elev_mean) and mean drainage path slope (dpsbar).

Our task is to train a regional hydrological model, i.e. one model for all catchments in the dataset. This means that we learn a single set of parameters, $\boldsymbol{\theta}$, of a model, M_θ, that minimizes the loss function, $\ell(\hat{\mathbf{y}}_{\mathbf{t,n}}, \mathbf{y_n})$, for all flow gauges, and thus accurately simulates discharge ($\hat{\mathbf{y}}_{\mathbf{t,n}}$) for all of the basins in our subset of CAMELS-GB:

$$\hat{\mathbf{y}}_{\mathbf{t},\mathbf{n}} = \mathrm{M}_{\theta}\left(\left[\mathbf{A}_{\mathbf{n}}, \mathbf{X}_{\mathbf{t}-\mathbf{k}+\mathbf{1},\mathbf{n}}, \ldots, \mathbf{X}_{\mathbf{t},\mathbf{n}}\right]; \boldsymbol{\theta}\right) \tag{3.1}$$

We train our model using the Nash Sutcliffe Efficiency (NSE) loss as our objective function (ℓ), as described in Kratzert *et al.* [2019e]. Other objective functions could be used, however, we use the same objective function as the conceptual models we compare against, in order to control the possible sources of performance differences. The NSE describes the squared error loss normalized by the total variance of the observations. In order to account for the fact that some basins will have lower variance than others, we follow Kratzert *et al.* [2019e] to normalize by basin-specific variance. This prevents the loss from being overly weighted towards high-variance catchments.

For this study we trained the models on the days from 1 January 1988 to 31 December 1997 and tested on a hold-out sample using the days from 1 January 1998 to 31 December 2008 (4018 days of test data). We withheld the years 1975 to 1980 from the training process to check the performance of the model during training (our validation set). This means that we have separate time periods for calibration (1988–1997; train period), and evaluation (1998–2008; hold-out test period). These train and test periods were chosen to facilitate the comparison with the study whose published results for four lumped hydrological models we use as a benchmark [Lane *et al.*, 2019]. For further analysis of the train and test periods please see Appendix Section A.1.

Our input data were taken from CAMELS-GB, described above [Coxon *et al.*, 2020b]. We used precipitation, potential evapotranspiration and temperature as dynamic inputs ($\mathbf{X}_{\mathbf{t},\mathbf{n}} = \left[\mathrm{p}_{t,n}, \mathrm{pet}_{t,n}, \mathrm{T}_{t,n}\right]$). We selected 21 individual features describing each catchment's topographic, soil, land-cover, and climatic properties as static inputs ($\mathbf{A}_{\mathbf{n}}$). These attributes were chosen to reflect hydrological information that the model can use to distinguish between catchment rainfall-runoff behaviours [Kratzert *et al.*, 2019e]. These catchment attributes are described in Table 3.1. For both LSTM models we pass the final hidden output through a fully connected (linear) layer. This final layer maps our hidden state vector to a scalar prediction ($\hat{\mathbf{y}}_{\mathbf{t},\mathbf{n}} \in \mathbb{R}$) for discharge at that gauge on that day. We give the models one year of daily dynamic data (365 input timesteps, $X_n = [\mathbf{X}_{\mathbf{t}-\mathbf{365},\mathbf{n}}, \ldots, \mathbf{X}_{\mathbf{t},\mathbf{n}}]$) to predict the final timestep of specific discharge ($\hat{\mathbf{y}}_{\mathbf{t},\mathbf{n}}$).

All national results shown below are calculated for the 518 gauges that are found in both the CAMELS GB data and the benchmark data. We then evaluate model performance on all of these basins for our test (evaluation) period (1998–2008). For each model (LSTM, EA LSTM) we also calculate the average of an ensemble of eight individually-

trained models with different random seeds. This strategy accounts for the random initialisation of the network and the stochastic nature of the optimisation algorithm. We used a hidden size (*hs*) of 64 and a final fully connected layer with a dropout rate of 0.4, which aims to avoid overfitting. Dropout works by randomly forcing certain weights in the network to zero ("dropping them out"), forcing the remaining weights to model the discharge without that extra information. This has been found to prevent weights 'fixing' the erroneous outputs of other weights, preventing co-adaptation of weights and, ultimately, encouraging the model to use a simpler and more robust representation of rainfall-runoff processes [Srivastava *et al.*, 2014a]. The hidden size determines the total number of parameters in the model. For the LSTM there are 23,361 trainable parameters, whereas the EA LSTM has 14,593 trainable parameters. These are trained on data from 669 catchments over 4018 timesteps (2,688,042 samples). Note that this is for a regional model and is independent of the number of catchments. Given that we train the LSTM on 669 catchments, we can interpret the LSTM as equivalent to using 35 parameters per catchment, with a median catchment area of 152 km^2. The EA LSTM has on the order of 22 parameters per catchment. We chose the hyper-parameters (dropout rate, hidden size - *hs*) based on analysis of the NSE performances, finding that the improvement of further model complexity (increased hidden size) was negligible after a hidden size of 64. The hidden size was also consistent with the choices made in previous studies [Kratzert *et al.*, 2019e]. We used the Adam optimisation algorithm [Kingma & Ba, 2014a] and stopped training after 30 epochs, after which there was no further improvement to the model. An epoch reflects a single pass of the training dataset through the model, such that every sample in the training dataset has been used to update the model weights. This reflects the fact that during the training of DL models, the data are often split into batches to allow large datasets to be read into memory. The LSTM ensemble took 10 hours to train. The EA LSTM ensemble took 96 hours to train. All models were trained on a machine with 188GB of RAM and a single NVIDIA V100 GPU.

3.2.4 Model Performance Comparisons

The LSTMs learn to represent hydrological processes directly from data. When the LSTMs perform well on hold-out test samples, a necessary (but not sufficient) conclusion is that the data contains useful information about the hydrological processes that translate inputs (precipitation) into outputs (discharge). The differences in model performance between the LSTMs and the benchmark hydrological models can be used to determine hydrological processes that are described by the input data, but not captured

or under-represented by the benchmark hydrological models.

Benchmark Models

In order to provide a reference for model performance statistics, we compare the performance of the LSTM based models against four lumped conceptual models from the FUSE framework [Clark *et al.*, 2008]. To be unbiased on the model calibration, we used simulated discharge time series from Lane *et al.* [2019] who calibrated and evaluated these four conceptual models on 1000 catchments across Great Britain. The four conceptual models used are: TOPMODEL [Beven & Kirkby, 1979], Variable Infiltration Capacity (VIC) [Liang, 1994], Precipitation-Runoff Modelling System (PRMS) [Leavesley *et al.*, 1983] and SACRAMENTO [Burnash *et al.*, 1973a]. These conceptual models are often used in operational settings, due to the relative ease of use and lower data requirements when compared with physically-based models. These conceptual models all explicitly maintain mass balance, and so assume no losses or gains of water other than flow from the catchment outlet or evaporation.

These conceptual models are all lumped models run at a daily time step. Each model is explicitly forced to close the water balance, limited by an upper limit of potential evapotranspiration for water losses. Every one of the conceptual models has a gamma distribution routing function. Furthermore, the four conceptual models do not include a snow routine nor a vegetation module [Clark *et al.*, 2008]. Sacramento has 5 stores and 12 parameters per catchment; both VIC and TOPMODEL have 2 stores, and 10 parameters; PRMS has 3 stores and 11 parameters. A more complete description of these benchmark models and the processes that they include can be found in Table 3 of Lane *et al.* [2019] and in Section 4 of [Clark *et al.*, 2008].

The benchmark study provides an assessment of conceptual model simulation performances across a large sample of GB catchments, and also quantifies uncertainty in hydrological simulations due to parameter uncertainty and model structural uncertainty [Lane *et al.*, 2019]. Parameter values for each conceptual model were selected from 10,000 simulations of multi-dimensional parameter space. The best-estimate model parameter values were selected from these 10,000 samples using the Nash-Sutcliffe Efficiency score. These best-fit parameters are used as a benchmark against which to compare the LSTM performance. To place the intercomparison in context, we critically reflect on the consistencies and differences between the different model configurations here.

First, the selection of model parameters differs between the LSTM and the concep-

tual models. The experimental design of the benchmarking study produced 10,000 samples of parameter values and Lane *et al.* [2019] provide the simulations given the best fitting parameters for future studies to employ as a benchmark. The LSTM parameters are optimised using stochastic gradient descent, choosing the best fitting set of parameters using the NSE score. While the method of choosing parameters differs, the objective function that determines the "best-fit" parameter values are the same for both the LSTMs and the conceptual models. Second, the calibration and evaluation data are the same. The calibration and evaluation of these models was performed using the same data from CAMELS-GB, i.e. the National River Flow Archive data [Centre for Ecology and Hydrology, 2016] for specific discharge (y_t), the Centre for Ecology and Hydrology Gridded Estimates of Areal Rainfall, CEH-GEAR, for precipitation [Tanguy *et al.*, 2014] and the Climate Hydrology and Ecology research Support System Potential Evapotranspiration (CHESS-PE) data set for PET [Robinson *et al.*, 2017]. The benchmark experiment selected the best-fitting parameter values using data from the period 1988–2008, and then evaluated their performance on data from 1993–2008 [Lane *et al.*, 2019]. Instead, we calibrate the LSTMs on data from 1988–1998 and then evaluate the LSTM performances for our hold-out evaluation period of 1998–2008. We recalculate the performance statistics of the benchmark conceptual models for this evaluation period, 1998–2008, using the published simulated time-series. Therefore, the LSTM is evaluated on out-of-sample (in time) data, whereas, the conceptual model parameters were calibrated on data included in the evaluation period (in-sample evaluation). Finally, it is worth noting that Lane *et al.* [2019] focused not only on model performances but also on parameter uncertainty. Uncertainty is an essential component of any modelling study, and our approach of training an ensemble of 8 models is one proposed method for dealing with uncertainty in LSTMs. For an analysis of model uncertainty with this method see Appendix Section A.3. For a more complete treatment and discussion of the different approaches for dealing with uncertainty using LSTMs see Klotz *et al.* [2020].

As with any benchmarking study, there are important caveats to the intercomparison of model results. Ultimately, the purpose of the comparison is: (i) to provide a reference for the diagnostic measures of LSTM performance, (ii) to identify the hydrological conditions where simulations differ, and (iii) use these insights to diagnose missing representations in the conceptual models. We agree with Lane *et al.* [2019] that "*benchmark [studies] provides a useful baseline for assessing more complex modelling strategies*" (p.4029), and we follow them in publishing the simulations and results of the LSTM models for future studies to use for comparison.

Evaluation Metrics

Each model produces a daily simulated discharge value at each station. Three example hydrographs are shown in Appendix Section A.2. The evaluation metrics described below evaluate the overall performance of each model to reproduce a specific aspect of the observed hydrograph. For the LSTM-based models the evaluation metrics are calculated given the average discharge of the ensemble. Since no single evaluation metric can fully capture the performance of streamflow simulations across all flow-regimes [Gupta *et al.*, 1998], we use a number of metrics to address the performance of models across the flow regime, outlined below.

We evaluate the goodness-of-fit of the LSTM based models and the conceptual models using six evaluation metrics. The Nash-Sutcliffe Efficiency (NSE) [Nash & Sutcliffe, 1970] score has been used in numerous studies and there is extensive literature discussing its strengths and weaknesses [Gupta *et al.*, 2009]. The NSE can be decomposed into three components, a correlation term, a bias term (BiasError) and a variability (SDError) term [Gupta *et al.*, 2009]. The bias term measures the error in predicting the mean flow. The variability term measures the error in predicting the standard deviation of discharge. We report results for the NSE and each of its three components.

To understand how well the LSTMs represent low, mean and high flows, we also consider the biases for different components of the flow duration curve. The low flow bias (%BiasFLV) is the diagnostic signature measure for long term base flow [Yilmaz *et al.*, 2008], and low flows are defined as those which are exceeded 70% of the time. For the middle of the flow duration curve we use the bias of the mid section of the flow duration curve, between the 20th and 70th percentiles (%BiasFMS). Finally, we also look at the bias of the high flows, considering the top 2% of flows (%BiasFHV).

3.3 Results

3.3.1 National Scale Model Performance

The LSTM and EA LSTM models produce accurate simulations across Great Britain when evaluated using a variety of metrics, with differing levels of performance improvement over the benchmark conceptual models (See Table 3.3).

Comparing the median NSE for all catchments, the LSTM ensemble (0.88) outperforms all other models, including the EA-LSTM ensemble (0.86). The slightly lower median NSE for the EA-LSTM models is consistent with results from previous studies [Kratzert

Table 3.3 | Summary of all goodness-of-fit metrics used to benchmark performance against the conceptual models for the validation period 1998–2008 on the 518 stations found in both CAMELS-GB data [Coxon *et al.*, 2020a] and the FUSE conceptual models [Lane *et al.*, 2019]. We have shown the median catchment score for the metric given the mean simulated discharge of our ensemble. Values that are not significantly different from the best model are highlighted in bold (α = 0.001).

	NSE	BiasError	SDError	Correlation	%BiasFMS	%BiasFLV	%BiasFHV
TOPMODEL	0.76	-0.04	-0.10	0.88	**5.70**	42.22	-13.04
ARNOVIC	0.78	0.06	-0.10	0.90	**2.25**	-60.34	-14.66
PRMS	0.77	0.03	**-0.03**	0.89	35.24	-315.25	-15.11
SACRAMENTO	0.80	**-0.01**	-0.07	0.90	27.91	-195.92	-16.19
EALSTM	0.86	**-0.02**	-0.10	0.94	-6.29	**23.61**	-10.81
LSTM	**0.88**	**-0.02**	-0.09	**0.94**	-3.67	**26.34**	**-9.09**

et al., 2019e]. The CDFs (cumulative distribution functions) of the NSE (Fig. 3.1a) show the entire distribution of LSTM scores is shifted towards better performances. The LSTM NSE scores are significantly different from all comparison models at α = 0.001 (Paired Wilcoxon signed-rank-test). We see the same pattern for the EA-LSTM models. The performance improvement at the tails is particularly pronounced. Neither the LSTM nor the EA-LSTM model have any station gauges with an NSE of less than zero.

As discussed in the methods, we can decompose the NSE into three components, bias (BiasError), correlation and error in predicting the variability of flows (SDError). The pattern of correlation scores closely follows the pattern of NSE, with the entire distribution of catchment correlation scores shifted towards improved performance. The CDFs in Fig. 3.1c show that the LSTM catchment bias scores are closer to zero than the benchmark models, which reflects the fact that the conceptual models are explicitly mass-conserving, whereas the LSTM models are not. The median variability error is negative (Fig. 3.1d), showing that the LSTMs tend towards underpredicting the variability of flows.

The LSTM shows a large performance improvement for low-flow bias score (%BiasFLV - Fig. 3.1e). The LSTM has lower median bias in the slope of the midsection of the flow duration curve (%BiasFMS) than all models except ARNOVIC. When we consider the CDFs, both LSTMs have shorter tails than the conceptual models, showing that a greater proportion of catchments have biases closer to zero. The high-flow biases (%BiasFHV) are relatively similar for all models, as shown by Fig. 3.1g).

The biases at different flow exceedances suggest that the conceptual models produce good simulations for the high flows, but are less able to simulate low flows. The LSTM shows a smaller performance decline at the low flows than the benchmark models and a competitive performance at high flows, suggesting that the LSTMs are robust to extreme conditions. We also note that the negative bias, for the midsection and the

upper-section of the flow duration curve, demonstrates that the LSTM model is conservative in its flow predictions, particularly in comparison to the other models.

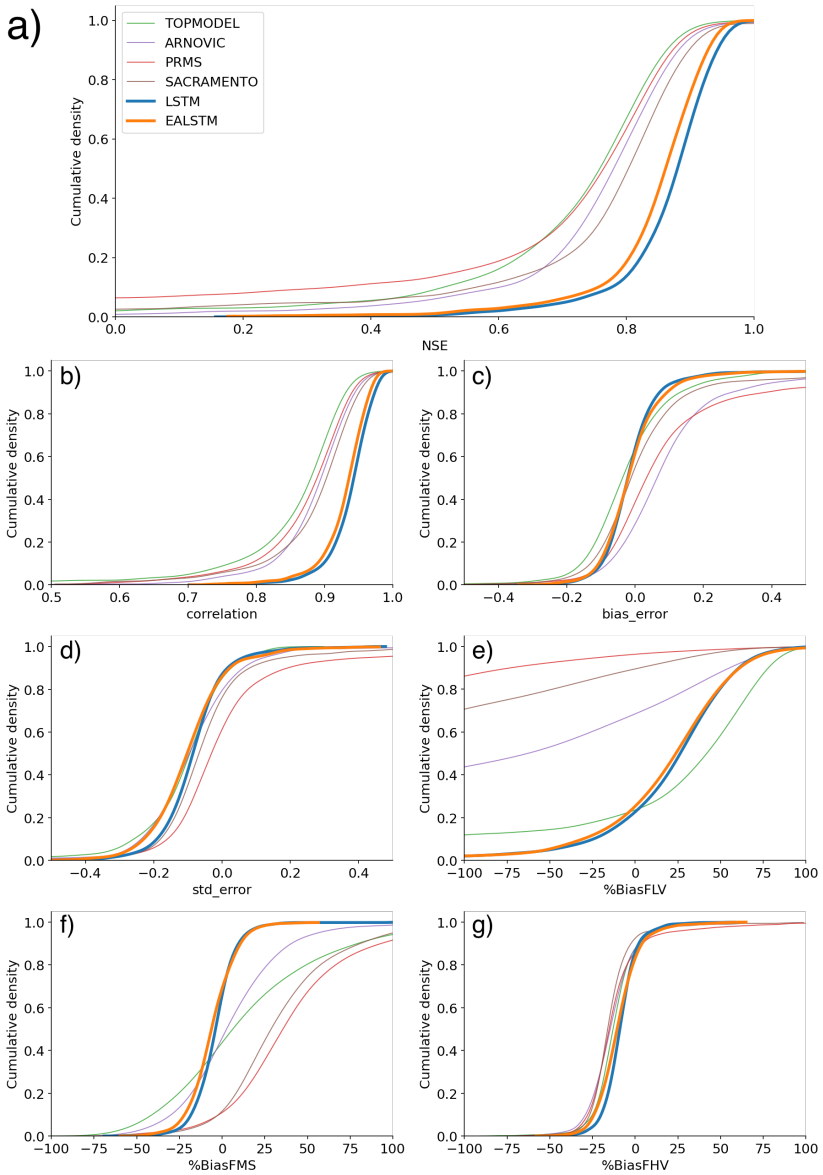


Figure 3.1 | Cumulative Distribution Functions (CDFs) of station goodness-of-fit metrics scores for each model. EALSTM (orange) and the LSTM (blue), and the conceptual models: TOPMODEL (green), VIC (red), PRMS (purple), Sacramento (brown) [Lane *et al.*, 2019]. Panels indicate distribution of station: a) NSE scores b) correlation scores c) bias error scores d) variability error scores e) low-flow bias scores f) mid-range of flow bias scores g) high-flow bias scores

3.3.2 Spatial Patterns of Performance

The LSTM demonstrates competitive simulation of discharge across Great Britain (see the spatial patterns of various performance metrics in Appendix Fig. 5). The EA LSTM has very similar spatial patterns to the LSTM, but shows a consistently worse performance than the LSTM across GB.

The benchmark conceptual models struggled when simulating discharge in catchments on the permeable bedrock in the South East of England and the mountainous catchments in the North East of Scotland [Lane *et al.*, 2019]. Performance metrics in the South East were lower due to poor simulation of variance and correlation, and in North-Eastern Scotland due to poor simulation of the mean flow conditions [Lane *et al.*, 2019]. We suggest that these differences in performance are due to the low rainfall and chalk aquifer in the South East of England, and to the lack of snow modules incorporated into the conceptual models for North East Scotland.

Interestingly, the LSTM simulates discharge less well in the South East of England relative to LSTM performance elsewhere in GB, particularly in the summer months (Fig. 3.2). Performances for all seasons are worse in the South East of England. This pattern is stronger in the summer months (JJA). The East-West gradient in model performances can be seen for all models, particularly in JJA. However, the range of errors is smaller for the LSTM based models when compared with the conceptual models.

The LSTM shows an underestimate of the variability and a cluster of high bias scores in the South East (Appendix Fig. 5). The LSTM both overestimated and underestimated mean flows in catchments in the South East region, explaining the relative under-performance in the composite metric (NSE) for the LSTM relative to the rest of GB.

Spatial patterns in the biases for different sections of the flow duration curve show that only the LSTMs demonstrate a consistent underprediction of the midsection slope of the flow duration curve (%BiasFMS). A steep slope in the midsection of the flow duration curve reflects a watershed having a “flashy” response [Yilmaz *et al.*, 2008], potentially due to low soil moisture capacity. Therefore, an underprediction of the midsection reflects an underestimation of the “flashiness” of the catchment. The LSTM %BiasFMS is largest for the South East of England. The LSTM shows improved performance compared to the benchmark models across GB, including these under performing regions, the South East of England, and North East Scotland.

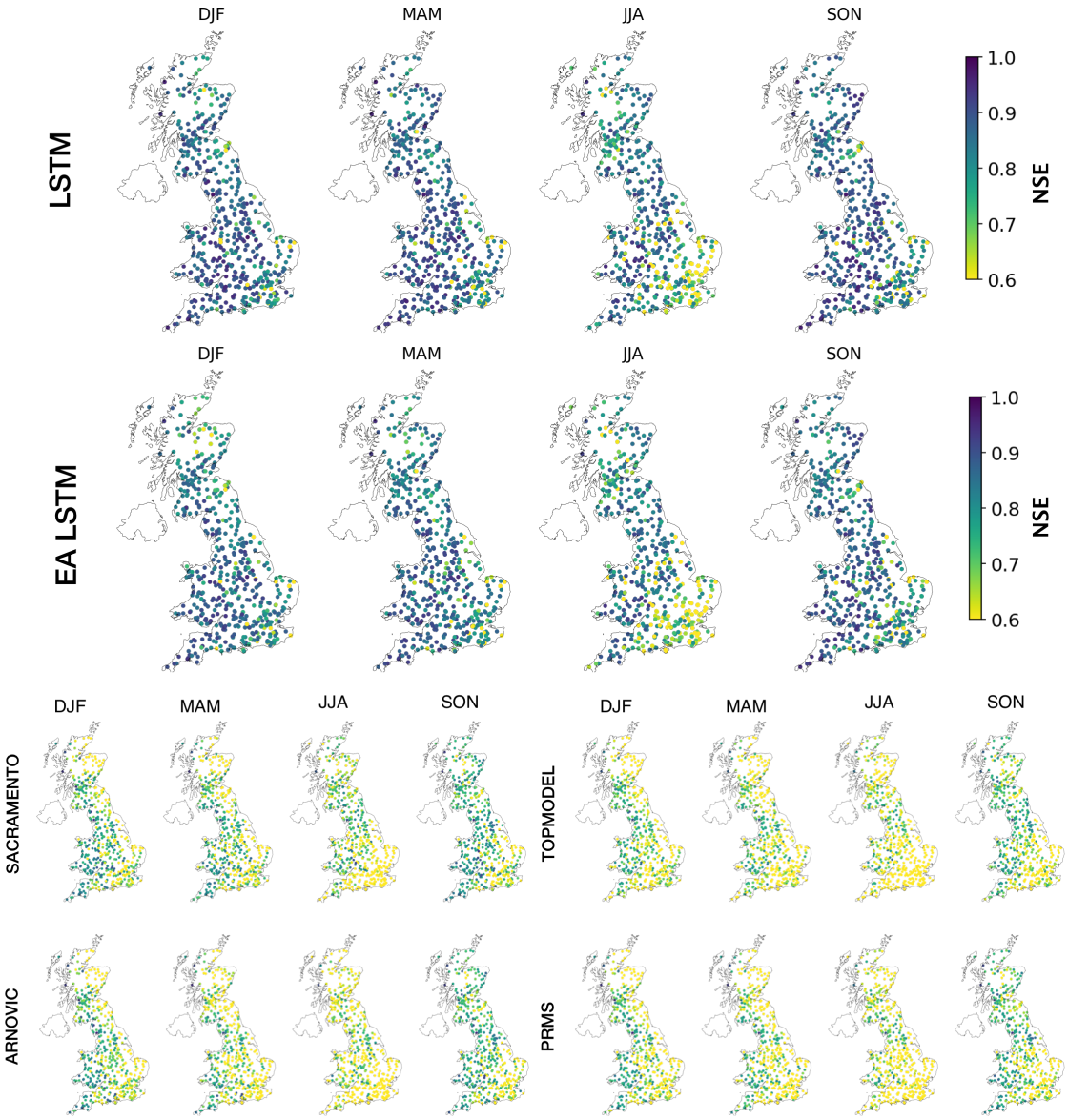


Figure 3.2 | Seasonal NSE patterns for the two LSTM based models (above) and the conceptual models (below). Each station in the evaluation data is shown as a point. The colour of the point reflects the NSE score. Brighter colors reflect lower NSE values, currently capped at a minimum of 0.7 NSE.

3.3.3 In what hydrological conditions do model performances differ?

Large sample studies allow us to detect catchment attributes that our models are (not) able to represent. In order to determine what the LSTM is capable of representing well we perform two analyses. Firstly, directly calculating the difference in NSE scores. Secondly, we correlate catchment attributes with model diagnostic scores.

The Δ_{mean}NSE is the mean difference between a reference model (LSTM) and the comparison model. The Δ_{median}NSE is the median difference. The mean differences between the LSTM station NSE and the other models is smallest for the EA-LSTM (Δ_{mean}NSE = 0.02). This is unsurprising given the very similar architectures of the two models. The differences for the conceptual models range from TOPMODEL (Δ_{mean}NSE = 0.15); ARNOVIC (Δ_{mean}NSE = 0.17); SACRAMENTO (Δ_{mean}NSE = 0.20), and PRMS (Δ_{mean}NSE = 0.43). While the mean performances show large differences, due to the presence of poorly performing stations, the median differences are smaller SACRAMENTO (Δ_{median}NSE = 0.07); ARNOVIC (Δ_{median}NSE = 0.09); PRMS (Δ_{median}NSE = 0.10) and TOPMODEL (Δ_{median}NSE = 0.10). Both summaries (median, mean) demonstrate that the LSTM offers a single model architecture that offers more accurate simulations than traditional hydrological models in a variety of hydrological conditions.

Spatially, the benchmark conceptual models struggled to produce good simulations in two geographical regions. These were in the South East of England and North East of Scotland. The performance improvement (ΔNSE) of the LSTM over the conceptual model was indeed largest in the South East of England and North East Scotland (see Fig. 3.3).

North East Scotland is one of the most mountainous regions of GB. The Cairngorm National Park and the North Pennines are the only areas of GB where snow processes are consistently important, owing to catchments having a higher elevation. There are 36 catchments in the CAMELS GB dataset with fraction of snow cover greater than 5%, and three are in the North Pennines, the other 33 are in the Cairngorm National Park. The results in Fig. 3.3 show that the LSTM exhibits a large performance improvement in these catchments, since ΔNSE is high. This is most likely due to the cell state being able to represent longer-term stores and fluxes of water, therefore capturing the melting snow processes. The conceptual models lack a snow module, and are therefore unable to capture snow melt or frozen ground processes, which are especially important in winter (DJF) and spring (MAM) [Lane *et al.*, 2019]. By contrast, what the LSTM performance shows is that data-driven models are able to flexibly incorporate snow processes in the catchments where they are required (NE Scotland) even when trained to produce one

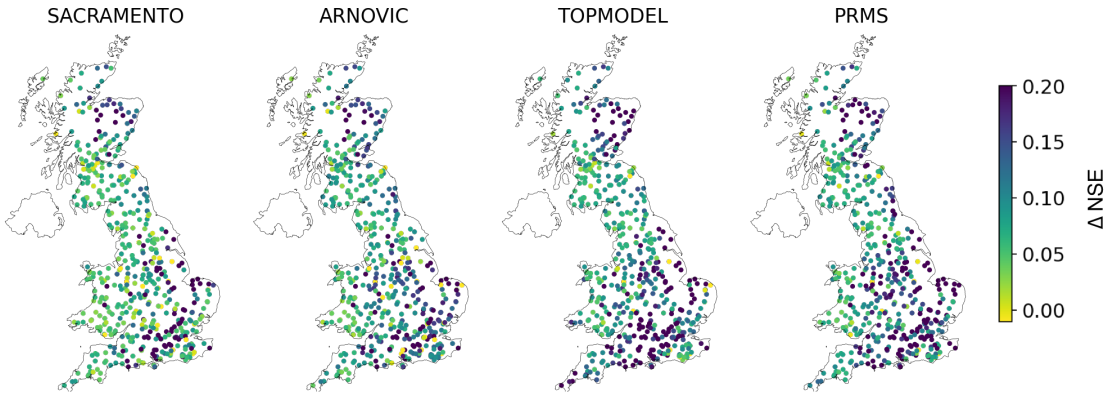


Figure 3.3 | The performance improvement of the LSTM relative to the four conceptual models, SACRAMENTO, ARNOVIC, TOPMODEL and PRMS. The difference in NSE is calculated by subtracting the conceptual model NSE from the LSTM NSE ($\Delta\mathrm{NSE} = NSE_{LSTM} - NSE_{conceptual}$). Each point represents a station and the colour reflects the performance improvement (measured by NSE) of the LSTM compared with the conceptual models. Positive values reflect stations where the LSTM outperforms the conceptual models.

set of weights. This flexibility is an important asset of data-driven approaches, since these hydrological processes do not need to be specified prior to model training, but can be learned from the available data.

The South East is a relatively dry area, with large chalk aquifers contributing to a high baseflow index and large urban and agricultural areas, contributing to a large anthropogenic signal in the hydrographs. Although the improvement in simulation accuracy compared to the conceptual models is large in the South East, the pattern of raw LSTM NSE shows that the LSTM still underperforms in the South East relative to elsewhere in GB. The seasonal patterns showed that the LSTMs performed worse in summer months, which is the drier period of the year. Consistent with this spatial pattern, the ratio of mean potential evapotranspiration to mean precipitation attribute (labelled "*aridity*" in the CAMELS GB dataset [Coxon *et al.*, 2020b]) is negatively correlated with model performance for all models (Fig. 3.4), although the magnitude of this association is smaller for the LSTM based models than the conceptual models.

We observe consistently poorer performance across all models, including the LSTMs, in drier hydrological conditions. This can be seen by the negative correlations between catchment P/PET (aridity) and model NSE scores (Figure 3.4).

The LSTM-based models show no significant correlation between baseflow index and model performances, in contrast with the other models. ARNOVIC also shows no significant correlation. ARNOVICs improved performance can be attributed to the non-linear relationship in the upper-storage, which means that the model will only produce

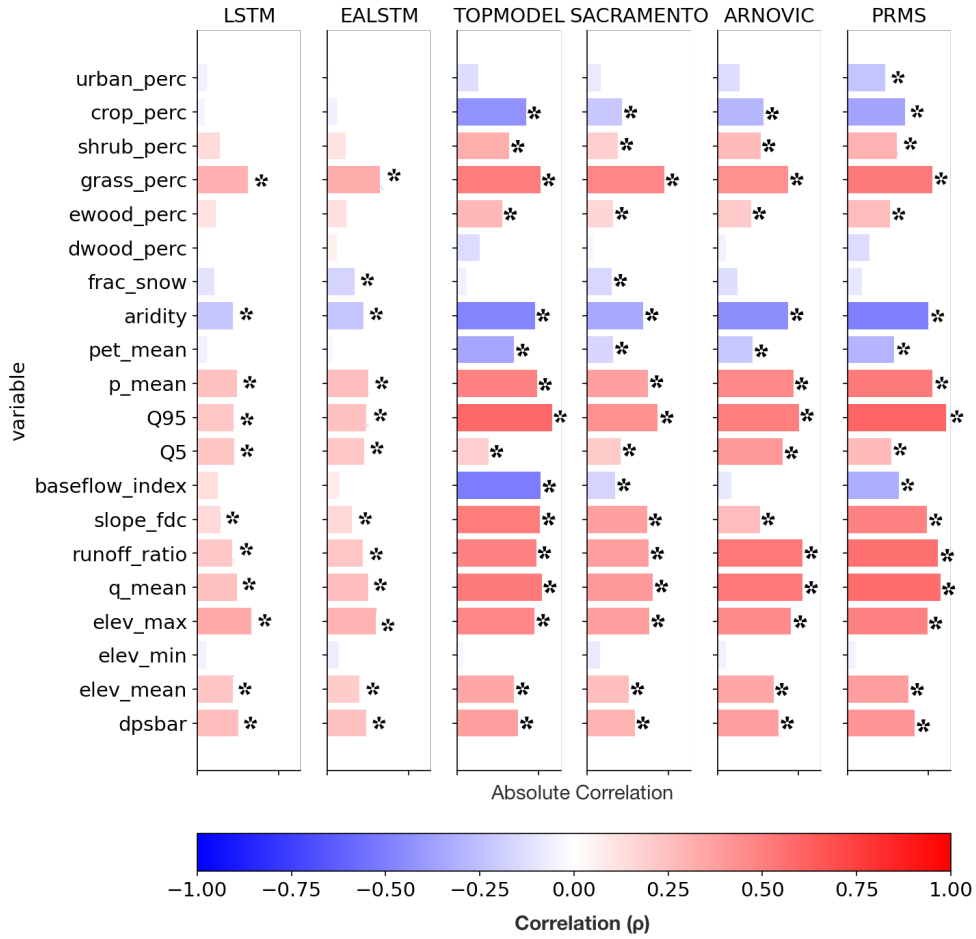


Figure 3.4 | Static features (rows) and their Spearman's Rank Correlation Coefficient with model (columns) NSE scores. The positive correlations are in blue, the negative correlations are in red. Pale bars show very low correlations. (*) indicates that the correlation is significant at the α=0.001 level. The first 6 features can be classified as landcover features. The next 4 features are climatic indices. The next 6 features are hydrologic attributes and the final 4 are topographic features. DPSBar refers to the mean drainage path slope, and reflects the average steepness of a catchment.

very fast responses when that storage is very close to full [Lane *et al.*, 2019].

The impact of water balance closure on simulation accuracy

One of the key hydrological conditions that hydrological models struggle with is the lack of closure of the catchment water balance. The conceptual models we test here explicitly maintain mass balance. They define the topographic surface water catchment as the surface over which water is conserved, i.e. the surface water catchment is not expected to leak, nor should any water enter the catchment other than through measured

precipitation. This will not then capture water losses or gains from undercatch, drifting snow, advection of fog, groundwater, or anthropogenic transfers into or out of the topographic catchment. Consequently, we would not expect the conceptual models to take account of catchments where the water balance (defined in the data) does not close. The LSTM, in contrast, is free to adjust to account for patterns in these anomalies. It is not yet possible to diagnose the origin of any such anomalies using the LSTM alone: they may arise from inter-catchment transfers (either through anthropogenic or groundwater processes), or data errors, among other reasons that the water balance might not be closed based on observations at the catchment scale. In spite of this, we expect that the LSTM will show improved performance in these catchments where there is no closure of the catchment water balance in the underlying dataset. Since we are calculating performance on out-of-sample timesteps, if the LSTM performance is improved, we can infer that the LSTM based model has learned to correct these inconsistencies in a way which is consistent between training and evaluation data, and is therefore adjusting the catchment water balance to better simulate the hydrograph.

We plot catchments on two dimensions (Fig. 3.5), their wetness index (P/PE) and the runoff coefficient (Q/P), to identify catchments where water is not conserved. Points above the horizontal line reflect catchments where the observed discharge is greater than the precipitation input to the catchment. This area of the graph represents catchments where the data has too little water to generate the observed runoff. Points below the curved line are where runoff deficits exceed total PET in a catchment. This area of the graph represents catchments where PET is not large enough to describe the water remaining after runoff is accounted for, i.e. the data has "excess" water (Fig. 3.5).

We tested whether the LSTM was better able to simulate discharge in catchments with "excess" water (i.e. the points below the curved lines in Fig. 3.5, which are then represented by the orange kernel density estimate in Fig. 3.6). As hypothesised, we find that the LSTM is more robust to these conditions and produces NSE scores that are comparable to the stations where the conceptual models perform best.

