Applications of information theory and machine learning for hydrologic modeling

Andrew Robert Bennett

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Bart Nijssen, Chair

Erkan Istanbulluoglu

Martyn P. Clark

Grey S. Nearing

Program Authorized to Offer Degree:

Civil and Environmental Engineering

University of Washington

Abstract

Applications of information theory and machine learning for hydrologic modeling

Andrew Robert Bennett

Chair of the Supervisory Committee:

Bart Nijssen

Civil and Environmental Engineering

An explosion of new data sources, expansion of computing resources, and theoretical advances in data science have spurred the rapid adaptation of data-driven methods in earth system science, including hydrology. In this dissertation I will describe three applications of data-driven methods with applications to hydrologic modeling. In chapter 2 I present a framework for hydrologic model intercomparison which examines process interactions within a process-based hydrologic model (PBHM). I show that taking a more holistic approach can shed light into the functioning of these complex models. In chapter 3 I couple machine learned representations of turbulent heat fluxes into a PBHM, and show that neural networks can provide better predictions and transferability than the process-based equations that are used in PBHMs. Building on this, in chapter 4 I use explainable AI (XAI) methods to examine what the neural network has learned. I find that the neural network is able to learn physically plausible relationships and can identify

how to partition between latent and sensible heat fluxes based only on short-term temporal data. I also show how we can use XAI to examine what neural networks have learned between sites. This method can uncover that certain sites can be used as predictors for many other sites, as well as that site specific traits such as vegetation type play a large role in the neural network's ability to generalize to sites it was not trained on. Finally, based on the findings of these three applications I discuss in Chapter 5 how data-driven techniques in general can contribute to improved hydrologic understanding.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I have been lucky to have a wide and illustrious cast of collaborators without whom this work would not be possible. First and foremost, I would like to thank Bart Nijssen for his superlative guidance. He has truly taught me to have an unwavering critical eye and has provided me a grounding in the practice of science that will remain invaluable for the rest of my life. Not to mention getting hot takes on Dutch licorice or obscure Wikipedia pages, but those may be for a different PhD.

To my committee members who have provided guidance and insight including Martyn Clark, Erkan Istanbulluoglu, Jessica Lundquist, and Grey Nearing. Your thoughts and insights continue to propel the hydrologic sciences forward. To the rest of the Computational Hydrology Group, I would not be here without you. From whiteboard sessions in trying to understand how to build datastructures to trying to get all of the lab dogs to pose for a quick picture, Wilson Ceramics Laboratory was a great place to "grow up" academically. On that thought, I also must acknowledge the amazing lab dogs – Chomsky, Jake, Leo, Mika, and Taco (and some cameos). And for B, the cat of my dreams.

To my people. Especially those I may have taken more from than I have given. My parents, Mike and Mary have always encouraged and supported me, and made my whole being possible with their care. To my old standbys, Brendan Carroll, Josh Goodwin, Andy Erickson, Max Pschorr, and Nick Pomplun, who have always been up for hosting me in an off-leg of a conference journey or just getting out there for a quick backpacking trip. To the newer ones, Trevor Daviscourt, Chris Kemly, Kate Leigh, Andra Bose, Carolyn Choudhary, Abhishek

Choudhary, Alex Griffiths, Ruben Conner, and Rhiannon Bronstein, thanks for letting me show up and clown around.

Joe Hamman was always ready to blow my mind with a tech demo of something when I was just trying to fill up my water bottle. Yifan Cheng was always there for building the most positive vibe and being willing to always talk through a complex issue and make it simple. Oriana Chegwidden provided the enthusiasm and inspiration for innumerable side projects and might was well have a doctorate in party planning. Abby Rhinehart reminded me of the beauty of stopping and looking at what's right beneath your feet, as well as being the instigator of too many adventures to count. To all the above for being able to nerd-snipe me repeatedly.

In random order, but with no less distinction, my development as a scientist and hydrologist would not be possible without input from Steven Weijs, Hoshin Gupta, Alden Sampson, Uwe Ehret, Wouter Knoben, Andy Wood, Tushar Khurana, Adi Stein, Katherine Evans, Joseph Kennedy, Liz Clark, Young Don-Choi, Naoki Mizukami, Nicoleta Cristea, Marketa McGuire, Jay Jay Billings, Diana Gergel, Ethan Gutmann, Yixin Mao.

I would