Interestingly, despite the performance improvement over the benchmark conceptual models the LSTMs continue to produce a performance decline in catchments with an imbalanced water balance (Fig. 3.6). This suggests that the LSTM models still struggle with water-limited and energy limited (low runoff coefficient and low wetness index) catchments. This could be because human management decisions that lead to abstractions are unpredictable without further dynamic inputs, such as timings of abstractions and effluent returns. Or else, that the underlying data does not contain sufficient geological information to describe the complex percolation and surface or subsurface con-

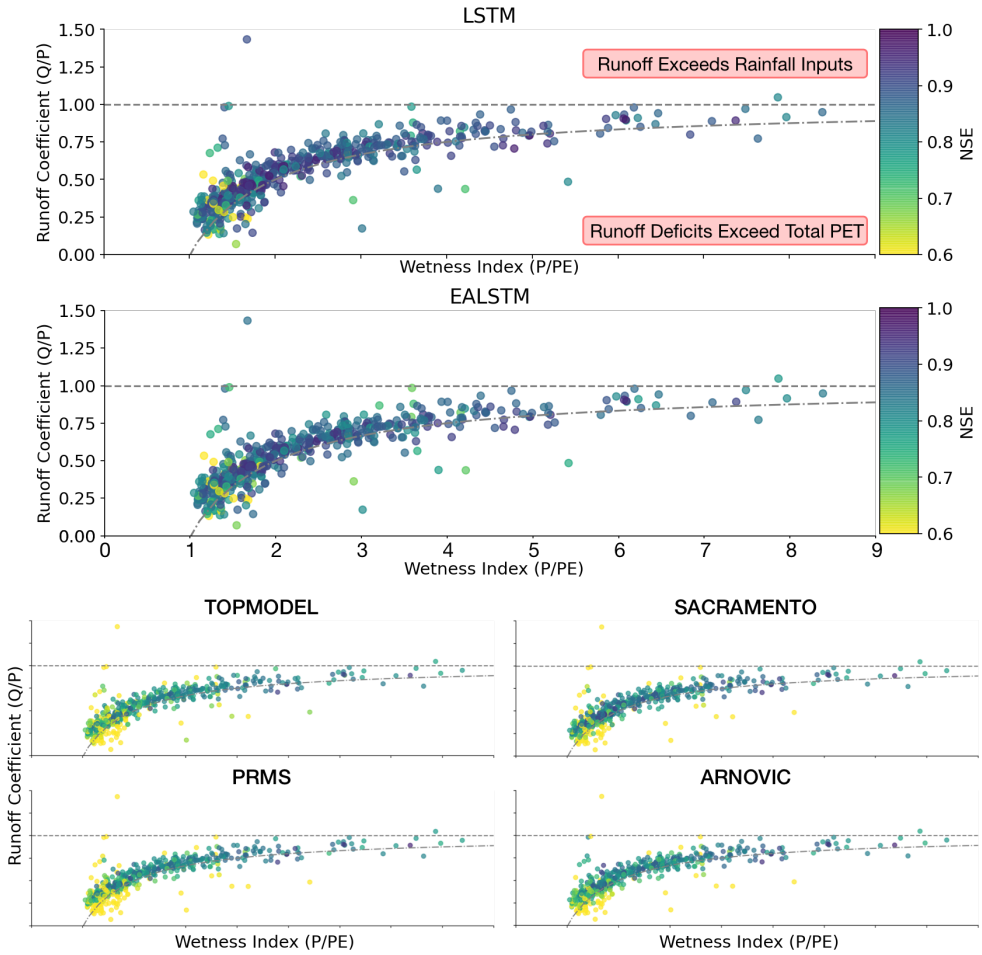


Figure 3.5 | Scatter plot for the relationship between the wetness index, runoff coefficient and the model NSE score. Each point is a catchment, coloured by the NSE score ranging from 0.8 (lighter) to 1.0 (darker). Points above the horizontal line reflect catchments where the observed discharge is greater than the precipitation input to the catchment. Points below the curved line are where runoff deficits exceed total PET in a catchment, therefore, there is "excess water" in the data, since PET cannot explain the leftover water after accounting for runoff.

nectivity pathways that cause a surface water catchment to leak.

Ultimately, the performance decline is less pronounced for the LSTM. The LSTM continues to produce simulations with NSE scores greater than 0.6. This suggests there remains information in the data that the LSTM is capable of using to maintain accurate simulations in out-of-sample conditions.

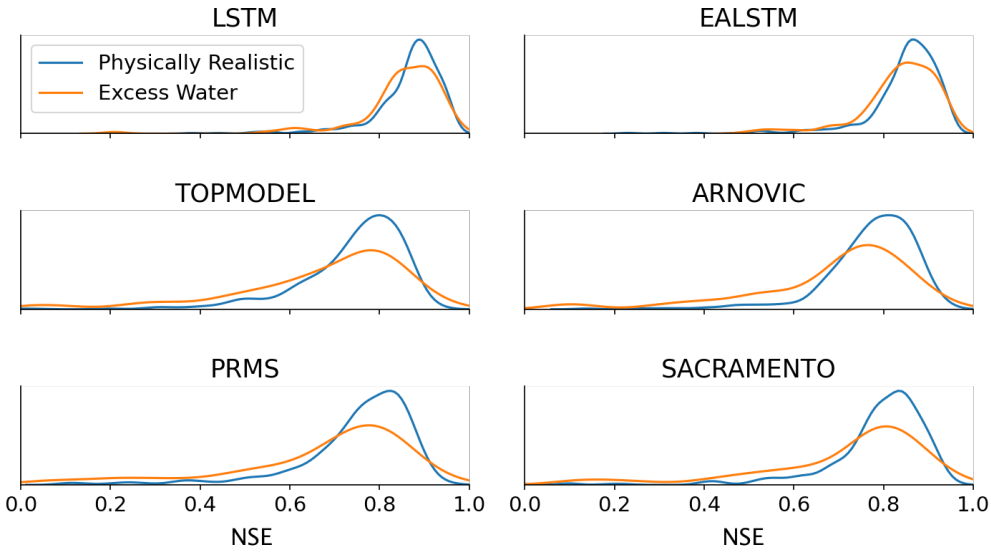


Figure 3.6 | Comparing the NSE results for catchments that have excess water, where runoff deficits exceed total PET (orange) to those catchments that have physically realistic conditions (blue). The orange line shows the histograms for stations that fall below the curved line in the Budyko analysis above (the runoff deficit exceeds total PET, therefore there is excess water in the model). The blue line shows the histograms for those stations between the two dashed lines.

3.4 Discussion

This study benchmarks the performance of the LSTM using four commonly-used conceptual models as a reference. The LSTM produced accurate simulations for a large number of catchments across Great Britain. The performance of the LSTM demonstrates that there is adequate information in the observational data to accurately simulate discharge behaviours across the various hydrological conditions found in Great Britain. The simulated time series and catchment error metrics can be found at: zenodo.org/record/4555820.

In the discussion that follows we return to our three research questions: (i) How well do LSTM-based models simulate discharge in Great Britain? (ii) How do LSTM-based model performances compare with the conceptual models used as a benchmark? (iii) Can we extract information from the spatial and temporal patterns in diagnostic measures?

3.4.1 Inter-Model Performances

The LSTM based models produce accurate simulations of discharge across GB, a temperate region. Two findings from this research confirm and extend the conclusions of

previous work. First, the LSTM consistently outperforms the EA-LSTM [Kratzert *et al.*, 2019e]. Secondly, both LSTM-based models demonstrate improved simulation accuracy for discharge modelling compared with the conceptual models we use as benchmarks.

How well do LSTM-based models simulate discharge in Great Britain?

It is worth noting that the LSTM and EA-LSTM also differ in terms of practical computational requirements. The LSTM trains much faster than the EA-LSTM. The LSTM will train 30 epochs in 1 hour, compared with 30 epochs in 10 hours for the EA-LSTM. This is due to the LSTM being an in-built Pytorch (v.1.7.1) function that makes use of CUDA optimised code (for running the models on a GPU). In contrast, the EA-LSTM relies on custom code without the CUDA enabled optimisations.

How does the LSTM performance compare with the conceptual models used as benchmark?

We have demonstrated that the LSTM is an effective model architecture for extracting information from hydro-meteorological data, providing a data-driven benchmark showing what is achievable given the information contained in available observation data from CAMELS-GB [Nearing *et al.*, 2021b]. The LSTMs demonstrate better performance on out-of-sample times than in-sample performance from the benchmark conceptual models.

There are obvious challenges with direct comparison of LSTM performance against the benchmark developed by Lane *et al.* [2019]. The first is that the LSTM is not constrained to maintain water mass balance, whereas the conceptual models discussed here are. Another challenge is that the method of optimisation used for choosing parameters in the LSTM (stochastic gradient descent) is different to the random-sampling and NSE selection criteria used to select the "best" model parameters for the conceptual models. The sampling process used by Lane *et al.* [2019] is explicitly for estimating uncertainty as well as providing a reference of conceptual model performances. Another difference is that the LSTM diagnostic scores are calculated on out-of-sample predictions, compared with the in-sample predictions for the benchmark conceptual models.

Finally, the LSTM-based models are trained on all basins, with a single set of weights for the whole of GB. Therefore, these LSTM models are regional models that are able to reproduce behaviours across Great Britain. In contrast, most hydrological models perform best when calibrated on individual basins [Beven, 2006a]. By contrast, LSTM-based models are most accurate when trained with as much data from as many catchments

as possible [Gauch *et al.*, 2021c]. It is important to interpret the number of parameters for each model type in light of this fact.

The catchments where the comparative performance difference is small, i.e. where the conceptual models perform almost as well as the LSTM, reflect areas where the conceptual models capture the majority of the information from the data, and the conceptual model well represents the hydrological process. This is the case in West Scotland, North West England & North Wales and North East England (see Appendix Fig. 6). The benchmark results are valuable in providing a reference point for us to assess the value of LSTM-based approaches. We welcome future studies using the LSTM simulations provided here and further explore performance differences and the limitations of DL methods across GB.

Can we extract information from the spatial and temporal patterns in diagnostic measures?

The LSTM shows the largest performance improvement over the conceptual models in the North West of Scotland and the South East of England. The performance differences in North West Scotland are very likely a result of the ability of the LSTM to learn a representation of snow processes from the input data, whereas, the conceptual models were simulating these catchments without a snow module.

Despite the performance improvement over conceptual models in the South East of England, the LSTM still struggles in the South East relative to elsewhere in GB. The South East is a relatively dry region compared to elsewhere in GB. It contains the highest proportion of catchments that fall below the dashed line in Fig. 3.5, and therefore stations where the surface water catchment is "leaky". Furthermore, there are underlying chalk aquifers which provide water storage and lateral transfers. We outline three hypotheses for why the LSTM performance may be lower in the South East compared with elsewhere in GB.

The first hypothesis is linked to the training of the LSTM based models. The LSTM shows a performance decline in drier conditions (Fig. 3.4, see "*aridity*"). This confirms the findings of other DL studies in the US, where the LSTM also struggled to reproduce hydrographs in drier conditions [Kratzert *et al.*, 2018, 2019e]. Basins that have long periods of low flow contain little information, since changing meteorological inputs co-occurs with very little change in the target discharge. Therefore, the physical process relating meteorological inputs to river discharge can only be inferred from those catchments with varying discharge. There is some evidence for this hypothesis. NSE scores

show positive correlations with increased discharge (at mean flow, Q5 and Q95), as well as increased NSE as rainfall increases (*p_mean*) (Fig. 3.4).

A related, but separate, hypothesis is that the use of NSE as an objective function fails to adequately weight performance in these low flow regimes (the NSE was the objective function across both the conceptual models and the DL models).

A final hypothesis is that groundwater dynamics and human abstractions, which influence catchments in the South East, are not well captured by the variables in CAMELS-GB. Hydrological processes are not simulated as effectively in "leaky" catchments compared to those catchments where the water balance can be closed with hydrometric data (Section 3.3.3), even using a very flexible and effective data-driven model that is not constrained to balance water (the LSTM). This suggests that the underlying data does not contain sufficient information to model the full range of processes that influence the hydrograph in these catchments, including groundwater and abstractions. The catchment averaged information on soil texture (sand-silt-clay) provides a coarse proxy for catchment porosity. Furthermore, further data, such as groundwater time-series, might be necessary to obtain more accurate discharge predictions. We suggest that different input data sets should be tested to try and improve LSTM performances enabling the LSTM to more properly account for the complex percolation and infiltration dynamics in these catchments.

In terms of the seasonal patterns in LSTM performances and the worse performances in summer, the above hypotheses also apply, since the summer is the driest season in GB. Despite this, the LSTM-based models have been able to use the information in the available data to better model summer (JJA) discharge than the benchmark models. As in the data-based mechanistic modelling framework [Young, 2003] the next stage for hydrologists is to search for a mechanistic interpretation of the learned model structure, also see Nearing *et al.* [2021b]. One possible mechanistic interpretation that warrants further exploration is the idea that the LSTM is capable of learning seasonally varying catchment “connectivity” [Bracken & Croke, 2007]. In winter, when soils are saturated, there are a greater number of pathways for water to enter river channels, and connectivity is high. In summer, however, there is greater resistance to water flow, since water can be absorbed and stored in drier soils, as found in Swiss catchments by van Meerveld *et al.* [2019], and connectivity is lower. Connectivity information could be represented by the hidden state ($\mathbf{h_t}$), or cell state vectors ($\mathbf{C_t}$). The proposed impact of catchment connectivity on the performance improvement of the LSTM based models is ultimately speculative, and future work will explore whether the LSTM has learned to represent the concept of connectivity and seasonally variable flow pathways.

In contrast with the benchmark conceptual models, the LSTM-based model NSE scores have no negative correlation with crop cover percentage (Fig. 3.4). It is possible that the LSTM has effectively used the cropland cover variable to improve its internal representation of hydrology in those catchments with a strong agricultural signal. In order to test this hypothesis, one could perform an ablation study, removing input features and determining the impact on model performances. Alternatively, sensitivity analysis could be used to determine the relative contribution of the input features to the discharge prediction, thus revealing what input features are important for the model simulations. We intend to pursue this idea in upcoming papers.

Ultimately, compared with the benchmark models, the LSTM shows robustness to catchment conditions associated with poor conceptual model performance. Dry catchments, catchments with a strong agricultural signal, and summer discharges are all strongly correlated with worse conceptual model performances. In contrast, the LSTM has good performance on out-of-sample times in these same conditions. There is therefore information that the LSTM has learned to generalise from the CAMELS-GB dataset that the conceptual models are not utilising. The experiments we present here demonstrate conditions in which we can (and cannot) improve our traditional hydrological models given the availability of high quality, large sample datasets [Beven, 2006b; Nearing *et al.*, 2020b].

3.5 Conclusions

In this study we have benchmarked the performance of two LSTM based models trained on 669 catchments across Great Britain. We have demonstrated that LSTM-based models trained on a large sample of catchment-averaged hydro-meteorological time-series produce accurate simulations across GB. There is clearly information available in CAMELS-GB for modelling diverse hydrological conditions, and the LSTM performances should be interpreted as a competitive reference for what simulation performance is possible on out-of-sample (in time) conditions. We trained an ensemble of LSTM-based models to account for random initialisation during the training process of these deep learning models, which also provides an estimate of prediction uncertainty (Appendix Section A.3). The ensemble mean simulation produces a median NSE score of 0.88 (LSTM) and 0.86 (EA LSTM), with no catchments scoring NSE below 0. These results are consistent with the findings from Kratzert *et al.* [2018] in a different geographical context.

We have explored the spatial and temporal patterns in LSTM and EA LSTM perfor-

mances, using the large-sample of catchments to better understand the conditions in which the LSTM-based models perform well, compared to themselves (LSTM in catchment A vs. LSTM in catchment B) and compared with traditional conceptual models. The results show that LSTM-based model performances are more robust to hydro-climatic conditions in the South East of England, in more arid catchments and in catchments where the water balance does not close. This suggests that there is more information in large-sample datasets such as CAMELS-GB than is captured by hydrological theory as encoded in the benchmark conceptual models. Further work remains to determine what information has been learned by these LSTM-based models, to use that information to improve hydrological theories, and feed them back, if possible, into further developments in conceptual and physically based models.

Relative to the LSTM-based model performances elsewhere in GB, the LSTM-based models continue to underperform in South East England relative to elsewhere in GB. Considering the catchment conditions that are associated with this pattern it is clear that all models struggle with drier conditions and catchments where the water balance does not close. It also seems possible that the training process fails to capture hydrological behaviours in drier catchments. There are a number of possible reasons. Firstly, changing meteorological conditions in dry catchments lead to little or no change in discharge (as would be the case in ephemeral streams). Alternatively, the LSTM architecture may not be capable of simulating both dry catchments and those with a higher runoff-ratio using just a single set of weights. Finally, the data may not contain sufficient information to capture the percolation and connectivity dynamics that drive hydrological behaviour in catchments with significant groundwater processes. Further studies will examine the internal representation of hydrological processes in these catchments to better understand what the LSTM has (not) learned about connectivity and groundwater processes.

This paper benchmarks LSTM performance across Great Britain using a new large-sample dataset, CAMELS-GB [Coxon *et al.*, 2020b], providing a reference for future hydrological modelling efforts. Furthermore, this manuscript outlines the hydrological conditions in which the LSTM-based models perform well and those conditions which are more difficult to model. We encourage future benchmarking studies to include LSTMs as a competitive model choice for simulating rainfall-runoff processes.

4 Concept Formation in Hydrological LSTMs

Contributions This chapter is largely based on the following submitted manuscript*

T Lees, S Reece, F Kratzert, D Klotz, M Gauch, J de Bruijn, R Kumar Sahu, P Greve, L Slater and SJ Dadson, 2021. *Hydrological Concept Formation inside Long Short-Term Memory (LSTM) networks*, **Hydrology and Earth System Sciences**, in review. Preprint 10.5194/hess-2021-566

Abstract. Neural networks have been shown to be extremely effective rainfall-runoff models, where the river discharge is predicted from meteorological inputs. However, the question remains, what have these models learned? Is it possible to extract information about the learned relationships that map inputs to outputs? And do these mappings represent known hydrological concepts? Small-scale experiments have demonstrated that the internal states of Long Short-Term Memory Networks (LSTMs), a particular neural network architecture predisposed to hydrological modelling, can be interpreted. By extracting the tensors which represent the learned translation from inputs (precipitation, temperature) to outputs (discharge), this research seeks to understand what information the LSTM captures about the hydrological system. We assess the hypothesis that the LSTM replicates real-world processes and that we can extract information about these processes from the internal states of the LSTM. We examine the cell-state vector, which represents the memory of the LSTM, and explore the ways in which the LSTM learns to reproduce stores of water, such as soil moisture and snow cover. We use a simple regression approach to map the LSTM state-vector to our target stores (soil moisture and snow). Good correlations ($R2 > 0.8$) between the probe outputs and the target variables of interest provide evidence that the LSTM contains information that reflects known hydrological processes comparable with the concept of variable-capacity soil moisture stores.

The implications of this study are threefold: 1) LSTMs reproduce known hydrological processes. 2) While conceptual models have theoretical assumptions embedded in the model a priori, the LSTM derives these from the data. These learned representations are

*with the following author contributions. Conceptualisation: TL, FK, DK, MG, SD, SR. Data curation: TL. Formal Analysis: TL. Methodology: TL, SR, FK, JDB, DK, MG. Visualisation: TL. Writing – original draft: TL. Writing – review and editing: TL, SR, FK, DK, MG, JDB, RKS, PG, LS.

interpretable by scientists. 3) LSTMs can be used to gain an estimate of intermediate stores of water such as soil moisture. While machine learning interpretability is still a nascent field, and our approach reflects a simple technique for exploring what the model has learned, the results are robust to different initial conditions and to a variety of benchmarking experiments. We therefore argue that deep learning approaches can be used to advance our scientific goals as well as our predictive goals.

4.1 Introduction

LSTMs have demonstrated state-of-the-art performance for rainfall-runoff modelling for a variety of locations and tasks [Frame *et al.*, 2021a; Kratzert *et al.*, 2018, 2019e; Lees *et al.*, 2021b; Ma *et al.*, 2020]. However, whether we can use these models to better interpret the hydrological system remains an open question. Given that LSTM-based models offer state-of-the-art hydrological performance, more research is required to better understand what conceptual structures the LSTM has learned and to diagnose potential gaps in our conceptual and process-based models, ultimately to stimulate innovation in hydrological theory.

The primary objective of this study is to test the hypothesis that the information stored in the LSTM state-vector reflects known hydrological concepts that are important for discharge generation, including soil water storage and snow processes. What have these models learned about the hydrological system that allows them to make highly accurate predictions? Can we interrogate the model to determine whether the LSTM has learned a physically realistic mapping from inputs to outputs? Being able to reason about the model and its behavior is a key component of dependable models. It allows researchers and practitioners to interrogate the model, making sure that it is giving the right results for the right reasons [Kirchner, 2006].

Deriving insights about the hydrological system has always been a goal of hydrological modelling [Beven, 2011b]. Peter Young's work on Data-Based Mechanistic modelling (DBM) emphasised the need to apply flexible data-driven models before then applying a mechanistic interpretation to the learned representation of these models [Young, 1998, 2003; Young & Beven, 1994]. Philosophically, this approach is similar to the one we take here, although the number of parameters in the DBM approach is much smaller. In an early application of neural networks to rainfall-runoff modelling, Wilby *et al.* [2003] sought to challenge preconceptions of neural network approaches as uninterpretable. They found that nodes in their Multi-Layer Perceptron corresponded to quickflow, base-

flow and soil saturation, and showed how the learned representation of deep learning models could be interpreted. They sought to determine whether neural networks were capable of reproducing both the outputs and internal functioning of conceptual hydrological models.

Recent studies call to more fully explore the potential for techniques from the fields of artificial intelligence and machine learning [Beven, 2020; Karpatne *et al.*, 2017; Reichstein *et al.*, 2019; Shen, 2018] by demonstrating predictive performance alongside interpretations of the model itself to improve our understanding of the modelled system. Several studies have suggested that LSTM rainfall-runoff models learn a generalizable representation of the underlying physical processes. This allows them to perform well in out-of-sample conditions, such as Prediction in Ungauged Basins (PUB) [Feng *et al.*, 2020b; Kratzert *et al.*, 2019c; Ma *et al.*, 2020], and unseen extreme events [Frame *et al.*, 2021a]. These results suggest that LSTMs have captured information that generalizes to these conditions, information that can help us improve hydrological theory and predictions.

Outside of hydrology, calls for interpreting machine learning and deep learning systems are getting louder and even generating legislative changes [European Union Digital Strategy, 2019; UK Statistics Authority, 2019]. Spiegelhalter [2020], for example, argues that as algorithmic decision support tools become widespread in everyday life, the ability to describe how predictions are made is essential for building trust in these systems. A large body of literature has arisen to: define interpretability [Doshi-Velez & Kim, 2017; Lipton, 2018; Ribeiro *et al.*, 2016a]; to measure how interpretable models aid human decision making [Chu *et al.*, 2020; Nguyen, 2018]; and to develop methods for interpreting models [Ghorbani & Zou, 2020; Lundberg & Lee, 2017; Olah *et al.*, 2018, 2020]. Our contribution draws on work from neuro-linguistic programming, where learned embeddings from models trained for speech-recognition tasks have been interpreted to better understand how parts-of-speech are recognised and used by LSTM models [Hewitt & Liang, 2019].

The exploration of the internal representations of LSTM based rainfall-runoff models is still at an early developmental stage. Kratzert *et al.* [2018] showed evidence that individual LSTM cells correlate with snow water content, although the model was only trained to predict discharge from meteorological inputs. Kratzert *et al.* [2019e] explored the learned embedding of catchment attributes, showing that an LSTM variant had learned to group the rainfall-runoff behaviours of hydrologically similar catchments. Using dimensionality reduction techniques, the static embedding (the output of the input

gate) was shown to reflect spatial and thematic groups of catchments that qualitatively correspond to catchments with similar hydrological behaviours. For two exemplary basins, Kratzert *et al.* [2019b] found correlations between the cell states of the LSTM and three hydrological states (upper zone storage, lower zone storage and snow depth) from the Sacramento + Snow-17 hydrological model [Burnash *et al.*, 1995]. All three of these studies introduced methods that can be used for exploring the internal representation of hydrological models, but there exists no comprehensive evaluation of the information stored in the cell state dimensions across a large sample of basins. Furthermore, these studies only compared individual memory cells with hydrological processes. The LSTM however, is not forced to store information about one process in a single memory cell but can distribute the information about hydrological processes across several cells. Therefore, we explore methods for extracting the information that is stored in the LSTM cell state, across all cells.

The aim of this research is to examine the internal functioning of the LSTM model. We explore the evolution of the LSTM state-vector and test whether information that reflects intermediate stores of water (soil moisture and snow depth) has been learned by the LSTM. This research is novel for providing a means of interpreting what information the LSTM rainfall-runoff model has encoded within its state-vector. To our knowledge, we are the first to apply techniques developed in machine learning interpretability and natural language processing research [Hewitt & Liang, 2019] to hydrology. We carry out a comprehensive evaluation of the LSTM cell states across a sample of 669 catchments in Great Britain [Lees *et al.*, 2021b]. This allows us to rigorously assess whether the LSTM has learned concepts that generalize over space. On this basis we devised several baseline experiments to provide evidence for an internal representation of hydrologically relevant processes. Furthermore, we consider information stored across all values in the LSTM state-vector, as opposed to identifying and focusing only on single values from within the cell state. This is important since there are no constraints forcing the LSTM to store information in individual cells.

4.2 Methods

In this study, we trained LSTM models using the same hyper-parameters as those trained in Lees *et al.* [2021b]. We offer a brief introduction to the state-space formulation of the LSTM [Kratzert *et al.*, 2019b] because it offers a clear explanation for why we explore the cell-state (c_t), since it reflects the state-vector of the LSTM.

The LSTM

Hydrological models are often formulated with a state-space based approach. This means that the states (s) at a specific time (t) depend on the input at time t (x_t), the model state in the previous timestep (s_{t-1}) and the model parameters (θ) [Kratzert *et al.*, 2019b].

$$s_t = g(i_t, s_{t-1}; \theta_j) \tag{4.1}$$

The model output (y_t, discharge) is a function of the states (s_t) and inputs (i_t) at that timestep, and the model parameters.

$$y_t = g(i_t, s_t; \theta_j) \tag{4.2}$$

Similarly, the LSTM can be formulated as:

$$c_t, h_t = f_{\text{LSTM}}(x_t, c_{t-1}, h_{t-1}; \theta_k) \tag{4.3}$$

$$y_t = f_{\text{Dense}}(h_t; \theta_l) \tag{4.4}$$

Where the state-vector (the "cell state" c_t) and output-vector (the "hidden state" h_t) of the LSTM at timestep t are a function of the current inputs (x_t, e.g. meteorological features and catchment attributes), the previous output and state (h_{t-1} and c_{t-1}) and some learnable parameters (θ_k). Similar to the state-update equations, the output of the model (y_t, e.g. the discharge) is a function of the output of the LSTM (h_t, which is a function of c_t) and some more (learnable) model parameters (θ_i).

The key difference between the LSTM and classical state models (e.g. conceptual and physical hydrology models) is that the LSTM can infer any process that is deducible from the data to solve the training task, while classical hydrological models are limited by the processes that are hard-coded in the model implementation.

In order for the LSTM models to produce accurate simulations of discharge across a variety of catchments, we hypothesise that the LSTM should have learned to represent hydrological processes and stores. We test whether the LSTM is able to recover intermediate stores of water by visualising the evolution of the LSTM cell-state and compare this to soil moisture and snow depth from ERA5-Land.

4.2.1 Experimental Design

We used the following experimental design to investigate the learned hydrological process understanding of LSTMs.

Following Lees *et al.* [2021b], we trained a single LSTM to predict runoff for 669 basins from the CAMELS-GB dataset [Coxon *et al.*, 2020b]. The input sequences are digested into the LSTM each consisting of one year's worth of daily data (365 timesteps). The model is forced by a set of meteorological variables (precipitation and temperature) and a series of static catchment attributes describing topography, climatic conditions, soil types and land cover classes. These static attributes are used to learn differences and similarities between catchments. For more details of the training procedure and for a comprehensive table listing all model inputs, we refer the reader to Table 3.1 from Lees *et al.* [2021b]. It is important to note that neither snow depth nor soil moisture were included as inputs or outputs during model training.

4.2.2 Probing

In the present context, a probe is a diagnostic device that is used on top of the trained LSTM model to examine the learned internal representation of the LSTM. In its simplest form, a probe is a linear regression model that connects the cell states to a given output. In a more complex form a probe might be realized in the form of a set of stacked multilayer perceptrons, or any other algorithm fit for regression tasks. As such, probes offer the opportunity to explore what the LSTM has learned during training, allowing us to use the LSTM to generate predictions of latent, intermediate variables. They also confirm whether our model has learned physically realistic mappings from inputs to outputs. To our knowledge, probes have not been used on hydrological LSTMs.

Probes have been used in natural language processing tasks to determine whether learned embeddings in deep learning models contain information that can be matched to semantically meaningful concepts [Hewitt & Liang, 2019]. The embeddings are used as inputs, and a probe is trained to map these embeddings onto properties, such as part-of-speech tags.

Since these probes are trained in a supervised way, we are currently limited to looking for known hydrological processes. Trained in this way, probes cannot be used to extract unknown information, since we require a target variable to fit the probe. This means that in the present study we are not looking for new hydrological understanding, or seeking to uncover as-of-yet undiscovered hydrological patterns, but explicitly looking for known physical processes in the learned LSTM representation. This paper

demonstrates a conceptual innovation moving the field towards extracting information from the LSTM, diagnosing whether these neural networks are learning physically realistic processes.

In this paper, we use the probe to explore whether the LSTM state-vector has information that is predictive of different latent hydrological variables, such as soil moisture at various depths and snow water equivalent. We begin with the simplest probe, a linear model. This encodes the strong assumption that latent variables can be extracted as a linear combination of the cell state values. While this is not necessarily the case, for the purposes of probe interpretation the simplicity of the linear model is preferred. It allows for intuitive explanations and simple visual analysis of results by exploring the weights of the linear model, and by interpreting the probe predictions. For our experiments, we fit one probe for all catchments, learning a set of weights and a bias term for each target variable (Figure 4.1c, d). That is, we hypothesize that there is a common set of weights that generalize to all basins in the training set.

As a control experiment, we ensure that we are not learning spurious relationships between the LSTM state and intermediate hydrological stores (soil moisture and snow depth) by designing two experiments to test that our findings are specific in time and space. We test both shuffling the target variables in space (i.e. breaking the spatial link between cell-states and the target variable) and shifting the target variable in time (breaking the temporal link between inputs and outputs), and find that these results show that the information contained in the cell-state vector is specific to the given location and time. These results can be found in Appendix Sect. A.5.

We train a linear probe parameterised by β to make predictions of our target storage variable ($\hat{s}$) for each catchment (i) and at each timestep ($\{1:T\}$).

$$\hat{s}_{i,\{1:T\}} = f_{\beta}(c_{i,\{1:T\}}) \tag{4.5}$$

Our linear model, f_{β}, is a penalized linear regression model. We use the elastic-net regularisation that combines the ℓ_1 penalty of lasso regression with the ℓ_2 penalty of ridge regression. The reason for choosing the elastic-net regularisation is that when we have correlated features in c_t, the lasso regression is likely to pick one of these at random, whereas the elastic-net regression will assign weights to both [Friedman *et al.*, 2010]. The lasso (ℓ_1 penalty) shrinks non-informative weights to zero, and the ridge (ℓ_2 penalty) ensures that correlated features are not randomly set to zero by the lasso ensures that the model is stable under rotation, so both are useful. The objective function

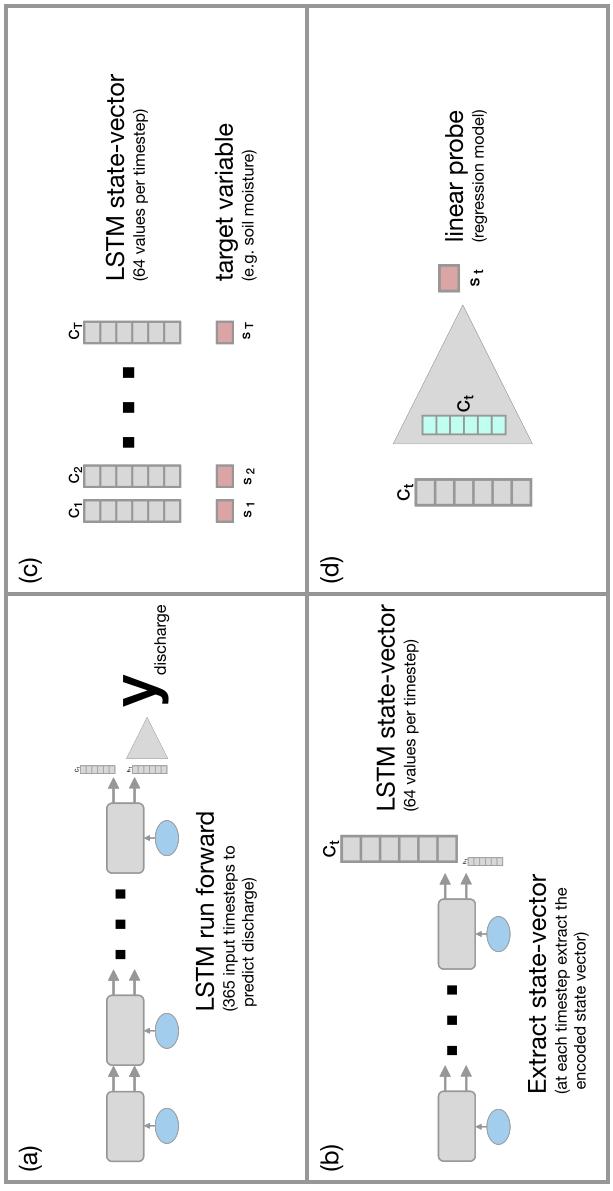


Figure 4.1 | An overview of the linear probe analysis. (a) Demonstrates the LSTM as an input-state-output model, where at each timestep, inputs (blue spheres, such as precipitation) are processed producing a prediction of discharge on the 365th day. This reflects how the LSTM is trained. (b) When forcing the model with already trained weights, we extract the LSTM state vector (c_t) at each timestep. (c) We then compile a dataset of inputs (the c_t vectors for each target timestep) and targets (s_t - the soil moisture measurements for the catchment) matched at each catchment and timestep. (d) Finally, we use this dataset to train a linear probe, a set of weights and a bias term for all catchments.

becomes:

$$\min_{\beta} \frac{1}{2n_{\text{samples}}} ||c_{i,t}\beta - y||_2^2 + \alpha\rho||\beta||_1 + \frac{\alpha(1-\rho)}{2}||\beta||_2^2 \tag{4.6}$$

We set the α parameter to 1.0 (describing the degree of shrinkage) and the ρ parameter to 0.15 (describing the degree of ℓ_1 loss relative to ℓ_2 loss). These parameters were set in order to give the best training-sample performance and ensure that non-informative weights are shrunk to zero.

4.2.3 ERA5-Land Data

In order to determine whether the information stored in the LSTM state-vector reflects known hydrological concepts, we used variables from the ERA5-Land dataset as the probe targets (Figure 4.1c, d). As designed, this protocol will determine whether the LSTM has learned consistent representations by comparing the probe output to commonly used soil-moisture products, including reanalysis data, rather than concepts directly from in-situ observations. Reanalysis observations were preferred because of the longer time series of available data, the gridded form meaning that a catchment-averaged soil moisture can be calculated, and because these products are globally available, meaning that this approach would generalise to LSTMs trained elsewhere. However, there are uncertainties associated with the soil moisture estimates from these gridded reanalysis products.

We used soil moisture and snow-depth data from ERA5-Land [Muñoz-Sabater *et al.*, 2021], a reprocessing of the land surface components of ERA5 forced by the ERA5 atmospheric model [Hersbach *et al.*, 2020]. ERA5-Land is a reanalysis product with 9 km grid spacing and an hourly temporal frequency. It is a global land surface reanalysis dataset for describing the water and energy cycles. Muñoz-Sabater *et al.* [2021] demonstrate that the ERA5-Land product has improved soil moisture and snow observations compared with ERA-Interim products when evaluated against in-situ soil moisture measurements. The results when compared against ERA5 are more variable, although ERA5-Land does show improvements in North America and small improvements in Europe. ERA5-Land does not assimilate soil moisture observations directly, rather the assimilation of observed meteorological data occurs only in the calculation of the atmospheric forcing variables from ERA5. Therefore, we can be confident that ERA5-Land provides a calculation of soil moisture independent of the observational, catchment-averaged datasets included in CAMELS-GB [Coxon *et al.*, 2020b] which are used to train the LSTM.

It is worth mentioning that the main focus of this study is to test whether the LSTM models intermediate stores of water (i.e. matches the relative changes in that unseen variable over time), and not necessarily to compare the values to the best-possible soil moisture (or snow) estimate/measurement.

The soil moisture data in ERA5-Land contains four layers, each of which is used as a target variable. The top layer, soil water volume level 1 (swvl1), is from 0-7cm, the second layer (swvl2) from 7-28cm, the third layer (swvl3) from 28-100cm and the final layer (swvl4) from 100-289cm. We therefore fit four separate probes, using each soil moisture layer as an independent target variable.

One drawback of using reanalysis soil moisture (ERA5-Land) is that we identify modelled soil moisture, rather than a directly observed soil moisture signal - that is, we discover whether our LSTM model has learned a process representation. To counter this, we also assess probe performances on alternative products such as ESA CCI Soil Moisture, a blended product combining multiple satellite-derived soil moisture estimates (see Appendix Sect. A.6). The ESA CCI dataset comes with its own caveats, namely that the measured soil moisture is restricted to the top-layer (5-10cm) and so we cannot observe whether deeper layers are represented by the internal state of the LSTM. Furthermore, the depth of the satellite derived estimate is itself a function of the water content and surface roughness. ESA CCI is not a direct observation either, but a blended estimate using radiative transfer functions and microwave-based estimations of soil moisture. We use ERA5-Land soil moisture and snow depth as our target variables for the results section of this paper. Further comparisons against ESA CCI Soil Moisture can be found in Appendix A.6.

We clipped the probe target variables (ERA5-Land soil water layers $1, \ldots, 4$ and snow depth) to the catchment shapefiles provided as part of CAMELS-GB dataset [Coxon *et al.*, 2020a] and calculated a catchment mean to produce a lumped catchment soil moisture time-series. This follows the methodology used to generate the CAMELS-GB meteorological forcing data. In order to train the linear probe we normalize the target data (ERA5-Land), centering the data using the mean target value across every catchment, and rescaling the data using the standard deviation of the target data. Both statistics are calculated in the training period.

4.3 Results

4.3.1 Soil Moisture Probe

We use the linear probe described above to see if a learned soil moisture signal for four soil-depths (from ERA5-Land) is present in the LSTM. The correlation between the inferred soil water volume and the target data from ERA5-Land is shown by the histograms in Fig. 4.2. We can see that on average the inferred soil moisture for the upper three soil layers has a high correlation with the soil moisture from ERA5-Land (median scores of 0.85, 0.90 and 0.89 for level 1, level 2 and level 3 respectively). The median catchment correlation coefficient for the fourth soil layer is less than the three upper layers (0.77). However, we would argue that for all four soil layers, the LSTM state seems to model the dynamics of the soil water content. For reference, we have run two experiments that act as a baseline (Appendix A.5).

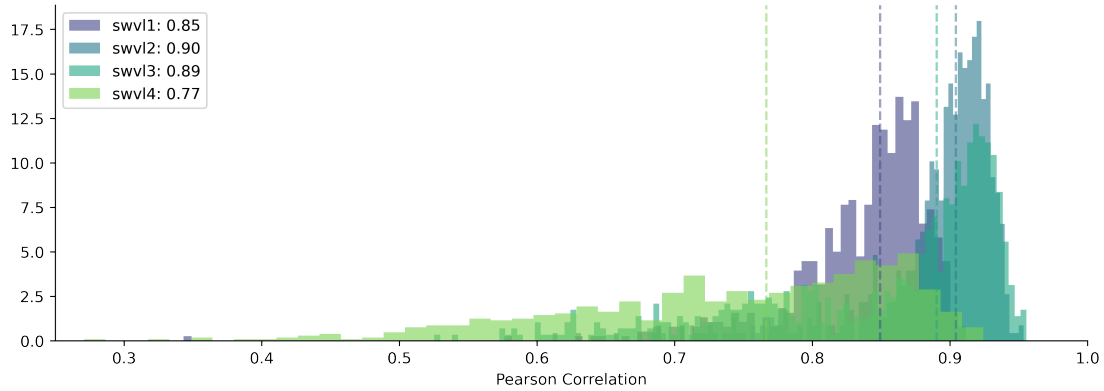


Figure 4.2 | Histogram of catchment correlation scores for each soil water volume level: soil water volume level 1 (swvl1) contains soil moisture from depths of 0–7cm, soil water volume level 2 (swvl2) contains soil moisture from depths of 7–28cm, soil water volume level 3 (swvl3) contains soil moisture from depths of 28–100cm, soil water volume level 4 (swvl4) contains soil moisture from depths of 100–289cm.

The time series plots in Fig. 4.3 show that the linear probe is able to capture the dynamics of the soil moisture values. However, modelling the catchment-specific offset in catchments with more or less saturated soils than the GB mean is difficult with the linear probe (e.g. Fig. 4.3b, a catchment with less saturated soils than the GB mean has an observed soil water volume less than the zero point, which describes the GB mean). By offset, we are referring to the point around which soil moisture fluctuates, the mean saturation level for that catchment. Looking at Fig. 4.4 confirms this hypothesis, since the catchments with the largest biases in probe outputs (blue circles) have wetter and

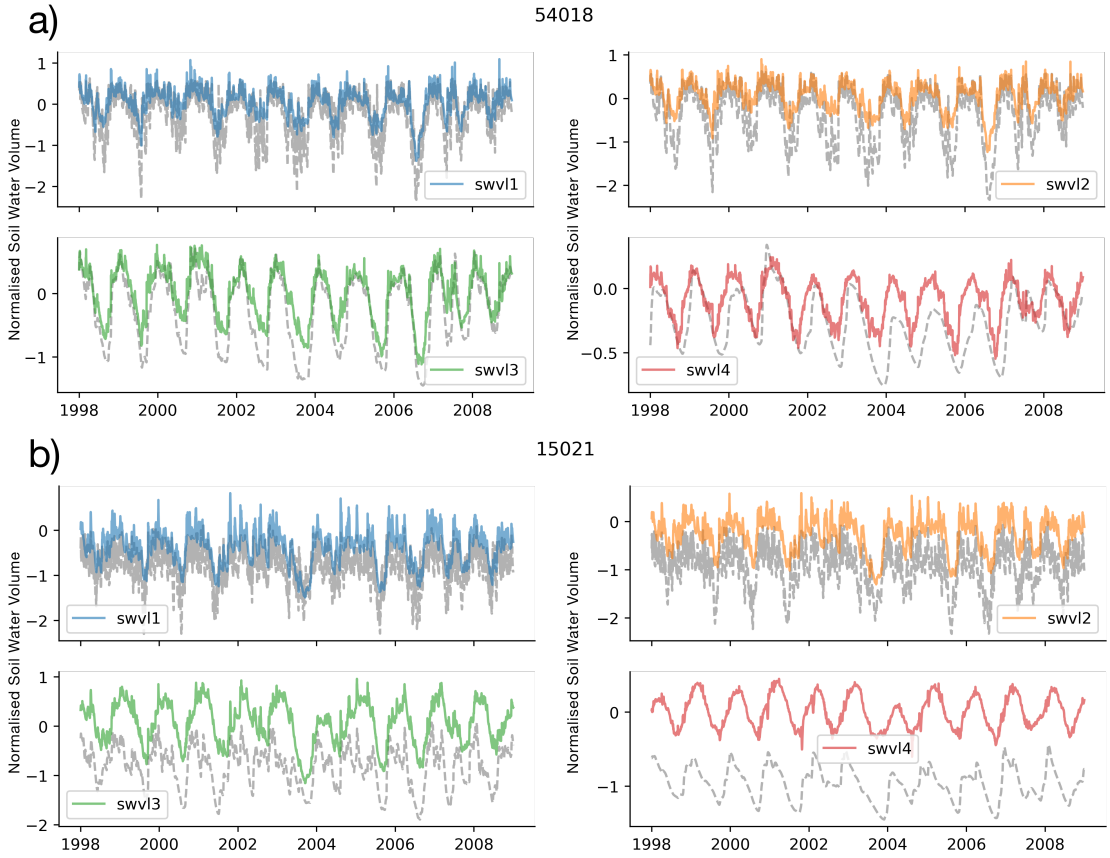


Figure 4.3 | Time series of probe predictions (coloured lines) compared with the target variables (grey dotted lines). We show two catchments here, (a) 54018 - Rea Brook at Hookagate and (b) 15021 - Burn at Burnham Overy (chosen to reflect both a catchment with a small bias and a catchment with a strong bias, see Appendix A.9 for more information on these two catchments) and four soil moisture levels, swvl1 (blue), swvl2 (orange), swvl3 (green), swvl4 (red). The probe captures the temporal dynamics of the soil moisture signals, but shows systematic bias, consistently predicting variability about zero, which defines the mean GB-wide soil moisture.

drier than average conditions, compared with the smallest biases (orange), which are concentrated in the middle of the distribution.

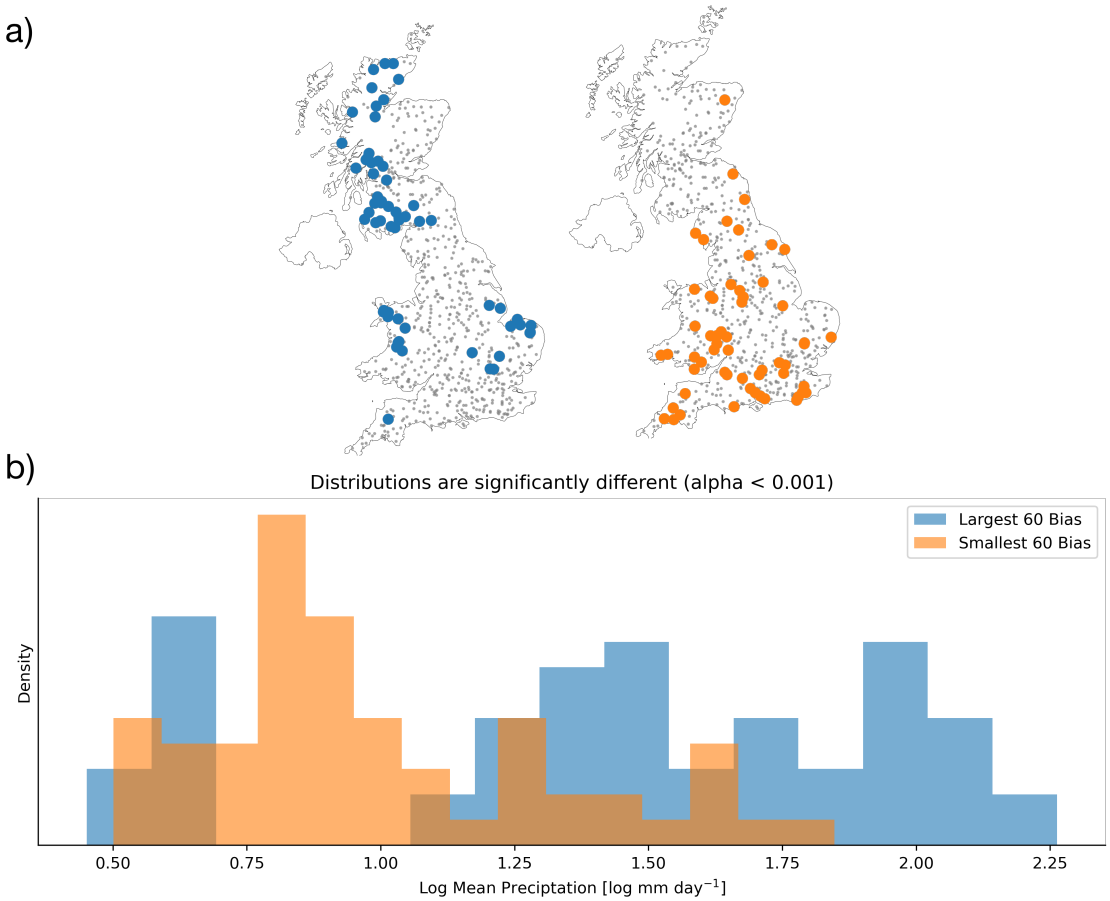


Figure 4.4 | a) The spatial location of the 60 catchments with the largest probe biases (blue) and smallest probe biases (orange) (b) The distribution of catchment log mean precipitation for the catchments with the largest and smallest probe biases. The largest 60 biases (blue) occur in catchments that are wetter or drier than average, as demonstrated by the gap in the middle of the distribution of mean catchment precipitations. By contrast the smallest 60 biases (orange) are mostly concentrated in the middle of the distribution, as we expect when using a simple linear model as the probe. These distributions are significantly different when using a 2-sample Kolmogorov-Smirnov test.

As we mentioned in Sect. 4.2.3, we centered and rescaled the ERA5-Land soil moisture data using the global mean and standard deviation. So a value of zero corresponds to the mean soil water volume across all catchments in the training period. This can be seen in catchment 15021, where the catchment specific soil moisture level is below zero (the grey dashed line is below zero), but the probe continues to predict values centered on zero. The dynamics remain well modelled (and therefore correlation scores are high),

but the probe identifies soil moisture anomalies rather than absolute values. Ultimately, we should expect this behaviour since we are fitting a single linear model with only one bias term. For alternative methods that model catchment-specific offsets, please see the experiment including a catchment-specific bias term (by including the gauge_id as input to the model, Appendix Sect A.8) and the experiments with a non-linear model (Appendix Sect. A.10). We return to this point in the discussion that follows.

4.3.2 Snow Depth Probe

Another process that influences river discharge is snow water storage. ERA5-Land offers a snow depth variable (m of water equivalent) that serves as a proxy of snow water equivalent. In order to determine whether the LSTM is representing snow processes in the state-vector, we use the probe analysis with ERA5-Land snow depth as our target variable.

Since snow processes are only significant in very few basins in Great Britain, we trained the probe on a subset of the basins shown in Figure 4.5c. These are defined as those catchments with a proportion of precipitation falling as snow greater than 5%. Figure 4.5b shows the probe output over one year for one station (Station 15025, Ericht at Craighall) and then Fig. 4.5a shows the probe output for that station and two other snowy catchments over the entire test period (1998 – 2008). They show clearly that the probe output correctly predicts very little variability in the summer period, when the snow will have melted. The winter peaks correspond to snow accumulation and snow melt. The median (over 33 catchments) Pearson correlation coefficient between the probe simulated snow depth and the ERA5-Land snow depth is 0.84, meaning that 84% of the snow variance can be described by the linear probe simulation.

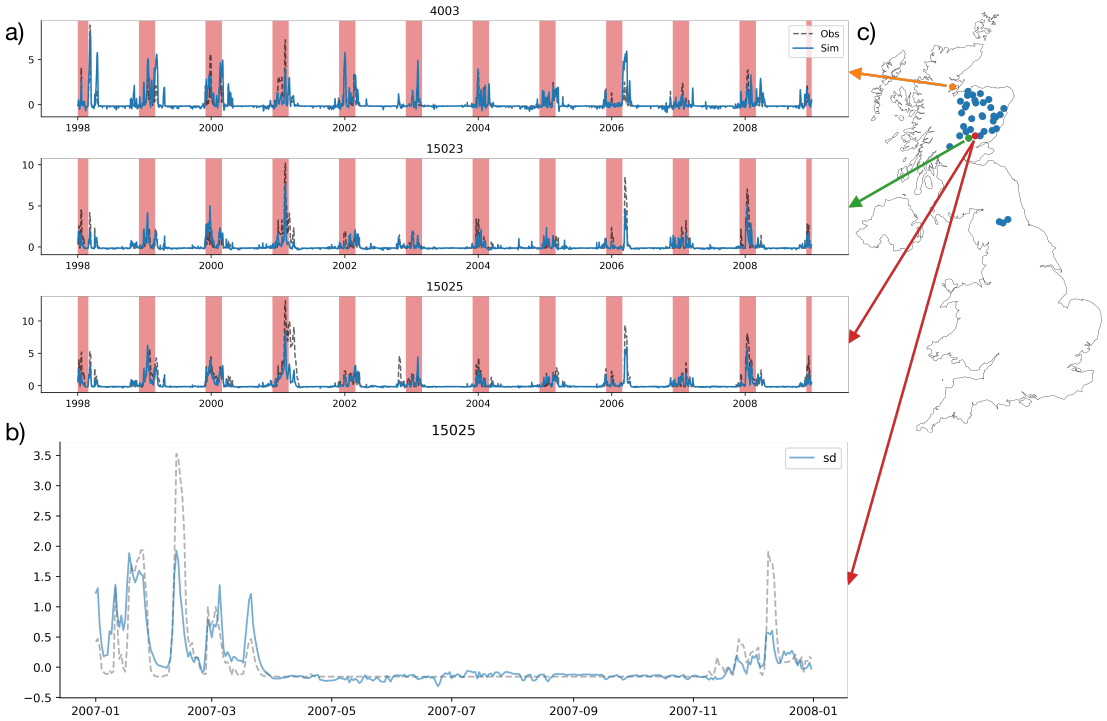


Figure 4.5 | (a) The probe simulation (blue line) plotted against the snow depth variable from ERA5-Land (black dashed line) showing the correlation between the cell-state dimension and the target variable for all years in the training dataset (1998–2008). The red-shaded regions represent winter months (December, January, February). Each subplot shows results for a selection of the snowy catchments as selected from (c). (b) Probe predictions for the snow depth target variable for a single snowy station, 15025 in the Cairngorms (the red point in (c)), for a single year 2007. (c) shows catchments with significant snow processes in Great Britain (defined as having a fraction of precipitation falling as snow greater than 5% using the CAMELS-GB frac_snow variable, which is the percentage of precipitation falling as snow) are concentrated in the Grampians range in North East Scotland and the Pennines in Northern England.

4.4 Discussion

4.4.1 The LSTM has learned physically realistic mappings

We find evidence that learned representations of catchment storages, as encoded by the LSTM state-vector, are predictive of a broad range of catchment hydrological storages. We have evidence that the LSTM, trained on a large dataset of 669 catchments, learns to model soil moisture processes and snow water processes internally. We have tested these findings with different initial conditions by initialising the LSTM with a different random seed, with different probe designs and with different data products. We find that our results are robust to these different setups. This is despite the LSTM never having seen these data, nor being constrained to model these processes. It is worth emphasising that the LSTM is trained to predict only discharge, using information from three meteorological drivers: temperature, precipitation and potential evaporation; as well as 21 static catchment attributes describing topographical conditions, climatological conditions, land cover types and soil texture (Lees *et al.* [2021b] Table 2). No direct information about snow accumulation and ablation, nor soil moisture is included in the training data. The LSTM has to learn these processes itself from the raw meteorological inputs and the catchment attributes, determining that these concepts are useful in solving its training task, i.e. predicting discharge. The states of the LSTM are learned through backpropagation, and therefore, these learned states are deemed useful by the model in the training process for minimising the error in the discharge signal. This finding offers evidence that the LSTM is learning a physically realistic mapping from inputs to outputs that corresponds with our understanding of the hydrological system. While there exists a large set of possible mappings from inputs to outputs, the LSTM converges on an answer that reflects our physical understanding. Although we only show results from one model initialisation here, these findings are robust to different initialisations (different random seeds).

We should not be entirely surprised by the finding that the LSTM has identified a physically realistic mapping from inputs to outputs. Neural networks are adept at learning the simplest solution to a given task. Since we are training the model to predict discharge in hundreds of basins across GB, the easiest solution is to learn the underlying physical relationships, since the alternative requires learning spurious correlations for all catchments. Nonetheless, by demonstrating that the LSTM learns physically realistic mappings we can imagine interesting opportunities for hydrologists to: (a) provide predictions of intermediate hydrological variables such as soil moisture and snow depth (b)

explore what information remains in the cell-states that has not already been identified as important for soil moisture or snow processes.

The former opportunity (a), that the probe analysis offers a means to predict intermediate variables, is interesting for two reasons. Firstly, the LSTM produces more accurate discharge simulations than any other hydrological model, and so we might expect that the intermediate variables are also better represented by the LSTM. Secondly, since the LSTM generalises to unseen basins [Frame *et al.*, 2021a; Kratzert *et al.*, 2019c], we might expect that the probe analysis also generalises to unseen basins. Exploring probe predictions in ungauged basins reflects an exciting area of future research. This could be done by learning the probe weights for certain basins, and then predicting on basins that have not been seen by the LSTM or probe before. The latter opportunity (b), that the LSTM allows us to learn something potentially "unknown", is an exciting aspect of using LSTM-based models for hydrological modelling. Exploring remaining values in the LSTM state-vector that have not already been assigned to a concept offers one promising avenue for identifying important processes or variables unaccounted for in the predictors (e.g. water management); and/or explaining the performance difference between the LSTM and traditional hydrological models. Once identified, we have the opportunity to incorporate this information back into traditional hydrological models. Alternatively, we can employ feature importance metrics, such as the integrated gradients method, to identify the signals that are most informative, and then reason about what these signals might represent. Such approaches may, for example, allow us to detect anthropogenic anomalies, such as reservoir operation decisions, changes in equipment, or biases in the observed data.

Since the LSTM is often the best performing rainfall-runoff model for discharge [Frame *et al.*, 2021a; Gauch *et al.*, 2020, 2021a; Kratzert *et al.*, 2018, 2019e], it makes sense to explore the soil moisture that the LSTM associates with a given level of discharge. Soil moisture is an important variable for understanding the runoff responses to precipitation events [Sklash & Farvolden, 1979], for assessing drought stress [Manning *et al.*, 2018], and for assessing the land-surface response to future climate change [Samaniego *et al.*, 2018]. We find that the LSTM is better able to model shallower layers of soil moisture than deeper layers (see Table 4.1). This likely reflects the fact that soil moisture in these shallower layers is more closely related to the discharge signal.

The probe analysis presented here represents one method for extracting intermediate hydrological concepts from the state vector of the LSTM. Our purpose of using such a probe is to interpret how information is processed by the LSTM. We recognise that we should not necessarily expect to use the probe output for more than interpreting the

Table 4.1 | Median catchment correlation scores over all 669 catchments for the training period for each soil water layer, 1 (0cm–7cm), 2 (7cm-29cm), 3 (29cm-100cm), 4 (100cm–289cm).

	Soil Layer 1	**Soil Layer 2**	**Soil Layer 3**	**Soil Layer 4**
Median Correlation	0.85	0.90	0.89	0.77

internal mappings of the LSTM. However, it is interesting to consider what catchment soil moisture anomalies the LSTM learns to associate with given discharge outputs. Fig. 4.6 shows the soil moisture anomalies as predicted by the probe.

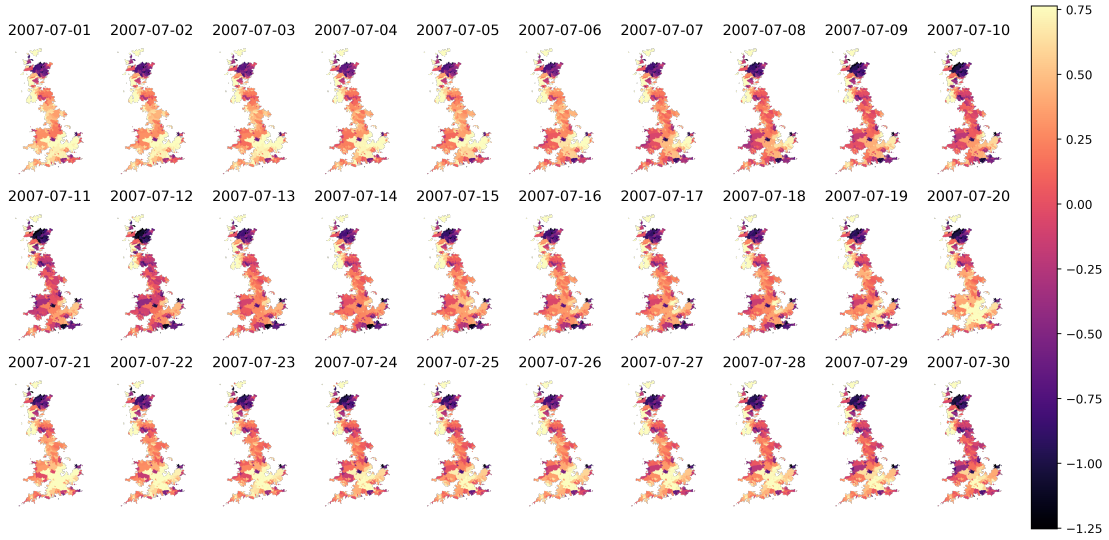


Figure 4.6 | The lumped (catchment averaged) soil moisture for 0–7cm at 30 daily timesteps covering most of July 2007.

We tested the probe against satellite derived soil moisture observations, using the blended active and passive ESA CCI Soil Moisture product to see if we get consistent results (Dorigo *et al.* [2017], see Appendix A.6). The results are similar to the results obtained when using the ERA5-Land data, with lower absolute correlation scores. Moreover, we see consistent performance declines when performing our two control experiments, i.e. shuffling the LSTM state vectors in space or shifting them in time (Appendix A.5), suggesting that soil information stored in the LSTM state vector is specific to each given catchment and timestep. We designed these control experiments to determine whether soil moisture time series can be reproduced from unrelated inputs. The results give evidence that the extracted signals are specific and therefore, that the results are unlikely due to chance.

Using alternative products such as ESA CCI Soil Moisture (see Appendix A.6), a blended

product combining multiple satellite-derived soil moisture estimates, comes with its own model assumptions and uncertainties. One key issue with remotely sensed soil moisture is that the measured soil moisture is restricted to the top-layer (5–10cm), and the depth of the satellite derived estimate is itself a function of the water content and surface roughness [Dorigo *et al.*, 2017]. Furthermore, it is difficult to measure and define catchment-scale soil moisture in the real world, but we are running our experiments using data lumped at the catchment scale. Neither reanalysis nor satellite-derived soil moisture reflects a true in-situ observation of catchment-scale soil moisture. Our approach was instead to let the LSTM learn the most effective mapping from inputs to outputs, and we interpret the intermediate state-vector, probing for the concepts that we expect to find given our expectations.

With respect to examining the learned representation of snow processes, one important aspect needs to be discussed. The snow probe was trained on catchments with a fraction of precipitation falling as snow greater than 5%. Training a linear probe, with a penalized squared error loss function (see Eq. 4.6), to predict snow depth on all 669 catchments caused the learned weights of the probe to be zero and no signal was detected. This is because only 33 of the catchments have significant snow processes. In contrast, when we trained the probe on catchments where we knew snow processes were occurring, we found that the snow processes were well captured by the variability in the cell state values. Note that the probe is still trained on the same LSTM. If we look at the cell-state values that had the largest weights in the linear probe, we can see that there exist snow-like signals which the probe incorporated into its representation of snow depth (Figure 4.7).

Our experiments outlined above explore whether known hydrological concepts (i.e. water stores) are captured by the LSTM. Our work here is the first step towards interpreting the internal dynamics of LSTM based models. Future work could consider how we might use feature attribution approaches (such as the integrated gradients method) to identify the most informative cell state values, or else examine the information contained within the states that do not correlate with snow or soil moisture processes.

4.4.2 The Catchment Biases in the Linear Probe

The probe predictions effectively model the dynamics of the soil moisture signals, however, they fail to predict absolute soil moisture levels, since they struggle to reproduce the catchment specific offsets that describe the mean catchment soil water volume. Two aspects of the training method for the probe need to be noted here. Firstly, we are train-

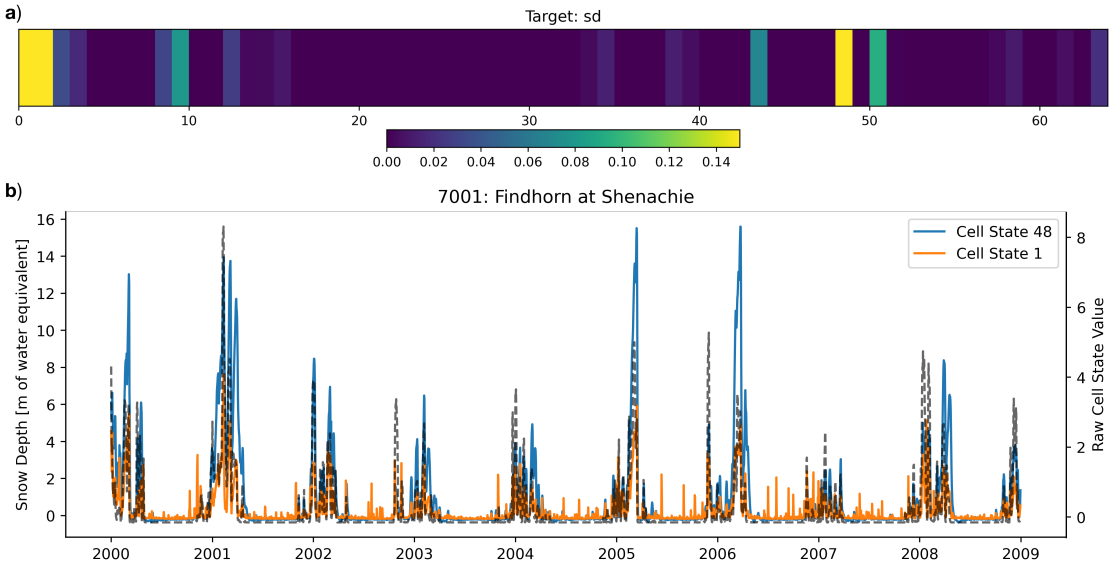


Figure 4.7 | (a) We visualise the most informative weights where lighter colours means larger weights assigned to that cell state dimension by the linear probe. (b) The largest 2 dimensions, in terms of their assigned weight, from the above plot are then over-plotted on the snow depth target timeseries (grey dashed line), showing that they really do contain information about the winter snow signal, particularly Cell State 48 which varies very little in the summer months.

ing one linear probe for all catchments. This means that we have a single set of weights that linearly map from cell-state to each soil moisture level in all of the basins. We tested training one probe on each catchment, which caused the offset to be well modelled by the unique bias for each probe (not shown). Thus the offset problem is not an issue when we train one probe on each catchment. The reason that we chose to use a single probe for all catchments was that we expect the learned concept of the soil and snow processes to be invariant over space, i.e. that the relationships between discharge, precipitation, and intermediate hydrological stores (snow and soil moisture) are encoded in the state-vector in a way that can be easily extracted by our probe.

In order to train the linear probe, we calculate a global normalisation, such that each catchment soil moisture is a value relative to the GB-wide mean. Since we are training a linear model, we should expect the observed behaviour of our probe, which captures the correlation and dynamics of the signal well, but learns a single "intercept" or bias-term for all catchments. This is because the optimisation of a linear model is a convex optimisation problem, where there is one global minimum. The model has a single intercept (bias) parameter for all catchments [James *et al.*, 2013]. The chosen bias term is the one that minimises the residual sum of squares, and this is the mean of the training target data (the probe target, such as soil moisture volume level 1). In Appendix Sect A.8 we

explore how incorporating a catchment specific intercept allows the linear model to accurately model the mean catchment specific offsets. We do this by augmenting the state vector with a one-hot encoding of gauge IDs. Furthermore, a non-linear probe is also able to more effectively model mean catchment soil saturation conditions (Appendix Sect A.10). Therefore, the information can be extracted from the LSTM, however, a linear probe is not sufficiently powerful to do so. That being said, the changes in relative soil saturation levels are clearly well modelled as demonstrated by Fig. 4.3 and the high correlation scores in Table 4.1 and Fig. 4.2.

4.4.3 Probes offer a means of interpreting the learned representation of the LSTM

The probe analysis that we have undertaken here provides a means of interpreting the internal states of the LSTM. We are interested in extracting the information content of the LSTM state-vector, allowing for the fact that this information may be stored across multiple cell state values.

There is no reason why the LSTM should store that information in a single cell, rather than distribute information across multiple cells. In fact, the LSTM could also model a process as the difference of two (or more cells) or any other combination of multiple different cells. Indeed it is likely that information is distributed due to the process of dropout. Using dropout randomly sets certain weights to zero during training in order to prevent the neural network overfitting, preventing the co-adaptation of weights such that one weight "corrects" the influence [Srivastava *et al.*, 2014b]. However, this also means that the network can potentially learn the same process in two different places in the network, since the network must be robust to those weights being "switched off" when dropped out.

The benefits of using the linear probe are twofold. The first is that our probe is relatively inflexible, and therefore, we can be confident that the probe is not overfitting to the targets [Hewitt & Liang, 2019]. Indeed, we perform a number of further experiments to check for spurious correlations, as described in Appendix Sect. A.5. Our experiments demonstrate that the findings show the information from the LSTM state-vector is both catchment and time-specific, i.e. that the correlations are significantly better than a sensible benchmark, ensuring results are robust and unlikely due to spurious correlations. This is an important result because it ensures that we are finding meaningful correlations between the information in the LSTM state-vector and the soil moisture target variable. The second benefit of using the linear probe is that we can easily interpret the

probe weights. However, there are limitations of this approach too. The linear probe limits the obtainable information to certain forms of information storage. In theory, it is possible for the LSTM to use information linearly for certain flow-regimes, and then use it differently for other flow regimes, even within a single basin. Thus, it is possible that the linear probe is unable to extract relevant information from the cell states that are being used by the LSTM. To test this we could use a fully connected neural network or any other non-linear regression model (see Appendix Sect. A.10). However, using a more complex model as a probe also comes with its own downsides. We lose the interpretability that comes from inspecting the probe weights and also increase the chance of getting false-positive results [Hewitt & Liang, 2019].

An open question remains, what other information is captured by the state-vector values that are not already assigned to a particular hydrological concept? Exploring these remaining cell states offers the potential to identify the underlying reason for the increased performance of the LSTM. Given that soil moisture and snow processes are already included in most hydrological models, the improvement of the LSTM over the conceptual and process-based models is unlikely to be a result of these processes we explore here. It remains possible that the LSTM is better able to learn the complex interaction of these processes with catchment-specific information, however, it is also possible that these remaining cell states contain information that describes other processes such as anthropogenic impacts on the hydrograph including withdrawals, transfers and reservoir management rules.

4.5 Conclusions

LSTM-based rainfall-runoff models offer good hydrological performance, however, interpretation and exploration of the concepts and structures that these models have learned is still in its infancy. In this paper we have explored the information captured by the LSTM state vector using tools from machine learning interpretability research.

We use a linear probe to map the state-vector onto a target variable, and find that there is sufficient information in the state-vector to represent the temporal dynamics in both soil moisture (at different levels) and snow depth from commonly applied data products. The state-vector of the underlying LSTM is trained only on meteorological forcings, static catchment attributes and asked to predict discharge.

Ultimately, these results suggest two key conclusions. First, the LSTM is learning a physically realistic mapping from meteorological inputs to discharge outputs. The con-

cept of a soil store and snowpack is encoded in most conceptual hydrological models [Beven, 2011b]. We therefore have evidence that in the UK the LSTM is learning to get the right results for physically-plausible reasons.

Finally, the conceptual approach that this paper has taken, using a linear probe, offers an effective method for extracting information from LSTM state vectors. Wherever LSTMs are applied, there is the possibility of exploring c_t in a similar way. This method has been applied in natural language processing, but as far as we are aware there are no applications of this method to LSTMs in Earth systems sciences.

5 Deep Learning for Vegetation Health Forecasting

Contributions This chapter is largely based on the following publications [*]

T Lees, G Tseng, SJ Dadson, C Atzberger, Alex Hernández, and S Reece 2019. *A Machine Learning Pipeline to Predict Vegetation Health*, **ICLR 2020 Workshop: Tackling Climate Change with Machine Learning**, Ethiopia, ICLR 2020 Workshop.

T Lees, G Tseng, C Atzberger, S Reece, and SJ Dadson, 2021. *Deep Learning for Vegetation Health Forecasting: A case study in Kenya*, **MDPI Remote Sensing**, https://doi.org/10.3390/rs14030698.

Abstract. East Africa has experienced a number of devastating droughts in recent decades, including the 2010/2011 drought. The National Drought Management Authority in Kenya relies on real-time information from MODIS satellites to monitor and respond to emerging drought conditions in the arid and semi-arid lands of Kenya. Providing accurate and timely information on vegetation conditions and health - and its probable near-term future evolution - is essential for minimising the risk of drought conditions evolving into disasters as the country's herders directly rely on the conditions of grasslands. Methods from the field of machine learning are increasingly being used in hydrology, meteorology and climatology. One particular method that has shown promise for rainfall-runoff modelling is the Long Short Term Memory (LSTM) network. In this study, we seek to test two LSTM architectures for vegetation health forecasting. We find that these models provide sufficiently accurate forecasts to be useful for drought monitoring and forecasting purposes, showing competitive performances with lower resolution ensemble methods and improved performances over a shallow neural network and a persistence baseline.

[*]with the following author contributions. Conceptualisation: TL, GT. Data curation: TL, GT, CA. Formal Analysis: TL. Modelling: TL, GT, AH. Methodology: TL, GT. Visualisation: TL. Writing – original draft: TL. Writing - review and editing: TL, GT, CA, SR, SJD.

5.1 Introduction

Drought is estimated to be one of the world's most costly hazards, accounting for 22% of damage from natural disasters [Wilhite *et al.*, 2007]. Droughts impact social and natural environments around the world [Van Loon, 2015; Vicente-Serrano *et al.*, 2010]. For example, in the twenty-first century alone there have been severe drought events on every continent such as the 2010 Russian Drought [Spinoni *et al.*, 2015], the 2011 Horn of Africa Drought [Nicholson, 2014], the 2013-2014 California drought [Swain *et al.*, 2014], the 2015-17 Southern African Drought [Baudoin *et al.*, 2017; Muller, 2018], the 2005 Amazon Drought Zeng *et al.* [2008] and the 2003 European Drought [García-Herrera *et al.*, 2010]). The most recent IPCC report outlines that globally, agricultural and ecological droughts are expected to increase (low to medium confidence, Arias *et al.* [2021]).

Vegetation health is a key drought indicator, and forms an important component of many drought early warning systems (EWS), as reviewed in Rembold *et al.* [2019]. The European Anomaly hotSpots of Agricultural Production(ASAP) system provides timely information about possible crop production anomalies based on a time series of satellite-based biophysical indicators for food insecure areas [Rembold *et al.*, 2019]. The European Drought Observatory combines precipitation, soil moisture and river flow metrics with a measure of vegetation health anomaly, using the remotely sensed fraction of photosynthetically active radiation (FAPAR) [Cammalleri & Vogt, 2019; Svoboda *et al.*, 2016]. Similarly the African Flood and Drought Monitor uses a 30-day moving average normalised difference vegetation index (NDVI) to create a percent of normal index [Sheffield *et al.*, 2014] The Kenyan National Drought Management Authority (NDMA) uses the remotely sensed Vegetation Condition Index (VCI), measuring the NDVI anomaly, to distribute emergency funds to drought affected counties [Klisch & Atzberger, 2016].

Unlike key hydro-meteorological variables that are forecast using numerical weather predictions, such as rainfall, temperature and soil moisture, vegetation health is not routinely forecast as vegetation properties are poorly parameterised in numerical weather prediction models. That being said, a range of mechanistic crop models do exist. For example, STICS [Brisson *et al.*, 2003], or WOFOST [Van Diepen *et al.*, 1989]. These models were first developed in the 1960s and continue to be improved today. However, they usually require detailed description of species, management, site and soil conditions, not always available in regions where access to in-situ data is problematic. Therefore, in this paper we consider the utility of using machine learning methods for predicting vegetation health, derived from earth observation (EO) data.

Artificial Neural Networks (ANN) have shown the ability to model complex and highly

nonlinear systems in situations where data is abundant. They have been used in hydrology and meteorology since the 1990s [Dawson & Wilby, 1998; Wilby *et al.*, 2003]. Deep learning techniques have been more readily applied to hydrological contexts in recent years, likely a result of increased availability data and improved computational capacities [Nearing *et al.*, 2021b; Shen, 2018].

In this paper we combine earth observation data and globally available weather simulation data (reanalysis data) with deep learning methods in order to explore the efficacy of a neural network based drought forecasting tool. A similar approach using gradient boosted regression trees was tested by Nay *et al.* [2018]. They sought to predict the enhanced vegetation index using spectral information from MODIS, land use information and autoregressive information.

In this paper we test the Long Short-Term Memory (LSTM) used elsewhere in hydrological studies for rainfall-runoff modelling [Gauch *et al.*, 2021a; Klotz *et al.*, 2020; Kratzert *et al.*, 2019e; Lees *et al.*, 2021a] driven by antecedent weather (derived from reanalysis products), antecedent vegetation health conditions (derived from earth observation products) and pixel-attributes (such as topography and land cover types) to predict future vegetation health one month ahead. We demonstrate the performance of our deep learning method using Kenya as our case study. Kenya is a climatologically diverse country with geographically varying vegetation health contexts. Furthermore, agricultural drought is a particularly pertinent issue in Kenya, where 80% of smallholder farmers rely on rain-fed agriculture for subsistence [FAO, 2019], where herders in the arid and semi-arid lands directly rely on grassland abundance, and recent decades in the wider East African region have experienced severe drought events in recent decades [Nicholson, 2017].

Our objectives for this research were twofold:

1. To test LSTM based models, used in other hydrological contexts, for predicting vegetation health.

2. To explore the models and demonstrate they learn physically realistic patterns.

We focused on Kenya for three reasons: (1) environmental diversity, (2) previous work done with the national agency as a collaborative stakeholder and (3) the prevalence and strengths of droughts in the country. Due to the diverse climatic regimes and agro-ecological regions in the country, there are different regimes in which to test our model. From the arid and semi-arid North and North East to the humid shores of Lake Victoria in the East. Secondly, because of the operational use of EO-based drought indicators

(VCI) by the National Drought Management Authority (NDMA) [Klisch & Atzberger, 2016]. For timely reaction to emerging drought conditions, the NDMA requests fast information with short lead times. If we can give advance warning (i.e. information about future conditions) to decision makers they can make their decisions in a more timely manner and ultimately we can help improve outcomes for people experiencing droughts. Finally, it is a region of the world that has experienced damaging droughts in previous decades. This also means that there are drought conditions to test our results on. If we can demonstrate skill then these models can be used in a region of the world where the benefit of the models can be most felt. We demonstrate the usefulness of this approach in Kenya, however the data sets we use for inputs and as our remotely-sensed target variable are globally available, and therefore, could be applied elsewhere.

5.2 Materials and Methods

5.2.1 Study Area

We evaluate our approach in Kenya, for the region bounded by latitudes (-5.202, 6.002 degrees North) and longitudes (33.501, 42.283 degrees East). This covers an area of 580,367km^2 and contains diverse hydro-climatic regimes. The arid and semi-arid lands cover 80% of the country's landmass and see on average less than 500mm of rainfall per year [Network, 2013]. The semi-humid areas are mostly mid-elevation areas where productivity is highly variable with maize being the dominant crop grown. The highly productive humid regions cover both sides of the Rift Valley, the valley floor, the humid lowlands on the shores of Lake Victoria and a thin strip on the coastal areas in the South East of the country (Network [2013] - see Fig. 5.1b). Some coastal areas receive over 500mm and the Western and central areas receive over 700mm of annual rainfall. The mean annual NDVI closely follows this pattern, showing a thin strip of green areas along the coast, either side of the Rift Valley and around the shores of Lake Victoria (see Fig. 5.1a). In contrast, the dry Turkana Channel shows low mean NDVI conditions.

When we aggregate the conditions at a pixel level to different counties we can see the diversity in conditions across the country. In Fig. 5.2 we consider the seasonality of precipitation and NDVI for four different counties. Marsabit is a dry county in the North West of Kenya. Kiambu is a central county, adjacent and North of Nairobi. This is a highly productive region of Kenya. Homa Bay is on the shores of Lake Victoria and the wettest of the four counties we display below. Finally, Kwale is in the far South East corner of Kenya, a coastal county.

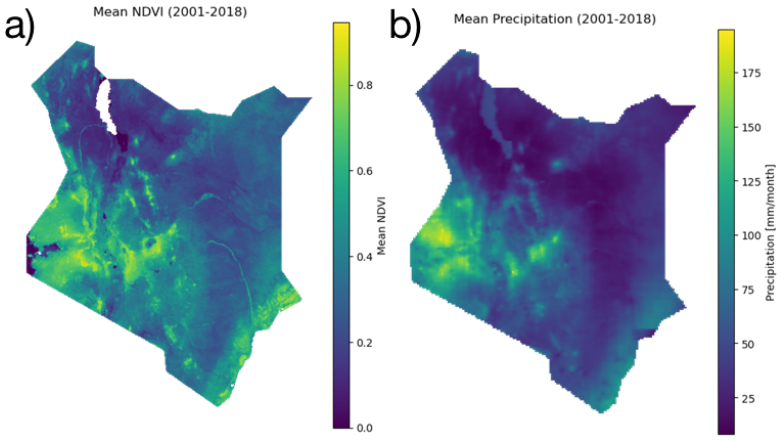


Figure 5.1 | Study region in Kenya shown in the raw spatial resolution of the underlying products, before preprocessing. (a) Colour scale shows the annual average of the spectral vegetation index, NDVI, calculated from MODIS Level 3 products, which is proportional to vegetation density. Lake Turkana has been masked from the image (b) Mean Precipitation. Colour scale shows mean monthly precipitation calculated from the CHIRPS dataset.

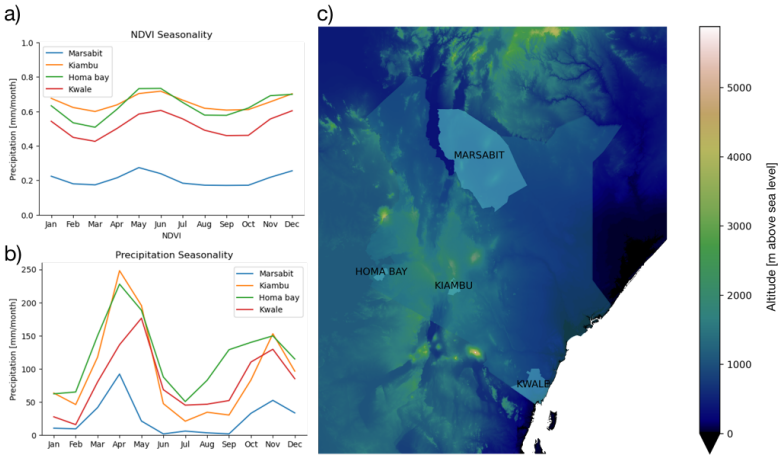


Figure 5.2 | (a) Mean NDVI conditions in four counties in Kenya (2002-2018). (b) Mean precipitation conditions in four counties in Kenya (2002-2018). (c) Spatial context of the four counties, Marsabit, Homa Bay, Kiambu and Kwale overlain on a map of the regional topography.

5.2.2 Data

Target Variable: Vegetation Condition Index

Remotely sensed vegetation has been used in many studies as a proxy for the agricultural drought conditions. The Normalised Difference Vegetation Index (NDVI) is a spectrally derived indicator of vegetation biomass and is commonly used for measuring agricultural drought conditions [Svoboda *et al.*, 2016]. The NDVI product used by the NDMA is derived from MOD13Q1 and MYD13Q1 NDVI collection 5 products and has been processed using the methods described by Klisch & Atzberger [2016]. These preprocessing steps involve the application of a modified Whittaker smoother in order to provide consistent, denoised images (removing clouds and poor atmospheric conditions) every 7 days. The data are processed to a 1km resolution and to a 250m resolution. Since we downscale the data to a lower resolution of the precipitation product, we use the 1km product here. This dataset is available from 2001 until today, for the analysis that follows we used 2001–2018.

The Vegetation Condition Index (VCI) is an anomaly index that ranks the observed NDVI for that pixel-time, when compared to all historical NDVI observations. The highest NDVI observation for that pixel-time in the historical dataset is given a VCI score of 100. The lowest NDVI observation for that pixel time is given a VCI score of 0. It is computed as in Kogan [1995]. In this way we are linearly scaling the observed NDVI between the minimum observed NDVI (0) and the maximum observed NDVI (100) for that pixel-time.

$$VCI_i = 100 * \frac{NDVI_i - NDVI_{min,i}}{NDVI_{max_i} - NDVI_{min,i}} \tag{5.1}$$

The VCI is calculated specifically for each pixel and each time. Therefore, the seasonality is removed, such that the VCI value reflects the current condition relative to prior conditions, scaled to between [0, 100]. One important thing to note is that values outside of the [0-100] range can be found if the observed NDVI is greater/smaller than the previous maximum/minimum value. We derived the VCI metrics, defining maximum and minimum NDVI values, from the timesteps in the historical period (2002 - 2015). This ensures consistency with the current approach used by NDMA.

To compare results with existing literature [Adede *et al.*, 2019] we have chosen to predict both the 3 month rolling average VCI (VCI3M), as well as the monthly VCI (VCI1M) one month ahead in time. The VCI3M metric reflects the most damaging agricultural droughts, those occurring over an entire season. Therefore, VCI3M is the metric used

for quantifying agricultural drought by the National Drought Management Authority (NDMA).

Input Variables

In Table 5.1 we list the variables that were included as input to the models. These variables are split into dynamic (X_dynamic) variables and static attributes for each pixel (X_static). We thereby condition the response of vegetation health (VCI) (y) to hydro-meteorological variables (X_dynamic) on static attributes (X_static) that don't change over time but vary spatially. The latter attributes include soil type and altitude as well as the mean conditions for each dynamic variable (hydro-meteorological signatures). The idea is that we communicate to the model how different locations will respond differently to dynamic forcings. Therefore, we are learning a mapping from the forcing variables (X_dynamic) to the target variable - VCI (y), which varies conditional upon the pixel attributes (X_static). These pixel attributes could, in principle, be measured anywhere globally, since the datasets that they are derived from are globally available. These input features were normalised by removing the mean and dividing by the standard deviation calculated in the training period for each pixel. In order to describe the changing conditions elsewhere we took a naiive approach. We provided the model with a spatial mean of each variable at that timestep. For example, there is one value describing the mean precipitation for each timestep across every pixel (*spatial_mean_precip*).

Table 5.1 | Variables included in the model. Note that we include 11 features for the month of the year, since including 12 would create features with perfect collinearity.

Feature	Model Usage	Source	Raw Spatial Resolution (prior to resmapling)
Precipitation	X_dynamic	CHIRPS [Funk *et al.*, 2015]	5km^2
2m Air Temperature (t2m)	X_dynamic	ERA5 [Hersbach *et al.*, 2020]	30km^2
Potential Evaporation (pev)	X_dynamic	ERA5 [Hersbach *et al.*, 2020]	30km^2
Evaporation (e)	X_dynamic	ERA5 [Hersbach *et al.*, 2020]	30km^2
Soil Moisture (4 Levels) swvl{1, …, 4}	X_dynamic	ERA5 [Hersbach *et al.*, 2020]	30km^2
Altitude	X_static	NASA SRTM [Farr *et al.*, 2007]	0.03km^2
Month of Year	X_dynamic	—	—
Vegetation Condition Index (VCI {1, ..., 3})	X_dynamic, y	MODIS Reflectances processed according to [Klisch & Atzberger, 2016]	1 km^2

ERA5 Reanalysis

ERA5 is ECMWF's most recent reanalysis product. Reanalysis data combine models and observations using data assimilation to provide the best estimate of hydro-meteorological variables over the Earth [Hersbach *et al.*, 2020]. ERA5 has global spatial coverage for 1979 till present at 137 pressure levels (vertical resolution) on an hourly time step. The spatial resolution is 0.31deg [Hersbach *et al.*, 2020]. For this study we use the spatial

fields for 2m Air Temperature, Potential Evaporation, Evaporation and Soil Moisture (Level 1 – 4).

As far as we are aware, ERA5 variables have not yet been validated over Kenya. ERA Interim, the precursor to ERA5 soil moisture has been shown to reproduce observed surface variability, however it overestimated soil moisture, especially in drylands [Albergel *et al.*, 2018]. ERA Interim Soil Moisture has been used by other studies in the region [Agutu *et al.*, 2021]. Tall *et al.* [2019] validated precipitation and incoming short-wave radiation in Burkina Faso using in-situ measurements. They found that ERA5 was better able to reproduce observations of both precipitation and incoming short-wave radiation than ERA Interim.

CHIRPS Precipitation

Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) is the precipitation product we used for our experiments [Funk *et al.*, 2015]. CHIRPS is a quasi-global (50^{o}S - 50^{o}N) daily precipitation product produced at 0.05deg spatial resolution. The data combines in-situ station observations and satellite precipitation estimates based on Cold Cloud Duration (CCD). The data has been validated against in-situ measurements and other gridded data products [Funk *et al.*, 2015]. Furthermore, it has been used extensively in drought-related studies in the region [Funk *et al.*, 2014, 2018; Uhe *et al.*, 2018].

NASA SRTM

We used the 1-arc second Shuttle Radar Topography Mission (SRTM) digital elevation model (DEM) to derive altitude as a static predictor variable in our experiments [Farr *et al.*, 2007].

5.2.3 Models

To forecast VCI, we utilise deep learning techniques often used by hydrologists for rainfall-runoff modelling, namely the Long Short Term Memory Network (LSTM) [Gauch *et al.*, 2020, 2021a; Kratzert *et al.*, 2018, 2019c; Lees *et al.*, 2021a]. These models can be considered data-driven state-space models [Kratzert *et al.*, 2019a], updating an internal state vector through time by incorporating new information at each timestep. The state vector is then used to make a prediction of the target variable at that time step.

The standard LSTM architecture was introduced by Hochreiter [1991], and has been described in a number of other hydrological contexts [Fang *et al.*, 2017; Kratzert *et al.*, 2018; Olah, 2016]. For the purposes of this paper we refer the reader to these references rather than repeating the standard LSTM conception here.

We also seek to test the Entity-Aware LSTM (EA LSTM) developed by [Kratzert *et al.*, 2019e], which fixes the input gate in time, only using static attributes (pixel specific properties) to determine which information flows from input data to the state-vector. This reduces the number of weights in the model and therefore should be considered a regularisation technique.

To contextualise and benchmark the LSTM based performances, we compare against a single hidden layer multi-layer perceptron (a fully connected neural network). This network is made up of an input layer, a hidden layer and an output layer. The hidden layer has a ReLU activation function and outputs a vector equal in size to the hidden size of the LSTMs (i.e. 64 values). We compare against this model as a simpler model architecture that does not take into account the inductive bias towards time-series data exhibited by the LSTM [Hochreiter, 1991].

We also compare model performances against a persistence baseline. This baseline predicts that the target month VCI (y_t) will be the same as the previous month VCI (y_{t-1}). This logic can be summarised as 'predict no change from the previous timestep'. We compare the models using both RMSE and R^2 for the monthly data in the hold-out test period (2016–2018).

5.2.4 Experimental Setup

All data has been spatially regridded to the same spatial resolution as the ERA5 product ($30 km^2$) using bilinear interpolation. The data is also temporally resampled to a monthly resolution. Therefore, we have one value for each variable, each pixel at each month. The analysis is conducted on monthly resolution data in order to account for the varying update frequencies of the input data and to ensure reasonable run times for Kenya-wide experiments.

We use the previous three months of dynamic data as input to our models. We use information from times t-3, t-2, t-1 as input to predict VCI at time t (one time-step ahead forecasting). We found that information prior to three months before the target time could be dropped without penalising performance, thereby speeding up training time and reducing the number of input features.

Models were trained using the smooth L1 loss function. The smooth L1 loss function,

also known as the Huber loss function, was used with $\delta = 1$. The smooth L1 loss is less sensitive to outliers than the mean squared error loss. This is because the error is linear when greater than δ, but squared between [-1, 1]. This has been shown to speed up training times and to prevent exploding gradients [Girshick, 2015]. Models were trained for a maximum of 100 epochs with a stopping criterion of 10 consecutive epochs without performance improvement (measured by the smooth L1 loss on the validation data).

$$\text{loss}(x, y) = \frac{1}{n} \sum_i z_i \tag{5.2}$$

where z_i is given by:

$$z_i = \begin{cases} 0.5\,(x_i - y_i)^2\,, & \text{if } |x_i - y_i| < 1 \\ |x_i - y_i| - 0.5, & \text{otherwise} \end{cases} \tag{5.3}$$

All models were trained on the period 2002 – 2015 and tested on the period 2016 – 2018. We split the training data into training and validation sets. We trained a single LSTM or EA LSTM model using data from all pixels. Therefore, we learn a single model for all of Kenya, testing the ability of these models to make predictions across diverse contexts. This is in contrast to other models that have sought to predict VCI in different regions of Kenya, which have traditionally refit each model to each location [Adede *et al.*, 2019].

5.2.5 Interpreting the Models

To provide a physical interpretation of model structure and results we took two approaches, detailed below. The first approach is similar to that taken by Kratzert *et al.* [2019e]. Using cluster analysis we interpret how the model has learned to utilise the static features - those features that vary over space but not time (Tab 5.1) - to group together similar vegetation health responses to the dynamic forcings. The second approach estimates the contribution of each input feature to the prediction for a given observation (x-y pair). This method is translated from game-theoretic approaches to machine learning [Lundberg & Lee, 2017] and gives a measure of feature importance with respect to a particular prediction.

Clustering Analysis of the Static Embedding Layer

The Entity Aware LSTM learns a mapping of the x_static data ($\mathbb{R}^{14}$) to the learned, higher-dimensional space ($\mathbb{R}^{64}$) (the "embedding") for each pixel. This reflects how the model prediction of VCI (y) is conditioned on the pixel attributes (x_{static}). The result is a set of weights (valued [0, 1]) controlling how much information is passed from the dynamic data to the cell memory ($C[t]$) passing through each cell of the LSTM sequence.

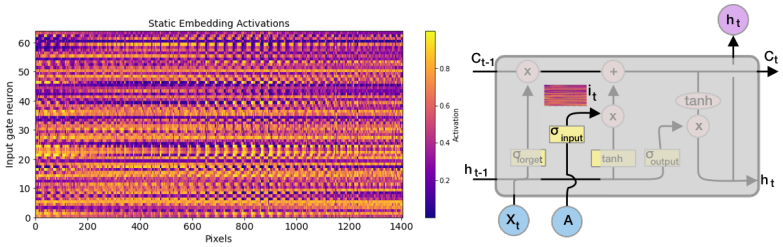


Figure 5.3 | The output of the static embedding from the EA LSTM input gate. Each pixel is represented as a column vector where the values ([0, 1]) describe how the model treats new information for that pixel. The schematic on the right shows a single EA LSTM cell and highlights the location in the model at which the static embedding is created. Schematic derived from Olah [2016]

What the model is learning therefore, is how the x_{static} data alters the response of vegetation health (VCI) to the dynamic data. This extends [Kratzert *et al.*, 2019e] by applying their methodology but in a different hydrological context. Rather than reflecting the grouping of hydrological catchments, we are grouping pixels to produce a data-driven classification of vegetation health regimes. These clusters highlight areas of similar vegetation health responses to dynamic forcing (e.g. precipitation).

By clustering the output of the static embedding layer ($\mathbb{R}^{64}$) we can visualise the areas where vegetation health responds similarly to inputs. These can be inspected visually to determine whether the spatial patterns that the model represents are qualitatively similar to known physical distributions of agro-ecological zones.

We test the hypothesis that the EA LSTM model is able to group together similar pixels. We expect that pixels with similar vegetation health responses use similar parts of the network.

Determining Feature Importance

We also estimate the relative importance of different pixel attributes. We calculate Shapley values [Shapley, 1953], which reflect the instance-wise contribution of each input feature to a specific prediction.

We first calculate the difference between the prediction for the specific instance we want to explain and a baseline prediction (the δy, or payout).

$$\delta y = \hat{y} - \hat{y}_{baseline} \tag{5.4}$$

We want to assign the δy to each of the input features (our dynamic and static input features). The output values can contribute both positively and negatively to the model's prediction.

We calculate Shapley values using the DeepLIFT method [Shrikumar *et al.*, 2017] implemented using the SHAP library [Lundberg & Lee, 2017]. Three inputs are needed to calculate the Shapley values:

1. a baseline output to compare our predictions to.
2. Our model prediction we want to explain.
3. The values for the features that we want to assign importance to.

For our baseline output we follow Lundberg & Lee [2017], and use the average of 100 sampled model outputs from the training dataset, as baseline model. Intuitively, we are asking: what would the model predict with no information, and how would the specific features in a datapoint change the model's prediction from the *no-information* prediction? For our analysis, a global sensitivity measure is calculated for each feature and each pixel by taking the average absolute Shapley value.

5.3 Results

5.3.1 Model Performances

Results comparing the LSTM approaches, our persistence baseline and a simple fully-connected feedforward neural network are summarised in Fig. 5.4 and Table 5.2. This shows the cumulative density functions of the models and the persistence baseline (BLINE). All three models were trained on the same input data and their performance evaluated on the hold-out test set, using all months from 2016–2018.

1. The LSTM and EA LSTM performances are extremely similar.
2. Both the EA LSTM and LSTM significantly outperform the persistence baseline.

3. The simple feedforward neural network performs very similarly to the persistence baseline.

4. All models predict the temporally smoother VCI3M better than VCI1M.

Table 5.2 | Average performance metric for each pixel in the domain and for each time in the hold-out test period (2016-2018). The best in class results are highlighted in bold.

	Error Metric	Persistence (BLINE)	Neural Network (LN)	LSTM	EA LSTM
VCI3M	RMSE	10.20	9.81	6.46	**5.88**
	R^2	0.86	0.88	**0.95**	**0.95**
VCI1M	RMSE	18.84	25.68	**13.23**	**13.23**
	R^2	0.66	0.14	**0.83**	0.82

The cumulative distribution functions (Fig. 5.4a) show the performance of the model RMSEs when predicting VCI3M. We can see that the LSTM models significantly outperform both the persistence baseline and the simple feedforward neural network. Interestingly the simple neural network seems to perform worse than the persistence baseline. This is shown by the long tail of model performances. The same is observed for R2 (Fig 5.4, b).

From the scatter plots shown in Fig 5.5, one can see that the predictions of the EA LSTM group are closely scattered around the 1-to-1 line, whereas both the LSTM and the simple neural network show some bias at low or high VCI values.

Spatial Distribution of Model Results

When we compare spatial performances against the persistence based model in Fig. 5.6 we can see that the EA LSTM performs well across Kenya. However, the baseline model fails in the highly productive regions, since the vegetation condition varies rapidly in time, such that the previous month is not as good a predictor of the following month in the more productive regions of Central Kenya, surrounding Nairobi, and near to Lake Victoria. Overall, the patterns show that the EA LSTM is able to significantly outperform the persistence baseline across Kenya.

In Fig. 5.7 we show model predictions for June 2018. What is interesting about June 2018 is that the persistence model significantly underpredicts the VCI. In contrast, with the information from the hydro-meteorological inputs, the LSTM based models were able to more accurately predict the degree of greening. The period starting in June 2018 was an extremely wet month across much of the country, with associated positive vegetation health anomalies [OCHA, 2018].

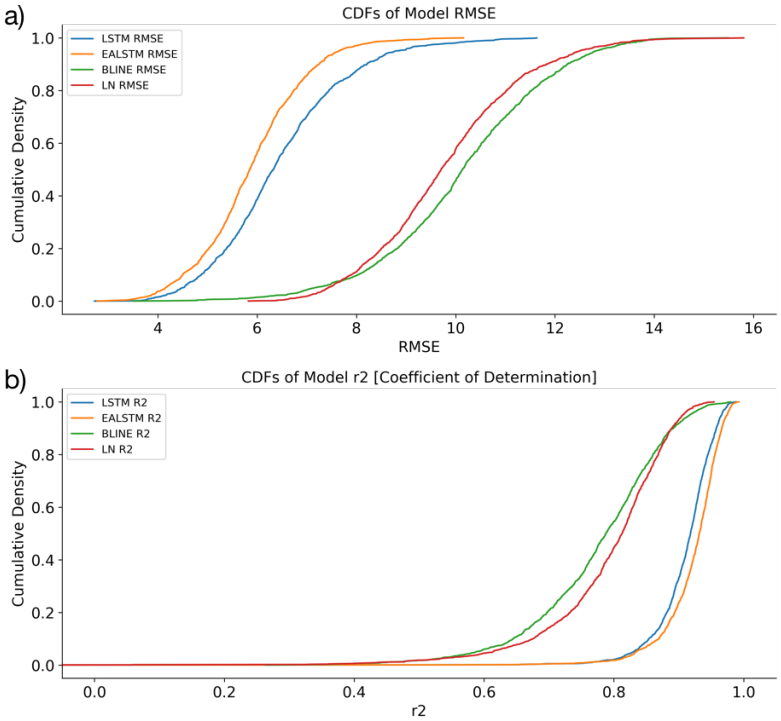


Figure 5.4 | Cumulative density functions of the VCI3M RMSE (a) and R2 (b) for the LSTM (blue), EA LSTM (orange), fully-connected neural network (LN - red) and the baseline persistence model (BLINE - green).

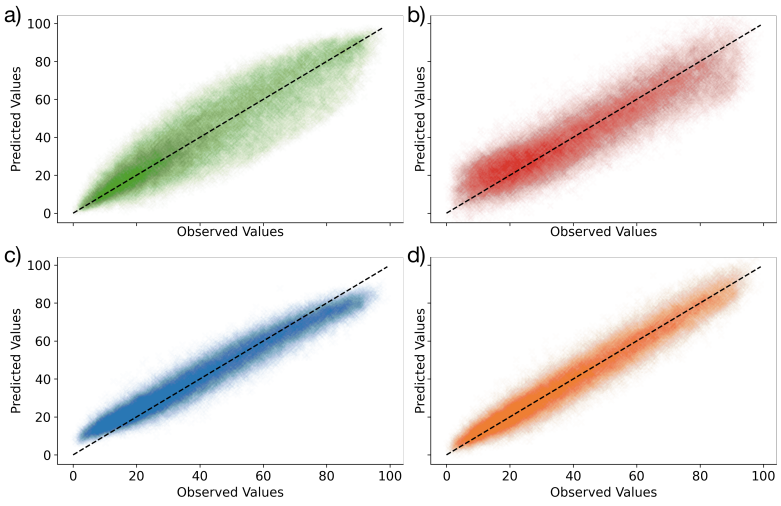


Figure 5.5 | Scatterplots of the observed VCI values against the predicted values (tested on hold out data from 2016–2018). The dashed line shows the 1:1 line. (a) shows the persistence baseline performances (green), (b) shows the linear network performances (red), (c) shows the LSTM performances (blue), (d) shows the EA LSTM performances (orange).

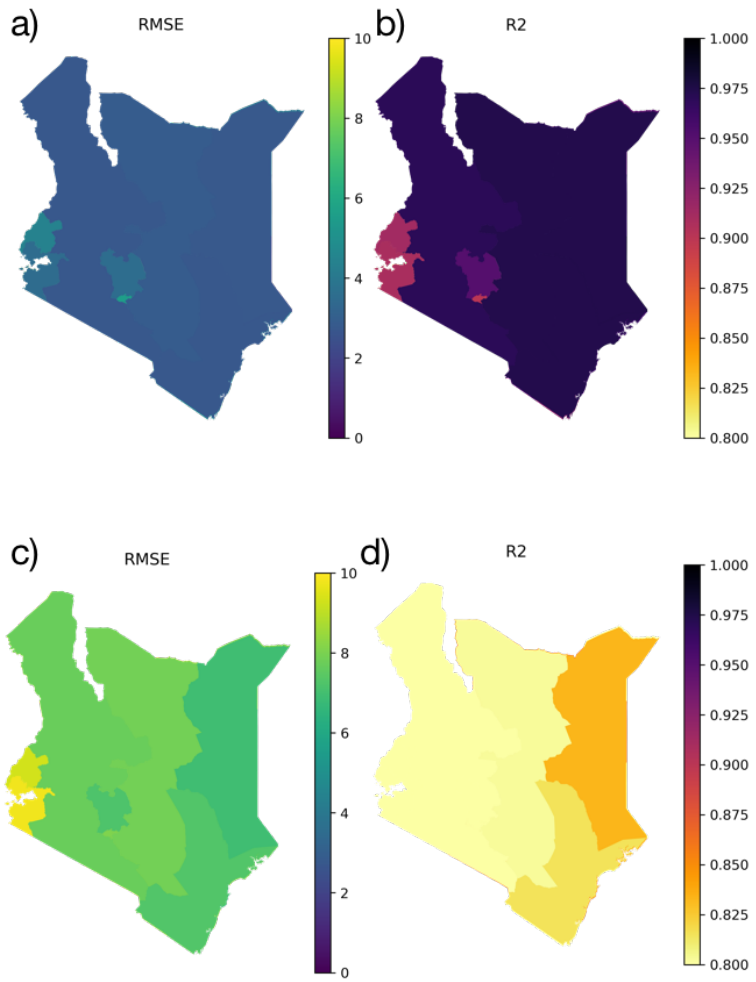


Figure 5.6 | (a) EA LSTM RMSE and (b) EA LSTM R2 averaged over each district in Kenya. (c) Persistence baseline RMSE and (d) Persistence baseline R2 averaged over each district in Kenya. Darker colours represent better performances. Note that LSTM errors are shown in the Appendices, but follow a very similar pattern to the EA LSTM errors.

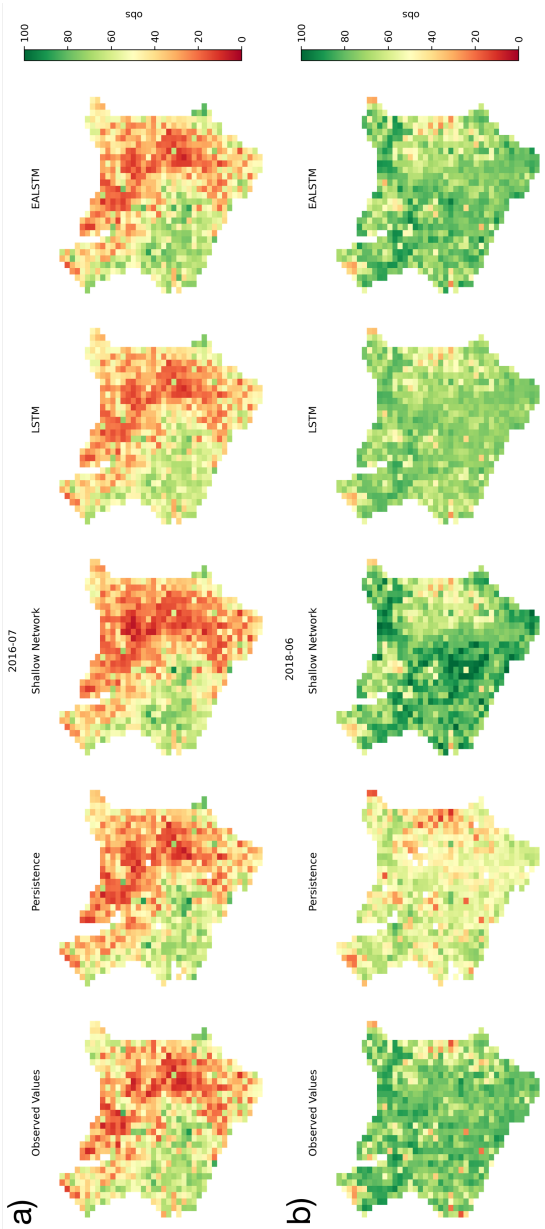


Figure 5.7 | Example VCI3M predictions on a pixel level for (a) and July 2016 a developing drought and (b) June 2018, showing rapid wetting and the associated vegetation response. All models were trained on identical data from 2001-2015.

Performance on drought classes

The models are trained as regression models, to predict the VCI as a continuous variable. However, when the VCI is used in operation by the NDMA and other drought agencies, the agencies are often interested in specific drought classes such as "extreme" or "severe" drought [Meroni *et al.*, 2019]. Drought classes and thresholds are also used in index insurances, where the correct modeling of the worst conditions matters most [Kenduiywo *et al.*, 2021]. To test the model performance across these drought classes we present the confusion matrix below, where we use the drought classes described by Klisch & Atzberger [2016], which are also used by the Kenyan NDMA. What we find is that the LSTM models are able to predict well across the distribution for VCI3M. The EA LSTM predicts the correct drought class with an accuracy of 78.3% and errors only occur between two adjacent classes. In the analysis that follows we only show the EA LSTM model performances.

The confusion matrix in Fig. 5.8 shows the EA LSTM performance across all drought classes. Overall, the EA LSTM is capable of detecting drought conditions one month ahead. Interestingly, the most critical drought class ("extreme") is predicted with higher accuracy compared to the other drought classes, except the class 5.

Table 5.3 | The Vegetation Deficit Index classes defined by Klisch & Atzberger [2016].

VCI3M Limits	Description	Value
$x < 10$	Extreme Vegetation Deficit	1
$10 <= x < 20$	Severe Vegetation Deficit	2
$20 <= x < 35$	Moderate Vegetation Deficit	3
$35 <= x < 50$	Normal Vegetation Conditions	4
$x <= 50$	Above normal Vegetation Condition	5

Comparison with state of the art

In addition to comparing against a persistence baseline, we compared our models to the NDMA model, developed by researchers at the NDMA [Adede *et al.*, 2019]. These published results were chosen as a useful comparison because we are directly predicting the same target variable, and because we are interested in the application of these methods to operational drought monitoring scenarios. Using indices derived from temperature, vegetation health, evapotranspiration, potential evapotranspiration and precipitation, the NDMA model consists of an ensemble of 111 fully-connected neural networks and

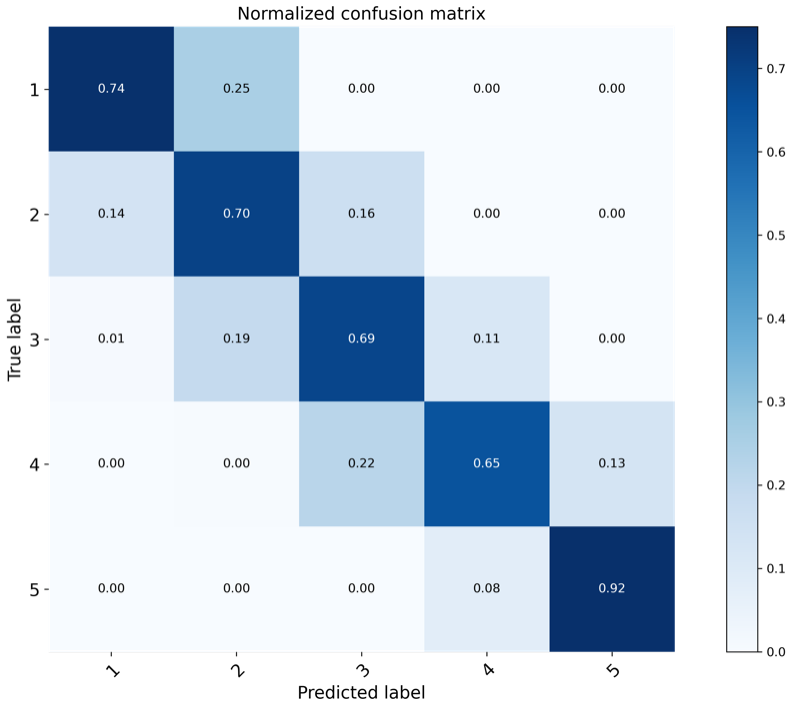


Figure 5.8 | Confusion matrix using data for the hold-out test set (2016-2018) showing the proportion of instances the EA LSTM predicts the VCI3M is within the bounds defined in Table 5.3.

support vector regressions.

Our results are presented in Table **??**. The only reported test results of R2 are from 2016–2017, and therefore, we only used those years and recalculated our VCI3M R2 scores. In addition, the NDMA model predicts a single VCI3M value per district and is trained on that district. To compare their model with ours, we took a district-wide mean of our pixel-wise predictions.

Our models are competitive with the NDMA ensemble model, but produce much more spatially granular predictions. This is particularly encouraging because during the 2016–2017 period being compared, significant droughts were recorded in Wajir [NDMA, 2017] and in Turkana and Marsabit Uhe *et al.* [2018]. These conditions represent the conditions in which we most want the models to perform well.

It is important to emphasise the differences between these models. The first is that Adede *et al.* [2019] train an ensemble of models, whereas we train one model. This offers the potential to provide confidence intervals, however the reported results show only the ensemble mean prediction. The second is that the spatial granularity of our predictions are much higher than the NDMA model, where predictions are produced at a district level. We therefore adapted our results to compare directly with the published

results. Finally, our model is applicable across the whole country, rather than being specifically trained for each of these four counties.

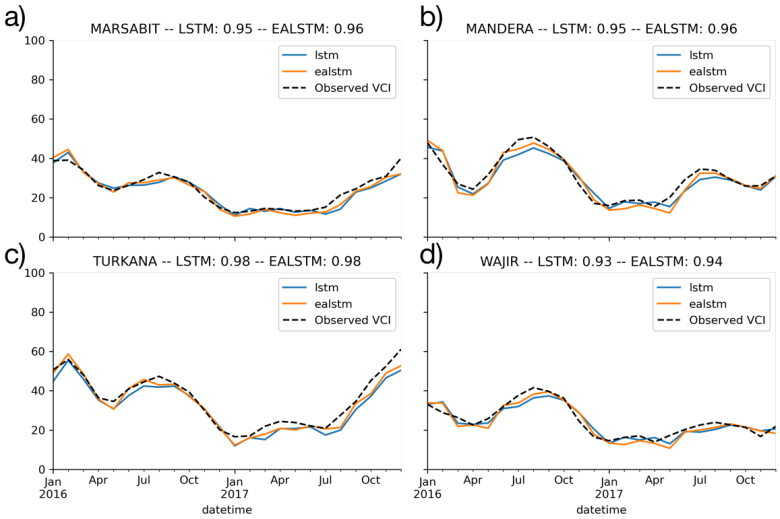


Figure 5.9 | Time series of VCI3M predictions for 2016-2017 in the four arid counties of (a) Marsabit (b) Mandera (c) Turkana and (d) Wajir. Observed values in black, the green line is the LSTM predictions, the orange line the EA LSTM predictions. Predictions were made one timesetep (one month) ahead.

5.3.2 Interpreting the Static Embedding

One appealing and unique feature of the EA LSTM is the static embedding layer. We learn a unique input-gate vector for each pixel in the data. These are the columns in the matrix in Fig. 5.3. Yellow colours indicate values close to 1 indicating that the cell is activated and all of the information from the dynamic data corresponding to this cell is contributing to the prediction of this pixel's VCI. These values are a function of the pixel's static data.

The static embedding permits us to test whether the model was learning physically realistic groupings of the pixels, based on their vegetation health response (VCI) to changes in hydro-meteorological variables. We do this by clustering the matrix in Fig. 5.3 using k-means clustering with a Euclidean distance criterion. This reduces the dimensionality of the $\mathbb{R}^{64}$ embeddings for each pixel to a lower dimensional representation. We chose k = 4 because there are roughly four (4) major bioclimatic zones in Kenya (Fig. 11 of Livelihood Zones p.15 or Fig. 5 of Agro-Ecological Zones p.10 [Network, 2013]). To test whether these groupings are physically realistic we plotted the pixel clusters geographically (Fig. 5.10).

In Fig. 5.10 we plot the output of the clustering of the learned model embedding. These clusters represent broad hydro-ecological zones in Kenya. The Turkana Channel region (shown in purple) is very arid and hot.

This method offers the opportunity to derive data-driven vegetation health regimes, showing the diversity of different vegetation responses to hydro-meteorological drivers. This analysis could be used to better understand the diversity of vegetation health responses in diverse hydro-climatic regimes in Kenya, and also help to identify areas that behave differently.

What this analysis offers most importantly, however, is evidence that the model is learning physically realistic groupings of vegetation health regimes. This offers an understanding that the model is not only an accurate predictor, but that it is making predictions for physically-plausible reasons that we can interrogate by visualising this learned matrix [Nearing *et al.*, 2021b].

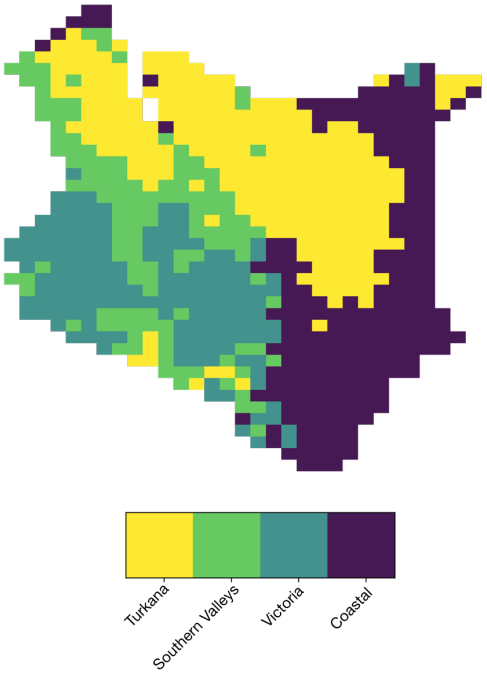


Figure 5.10 | The output of clustering the static embeddings from the Entity Aware LSTM (using k-means, where $k = 4$), plotted geographically. The color of the pixel corresponds with the cluster that the pixel has been assigned to. The map is relatively insensitive to the choice of k; there remain four broad clusters, and when $k = 3$ the coastal and Turkana region are grouped.

5.3.3 Measuring the contribution of Dynamic Features

We calculated the mean absolute shapley value [Lundberg & Lee, 2017] to give an estimate of the global feature importance. This allows us to identify which features are significantly contributing to the predictions.

The autoregressive component is extremely important for the models (Fig. 5.11). This reflects the behaviour of 3-month moving average vegetation health anomalies (VCI3M). They tend to persist from month to month, exhibiting a high degree of autocorrelation. However, the model significantly outperforms the persistence based model by incorporating new information from the other hydro-meteorological input features (X_dynamic) as shown in Fig. 5.4.

Fig. 5.11 shows that other dynamic features also contribute to the predictions, albeit a smaller absolute contribution than the autoregressive component. Precipitation and evaporation features offer additional information to the model beyond this autoregressive component. This makes sense, since the model will need to learn a rudimentary approximation of the water balance, using the precipitation to describe water entering the land-surface and evaporation describing water leaving the land surface. This learned water balance is useful for the EA LSTM to estimate vegetation health. It is also interesting that soil water volume - level 2 (swvl2) is the most informative soil water volume metric. Level 2 corresponds with water between 7 - 28cm.

Ultimately, these figures give us faith that the model is learning physically realistic signals, incorporating information from precipitation and evaporation to calculate the contribution of available water to the vegetation health.

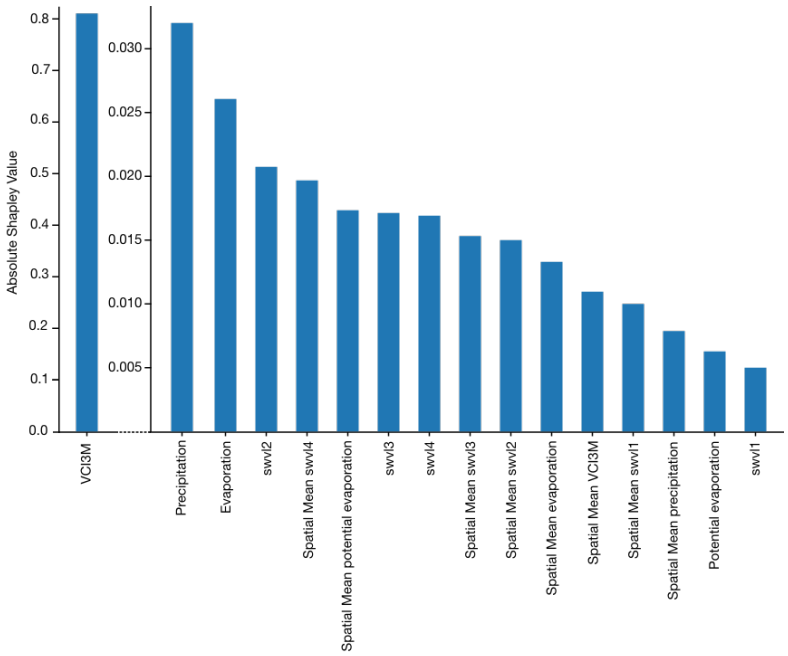


Figure 5.11 | The mean absolute Shapley value for each input feature, over all of the input times, t-3, t-2, t-1. Shapley values can be positive and negative, and therefore we use the absolute value to determine the overall contribution of that feature. The lagged target feature (VCI3M) is the autoregressive component in the model. The soil water volumes have been abbreviated to swvl{1, ..., 4}. The spatial mean features describe the mean value for all pixels in the training data at that timestep. In order to show all features on the same axis, and to compare the contribution from the non-autoregressive dynamic features, we have separated the autoregressive feature with a separate axis.

5.4 Discussion and Conclusion

In this study we have explored the efficacy of two LSTM-based architectures for vegetation health forecasting. We find that these recurrent neural network architectures outperform a shallow neural network and the persistence baseline. The most likely explanation for the improved performance of the LSTM compared with the shallow neural network is the inductive bias towards modelling time-series [Hochreiter, 1991]. The LSTMs are competitive with the published results for the ensemble of models developed by the NDMA [Adede *et al.*, 2019], while also having significantly greater spatial resolution. While there are difficulties in directly comparing between different methods, because of differences in input datasets, training periods and experimental designs, it is important to place these results in the broader context of previously completed research. We test our models across the whole of Kenya, and demonstrate the LSTM models are able to make accurate predictions across the country, despite the diversity of environmental conditions.The forecasts are also robust across the distribution of VCI values. We tested the model's ability to predict the VCI classes defined by Klisch & Atzberger [2016]. We are able to perform well across the distribution and also for the most relevant drought classes "severe" and "extreme vegetation deficit" that trigger payouts under NDMA's Disaster Contingency Funds and the Human Safety Net Program.

The results of this study demonstrate that methods used elsewhere in hydrology offer potential for vegetation health forecasting, and in turn offer opportunities for EO-based drought early warning systems.

We have also demonstrated that deep learning models offer the potential to combine information from earth observation data with meteorological variables (such as precipitation and temperature) from reanalysis data. These datasets are globally available, and demonstrate the power of deep learning techniques for deriving accurate forecasts of variables of interest (in our case vegetation health). This can be particularly valuable in cases where we lack the physical intuition to accurately model these variables using process-based models, as is the case for vegetation health. It is also valuable because it allows us to generate accurate forecasts in regions of the world where in-situ data is difficult to access or even non-existent. Moreover, it is easy to envision including either numerical weather forecasts or (user-defined) weather scenarios into our framework.

One of the key criticisms of deep learning methods is the fact that the models are often uninterpretable [Nearing *et al.*, 2021b]. We attempted to address this shortcoming by using two methods to determine what the LSTM-based models were learning.

The first was the clustering methodology proposed by [Kratzert *et al.*, 2019e]. Using the learned representation of the pixel attributes, we were able to produce spatial clusters of distinct vegetation health regimes. These data-driven clusters of vegetation health regimes are qualitatively realistic and provide evidence that the model is learning sensible relationships that can be interpreted physically. Secondly, we used gradient based methods to estimate Shapley values [Shrikumar *et al.*, 2017], determining the relative contribution of input features to a specific prediction. These show that the autoregressive component contributed the largest amount of information to the EA LSTM. It is clear that precipitation and evaporation also contributed, however, the importance of these variables was small when compared with the previous month's vegetation condition. Further research should consider how the sensitivity to different features varies with different lag-times and across space.

We were able to use the shapley values to make practical decisions about our experiments and model architecture. We experimented with feeding a longer time-series to the model, but using the calculated shapley values we determined that the LSTMs did not use information from more than 3 months before the prediction date. Reducing the length of the input time-series reduced training time without penalising performance.

Exploring methods for interpreting deep learning time-series models is an area of active research [Lim & Zohren, 2021]. In this study we have demonstrated two approaches that help us to understand what the model is learning and there is good evidence that these relationships are physically realistic. One might also want to determine the contribution of features over time, to interpret the seasonal patterns, and the contribution over space, to interpret the spatial patterns. Moreover, alternative feature sensitivity metrics exist, such as integrated gradients [Sundararajan *et al.*, 2017]. Future work should consider how robust feature importance scores are to different calculation methods.

There are a number of avenues for further exploration. Firstly, as we are most interested in drought conditions (VCI < 35) we can experiment with weighting these observations more heavily, forcing the model to pay more attention to these observations when performing the backward pass. Secondly, it would be interesting to explore the use of hydro-meteorological forecasts (rainfall and temperature) as inputs to the model. The use of (probabilistic or scenario-based) forecasts would potentially also offer the possibility of extending the forecast horizon beyond the current 1-month limit. This offers two benefits, firstly, it provides an estimate of model uncertainty for the one month forecasts. Secondly, we are able to make forecasts over a longer time horizon.

This study has demonstrated the ability of recurrent deep learning models to pre-

dict a widely used agricultural drought index (VCI) derived from earth observation data. This confirms and extends other studies that show LSTM models are effective for hydrological modelling [Fang *et al.*, 2017; Kratzert *et al.*, 2018, 2019e] and for combining the information from earth observation datasets with weather data [Shen, 2018]. We have explored two methods for interpreting these deep learning models and find that the results suggest we are learning physically realistic signals. Kenya has been struck by drought conditions in recent decades. In this manuscript, we have presented evidence that our approach has the potential to provide useful information to drought management authorities in a timely manner. Furthermore, because the method, datasets and overall approach is location agnostic, there is the opportunity to test a similar approach elsewhere.

6 Concluding Discussion

The primary aim of this thesis was to evaluate the application of a particular deep learning approach, the LSTM, to hydrological modelling, including both rainfall runoff modelling and vegetation health modelling. This is important for progressing our predictive and scientific goals, and this DPhil was able to both determine the accuracy of the LSTM in comparison with other models, and also evaluate different approaches for interpreting what information the LSTM has learned. This dissertation has sought to contribute to the rapidly growing body of literature that seeks to test new approaches from the field of machine learning to earth systems modelling, and has sought to develop new techniques for interpreting what these models have learned. The conclusion that follows first provides a summary of each chapter, bringing these findings together into a set of concluding remarks. Finally, this dissertation closes with an outline of future work.

6.1 Overarching Remarks

Through this DPhil I have explored how to utilise LSTM-based models to advance our predictive goals (getting accurate simulations) but also our scientific ones, using these models to derive insights about the underlying system. In this dissertation, I have sought to inspect the LSTM and therefore, continue the project of better understanding the mechanisms by which these models make accurate predictions.

In this dissertation, I have tested the LSTM, a general purpose recurrent neural network, in two different hydrological contexts, rainfall runoff modelling in Great Britain and vegetation health monitoring in Kenya. The LSTM has a number of characteristics that make it particularly suited to these modelling problems. The main characteristic that are encoded in the model architecture is the importance of an unobserved state and the ability to flexibly incorporate new information that arrives at each timestep.

This thesis has contributed to advancing the predictive goals of hydrological modelling. The LSTM and its variants offer more accurate simulations in unobserved test scenarios, when compared against traditional conceptual models, even when those conceptual models are evaluated in sample (Chapter 3). This performance improvement is robust to a variety of metrics, although there remains interesting geographical variation in the absolute performance improvement of the LSTM relative to the conceptual benchmarks. The same is true for the forecasting of vegetation health, the LSTM offers a flexible means to accurately combine diverse data sources to make useful predictions

of vegetation health. Large sample benchmarks in Great Britain offer a starting point for future modelling studies to compare results against.

This thesis has also advanced the scientific goals of hydrology, demonstrating different methods for model interpretability with LSTM-based models. Despite claims to the contrary, the information encoded by the learned LSTM can be usefully extracted and turned into human-interpretable concepts. Throughout the chapters, I have explored various methods for interpreting the LSTM including looking at characteristics that are associated with performance differences (Chapter 3), looking at the sensitivity of the outputs to different inputs (Chapter 5), clustering the internal embeddings of a more interpretable architecture (Chapter 5) and finally using a probe to map model weights directly onto hydrological concepts (Chapter 4).

6.2 Chapter summaries

6.2.1 Chapter 3: Benchmarking LSTMs in Great Britain

This chapter sought to test LSTM-based models for simulating rainfall runoff behaviours in Great Britain. The LSTM and the EA LSTM were tested against a suite of benchmark model performances from commonly used conceptual models, previously calibrated in a large sample benchmarking study performed by Lane *et al.* [2019]. The LSTM-based models were trained as a national model of Great Britain, with a single set of weights calibrated for all 669 catchments in the CAMELS GB dataset [Coxon *et al.*, 2020b]. Across a variety of metrics the LSTM outperformed the benchmark models, producing an average NSE score of 0.88 (LSTM) and 0.86 (EA LSTM) respectively.

Using the differences in model performances Chapter 3 sought to determine whether there are certain types of catchments that are consistently modeled better by a class of models (LSTMs or conceptual models). I found that the performance improvement of the LSTM was small compared with the conceptual models in wetter regions (WS, NWENW, NEE) and large in more arid regions, or areas with significant snow processes. As expected, part of the performance improvement in the SE can be explained by the conceptual models needing to close the water balance, when the underlying data does not have closure due to uncertainties in inputs and flow estimates. I also found that training an ensemble of LSTM based models with different random seeds produces estimates of predictive uncertainty. The resulting outputs suggested that the greatest uncertainty in predictions occurred in the South East of England, where there are groundwater processes and human abstractions.

The simulation results and catchment error metrics have been published on zenodo (https://doi.org/10.5281/zenodo.4555820, last accessed: 01/02/2022) to provide a benchmark for future work looking to develop a national rainfall runoff model.

6.2.2 Chapter 4: Concept formation in Hydrological LSTMs

After demonstrating the efficacy of regional LSTM-based models for simulating rainfall runoff processes in Great Britain in Chapter 3, Chapter 4 sought to better determine what the LSTM had learned about the hydrological system. One method of ML interpretability involves interpreting the weights of deep learning models, finding a mapping from those weights to human-interpretable concepts. Chapter 4 explored the idea of a "probe", a learned mapping from the internal weights of the LSTM state vector to known hydrological storages, snow and soil moisture.

Chapter 4 took the LSTM models from Chapter 3 and extracted the state vector for the test period, creating a dataset of the LSTM state vector and estimated hydrological stores for the associated catchments and times. The key innovation over previous work was that we do not assume that information must be stored in a single cell, rather we consider that information can be distributed across the cell state. This is important because there are no constraints forcing process information to be stored in a single cell. By comparing our predicted hydrological storages with the external data describing soil moisture and snow water content, we were able to demonstrate that the LSTM learns a physically realistic mapping from inputs (rainfall, temperature) to outputs (discharge), since the cell state contains information that closely corresponds to these hydrological stores.

The most important contribution of this chapter is the development of a general purpose framework to extract information from the weights in the LSTM cell state.

6.2.3 Chapter 5: LSTMs for Vegetation Health Forecasting

Chapters 3 and 4 evaluated the LSTM for rainfall-runoff simulation, determining that the LSTM is an effective model for this task, and then sought to interpret what has been learned by the model. In Chapter 5 we sought to apply the LSTM-based approach to a related system where the target variable is no longer a flux (discharge), but one of the stores (vegetation health). Satellite derived vegetation health is used as an indicator for drought conditions by the Kenyan National Drought Management Authority, and we apply the LSTM to this domain in order to evaluate the potential for this model to be used in an operational setting. I developed an open source pipeline during this DPhil

(https://github.com/esowc/ml_drought) for downloading the data, preprocessing the data and training the models. The LSTMs that we trained made predictions on a pixel-level for the whole of Kenya and was shown to outperform two benchmark models, demonstrating performances competitive with previously published approaches from the National Drought Management Authority.

Chapter 5 used two interpretability techniques to determine what the EA LSTM had learned. The first used the interpretability offered by the static input gate, clustering pixels by their responses to hydro-climatic inputs, and secondly using sensitivity analysis techniques to determine the relative contribution of input features to the target variable. Using this information, it was demonstrated that the EA LSTM learns to cluster pixels in a geographically meaningful way, separating the arid Turkana Channel from the Coastal region, the tropical Victoria region and the semi-tropical valleys of Eastern Kenya.

There is enormous potential for future work building on where this dissertation has ended, both in terms of evaluating forecast performance and also in terms of exploring model interpretability techniques. The next section offers a useful starting point for this future work.

6.3 Outlook & Future work

This DPhil has developed the application of data-driven models to hydrological problems. I have explored rainfall runoff applications, as well as operationally important drought monitoring situations. While progress has been made, this research can be developed further and I outline possible directions, some of which are being currently explored.

6.3.1 How do we encode physical theories within data-driven methods, and to what extent do they add value?

"*It is almost axiomatic that we need "physically based" models in order to make reliable predictions beyond the range of prior observations. However, the key question is not whether models of hydrologic systems should be physically based; instead, the question is how they should be based on physics. The physical laws governing water movement at small scales have been understood for decades. What we still don't understand well enough is how to apply these physical laws to systems that are complex, heterogeneous on all scales, and often poorly characterized by direct measurement.*" [Kirchner, 2006](p.2)

The general purpose deep recurrent neural networks tested in this DPhil have demonstrated state of the art hydrologic forecasts [Kratzert *et al.*, 2018, 2019e; Lees *et al.*, 2021a; Shalev *et al.*, 2019] and vegetation health forecasts (Chapter 5). It remains an open question as to whether the hydrological community can improve the performance, interpretability and usefulness of these models by encoding prior (physical) information into the architecture of data driven models. Designing and testing appropriate constraints for the DL models that encourage them to learn physically realistic mappings is a worthy area of further study, likely to prevent the finding of spurious local minima.

An interesting example of such an approach is the Mass Conserving LSTM [Hoedt *et al.*, 2021]. This model architecture encodes mass conserving properties, such that the cell state values in the LSTM can be interpreted as representing *mm* of water. This offers benefits for interpretation, where our probe based approach from Chapter 4 could be used to interpret the learned state values. However, it is unknown whether this aids performance, Hoedt *et al.* [2021] found that predictions at the tails were improved in the mass conserving LSTM compared with the LSTM, however, Frame *et al.* [2021a] found the opposite for out of sample extremes.

Another example of encoding prior physical information is utilising upstream measurements and forecasts in the prediction of discharge at a point, thus encoding the sub basins structure into the model architecture [Moshe *et al.*, 2020]. This architecture is shown to avoid overfitting by sharing weights across connected sub basins, and to improve performance in situations where data is scarce.

Recent advances in DL show how physics based model priors can be incorporated within the NN cost function [Thuerey *et al.*, 2021]. Furthermore, partial derivatives can be efficiently calculated through the NN backpropagation mechanism, offering the potential to combine DL approaches and traditional physically-based approaches.

6.3.2 Why is the LSTM outperforming traditional methods? What extra information is being encoded?

"*[O]ur 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*" [Nearing *et al.*, 2021b](p4)

This is perhaps the most important scientific question in hydrology in the coming years: how can we extract the learned information from the network in a useful way?

Chapter 4 explored a method for diagnosing what information has been learned by the LSTM. This led to the determination that the LSTM state vector really was learning

a hydrologically realistic mapping from inputs to outputs, determining that information needed to be stored over longer term time horizons in a mechanism consistent with snow and soil moisture processes. However, this information is already encoded in conceptual hydrological models. What other information is being utilised by the LSTM? How might we use this information to improve process information and understanding [Beven, 2020]? The probe technique applied in Chapter 4 offers the potential to explore remaining information stored in other cell states, moving us towards answering these questions.

An alternative approach may follow an approach developed for high energy physics [de Oliveira *et al.*, 2016]. Using a convolutional neural network architecture for image recognition, they calculated the pixels in an image that most caused neurons in the network to fire. They created a mask for those areas already predicted by their known physical laws. The remaining pixels that are important for the neural network but not described by the physical laws represent features that the network has found to be important for it's accuracy, but not currently used by their physical theories. The same approach could be applied to hydrology by focusing on the catchments and times where the physical laws as encoded in our process based models most diverge from our observations.

A possible example finding would be that the model has learned to diagnose inter-catchment transfers of water from the data alone. This deserves further attention. Combining the network architecture of [Moshe *et al.*, 2020], outlined in the previous area for future work, with further analysis of the model weights may help to find which catchments have significant inter-catchment transfers and the size of those transfers.

A further experiment that we can imagine given this research question is whether we can extract from the LSTM a simplified representation of the catchment, and explore how this simplified representation corresponds to the findings of other large sample studies looking at which particular models perform best in which scenario (e.g. MARMOT Knoben *et al.* [2019]).

6.3.3 How well do findings generalise to other geographical contexts? How can we maximise predictive accuracy in small-data scenarios?

Predictions in ungauged basins is a key challenge in hydrology [Blöschl *et al.*, 2019]. While LSTMs showed good generalisation capabilities in CONUS and GB when trained on data that is geographically proximate, often we want to make predictions in locations where complete datasets like CAMELS [Addor *et al.*, 2017] do not exist. For example,

good discharge forecasts would be useful to dam operators and water resource managers in Kenya. Is it possible to learn general hydrological patterns in CONUS and then make forecasts in Kenya when we might only have a small amount of data for fine tuning of model weights?

This touches on a larger question that this dissertation approached, but further work is required: what are the limits of DL in hydrological modelling?

The answer is that we do not know, yet we stand at the foothills of large potential advances in our understanding of the world we live in, and much of the evidence suggests that DL offers a powerful tool for predictive accuracy and knowledge extraction.

6.3.4 How can we use the LSTM for testing scenarios?

"*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 hypotheses. 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*" [Nearing *et al.*, 2021b](p5)

Since the LSTM offers a hydrological model with unparalleled accuracy, one avenue for further research is to test to what extent these models can be used to simulate different scenarios. The LSTM we trained in Chapters 3 and 4 both contain attributes, such as land cover types, assumed to be static over time. That assumption could be relaxed by including dynamic land cover variables. Once the model has been trained, one could test how the catchment responds to varying this land cover parameter. One possible experiment may involve comparing the results of an experiment run on a process based models, such as JULES. Buechel *et al.* [2022] explored the impact of changing tree cover on catchments in GB. It would be interesting to compare and contrast the findings from the LSTM with the physically based model.

6.3.5 How can we incorporate forecast information into LSTMs and use the architecture for making multi-timestep forecasts?

In Chapter 5 we explored the LSTM architecture for one-step-ahead forecasts of vegetation condition index. Architectures exist that stack multiple LSTM based models to incorporate different input datasets and pass encoded historical information to future

forecasts. Further research should consider these model architectures [Sutskever *et al.*, 2014] in the context of forecasting vegetation health and other hydrological systems. Providing a skillful forecast for the next 3 months could massively improve the ability of agencies like the National Drought Management Authority to respond in a timely manner to drought scenarios, thus improving outcomes with respect to drought natural hazards.

Related to this, implementing this algorithm in a real world setting in partnership with the National Drought Management Authority to test its usefulness is a necessary next step to apply this research.

This DPhil has tested a DL architecture, the LSTM, specifically designed for working with recurrent inputs, such as time series data. Importantly for hydrological modelling, this architecture can be interpreted as a state-space model, with a representation of the state of the hydrological system (the LSTM cell-state) and a data-driven process for determining the fluxes of information into and out of that state. This input-state-output relationship corresponds closely to our perceptual model of hydrolgoical systems, and the state-of-the-art performance of the LSTM empirically demonstrates the utility of this approach. Referring back to the original research questions, this DPhil has demonstrated that the LSTM can accurately model hydrological systems, characterised by a hidden state with long-term time dependencies. Furthermore, we have tested and demonstrated methods for extracting hydrologically relevant information from the LSTM, such as the ability to extract the most important features from the LSTM cell-state, which has shed light on what has been learned about rainfall-runoff and land surface systems.

Ultimately, the promise of DL models is that they can be used to make predictions in a wide range of scenarios, and that they can be used to extract knowledge from large samples of data. The LSTM architecture as tested in this DPhil is not constrained to make physically reasonable predictions, however, a physically realistic mapping from inputs to outputs is found by the LSTM. This likely reflects an inductive bias towards simpler solutions, and by finding a simple mapping that matches the data we likely find the physically realistic mapping. Future work needs to consider how we can encode our physical laws into DL models in a way that does not impede the ability to make accurate predictions. This is a key test for our theories, with performant LSTM-based hydrological models as a competitive benchmark for further development and improvement in hydrological modelling. Combining these two worlds of physically based modelling and DL approaches offers a powerful new chapter for hydrological modelling.

Appendix 1

Contributions This chapter corresponds to the Supplementary Information submitted with the following publication

T Lees, M Buechel, B Anderson, L Slater, S Reece, G Coxon and SJ Dadson, 2021. *Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models*, **Hydrology and Earth System Sciences**, accepted. Preprint 10.5194/hess-25-5517-2021

A.1 Comparison of the Train and Test Periods

The calibration (train) period and the evaluation (test) period are similar in terms of their predictability, although the evaluation period was slightly less predictable, as can be seen in the shifting of the two baseline model distributions towards lower NSE values (see Fig. 1). We used two baseline models to test how "predictable" the catchment hydrographs are in these two time periods. Climatology makes a prediction based on the mean discharge for that day of the year. Persistence is equivalent to predicting yesterday's value today, predicting the future will be the same as the past. Fig. 1 shows that the processes are largely stationary, and the period we use for calibration is similar to the period we use for evaluation. Indeed, the period we use for calibration is slightly easier to predict than the test period, since the benchmark models perform better, i.e. the distribution of catchment NSE scores is shifted towards higher NSE scores during the train period. Furthermore, the conditions for precipitation, PET, temperature and specific discharge are very similar between the train and test period. The temperatures have warmed slightly and there are slightly more days with zero precipitation, however, it is unlikely that such small changes have impacted the ability of the DL model to generalize. Discharge has risen slightly in the period of interest, across Great Britain.

A.2 Model Hydrographs

We illustrate the model predictions by showing the hydrographs for three stations from the Thames, the Severn and the Tay, as the largest rivers having at least part of their catchment in England, Wales and Scotland respectively.

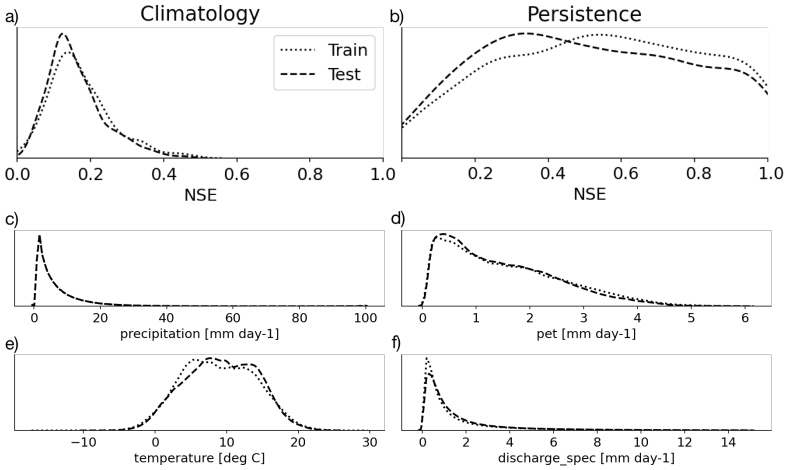


Figure 1 | Kernel Density Estimates (KDE) of NSE scores for two baseline models (above), Climatology (a) (calculated as the mean discharge for that day-of-year for each site) and Persistence (b) (calculated as the discharge shifted one day into the future, so yesterdays discharge is a prediction of today). Below, Kernel Density Estimates are provided for hydro-meteorological variables, precipitation (c), potential evaporation (pet) (d), temperature (e) and specific discharge (f) in the training period (1980–1997, dotted line) and the test period (1998–2008, dashed line). Climatology represents the mean conditions for that day of the year. Persistence reflects predicting yesterday's values today, i.e. predicting no change from yesterday. These give an overview of how "predictable" a time period is, since if these baseline models perform well, it will be easier to score at least as well as the baseline.

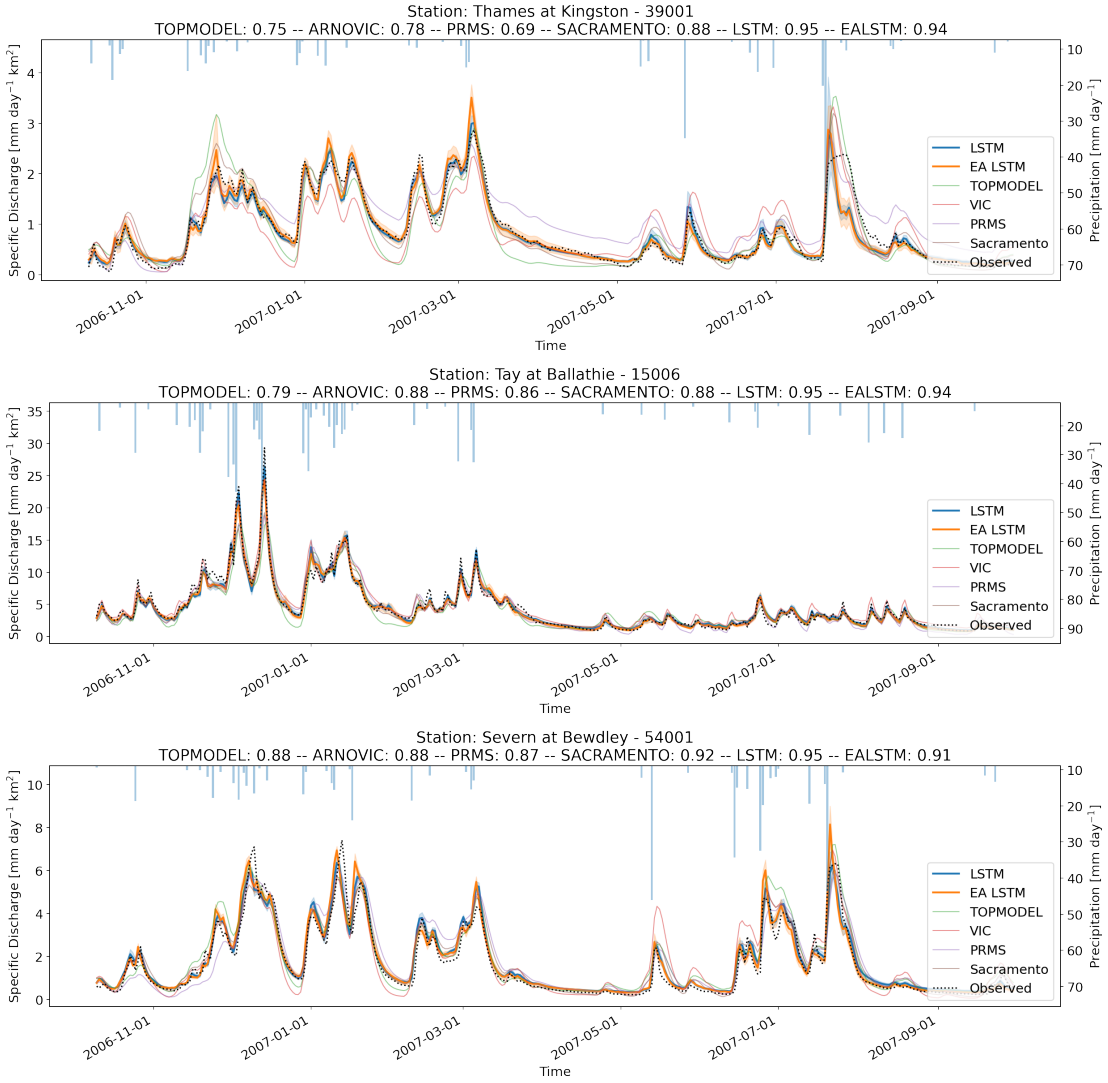


Figure 2 | Hydrographs for the Thames at Kingston (Station 39001), the Tay at Ballathie (Station 15006) and the Severn at Bewdley (Station 54001), for the hydrological year from October 2006 – September 2007. The model performances displayed in the header reflect the performance of each model on the entire test period (1998–2008), not just the displayed period. The observed discharge, from [Coxon *et al.*, 2020a], is shown as a dotted black line. The bars reflect catchment averaged precipitation with the axis shown on the right side. The LSTM and EA LSTM simulations are shown in blue and orange respectively. Conceptual model simulations for Sacramento (brown), VIC (red), PRMS (purple) and TOPMODEL (green) are taken from published timeseries from Lane *et al.* [2019].

A.3 Model Uncertainty

Uncertainty is present in all rainfall-runoff models. Model uncertainty has three main sources: (i) uncertainties in the observed data used to calibrate (train) hydrological models [McMillan *et al.*, 2010]; (ii) uncertainties in model structure [Fenicia *et al.*, 2014; Krueger *et al.*, 2010]; and (iii) uncertainties in model parameters [Arsenault *et al.*, 2014; Beven & Freer, 2001b; Gupta *et al.*, 2009]. Parameter uncertainty can be evaluated by using an uncertainty evaluation framework [Beven & Binley, 2014], often involving a sampling strategy. Model structural uncertainty is often estimated within multi-model frameworks, such as the Modular Modelling System [Leavesley *et al.*, 1996] or the Framework for Understanding Structural Errors (FUSE) [Clark *et al.*, 2008]. Uncertainties in observations can be estimated and accounted for by using multiple forcing products [Kratzert *et al.*, 2021] or by resampling the input data. This study addresses predictive uncertainty in the LSTM-based models by using an ensemble of 8 LSTM models trained with different random seeds, representing different starting conditions for the training process.

The results in the main text, unless otherwise specified, show diagnostic scores given the ensemble mean discharge. Here, we discuss the ensemble range and the uncertainty that this represents. The ensemble is produced by different random seeds, and therefore different starting parameters used during the training process. The mean catchment ensemble variability is 0.16 $mm^3\ day^{-1}$. The median is 0.12 $mm^3\ day^{-1}$. However, model uncertainties and their relationship with catchment attributes are in accordance with our hydrological intuition. For example, we see increasing uncertainty at increased streamflow (Fig. 3). Furthermore, by normalising for mean catchment discharge we can calculate ensemble standard deviation as a ratio of total discharge. This coefficient of variability is greatest in the South East of England (Fig. 4). A more principled treatment of uncertainty, which benchmarks various methods for using DL models to directly simulate a distribution can be found in Klotz *et al.* [2020].

A.4 Spatial Performances of Error Metrics

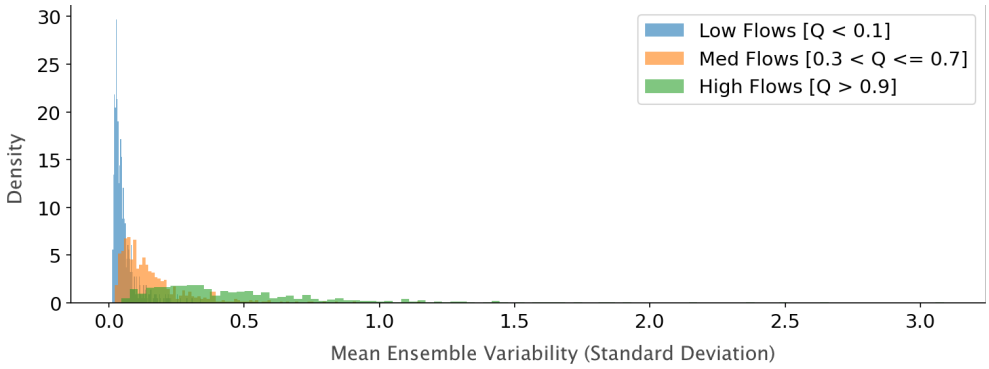


Figure 3 | Histogram of raw station averaged variability (standard deviation) across ensemble members. The blue histogram reflects the variability in simulations where observed discharge is lower than the 10th percentile The green histogram shows variability for only those times where observed discharge is greater than the 90th percentile The orange histogram shows variability for all times when the observed discharge is between the 30th and 70th percentile

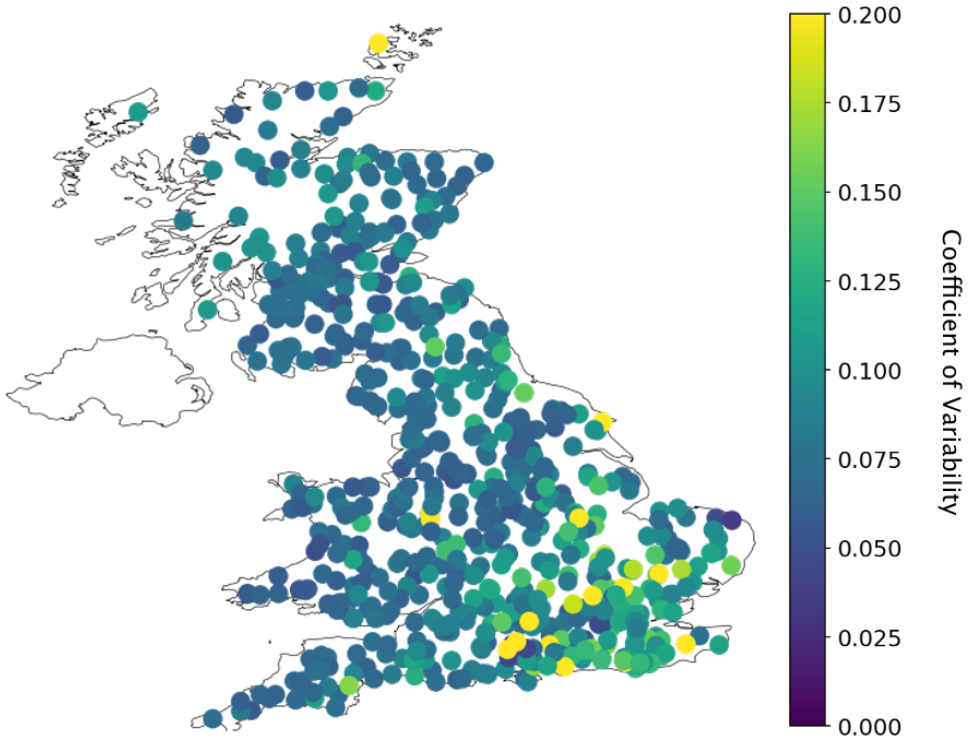
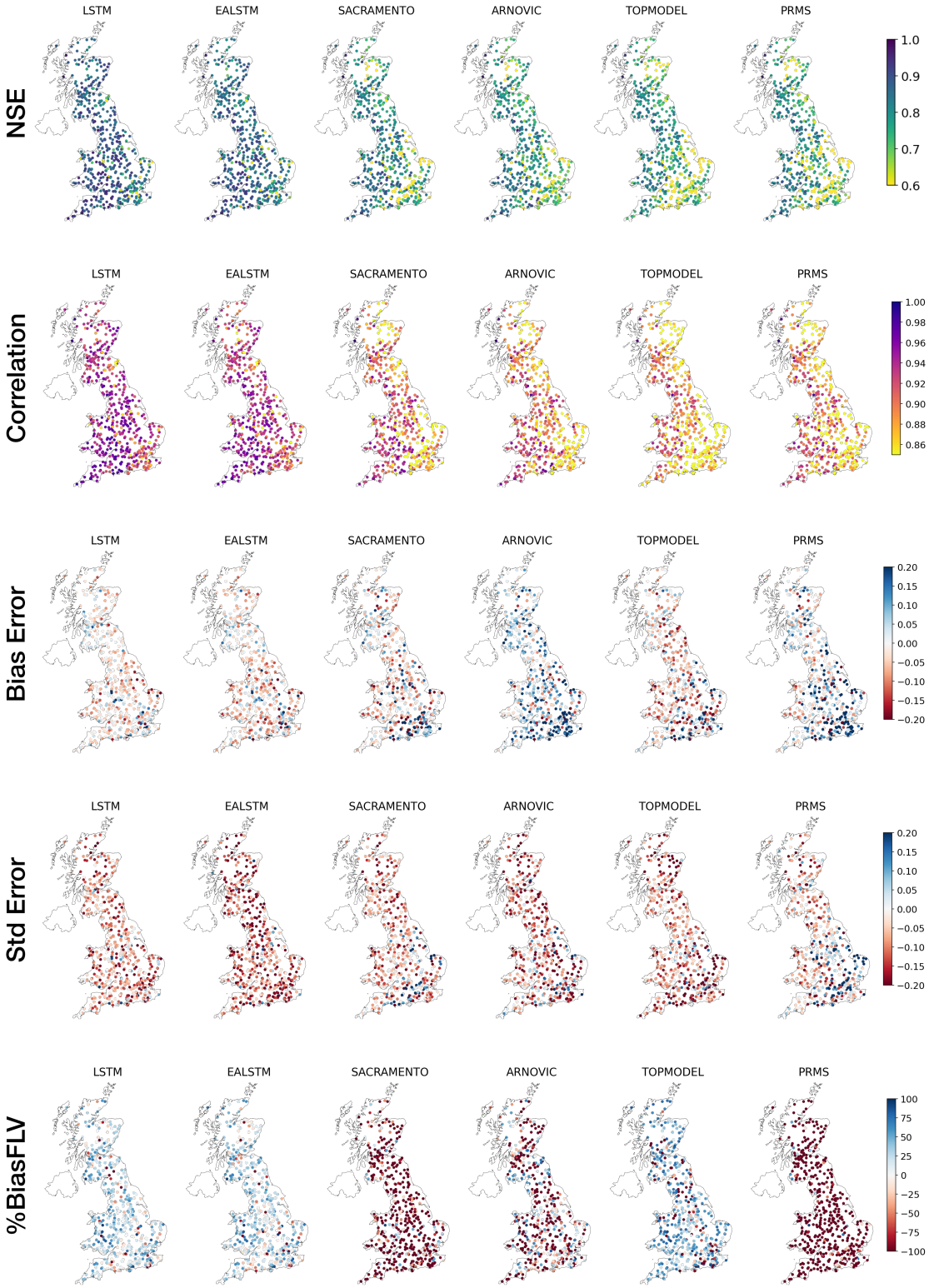


Figure 4 | Spatial Patterns of normalised catchment averaged variability (standard deviation) of ensemble predictions. Brighter colours reflect greater variability across members of the ensemble of LSTMs.

A.4. Spatial Performances of Error Metrics



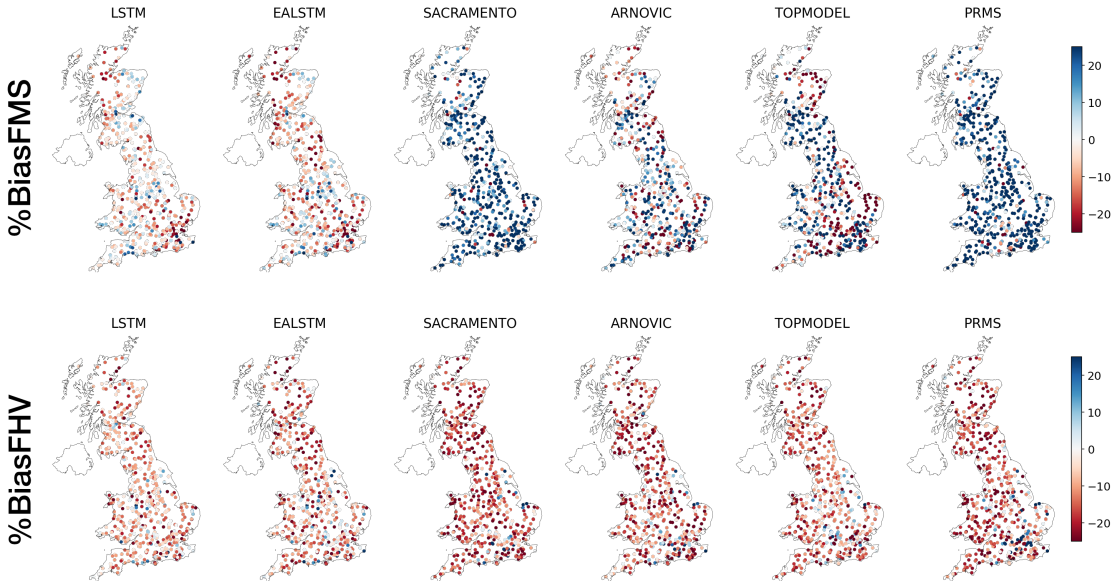


Figure 5 | Spatial Patterns of different performance metrics. Each point is a single station-gauge, and the point is coloured according to the performance metric. For performance metrics with a diverging score (above and below an optimum, e.g. Bias Error) more intense colours represent worse performance. Red represents an under-prediction, blue an over-prediction. For scores which are increasing (e.g. NSE, Correlation), darker colours reflect improved performance.

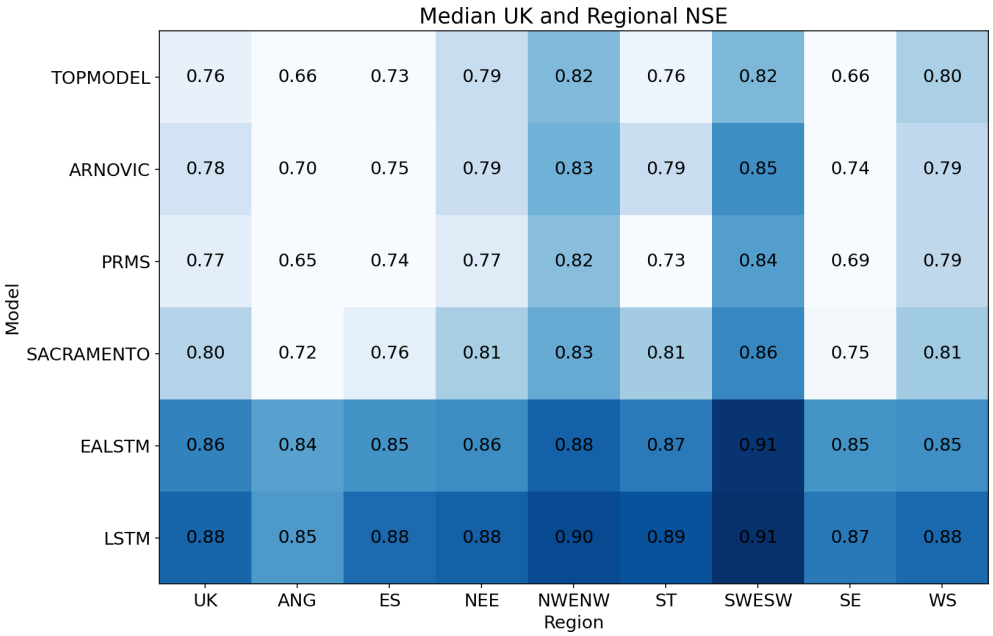


Figure 6 | Median NSE scores for eight Great Britain river basin regions. The regions are based on the UKCP09 river basins [Murphy *et al.*, 2009] aggregated from 21 river basin districts to eight regions. The leftmost column is the median score for all GB catchments, which is the same as in Table 3 in the main text. It is included here for reference.

Appendix 2

Contributions This chapter corresponds to the Supplementary Information submitted with the following publication

T Lees, S Reece, F Kratzert, D Klotz, M Gauch, J De Bruijn, R Kumar Sahu, P Greve, L Slater and SJ Dadson, 2021. *Hydrological Concept Formation inside Long Short-Term Memory (LSTM) networks*, **Hydrology and Earth System Sciences**, in review. Preprint 10.5194/hess-2021-566

A.5 Control Experiments

We want to be sure that the signals found by the probe are specific to a given catchment and time, and control for false positives (finding a strong correlation between the LSTM state variable and the intermediate target variable of interest). Testing for the probe confounder problem, which describes "when the probe is able to detect and combine disparate signals, some of which are unrelated to the property we care about" [Hewitt & Liang, 2019] requires that we ensure that our detected signals are specific to the catchments (spatial specificity) and times (temporal specificity). Our hypothesis is that the LSTM is learning information that is specific to soil-moisture and snow processes for a given catchment at a given time. In order to test this hypothesis, we designed two control experiments. The first involves spatial shuffling, where we take the LSTM state-vector from Catchment A and test whether the information content is sufficient to accurately model Soil Moisture in Catchment B. Spatial shuffling tests for spatial specificity and can be seen in Figure 7.

The second experiment shifted the data in time, testing for temporal specificity. We tested the performance of the probe with inputs shifted by 180 days, breaking the temporal link between inputs and outputs (Fig. 8).

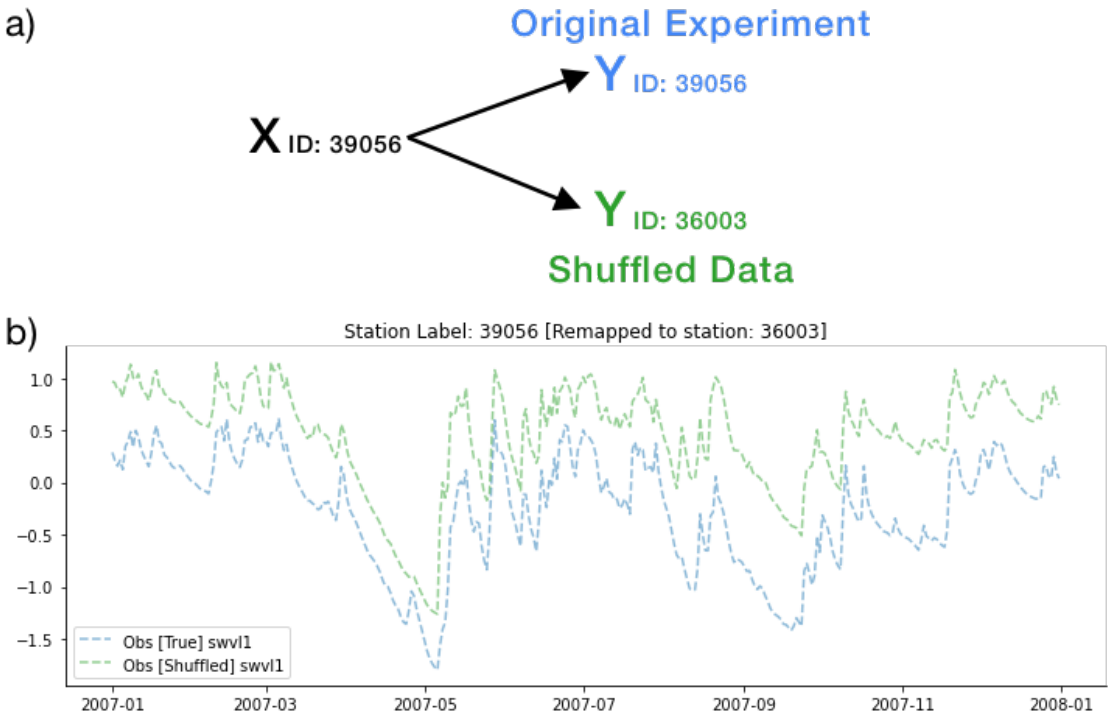


Figure 7 | (a) Conceptual diagram explaining spatial shuffling. In the original experiment, c_t from Gauge ID 39056 is linked to the time-series for that catchment 39056. The shuffled target instead asks the probe to detect the target variable time-series (belonging to Gauge ID 36003) using the c_t vector from Gauge ID 39056. (b) Example time series from shuffled basins, where station 39056 is the original experiment and station 36003 is the shuffled experiment. It is worth noting that the soil moisture measurements are highly correlated in space (shown by the blue and green lines following each other), and in some instances multiple basins may fall within the same ERA5-Land pixel, in which case the information content will be the same.

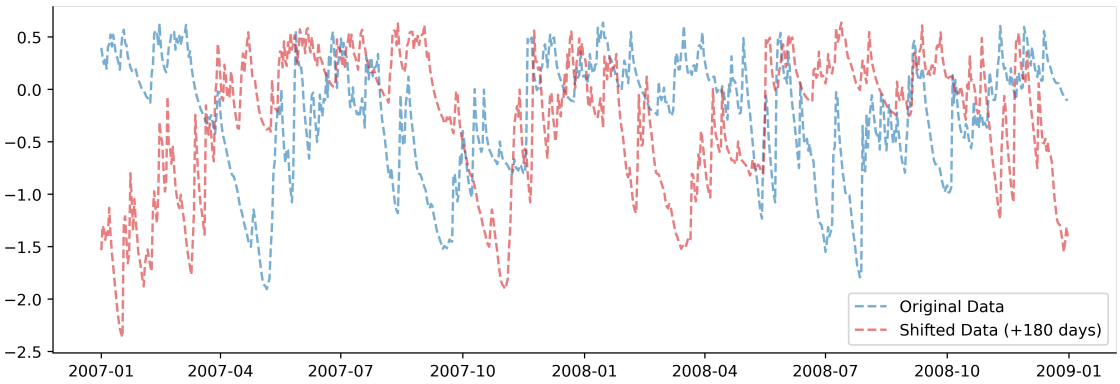


Figure 8 | Shifting the input data by 180 days in time breaks the temporal link between the inputs (shown here) and the target variables (not shown). We fit the probe on the shifted data to determine how much less information it contains when compared with the original data.

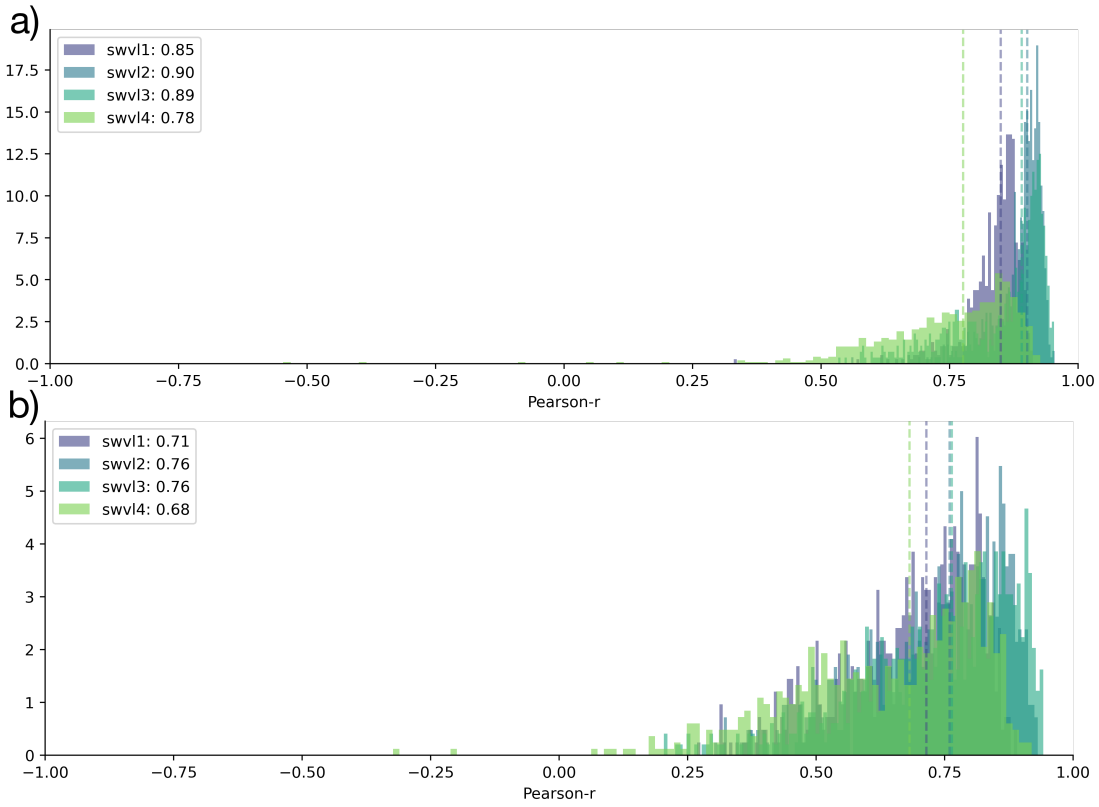


Figure 9 | Shuffling in space. (a) Histograms of the original experiments with unshuffled data, repeated from Fig. 4.2. (b) Histograms of catchment correlation scores after having trained the probes on data shuffled in space. The performances decline quite significantly when compared with the original experiments.

The results show that the probe performances declined for both experiments, shifting the inputs in time and shuffling the inputs in space (Figure 9. 10). This suggests that the information captured by the LSTM state-vector is specific to the catchment and time. The original experiments had correlation scores of: 0.88 for swvl1; 0.90 for swvl2; 0.90 for swvl3; and 0.84 for swvl4 respectively. When shifting in space, these declined to 0.72, 0.76, 0.76, 0.68 for each target variable. When shifting in time, these declined to 0.39, 0.26, 0.63, 0.50. It is likely that the larger performance drop for shifting in time is because of the high degree of spatial correlation in the catchment-averaged soil moisture time series. Ultimately, these experiments give us a high degree of confidence that the observed correlations are unlikely due to chance, and that the information that the probe is extracting from the LSTM state vector reflects the hydrological variables we compare against.

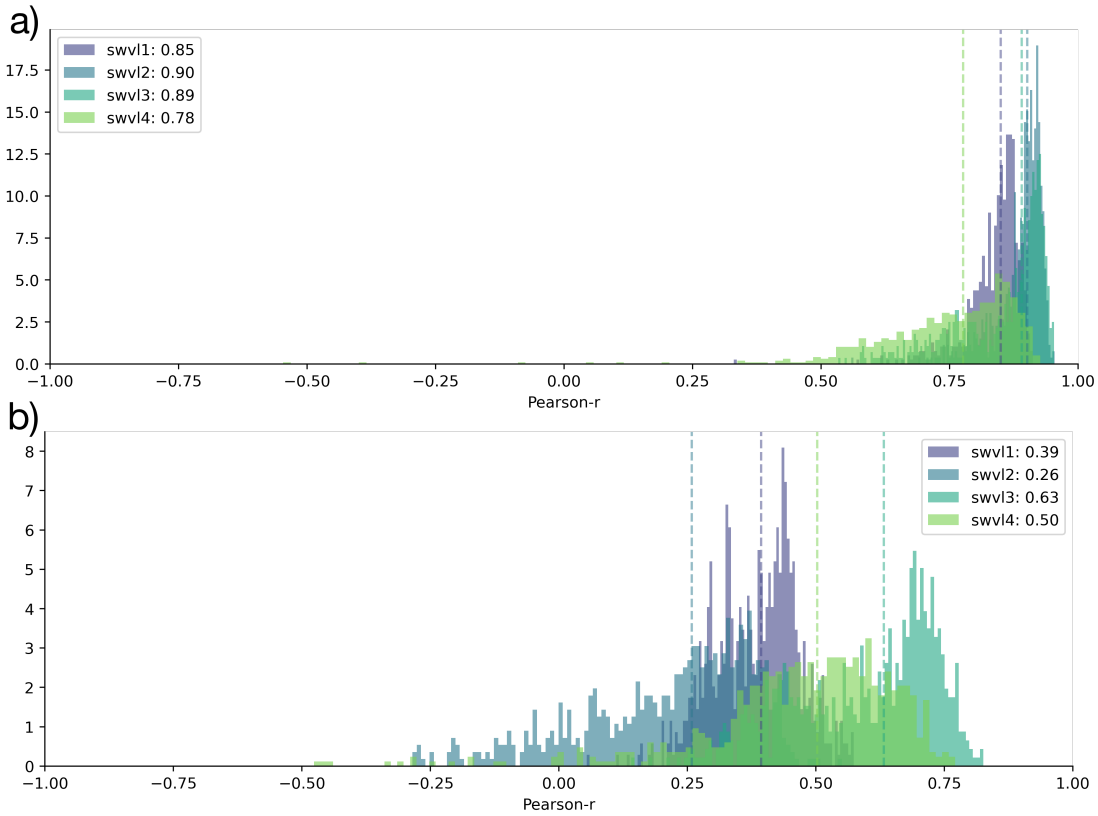


Figure 10 | Shifting in time. (a) Histograms of the original experiments with unshuffled data, repeated from Fig. 4.2. (b) Histograms of catchment correlation scores after having trained the probes on data shifted in time.

A.6 Probing the ESA CCI Soil Moisture

We also tested using the ESA CCI Soil Moisture as our target variable [Dorigo *et al.*, 2017; Gruber *et al.*, 2019]. For this analysis we use the blended product combining active and passive satellite-based sensors (the combined product). The data is a globally available long-term daily satellite soil moisture product that covers the period from 1978–2015 at a 0.25° resolution [Dorigo *et al.*, 2017]. The daily estimate is noisy and therefore, we smoothed the daily data using a 7-day moving average window. The product is of lower spatial and temporal resolution than ERA5-Land, however, it provides an independent estimate of soil moisture based on satellite-derived soil moisture estimates. Neither product is an in situ observation and so both rely on modelled assumptions.

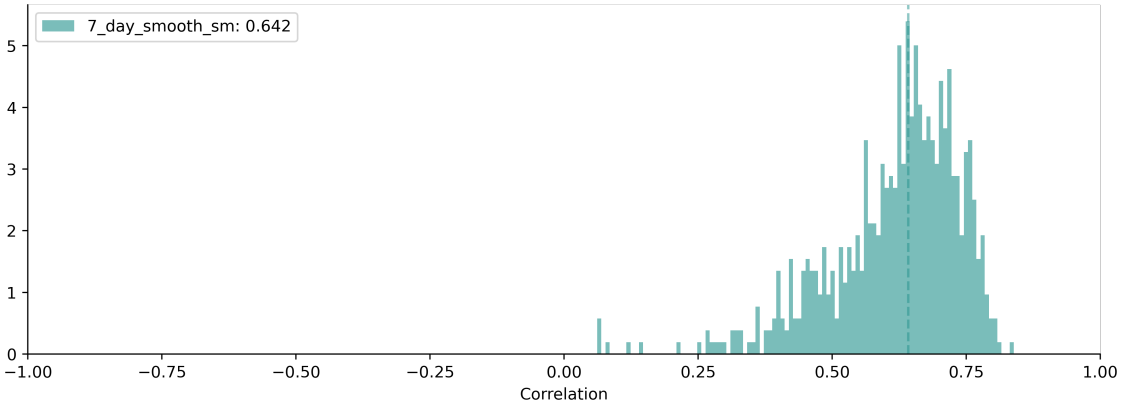


Figure 11 | Histogram of catchment correlation scores for the probe-simulated soil moisture and the 7 day smoothed ESA-CCI Soil Moisture. The median score is 0.64.

The probe captures the temporal dynamics of the soil moisture signals, but struggles to reproduce the catchment specific variability, failing to match the peaks and troughs associated with the target variable (ESA CCI soil moisture). The overall results indicate that the LSTM probes are less able to reproduce ESA CCI Soil Moisture than the signals found in ERA5-Land. Ultimately, we chose to use the ERA5-Land results because of the higher spatial resolution, and therefore, increased specificity of catchment averaged soil moisture values. We can see the intercomparison of these factors in Fig. 14, and Fig. 14.

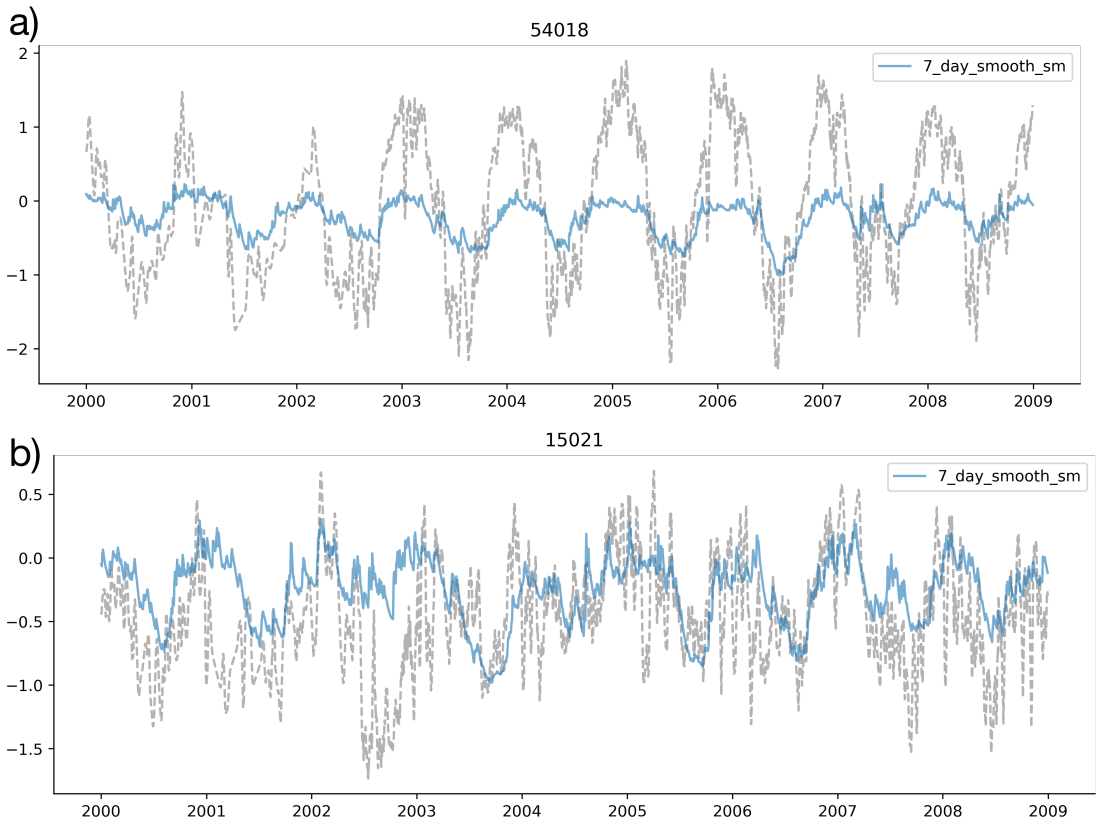


Figure 12 | Time series of probe predictions (coloured lines) compared with the target variables (grey dotted lines). We show two catchments here, 54018 and 15021 for the ESA CCI Soil Moisture Products which is the same as Fig. 4.3

A.7 How Similar are the ESA CCI Soil Moisture and the ERA5-Land Soil Moisture?

We show the correlation between catchment-averaged soil moisture from the two products in Fig. 13. There is a median Pearson correlation of 0.50 between the products. The underlying spatial resolution of the two products, and the median spatial patterns can be seen in Fig. 14. This demonstrates that the ESA CCI soil moisture and the ERA5-Land soil water volume level 1 have similar spatial patterns, with low soil saturation in the Scottish Highlands, saturated soils on the Scottish West Coast and relatively unsaturated soils in Central and Eastern England. However, the spatial resolution for ERA5-Land is much higher, with 6 times as many pixels as the ESA CCI data.

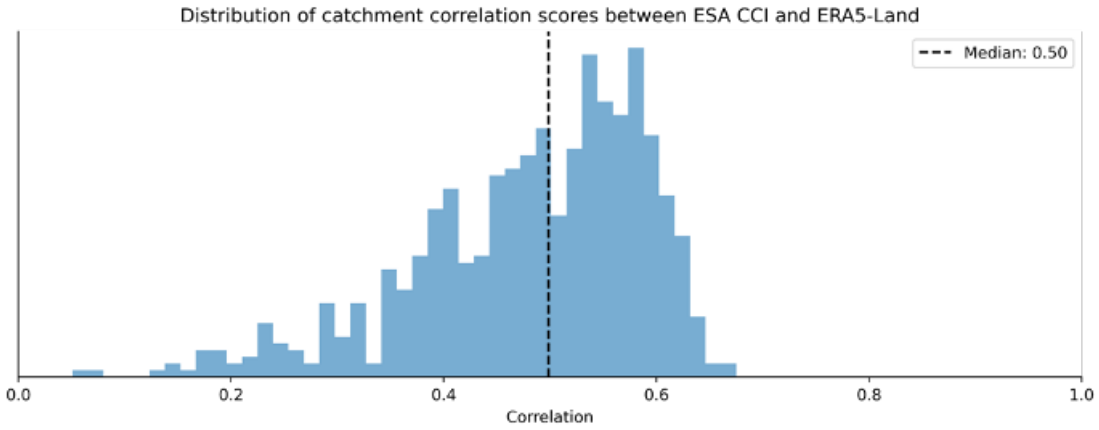


Figure 13 | Histogram showing the distribution of correlations between ESA CCI Soil Moisture and the ERA5-Land Soil Moisture Products (Soil Water Volume 1 - 0–7cm) for each catchment.

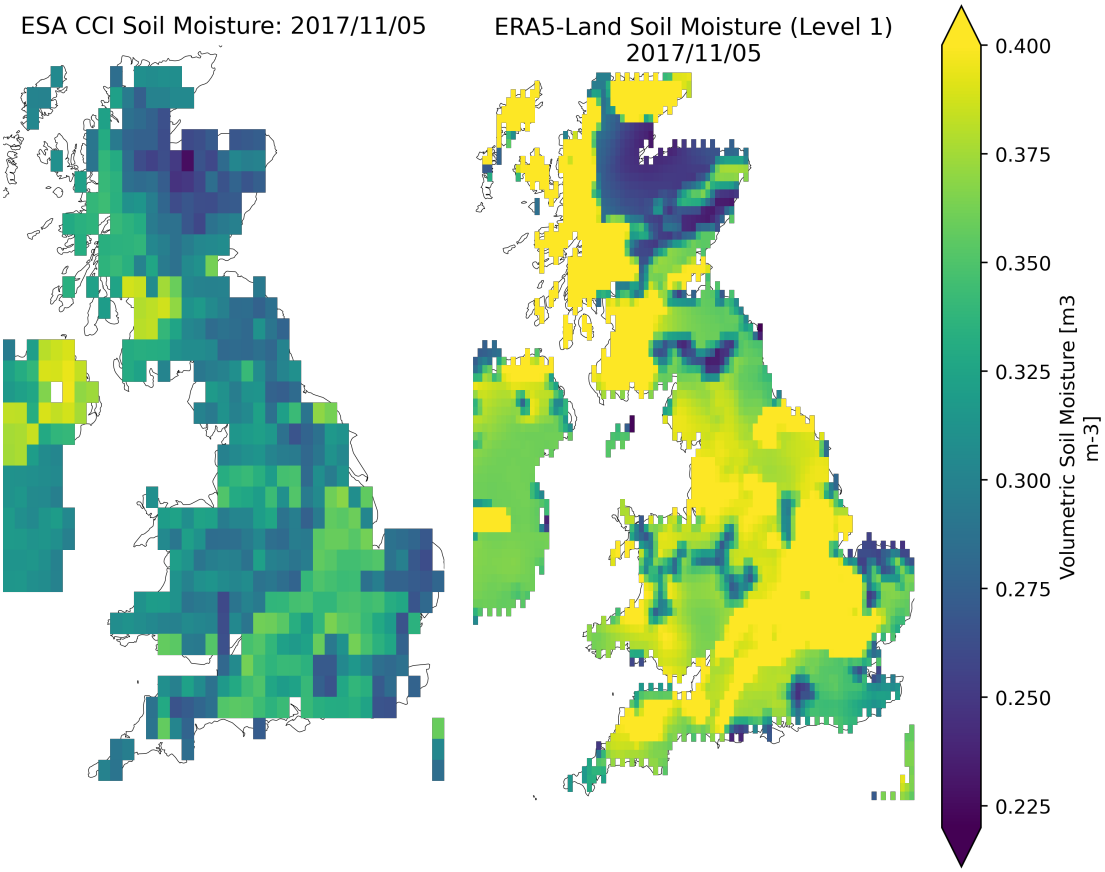


Figure 14 | The spatial resolution of the two soil moisture products (ESA CCI Combined left, ERA5-Land Soil Water Volume Level 1 right). The ERA5-Land data (87 x 93) contains roughly 6 times as many pixels as the ESA CCI soil moisture (35 x 37).

A.8 Investigating the Catchment Specific Probe Offsets

As discussed in Sect. 4.4.2, a simple linear model is not able to predict catchment-specific offsets because it is constrained to model a single bias term for each catchment. In order to minimise the residual sum of squares, the optimum solution is the mean of the target data in the training period. Given that we normalize the target data, this global mean is equal to zero. Therefore, the learned bias in the linear probe will be zero and we cannot expect the probe to reproduce catchment specific offsets. One simple solution to this problem is to include one hot encoded information for each catchment as input features to the regression. We augment our input state dimensions (64) with a binary encoding representing which catchment that data point is drawn from (Fig. 15). Then the linear model can learn a catchment-specific bias, which will be the mean value for that catchment.

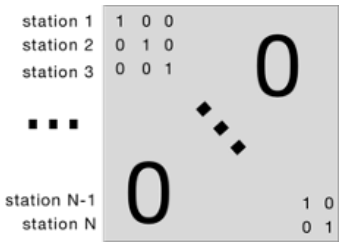


Figure 15 | A diagrammatic explanation of the one hot encoding vectors, where each station is given a unique encoding by using an identity matrix of size N (number of stations). Each gauging station is encoded by a vector of 1s and zeros (a vector of length N). We then append this vector of OHE data (X_i^{OHE}) to each state vector (c_i) to create an augmented input to the linear probe, X_i^*.

We train a linear probe (f_β) using augmented input data (X_i^*), combining the state vector from the LSTM (c_i) with an encoding of the gauging station number (X_i^{OHE}). This is then included as the input to the probe to produce a predicted output ($\hat{s}_i$).

$$X_i^* = [c_i, X_i^{OHE}] \tag{A.1}$$

$$\hat{s}_i = f_\beta(X_i^*) \tag{A.2}$$

In Fig. 16 we can see that including the augmented input variables hugely reduces the biases in the probe outputs, and largely solves the catchment offsets, as expected. This result is expected because the linear probe now can learn a catchment-specific intercept term (a catchment specific bias) that adjusts the probe outputs to the appropriate mean saturation conditions.

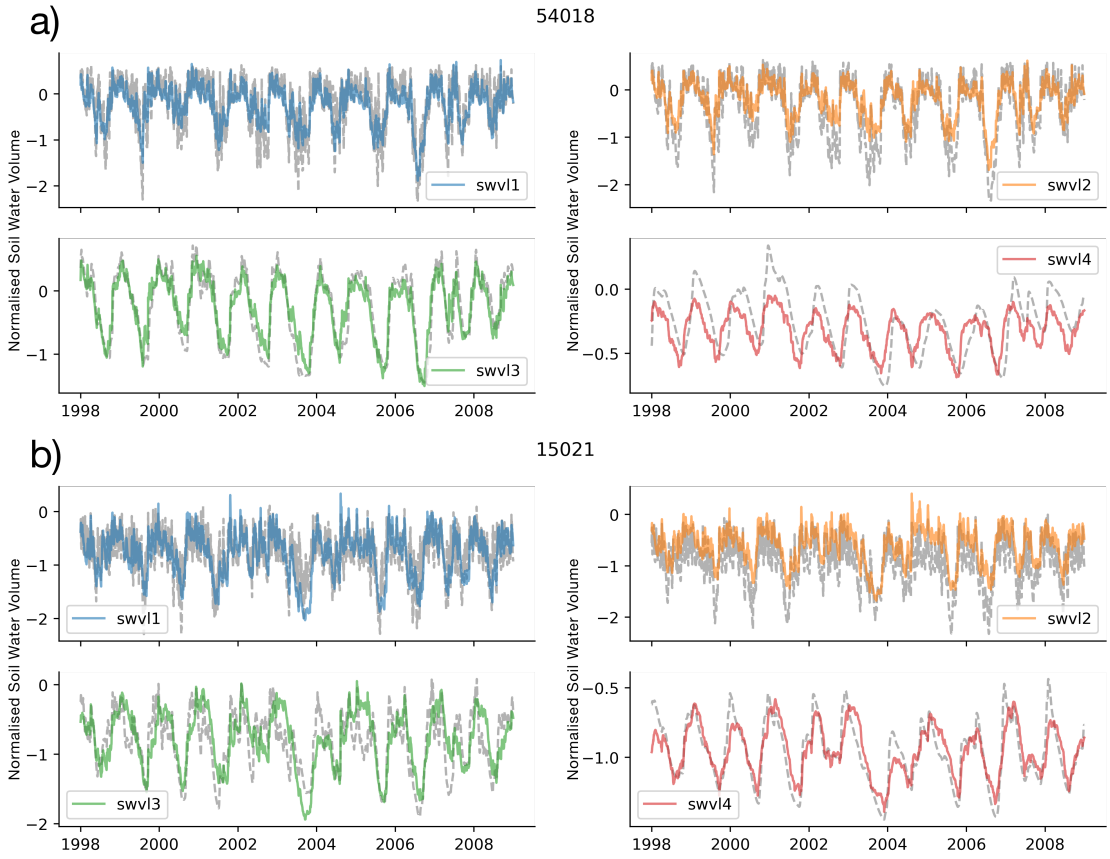


Figure 16 | Time series of probe predictions with augmented inputs, including one-hot encodings of the gauge station ID (coloured lines) compared with the target variables (grey dotted lines). We show two catchments here, 52010 - Rea Brook at Hookagate and 15021 - Burn at Burnham Overy. The results show that we learn a catchment specific weight for adjusting the predictions and address the biases for each catchment.

If we observe these patterns over all catchments we can see that the performance improvement for different soil moisture levels is marked for RMSE (Fig. 17a) but is small for correlation (Fig. 17b). This is because the modelled dynamics are very similar (high correlations), but accurately predicting the mean catchment saturation at different soil levels greatly reduces the absolute squared error.

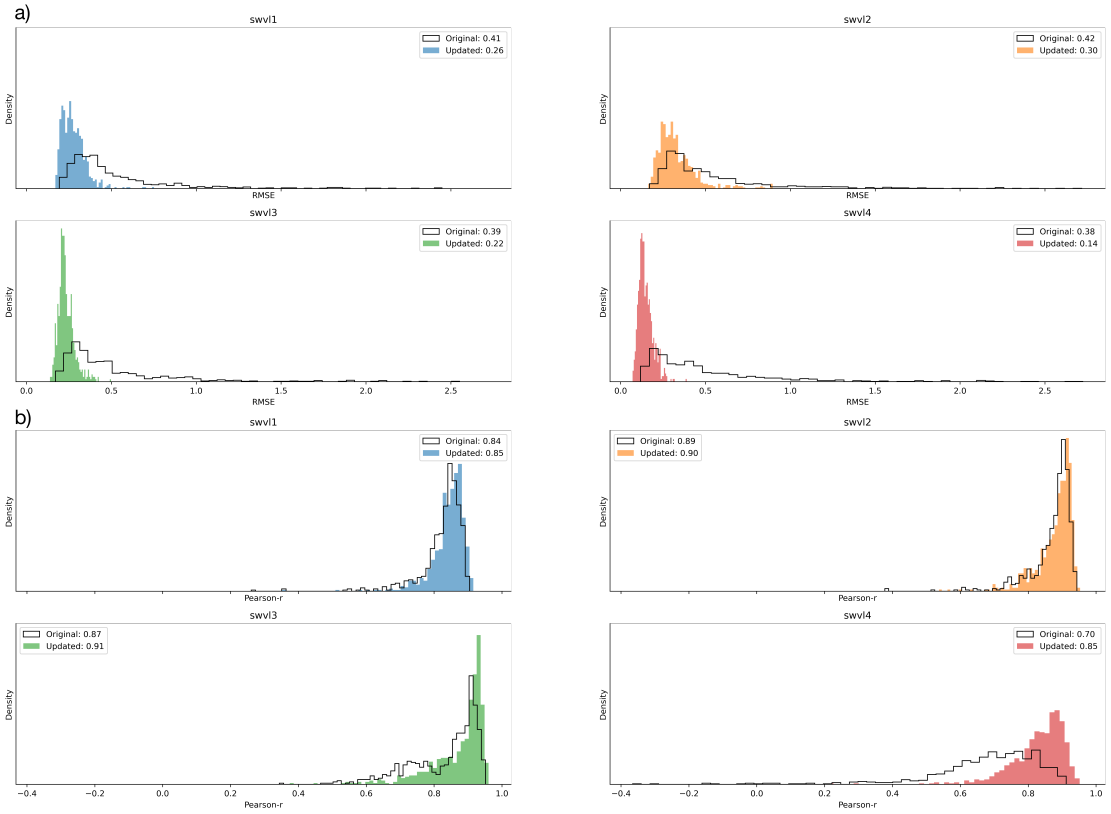


Figure 17 | Histograms showing the distribution of catchment error metrics. The original linear regression model is shown as a solid black line. The coloured histograms correspond to the updated model with augmented inputs (using one hot encoded data) for each soil water level. Subplot (a) contains histograms showing the root mean squared error (RMSE) metric is much improved when modelling with the one hot encoded data. Subplot (b) contains histograms showing the correlation metric, showing that the dynamics are well modelled by the original linear regression, since the different biases do not influence the correlation metric.

A.9 Spatial Context of Demonstration Basins

In the main body of text we have shown soil moisture timeseries for two catchments that were selected to demonstrate a well-performing catchment and a catchment with

a significant bias to demonstrate the problem with modelling catchment offsets. They were randomly chosen from the lower tercile of the RMSE error distribution and the upper tercile. The size, wetness and elevation of these two catchments, Gauge 54018: Rea Brook at Hookagate and Gauge 15021: Lunan Burn at Mill Bank, are described by Fig. 18.

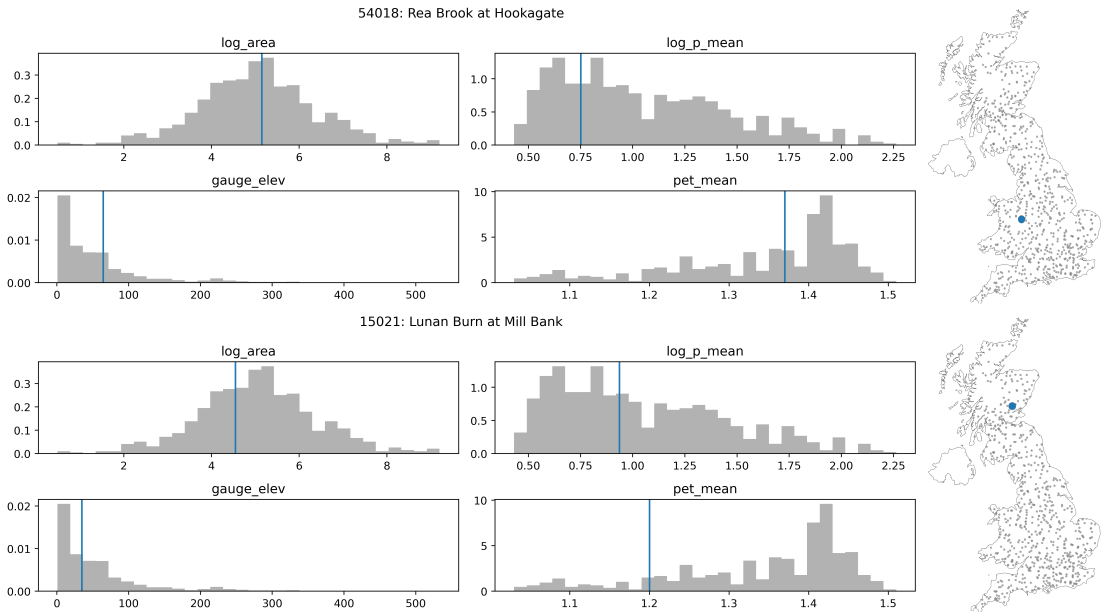


Figure 18 | The spatial context of the two demonstration catchments shown in the Figures in the sections above. For both Gauge ID: 54018, Rea Brook at Hookagate (a), and Gauge ID: 15021, Lunan Burn at Mill Bank (b), we show the log area, the log mean precipitation, the gauge elevation and the mean potential evapotranspiration, as well as the location of the station on the map on the right.

A.10 Non Linear Probe Results

We initially explored a linear probe for the simplicity of interpretation. However, there is no reason why we could not also use a non-linear model to extract non-linear patterns from the cell states. We tested a simple two layer neural network. We trained a fully connected network with hidden sizes of 20 and 10 for the first and second layer respectively. These layer sizes were chosen using a grid search algorithm, where we determined that this was the optimum layer size for minimizing the loss function (RMSE). We used the Rectified Linear Unit (ReLU) as our activation function which was chosen because it is known to reduce the risk of vanishing gradients [Nair & Hinton, 2010], and trained our probe using stochastic gradient descent with the Adam optimizer [Kingma

& Ba, 2014b].

The correlation scores are similar, as shown by the overall fit between the predicted time series and the observed timeseries in Fig. 19. The non-linear probe, unlike the linear probe, is able to model catchment specific offsets most obvious when looking at catchment 15021 and soil water volume level 4 (lower figure, lower-right subplot Fig. 19). This pattern is repeated across other catchments, and the offset problem is much reduced for the non-linear model. This suggests that while the information may not be linearly extracted, a non-linear combination of the cell states does contain the information for the catchment specific offsets. Further research will consider various probe architectures and the other hypotheses outlined above to more fully explore whether catchment offset information is captured by the LSTM state-vector.

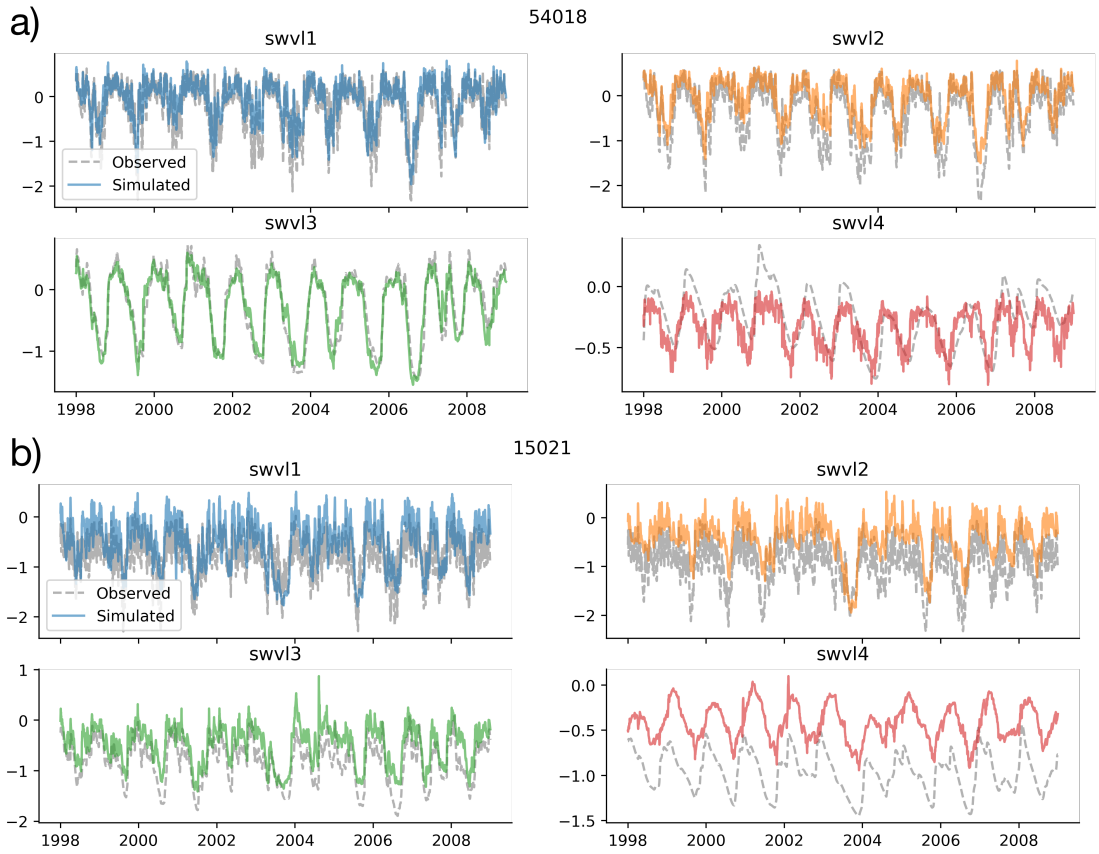


Figure 19 | Time series of non-linear probe predictions (coloured lines) compared with the target variables (grey dotted lines). We show two catchments here, 54018 and 15021 (the same as Fig. 4.3) and four soil moisture levels, swvl1 (blue), swvl2 (orange), swvl3 (green), swvl4 (red). The non-linear probe shows reduced probe offset errors, showing that the information is contained but the linear probe is by definition not able to model catchment-specific intercept terms to account for unique biases.

Acknowledgements

Completing my DPhil at the University of Oxford has been an exceptional privelege.

First, I want to thank my supervisors, Professor Simon Dadson and Doctor Steven Reece. As I enter the next stage of my life and career I hope that we will remain in touch, since I greatly value the mentorship and guidance that your experiences have to offer.

To Simon, thank you for your direction, foresight and the mischevious glint after one of your sporting analogies. Thank you for accepting me onto your project with the NERC Environmental Research DTP. I am lucky to have worked with you since 2012, and have surely become the person I am today during these years. To Steve, thank you for your irreverence, youthfulness and your constant support. I will never forget the days spent in the lab at Eagle House, and the conversations that spanned machine learning, environmental applications and the hallmarks of a good life.

I would like to thank the team at ECMWF who I had the pleasure of completing a summer project with during the ECMWF Summer of Weather Code. Particular thanks go to Gabriel Tseng, my project partner, whose guidance propelled my software engineering abilities and developed my ability to work with deep learning models. I hope that we will remain in contact, since those hours spent working together were some of the fastest learning experiences of my life.

Further gratitude needs to be extended to the NeuralHydrology team, Frederik Kratzert, Daniel Klotz and Martin Gauch. Your codebase has been central to my work in the past few years and I have learned a lot about LSTMs from your work. Even more valuable have been the conversations that we have shared, on papers, ideas and discussing comments from reviewers. I am grateful that you took the time to welcome me into your work and I hope that we have the opportunity to work together again in the future.

I would like to thank the IIASA YSSP team for the support on my summer project. It is unfortunate that the YSSP was remote due to the COVID pandemic, however, the weekly meetings with Jens de Bruijn, Reetik Sahu Kumar and Peter Greve constantly challenged me to justify my work, design better experiments and to extend my knowledge of the field.

I have been blessed with exceptional friends throughout my time at Oxford. I love the city itself, but that is entirely due to the memories created here with many friends and colleagues. I am lucky to have drawn lessons from innumerable sessions and competitions with the various sports teams and clubs I have been involved with: OUMPA, OUCCC, OUAC, OUSC, OUFC and ChChFC. Thank you to Vincent's Club, the college that chose me. To Crystal Palace, staying in the Premier League for the past 8 years has brought a degree of emotional stability that I never thought I could associate with you.

Funding

I gratefully acknowledge funding from

- The UK Natural Environmental Research Council NERC under grant number NE/L002612/1.
- The Copernicus Programme (EU Commission) through the ECMWF Summer of Weather Code 2019.
- IIASA's United Kingdom National Member Organisation (NMO) for funding my Summer at IIASA 2021.

References

M Abadi, A Agarwal, P Barham, E Brevdo, Z Chen, C Citro, GS Corrado, A Davis, J Dean, M Devin *et al.* (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*. 6

M Abbott, J Bathurst, J Cunge, P O'connell & J Rasmussen (1986a). An introduction to the european hydrological system—systeme hydrologique europeen,"she", 2: Structure of a physically-based, distributed modelling system. *Journal of hydrology*, **87**, 61–77. 2, 8

MB Abbott, JC Bathurst, JA Cunge, PE O'Connell & J Rasmussen (1986b). An introduction to the european hydrological system—systeme hydrologique europeen,"she", 1: History and philosophy of a physically-based, distributed modelling system. *Journal of Hydrology*, **87**, 45–59. 8

MB Abbott *et al.* (1991). *Hydroinformatics: information technology and the aquatic environment.*. Avebury Technical. 13

RJ Abrahart, LM See & DP Solomatine (2008). *Practical hydroinformatics: computational intelligence and technological developments in water applications*, vol. 68. Springer Science & Business Media. 13

RJ Abrahart, F Anctil, P Coulibaly, CW Dawson, NJ Mount, LM See, AY Shamseldin, DP Solomatine, E Toth & RL Wilby (2012). Two decades of anarchy? emerging themes and outstanding challenges for neural network river forecasting. *Progress in Physical Geography*, **36**, 480–513. 14

TE Adams & TC Pagano (2016). *Flood forecasting: A global perspective*. Academic Press. 8

N Addor & L Melsen (2019). Legacy, rather than adequacy, drives the selection of hydrological models. *Water Resources Research*, **55**, 378–390. 42

N Addor, AJ Newman, N Mizukami & MP Clark (2017). The camels data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences (HESS)*, **21**, 5293–5313. 44, 122

C Adede, R Oboko, PW Wagacha & C Atzberger (2019). A mixed model approach to vegetation condition prediction using artificial neural networks (ann): case of kenya's operational drought monitoring. *Remote Sensing*, **11**, 1099. 96, 100, 107, 108, 114

HA Afan, A Ibrahem Ahmed Osman, Y Essam, AN Ahmed, YF Huang, O Kisi, M Sherif, A Sefelnasr, Kw Chau & A El-Shafie (2021). Modeling the fluctuations of groundwater level by employing ensemble deep learning techniques. *Engineering Applications of Computational Fluid Mechanics*, **15**, 1420–1439. 24

N Agutu, J Awange, A Zerihun, C Ndehedehe, M Kuhn & Y Fukuda (2017). Assessing multi-satellite remote sensing, reanalysis, and land surface models' products in characterizing agricultural drought in east africa. *Remote sensing of environment*, **194**, 287–302. 36

N Agutu, C Ndehedehe, J Awange, F Kirimi & M Mwaniki (2021). Understanding uncertainty of model-reanalysis soil moisture within greater horn of africa (1982-2014). *Journal of Hydrology*, 127169. 98

C Albergel, E Dutra, S Munier, JC Calvet, J Munoz-Sabater, Pd Rosnay & G Balsamo (2018). Era-5 and era-interim driven isba land surface model simulations: which one performs better? *Hydrology and Earth System Sciences*, **22**, 3515–3532. 98

C Alvarez-Garreton, PA Mendoza, JP Boisier, N Addor, M Galleguillos, M Zambrano-Bigiarini, A Lara, G Cortes, R Garreaud, J McPhee *et al.* (2018). The camels-cl dataset: catchment attributes and meteorology for large sample studies-chile dataset. *Hydrology and Earth System Sciences*, **22**, 5817–5846. 44

A Anshuka, van FF Ogtrop & RW Vervoort (2019). Drought forecasting through statistical models using standardised precipitation index: a systematic review and meta-regression analysis. *Natural Hazards*, **97**, 955–977. 13, 14

P Arias, N Bellouin, E Coppola, R Jones, G Krinner, J Marotzke, V Naik, M Palmer, GK Plattner, J Rogelj *et al.* (2021). Climate change 2021: The physical science basis. contribution of working group14 i to the sixth assessment report of the intergovernmental panel on climate change; technical summary. 27, 35, 92

R Arsenault, A Poulin, P Côté & F Brissette (2014). Comparison of stochastic optimization algorithms in hydrological model calibration. *Journal of Hydrologic Engineering*, **19**, 1374–1384. 128

MA Baudoin, C Vogel, K Nortje & M Naik (2017). Living with drought in south africa: lessons learnt from the recent el niño drought period. *International journal of disaster risk reduction*, **23**, 128–137. 35, 92

A Belayneh, J Adamowski, B Khalil & B Ozga-Zielinski (2014). Long-term spi drought forecasting in the awash river basin in ethiopia using wavelet neural network and wavelet support vector regression models. *Journal of Hydrology*, **508**, 418–429. 13

A Belayneh, J Adamowski, B Khalil & J Quilty (2016). Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction. *Atmospheric research*, **172**, 37–47. 13

M Belkin, D Hsu, S Ma & S Mandal (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, **116**, 15849–15854. 33

Y Bengio, P Simard & P Frasconi (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, **5**, 157–166. 19, 45

Y Bengio, A Courville & P Vincent (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, **35**, 1798–1828. 5

MJ Best, M Pryor, DB Clark, GG Rooney, RLH Essery, CB Ménard, JM Edwards, MA Hendry, A Porson, N Gedney, LM Mercado, S Sitch, E Blyth, O Boucher, PM Cox, CSB Grimmond & RJ Harding (2011). The joint uk land environment simulator (jules), model description - part 1: Energy and water fluxes. *Geoscientific Model Development*, **4**, 677–699, 10.5194/gmd-4-677-2011. 8

K Beven (2006a). A manifesto for the equifinality thesis. *Journal of hydrology*, **320**, 18–36. 62

K Beven (2006b). Searching for the holy grail of scientific hydrology: Q t=(s, r, δt) a as closure. *Hydrology and earth system sciences*, **10**, 609–618. 65

K Beven (2016). Facets of uncertainty: epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. *Hydrological Sciences Journal*, **61**, 1652–1665. 10

K Beven (2020). Deep learning, hydrological processes and the uniqueness of place. *Hydrological Processes*, **34**, 3608–3613, 10.1002/hyp.13805. 41, 69, 122

K Beven & A Binley (1992). The future of distributed models: model calibration and uncertainty prediction. *Hydrological processes*, **6**, 279–298. 10

K Beven & A Binley (2014). Glue: 20 years on. *Hydrological processes*, **28**, 5897–5918. 10, 128

K Beven & J Freer (2001a). A dynamic topmodel. *Hydrological processes*, **15**, 1993–2011. 10, 11

K Beven & J Freer (2001b). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the glue methodology. *Journal of hydrology*, **249**, 11–29. 31, 128

KJ Beven (2000). Uniqueness of place and process representations in hydrological modelling. *Hydrology and Earth System Sciences*, **4**, 203–213, 10.5194/hess-4-203-2000. 9, 31

KJ Beven (2011a). *Rainfall-runoff modelling: the primer*. John Wiley & Sons. 1, 9, 10, 16, 31, 41, 42

KJ Beven (2011b). *Rainfall-runoff modelling: the primer*. John Wiley & Sons. 1, 2, 7, 10, 12, 17, 68, 90

KJ Beven & MJ Kirkby (1979). A physically based, variable contributing area model of basin hydrology/un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrological Sciences Journal*, **24**, 43–69. 2, 7, 41, 49

SJ Birkinshaw, P James & J Ewen (2010). Graphical user interface for rapid set-up of shetran physically-based river catchment model. *Environmental Modelling & Software*, **25**, 609–610. 41

G Blöschl, MF Bierkens, A Chambel, C Cudennec, G Destouni, A Fiori, JW Kirchner, JJ McDonnell, HH Savenije, M Sivapalan *et al.* (2019). Twenty-three unsolved problems in hydrology (uph)–a community perspective. *Hydrological sciences journal*, **64**, 1141–1158. 25, 122

D Booker & R Woods (2014). Comparing and combining physically-based and empirically-based approaches for estimating the hydrology of ungauged catchments. *Journal of Hydrology*, **508**, 227–239. 41

LJ Bouaziz, F Fenicia, G Thirel, de T Boer-Euser, J Buitink, CC Brauer, J De Niel, BJ Dewals, G Drogue, B Grelier *et al.* (2021). Behind the scenes of streamflow model performance. *Hydrology and Earth System Sciences*, **25**, 1069–1095. 8

LJ Bracken & J Croke (2007). The concept of hydrological connectivity and its contribution to understanding runoff-dominated geomorphic systems. *Hydrological Processes*, **21**, 1749–1763, https://doi.org/10.1002/hyp.6313. 64

N Brisson, C Gary, E Justes, R Roche, B Mary, D Ripoche, D Zimmer, J Sierra, P Bertuzzi, P Burger *et al.* (2003). An overview of the crop model stics. *European Journal of agronomy*, **18**, 309–332. 8, 92

M Buechel, L Slater & S Dadson (2022). Hydrological impact of widespread afforestation in great britain using a large ensemble of modelled scenarios. *Communications Earth & Environment*, **3**, 1–10. 123

R Burnash, R Ferral & R McGuire (1973a). A generalised streamflow simulation system–conceptual modelling for digital computers. joint federal and state river forecast center. Tech. rep., Sacramento, Technical Report. 49

R Burnash *et al.* (1995). The nws river forecast system-catchment modeling. *Computer models of watershed hydrology.*, 311–366. 70

RJ Burnash, RL Ferral & RA McGuire (1973b). *A generalized streamflow simulation system: Conceptual modeling for digital computers*. US Department of Commerce, National Weather Service, and State of California 7

C Cammalleri & JV Vogt (2019). Non-stationarity in modis fapar time-series and its impact on operational drought detection. *International Journal of Remote Sensing*, **40**, 1428–1444, 10.1080/01431161.2018.1524603. 92

C Cammalleri, P Barbosa & J Vogt (2020). Evaluating simulated daily discharge for operational hydrological drought monitoring in the global drought observatory (gdo). *Hydrological Sciences Journal*, **65**, 1316–1325. 36

N Cartwright & E McMullin (1984). How the laws of physics lie. 11

Centre for Ecology and Hydrology (2016). 50

J Chadalawada, H Herath & V Babovic (2020a). Hydrologically informed machine learning for rainfall-runoff modeling: A genetic programming-based toolkit for automatic model induction. *Water Resources Research*, **56**, e2019WR026933. 13

J Chadalawada, H Herath & V Babovic (2020b). Hydrologically informed machine learning for rainfall-runoff modeling: A genetic programming-based toolkit for automatic model induction. *Water Resources Research*, **56**, e2019WR026933. 41

VB Chagas, PL Chaffe, N Addor, FM Fan, AS Fleischmann, RC Paiva & VA Siqueira (2020). Camels-br: hydrometeorological time series and landscape attributes for 897 catchments in brazil. *Earth System Science Data*, **12**, 2075–2096. 44

E Chu, D Roy & J Andreas (2020). Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248*. 69

D Clark, L Mercado, S Sitch, C Jones, N Gedney, M Best, M Pryor, G Rooney, R Essery, E Blyth *et al.* (2011). The joint uk land environment simulator (jules), model description—part 2: carbon fluxes and vegetation dynamics. *Geoscientific Model Development*, **4**, 701–722. 2

M Clark & S Khatami (2021a). The evolution of water resources research. 34

M Clark & S Khatami (2021b). The evolution of water resources research. *Eos*, 10.1029/2021EO155644. 42

MP Clark, AG Slater, DE Rupp, RA Woods, JA Vrugt, HV Gupta, T Wagener & LE Hay (2008). Framework for understanding structural errors (fuse): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, **44**. 11, 42, 49, 128

MP Clark, B Nijssen, JD Lundquist, D Kavetski, DE Rupp, RA Woods, JE Freer, ED Gutmann, AW Wood, LD Brekke *et al.* (2015). A unified approach for process-based hydrologic modeling: 1. modeling concept. *Water Resources Research*, **51**, 2498–2514. 11

G Coxon, N Addor, J Bloomfield, J Freer, M Fry, J Hannaford, N Howden, R Lane, M Lewis, E Robinson, T Wagener & R Woods (2020a). Catchment attributes and hydro-meteorological timeseries for 671 catchments across great britain (camels-gb). 10.5285/8344e4f3-d2ea-44f5-8afa-86d2987543a9. 43, 52, 76, 127

G Coxon, N Addor, JP Bloomfield, J Freer, M Fry, J Hannaford, NJ Howden, R Lane, M Lewis, EL Robinson *et al.* (2020b). Camels-gb: Hydrometeorological time series and landscape attributes for 671 catchments in great britain. *Earth System Science Data Discussions*, 1–34. 44, 47, 57, 66, 72, 75, 118

SM Crooks, AL Kay, HN Davies & VA Bell (2014). From catchment to national scale rainfall-runoff modelling: Demonstration of a hydrological modelling framework. *Hydrology*, **1**, 63–88, 10.3390/hydrology1010063. 8, 41

SJ Dadson, F Hirpa, P Thomson & M Konar (2019). Monitoring and modelling hydrological processes. *Water Science, Policy, and Management: A Global Challenge*, 117–137. 2, 7

T Daniell (1991). Neural networks. applications in hydrology and water resources engineering. In *National Conference Publication- Institute of Engineers. Australia*. 14, 41

C Dawson & R Wilby (2001). Hydrological modelling using artificial neural networks. *Progress in physical Geography*, **25**, 80–108. 14, 18

CW Dawson & R Wilby (1998). An artificial neural network approach to rainfall-runoff modelling. *Hydrological Sciences Journal*, **43**, 47–66. 13, 41, 93

de L Oliveira, M Kagan, L Mackey, B Nachman & A Schwartzman (2016). Jet-images—deep learning edition. *Journal of High Energy Physics*, **2016**, 1–32. 122

J Deng, W Dong, R Socher, LJ Li, K Li & L Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255, Ieee. 6

W Dorigo, W Wagner, C Albergel, F Albrecht, G Balsamo, L Brocca, D Chung, M Ertl, M Forkel, A Gruber *et al.* (2017). Esa cci soil moisture for improved earth system understanding: State-of-the art and future directions. *Remote Sensing of Environment*, **203**, 185–215. 85, 86, 136

F Doshi-Velez & B Kim (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. 69

S Duan, P Ullrich & L Shu (2020). Using convolutional neural networks for streamflow projection in california. *Front. Water 2: 28. doi: 10.3389/frwa*. 41

A Elshorbagy, G Corzo, S Srinivasulu & D Solomatine (2010). Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology-part 2: Application. *Hydrology and Earth System Sciences*, **14**, 1943–1961. 41

European Union Digital Strategy (2019). Ethics guidelines for trustworthy ai. 69

K Fang, C Shen, D Kifer & X Yang (2017). Prolongation of smap to spatiotemporally seamless coverage of continental u.s. using a deep learning neural network. *Geophysical Research Letters*, **44**, 11,030–11,039, https://doi.org/10.1002/2017GL075619. 3, 24, 28, 99, 116

K Fang, M Pan & C Shen (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**, 2221–2233. 41

K Fang, D Kifer, K Lawson & C Shen (2020). Evaluating the potential and challenges of an uncertainty quantification method for long short-term memory models for soil moisture predictions. *Water Resources Research*, **56**, e2020WR028095. 41

FAO (2019). Kenya at a glance. 93

TG Farr, PA Rosen, E Caro, R Crippen, R Duren, S Hensley, M Kobrick, M Paller, E Rodriguez, L Roth *et al.* (2007). The shuttle radar topography mission. *Reviews of geophysics*, **45**. 97, 98

D Feng, K Fang & C Shen (2020a). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resources Research*, **56**, e2019WR026793. 41

D Feng, K Fang & C Shen (2020b). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resources Research*, **56**, e2019WR026793. 69

F Fenicia, D Kavetski, HH Savenije, MP Clark, G Schoups, L Pfister & J Freer (2014). Catchment properties, function, and conceptual model representation: is there a correspondence? *Hydrological Processes*, **28**, 2451–2467. 128

A Fisher, C Rudin & F Dominici (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, **20**, 1–81. 29

J Frame, F Kratzert, D Klotz, M Gauch, G Shelev, O Gilon, LM Qualls, HV Gupta & GS Nearing (2021a). Deep learning rainfall-runoff predictions of extreme events. *Hydrology and Earth System Sciences Discussions*, 1–20. 27, 68, 69, 84, 121

JM Frame (2022). On mass conservation for predicting the long term water balance of the rainfall runoff process. `https://github.com/jmframe/mclstm_2021_mass_balance`. 27

JM Frame, F Kratzert, A Raney, M Rahman, FR Salas & GS Nearing (2021b). Post-processing the national water model with long short-term memory networks for

streamflow predictions and model diagnostics. *JAWRA Journal of the American Water Resources Association*, **57**, 885–905. 26, 30

M Franchini & M Pacciani (1991). Comparative analysis of several conceptual rainfall-runoff models. *Journal of hydrology*, **122**, 161–219. 9

RA Freeze & R Harlan (1969). Blueprint for a physically-based, digitally-simulated hydrologic response model. *Journal of hydrology*, **9**, 237–258. 2, 8

J Friedman, T Hastie, R Tibshirani *et al.* (2001). *The elements of statistical learning*, vol. 1. Springer series in statistics New York. 6, 31, 32, 33

J Friedman, T Hastie & R Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, **33**, 1. 73

K Fung, Y Huang, C Koo & Y Soh (2020). Drought forecasting: A review of modelling approaches 2007–2017. *Journal of Water and Climate Change*, **11**, 771–799. 13, 14

C Funk, A Hoell, S Shukla, I Blade, B Liebmann, JB Roberts, FR Robertson & G Husak (2014). Predicting east african spring droughts using pacific and indian ocean sea surface temperature indices. *Hydrology and Earth System Sciences*, **18**, 4965–4978. 98

C Funk, P Peterson, M Landsfeld, D Pedreros, J Verdin, S Shukla, G Husak, J Rowland, L Harrison, A Hoell *et al.* (2015). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. *Scientific data*, **2**, 1–21. 97, 98

C Funk, L Harrison, S Shukla, C Pomposi, G Galu, D Korecha, G Husak, T Magadzire, F Davenport, C Hillbruner *et al.* (2018). Examining the role of unusually warm indo-pacific sea-surface temperatures in recent african droughts. *Quarterly Journal of the Royal Meteorological Society*, **144**, 360–383. 98

Y Gal & Z Ghahramani (2015). Dropout as a bayesian approximation: representing model uncertainty in deep learning. arxiv. *arXiv preprint arxiv:1506.02142*. 26, 33

R García-Herrera, J Díaz, RM Trigo, J Luterbacher & EM Fischer (2010). A review of the european summer heat wave of 2003. *Critical Reviews in Environmental Science and Technology*, **40**, 267–306. 35, 92

JS Garofolo, LF Lamel, WM Fisher, JG Fiscus & DS Pallett (1993). Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, **93**, 27403. 6

M Gauch, & J Lin (2020). A data scientist's guide to streamflow prediction. 84, 98

M Gauch, F Kratzert, D Klotz, G Nearing, J Lin & S Hochreiter (2021a). Rainfall–runoff prediction at multiple timescales with a single long short-term memory network. *Hydrology and Earth System Sciences*, **25**, 2045–2062. 3, 27, 42, 84, 93, 98

M Gauch, J Mai & J Lin (2021b). The proper care and feeding of camels: How limited training data affects streamflow prediction. *Environmental Modelling & Software*, **135**, 104926. 13

M Gauch, J Mai & J Lin (2021c). The proper care and feeding of camels: How limited training data affects streamflow prediction. *Environmental Modelling & Software*, **135**, 104926. 41, 42, 63

FA Gers, J Schmidhuber & F Cummins (2000). Learning to forget: Continual prediction with lstm. *Neural computation*, **12**, 2451–2471. 20

G Ghiggi, V Humphrey, SI Seneviratne & L Gudmundsson (2019). Grun: an observation-based global gridded runoff dataset from 1902 to 2014. *Earth System Science Data*, **11**, 1655–1674. 37

G Ghiggi, V Humphrey, S Seneviratne & L Gudmundsson (2021). G-run ensemble: A multi-forcing observation-based global runoff reanalysis. *Water Resources Research*, **57**, e2020WR028787. 37

A Ghorbani & J Zou (2020). Neuron shapley: Discovering the responsible neurons. *arXiv preprint arXiv:2002.09815*. 69

R Girshick (2015). Fast r-cnn. arxiv 2015. *arXiv preprint arXiv:1504.08083*. 100

W Gong, HV Gupta, D Yang, K Sricharan & AO Hero III (2013). Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach. *Water resources research*, **49**, 2253–2273. 10

A Gruber, T Scanlon, van der R Schalie, W Wagner & W Dorigo (2019). Evolution of the esa cci soil moisture climate data records and their underlying merging methodology. *Earth System Science Data*, **11**, 717–739. 38, 136

L Gudmundsson & SI Seneviratne (2015). Towards observation-based gridded runoff estimates for europe. *Hydrology and Earth System Sciences*, **19**, 2859–2879. 37

HV Gupta, S Sorooshian & PO Yapo (1998). Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, **34**, 751–763, 10.1029/97WR03495. 51

HV Gupta, H Kling, KK Yilmaz & GF Martinez (2009). Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *Journal of hydrology (Amsterdam)*, **377**, 80–91. 51, 128

HV Gupta, C Perrin, G Blöschl, A Montanari, R Kumar, M Clark & V Andréassian (2014). Large-sample hydrology: a need to balance depth with breadth. *Hydrology and Earth System Sciences*, **18**, 463–477. 42

AH Halff, HM Halff & M Azmoodeh (1993). Predicting runoff from rainfall using neural networks. In *Engineering hydrology*, 760–765, ASCE. 14, 41

HMVV Herath, J Chadalawada & V Babovic (2020). Hydrologically informed machine learning for rainfall-runoff modelling: Towards distributed modelling. *Hydrology and Earth System Sciences Discussions*, 1–42. 41

H Hersbach, B Bell, P Berrisford, S Hirahara, A Horányi, J Muñoz-Sabater, J Nicolas, C Peubey, R Radu, D Schepers, A Simmons, C Soci, S Abdalla, X Abellan, G Balsamo, P Bechtold, G Biavati, J Bidlot, M Bonavita, GD Chiara, P Dahlgren, D Dee, M Diamantakis, R Dragani, J Flemming, R Forbes, M Fuentes, A Geer, L Haimberger, S Healy, RJ Hogan, E Hólm, M Janisková, S Keeley, P Laloyaux, P Lopez, C Lupu, G Radnoti, de P Rosnay, I Rozum, F Vamborg, S Villaume & JN Thépaut (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146**, 1999–2049, 10.1002/qj.3803. 75, 97

J Hewitt & P Liang (2019). Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*. 69, 70, 72, 88, 89, 132

GE Hinton, N Srivastava, A Krizhevsky, I Sutskever & RR Salakhutdinov (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*. 10, 33

S Hochreiter (1991). Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, **91**. 2, 20, 41, 45, 99, 114

S Hochreiter & J Schmidhuber (1997). Lstm can solve hard long time lag problems. In *Advances in neural information processing systems*, 473–479. 19

S Hochreiter, Y Bengio, P Frasconi, J Schmidhuber *et al.* (2001). Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. 41

PJ Hoedt, F Kratzert, D Klotz, C Halmich, M Holzleitner, G Nearing, S Hochreiter & G Klambauer (2021). Mc-lstm: Mass-conserving lstm. 27, 42, 121

M Hrachowitz, H Savenije, G Blöschl, J McDonnell, M Sivapalan, J Pomeroy, B Arheimer, T Blume, M Clark, U Ehret *et al.* (2013). A decade of predictions in ungauged basins (pub)—a review. *Hydrological sciences journal*, **58**, 1198–1255. 1, 10, 15, 25, 28, 31, 33

C Huntingford, ES Jeffers, MB Bonsall, HM Christensen, T Lees & H Yang (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, **14**, 124007. 2, 41

A Jain & S Srinivasulu (2004). Development of effective and efficient rainfall-runoff models using integration of deterministic, real-coded genetic algorithms and artificial neural network techniques. *Water Resources Research*, **40**. 18

G James, D Witten, T Hastie & R Tibshirani (2013). *An introduction to statistical learning*, vol. 112. Springer. 87

S Jiang, Y Zheng & D Solomatine (2020). Improving ai system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning. *Geophysical Research Letters*, **47**, e2020GL088229. 42

JW Jones, JM Antle, B Basso, KJ Boote, RT Conant, I Foster, HCJ Godfray, M Herrero, RE Howitt, S Janssen *et al.* (2017). Brief history of agricultural systems modeling. *Agricultural systems*, **155**, 240–254. 8

C Jothityangkoon, M Sivapalan & D Farmer (2001). Process controls of water balance variability in a large semi-arid catchment: downward approach to hydrological model development. *Journal of hydrology*, **254**, 174–198. 17

M Jung, M Reichstein & A Bondeau (2009). Towards global empirical upscaling of fluxnet eddy covariance observations: validation of a model tree ensemble approach using a biosphere model. *Biogeosciences*, **6**, 2001–2013. 38

M Jung, C Schwalm, M Migliavacca, S Walther, G Camps-Valls, S Koirala, P Anthoni, S Besnard, P Bodesheim, N Carvalhais *et al.* (2020). Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the fluxcom approach. *Biogeosciences*, **17**, 1343–1365. 37

A Karpatne, G Atluri, JH Faghmous, M Steinbach, A Banerjee, A Ganguly, S Shekhar, N Samatova & V Kumar (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on knowledge and data engineering*, **29**, 2318–2331. 69

D Kavetski, G Kuczera & SW Franks (2006). Calibration of conceptual hydrological models revisited: 1. overcoming numerical artefacts. *Journal of Hydrology*, **320**, 173–186, the model parameter estimation experiment, https://doi.org/10.1016/j.jhydrol.2005.07.012. 7

BK Kenduiywo, MR Carter, A Ghosh & RJ Hijmans (2021). Evaluating the quality of remote sensing products for agricultural index insurance. *PloS one*, **16**, e0258215. 107

H Kim (2017). Global soil wetness project phase 3 atmospheric boundary conditions (experiment 1). *Data Integration and Analysis System (DIAS), Data set, https://doi.org/10.20783/DIAS*, **501**. 37

DP Kingma & J Ba (2014a). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 48

DP Kingma & J Ba (2014b). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 143

JW Kirchner (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, **42**. 7, 9, 15, 31, 68, 120

JW Kirchner (2009). Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward. *Water Resources Research*, **45**. 2

A Klisch & C Atzberger (2016). Operational drought monitoring in kenya using modis ndvi time series. *Remote Sensing*, **8**, 267. 36, 92, 94, 96, 97, 107, 114

D Klotz, F Kratzert, M Gauch, AK Sampson, G Klambauer, S Hochreiter & G Nearing (2020). Uncertainty estimation with deep learning for rainfall-runoff modelling. *arXiv preprint arXiv:2012.14295*. 26, 28, 42, 50, 93, 128

WJ Knoben, JE Freer, KJ Fowler, MC Peel & RA Woods (2019). Modular assessment of rainfall-runoff models toolbox (marrmot) v1. 2: an open-source, extendable frame-

work providing implementations of 46 conceptual hydrologic models as continuous state-space formulations. *Hydrology and Earth System Sciences*. 8, 42, 122

FN Kogan (1995). Application of vegetation index and brightness temperature for drought detection. *Advances in space research*, **15**, 91–100. 96

R Kohavi *et al.* (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, vol. 14, 1137–1145, Montreal, Canada. 32

F Kratzert, D Klotz, C Brenner, K Schulz & M Herrnegger (2018). Rainfall–runoff modelling using long short-term memory (lstm) networks. *Hydrology and Earth System Sciences*, **22**, 6005–6022, 10.5194/hess-22-6005-2018. 3, 20, 24, 25, 27, 41, 42, 45, 63, 65, 68, 69, 84, 98, 99, 116, 121

F Kratzert, M Herrnegger, D Klotz, S Hochreiter & G Klambauer (2019a). *NeuralHydrology – Interpreting LSTMs in Hydrology*, 347–362. Springer International Publishing, Cham, 10.1007/978-3-030-28954-6_19. 27, 30, 34, 98

F Kratzert, M Herrnegger, D Klotz, S Hochreiter & G Klambauer (2019b). Neuralhydrology–interpreting lstms in hydrology. In *Explainable AI: Interpreting, explaining and visualizing deep learning*, 347–362, Springer. 22, 70, 71

F Kratzert, D Klotz, M Herrnegger, AK Sampson, S Hochreiter & GS Nearing (2019c). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, **55**, 11344–11354, 10.1029/2019WR026065. 3, 20, 69, 84, 98

F Kratzert, D Klotz, M Herrnegger, AK Sampson, S Hochreiter & GS Nearing (2019d). Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resources Research*, **55**, 11344–11354, 10.1029/2019WR026065. 25, 27, 33

F Kratzert, D Klotz, G Shalev, G Klambauer, S Hochreiter & G Nearing (2019e). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, **23**, 5089–5110, 10.5194/hess-23-5089-2019. 3, 20, 24, 25, 27, 30, 31, 42, 44, 45, 46, 47, 48, 51, 62, 63, 68, 69, 84, 93, 99, 100, 101, 115, 116, 121

F Kratzert, D Klotz, S Hochreiter & GS Nearing (2021). A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling. *Hydrology and Earth System Sciences*, **25**, 2685–2703. 27, 28, 30, 128

A Krizhevsky, I Sutskever & GE Hinton (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, **25**, 1097–1105. 5

T Krueger, J Freer, JN Quinton, CJ Macleod, GS Bilotta, RE Brazier, P Butler & PM Haygarth (2010). Ensemble evaluation of hydrological model hypotheses. *Water Resources Research*, **46**. 128

DN Kumar, KS Raju & T Sathish (2004). River flow forecasting using recurrent neural networks. *Water resources management*, **18**, 143–161. 19

RA Lane, G Coxon, JE Freer, T Wagener, PJ Johnes, JP Bloomfield, S Greene, CJ Macleod & SM Reaney (2019). Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in great britain. *Hydrology and Earth System Sciences*, **23**, 4011–4032. 11, 34, 42, 43, 47, 49, 50, 52, 53, 54, 56, 58, 62, 118, 127

XH Le, HV Ho, G Lee & S Jung (2019). Application of long short-term memory (lstm) neural network for flood forecasting. *Water*, **11**, 1387. 41

G Leavesley, R Lichty, B Troutman & L Saindon (1983). Precipitation-runoff modelling system: user's manual. report 83–4238. *US Geological Survey Water Resources Investigations*, **207**. 49

G Leavesley, S Markstrom, M Brewer & R Viger (1996). The modular modeling system (mms)—the physical process modeling component of a database-centered decision support system for water and power management. *Water, Air, & Soil Pollution*, **90**, 303–311. 128

CS Lee, E Sohn, JD Park & JD Jang (2019). Estimation of soil moisture using deep learning based on satellite data: A case study of south korea. *GIScience & Remote Sensing*, **56**, 43–67. 24

T Lees, M Buechel, B Anderson, L Slater, S Reece, G Coxon & SJ Dadson (2021a). Benchmarking data-driven rainfall–runoff models in great britain: a comparison of long short-term memory (lstm)-based models with four lumped conceptual models. *Hydrology and Earth System Sciences*, **25**, 5517–5534. 93, 98, 121

T Lees, M Buechel, B Anderson, L Slater, S Reece, G Coxon & SJ Dadson (2021b). Benchmarking data-driven rainfall-runoff models in great britain: A comparison of lstm-

based models with four lumped conceptual models. *Hydrology and Earth System Sciences Discussions*, 1–41. 68, 70, 72, 83

X Liang (1994). A two-layer variable infiltration capacity land surface representation for general circulation models. *PhD Thesis*. 7, 41, 49

B Lim & S Zohren (2021). Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, **379**, 20200209. 115

lin K Hsu, HV Gupta & S Sorooshian (1997). Application of a recurrent neural network to rainfall-runoff modeling. In *Proceedings of the 1997 24th Annual Water Resources Planning and Management Conference*. 19

G Lindström, C Pers, J Rosberg, J Strömqvist & B Arheimer (2010). Development and testing of the hype (hydrological predictions for the environment) water quality model for different spatial scales. *Hydrology research*, **41**, 295–319. 8

ZC Lipton (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, **16**, 31–57. 69

SM Lundberg & SI Lee (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, 4768–4777. 29, 69, 100, 102, 112

K Ma, D Feng, K Lawson, WP Tsai, C Liang, X Huang, A Sharma & C Shen (2020). Transferring hydrologic data across continents – leveraging us data to improve hydrologic prediction in other countries. *Earth and Space Science Open Archive*, 28, 10.1002/essoar.10504132.1. 68, 69

C Manning, M Widmann, E Bevacqua, AF Van Loon, D Maraun & M Vrac (2018). Soil moisture drought in europe: a compound event of precipitation and potential evapotranspiration on multiple time scales. *Journal of Hydrometeorology*, **19**, 1255–1271. 84

SL Markstrom, RS Regan, LE Hay, RJ Viger, RM Webb, RA Payn & JH LaFontaine (????). Prms-iv, the precipitation-runoff modeling system, version 4. *USGS Publications Wareouse*. 7

B Martens, DG Miralles, H Lievens, R Van Der Schalie, RA De Jeu, D Fernández-Prieto, HE Beck, WA Dorigo & NE Verhoest (2017). Gleam v3: Satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development*, **10**, 1903–1925. 38

T Marthews, FE Otto, D Mitchell, SJ Dadson & RG Jones (2015). 17. the 2014 drought in the horn of africa: Attribution of meteorological drivers. *Bulletin of the American Meteorological Society*, **96**, S83–S88. 13

RM Maxwell, SJ Kollet, SG Smith, CS Woodward, RD Falgout, IM Ferguson, C Baldwin, WJ Bosl, R Hornung & S Ashby (2009). Parflow user's manual. *International Ground Water Modeling Center Report GWMI*, **1**, 129. 41

WS McCulloch & W Pitts (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, **5**, 115–133. 6

TB McKee, NJ Doesken, J Kleist *et al.* (1993). The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology*, vol. 17, 179–183, Boston. 13

H McMillan, J Freer, F Pappenberger, T Krueger & M Clark (2010). Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. *Hydrological Processes: An International Journal*, **24**, 1270–1284. 128

HK McMillan, IK Westerberg & T Krueger (2018). Hydrological data uncertainty and its implications. *Wiley Interdisciplinary Reviews: Water*, **5**, e1319. 11

M Meroni, D Fasbender, F Rembold, C Atzberger & A Klisch (2019). Near real-time vegetation anomaly detection with modis ndvi: Timeliness vs. accuracy and effect of anomaly computation options. *Remote sensing of environment*, **221**, 508–521. 107

A Mishra & V Desai (2006). Drought forecasting using feed-forward recursive neural network. *ecological modelling*, **198**, 127–138. 13

AK Mishra & VP Singh (2011). Drought modeling–a review. *Journal of Hydrology*, **403**, 157–175. 14

C Molnar (2020). *Interpretable machine learning*. Lulu. com. 28

Z Moshe, A Metzger, G Elidan, F Kratzert, S Nevo & R El-Yaniv (2020). Hydronets: Leveraging river structure for hydrologic modeling. *arXiv preprint arXiv:2007.00595*. 121, 122

M Muller (2018). Cape town's drought: don't blame climate change. 35, 92

TJ Mulvaney (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. *Proceedings of the institution of Civil Engineers of Ireland*, **4**, 19–31. 7, 13

J Muñoz-Sabater, E Dutra, A Agustí-Panareda, C Albergel, G Arduini, G Balsamo, S Boussetta, M Choulga, S Harrigan, H Hersbach *et al.* (2021). Era5-land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data Discussions*, 1–50. 75

J Murphy, D Sexton, G Jenkins, B Booth, C Brown, R Clark, M Collins, G Harris, E Kendon, R Betts, S Brown, K Humphrey, M McCarthy, R McDonald, A Stephens, C Wallace, R Warren, R Wilby & R Wood (2009). Uk climate projections science report: Climate change projections. *Met Office Hadley Centre*, © Crown Copyright 2009. The UK Climate Projections data have been made available by the Department for Environment, Food and Rural Affairs (Defra) and Department for Energy and Climate Change (DECC) under licence from the Met Office, Newcastle University, University of East Anglia and Proudman Oceanographic Laboratory. These organisations accept no responsibility for any inaccuracies or omissions in the data, nor for any loss or damage directly or indirectly caused to any person or body by reason of, or arising out of, any use of this data. 131

V Nair & GE Hinton (2010). Rectified linear units improve restricted boltzmann machines. In *Icml*. 143

JE Nash & JV Sutcliffe (1970). River flow forecasting through conceptual models part i—a discussion of principles. *Journal of hydrology*, **10**, 282–290. 51

J Nay, E Burchfield & J Gilligan (2018). A machine-learning approach to forecasting remotely sensed vegetation health. *International journal of remote sensing*, **39**, 1800–1816. 93

NDMA (2017). Drought early warning bulletin for april 2017. 108

B Neal (2019). On the bias-variance tradeoff: textbooks need an update. *arXiv preprint arXiv:1912.08286*. 33

GS Nearing, Y Tian, HV Gupta, MP Clark, KW Harrison & SV Weijs (2016). A philosophical basis for hydrological uncertainty. *Hydrological Sciences Journal*, **61**, 1666–1678. 10

GS Nearing, BL Ruddell, AR Bennett, C Prieto & HV Gupta (2020a). Does information theory provide a new paradigm for earth science? hypothesis testing. *Water Resources Research*, **56**, e2019WR024918. 1, 41, 43

GS Nearing, BL Ruddell, AR Bennett, C Prieto & HV Gupta (2020b). Does information theory provide a new paradigm for earth science? hypothesis testing. *Water Resources Research*, **56**. 10, 65

GS Nearing, D Klotz, AK Sampson, F Kratzert, M Gauch, JM Frame, G Shalev & S Nevo (2021a). Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks. *Hydrology and Earth System Sciences Discussions*, 1–25. 28

GS Nearing, F Kratzert, AK Sampson, CS Pelissier, D Klotz, JM Frame, C Prieto & HV Gupta (2021b). What role does hydrological science play in the age of machine learning? *Water Resources Research*, **57**, e2020WR028091. 26, 27, 28, 29, 30, 33, 43, 62, 64, 93, 110, 114, 121, 123

F Network (2013). Kenya food security: In brief. 94, 110

D Nguyen (2018). Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1069–1078. 69

SE Nicholson (2014). A detailed look at the recent drought situation in the greater horn of africa. *Journal of Arid Environments*, **103**, 71–79. 35, 92

SE Nicholson (2017). Climate and climatic variability of rainfall over eastern africa. *Reviews of Geophysics*, **55**, 590–635, https://doi.org/10.1002/2016RG000544. 93

MA Nielsen (2015). *Neural networks and deep learning*, vol. 25. Determination press San Francisco, CA. 5, 20

V Nourani, AH Baghanam, J Adamowski & O Kisi (2014). Applications of hybrid wavelet–artificial intelligence models in hydrology: a review. *Journal of Hydrology*, **514**, 358–377. 41

OCHA (2018). Ocha flash update 6: Floods in kenya: 7 june 2018 - kenya. 103

C Olah (2016). Understanding LSTM Networks – colah's blog. 21, 99, 101

C Olah, A Satyanarayan, I Johnson, S Carter, L Schubert, K Ye & A Mordvintsev (2018). The building blocks of interpretability. *Distill*, **3**, e10. 69

C Olah, N Cammarata, L Schubert, G Goh, M Petrov & S Carter (2020). Zoom in: An introduction to circuits. *Distill*, **5**, e00024–001. 69

N Oreskes (2003). The role of quantitative models in science naomi oreskes. *Models in ecosystem science, edited by: Canham, CD, Cole, JJ, and Lauenroth, WK*, 13–31. 15

N Oreskes, K Shrader-Frechette & K Belitz (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, **263**, 641–646. 1, 7, 11

R Orth *et al.* (2021). Global soil moisture data derived through machine learning trained with in-situ measurements. *Scientific Data*, **8**, 1–14. 38

T Parr, K Turgutlu, C Csiszar & J Howard (????). Beware default random forest importances. `https://explained.ai/rf-importance/`. 29

R Parry (2003). Episteme and techne. *Stanford Encyclopedia of Philosophy*. 1

R Pascanu, T Mikolov & Y Bengio (2013). On the difficulty of training recurrent neural networks. In *International conference on machine learning*, 1310–1318, PMLR. 20

A Paszke, S Gross, S Chintala, G Chanan, E Yang, Z DeVito, Z Lin, A Desmaison, L Antiga & A Lerer (2017). Automatic differentiation in pytorch. *arxiv*. 6

MC Peel & TA McMahon (2020). Historical development of rainfall-runoff modeling. *Wiley Interdisciplinary Reviews: Water*, **7**, e1471. 41

C Perrin, C Michel & V Andréassian (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of hydrology*, **279**, 275–289. 8

E Peters, P Torfs, HA Van Lanen & G Bier (2003). Propagation of drought through groundwater—a new approach using linear reservoir theory. *Hydrological processes*, **17**, 3023–3040. 36

S Ravuri, K Lenc, M Willson, D Kangin, R Lam, P Mirowski, M Fitzsimons, M Athanassiadou, S Kashem, S Madge *et al.* (2021). Skillful precipitation nowcasting using deep generative models of radar. *arXiv preprint arXiv:2104.00954*. 24

M Reichstein, G Camps-Valls, B Stevens, M Jung, J Denzler, N Carvalhais & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, **566**, 195–204, 10.1038/s41586-019-0912-1. 2, 6, 12, 28, 29, 37, 41, 69

F Rembold, M Meroni, F Urbano, G Csak, H Kerdiles, A Perez-Hoyos, G Lemoine, O Leo & T Negre (2019). Asap: A new global early warning system to detect anomaly hot spots of agricultural production for food security analysis. *Agricultural systems*, **168**, 247–257. 92

R Remesan & J Mathew (2015). Hydroinformatics and data-based modelling issues in hydrology. In *Hydrological Data Driven Modelling*, 19–39, Springer. 14

MT Ribeiro, S Singh & C Guestrin (2016a). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144. 29, 69

MT Ribeiro, S Singh & C Guestrin (2016b). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*. 29

E Robinson, E Blyth, D Clark, E Comyn-Platt, J Finch & A Rudd (2017). Climate hydrology and ecology research support system meteorology dataset for great britain (1961-2015) [chess-met] v1.2. 10.5285/b745e7b1-626c-4ccc-ac27-56582e77b900. 50

F Rosenblatt (1957). *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory. 6

DE Rumelhart, GE Hinton & RJ Williams (1986). Learning representations by back-propagating errors. *nature*, **323**, 533–536. 6, 19

L Samaniego, S Thober, R Kumar, N Wanders, O Rakovec, M Pan, M Zink, J Sheffield, EF Wood & A Marx (2018). Anthropogenic warming exacerbates european soil moisture droughts. *Nature Climate Change*, **8**, 421–426. 84

W Samek, G Montavon, A Vedaldi, LK Hansen & KR Müller (2019). *Explainable AI: interpreting, explaining and visualizing deep learning*, vol. 11700. Springer Nature. 28

AM Schäfer & HG Zimmermann (2006). Recurrent neural networks are universal approximators. In *International Conference on Artificial Neural Networks*, 632–640, Springer. 23

J Schmidhuber (2015). Deep learning in neural networks: An overview. *Neural networks*, **61**, 85–117. 2, 6, 33

L See, A Jain, C Dawson & R Abrahart (2009). Visualisation of hidden neuron behaviour in a neural network rainfall-runoff model. In *Practical Hydroinformatics*, 87–99, Springer. 18

TJ Sejnowski (2018). *The deep learning revolution*. MIT press. 6

TJ Sejnowski (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*, **117**, 30033–30038. 5

SI Seneviratne (2012). Historical drought trends revisited. *Nature*, **491**, 338–339. 36

G Shalev, R El-Yaniv, D Klotz, F Kratzert, A Metzger & S Nevo (2019). Accurate hydrologic modeling using less information. *arXiv preprint arXiv:1911.09427*. 121

LS Shapley (1953). Stochastic games. *Proceedings of the national academy of sciences*, **39**, 1095–1100. 101

J Sheffield & EF Wood (2012). *Drought: past problems and future scenarios*. Routledge. 35

J Sheffield, EF Wood, N Chaney, K Guan, S Sadri, X Yuan, L Olang, A Amani, A Ali, S Demuth *et al.* (2014). A drought monitoring and forecasting system for sub-sahara african water resources and food security. *Bulletin of the American Meteorological Society*, **95**, 861–882. 92

C Shen (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, **54**, 8558–8593, 10.1029/2018WR022643. 2, 5, 6, 41, 42, 69, 93, 116

C Shen, E Laloy, A Albert, FJ Chang, A Elshorbagy, S Ganguly, Kl Hsu, D Kifer, Z Fang, K Fang *et al.* (2018). Hess opinions: Deep learning as a promising avenue toward knowledge discovery in water sciences. *Hydrology and Earth System Sciences Discussions*, **2018**, 1–21. 2, 6, 45

LK Sherman (1932). Streamflow from rainfall by the unit-graph method. *Eng. News Record*, **108**, 501–505. 13

A Shrikumar, P Greenside & A Kundaje (2017). Learning important features through propagating activation differences. In *International Conference on Machine Learning*, 3145–3153, PMLR. 102, 115

M Sivapalan, G Blöschl, L Zhang & R Vertessy (2003a). Downward approach to hydrological prediction. 7

M Sivapalan, K Takeuchi, S Franks, V Gupta, H Karambiri, V Lakshmi, X Liang, J McDonnell, E Mendiondo, P O'connell *et al.* (2003b). Iahs decade on predictions in ungauged

basins (pub), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological sciences journal*, **48**, 857–880. 25

MG Sklash & RN Farvolden (1979). The role of groundwater in storm runoff. *Journal of Hydrology*, **43**, 45–65. 84

LJ Slater, B Anderson, M Buechel, S Dadson, S Han, S Harrigan, T Kelder, K Kowal, T Lees, T Matthews *et al.* (2020). Nonstationary weather and water extremes: a review of methods for their detection, attribution, and management. *Hydrology and Earth System Sciences Discussions*, 1–54. 15, 27

D Solomatine, LM See & R Abrahart (2009). Data-driven modelling: concepts, approaches and experiences. *Practical hydroinformatics*, 17–30. 12

DP Solomatine & KN Dulal (2003). Model trees as an alternative to neural networks in rainfall—runoff modelling. *Hydrological Sciences Journal*, **48**, 399–411. 13

D Spiegelhalter (2020). Should we trust algorithms? *Harvard Data Science Review*, **2**. 69

J Spinoni, G Naumann, JV Vogt & P Barbosa (2015). The biggest drought events in europe from 1950 to 2012. *Journal of Hydrology: Regional Studies*, **3**, 509–524. 35, 92

N Srivastava, G Hinton, A Krizhevsky, I Sutskever & R Salakhutdinov (2014a). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, **15**, 1929–1958. 48

N Srivastava, G Hinton, A Krizhevsky, I Sutskever & R Salakhutdinov (2014b). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, **15**, 1929–1958. 88

M Sundararajan & A Najmi (2020). The many shapley values for model explanation. In *International Conference on Machine Learning*, 9269–9278, PMLR. 29

M Sundararajan, A Taly & Q Yan (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 3319–3328, PMLR. 29, 30, 115

I Sutskever, O Vinyals & QV Le (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112. 6, 124

MD Svoboda, BA Fuchs *et al.* (2016). *Handbook of drought indicators and indices*. World Meteorological Organization Geneva, Switzerland. 92, 96

DL Swain, M Tsiang, M Haugen, D Singh, A Charland, B Rajaratnam & NS Diffenbaugh (2014). The extraordinary california drought of 2013/2014: Character, context, and the role of climate change. *Bulletin of the American Meteorological Society*, **95**, S3. 35, 92

M Tall, C Albergel, B Bonan, Y Zheng, F Guichard, MS Dramé, AT Gaye, LO Sintondji, FC Hountondji, PM Nikiema *et al.* (2019). Towards a long-term reanalysis of land surface variables over western africa: Ldas-monde applied over burkina faso from 2001 to 2018. *Remote Sensing*, **11**, 735. 98

M Tanguy, H Dixon, I Prosdocimi, DG Morris & VDJ Keller (2014). Gridded estimates of daily and monthly areal rainfall for the united kingdom (1890-2012) [ceh-gear]. 10.5285/5dc179dc-f692-49ba-9326-a6893a503f6e. 50

Y Tao, K Hsu, A Ihler, X Gao & S Sorooshian (2018). A two-stage deep neural network framework for precipitation estimation from bispectral satellite information. *Journal of Hydrometeorology*, **19**, 393–408. 24

J Teng, AJ Jakeman, J Vaze, BF Croke, D Dutta & S Kim (2017). Flood inundation modelling: A review of methods, recent advances and uncertainty analysis. *Environmental modelling & software*, **90**, 201–216. 35

J Thielen, J Bartholmes, MH Ramos & Ad Roo (2009). The european flood alert system–part 1: concept and development. *Hydrology and Earth System Sciences*, **13**, 125–140. 8

N Thuerey, P Holl, M Mueller, P Schnell, F Trost & K Um (2021). *Physics-based Deep Learning*. WWW. 121

E Todini (2007). Hydrological catchment modelling: past, present and future. *Hydrology and Earth System Sciences*, **11**, 468–482. 15

G Tramontana, M Jung, CR Schwalm, K Ichii, G Camps-Valls, B Ráduly, M Reichstein, MA Arain, A Cescatti, G Kiely *et al.* (2016). Predicting carbon dioxide and energy fluxes across global fluxnet sites with regression algorithms. *Biogeosciences*, **13**, 4291–4313. 38

P Uhe, S Philip, S Kew, K Shah, J Kimutai, E Mwangi, van GJ Oldenborgh, R Singh, J Arrighi, E Jjemba *et al.* (2018). Attributing drivers of the 2016 kenyan drought. *International Journal of Climatology*, **38**, e554–e568. 13, 98, 108

UK Statistics Authority (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of ai systems in the public sector. *Available at SSRN 3403301*. 69

Cv Van Diepen, J Wolf, H Van Keulen & C Rappoldt (1989). Wofost: a simulation model of crop production. *Soil use and management*, **5**, 16–24. 8, 92

AF Van Loon (2015). Hydrological drought explained. *Wiley Interdisciplinary Reviews: Water*, **2**, 359–392. 35, 92

van HJI Meerveld, JW Kirchner, MJP Vis, RS Assendelft & J Seibert (2019). Expansion and contraction of the flowing stream network alter hillslope flowpath lengths and the shape of the travel time distribution. *Hydrology and Earth System Sciences*, **23**, 4825–4834, 10.5194/hess-23-4825-2019. 64

G Van Rossum *et al.* (2007). Python programming language. In *USENIX annual technical conference*, vol. 41, 36. 46

A Vehtari, A Gelman & J Gabry (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, **27**, 1413–1432. 32

SM Vicente-Serrano, S Beguería & JI López-Moreno (2010). A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *Journal of climate*, **23**, 1696–1718. 92

S Wachter, B Mittelstadt & C Russell (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, **31**, 841. 29

SV Weijs & BL Ruddell (2020). Debates: Does information theory provide a new paradigm for earth science? sharper predictions using occam's digital razor. *Water resources research*, **56**. 10, 33

J Wesemann, M Herrnegger & K Schulz (2018). Hydrological modelling in the anthroposphere: predicting local runoff in a heavily modified high-alpine catchment. *Journal of Mountain Science*, **15**, 921–938. 8

R Wilby, R Abrahart & C Dawson (2003). Detection of conceptual model rainfall—runoff processes inside an artificial neural network. *Hydrological Sciences Journal*, **48**, 163–181. 14, 15, 28, 41, 68, 93

DA Wilhite, MD Svoboda & MJ Hayes (2007). Understanding the complex impacts of drought: A key to enhancing drought mitigation and preparedness. *Water resources management*, **21**, 763–774. 92

Y Yao, L Rosasco & A Caponnetto (2007). On early stopping in gradient descent learning. *Constructive Approximation*, **26**, 289–315. 32

KK Yilmaz, HV Gupta & T Wagener (2008). A process-based diagnostic approach to model evaluation: Application to the nws distributed hydrologic model. *Water Resources Research*, **44**, https://doi.org/10.1029/2007WR006716. 51, 54

P Young (1998). Data-based mechanistic modelling of environmental, ecological, economic and engineering systems. *Environmental Modelling & Software*, **13**, 105–122. 41, 68

P Young (2003). Top-down and data-based mechanistic modelling of rainfall–flow dynamics at the catchment scale. *Hydrological processes*, **17**, 2195–2217. 2, 16, 18, 28, 31, 41, 64, 68

P Young & A Chotai (2001). Data-based mechanistic modeling, forecasting, and control. *IEEE Control systems magazine*, **21**, 14–27. 17

PC Young (2002). Advances in real–time flood forecasting. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, **360**, 1433–1450. 17

PC Young & KJ Beven (1994). Data-based mechanistic modelling and the rainfall-flow non-linearity. *Environmetrics*, **5**, 335–363. 41, 68

N Zeng, JH Yoon, JA Marengo, A Subramaniam, CA Nobre, A Mariotti & JD Neelin (2008). Causes and impacts of the 2005 amazon drought. *Environmental Research Letters*, **3**, 014002. 35, 92

284 references in total.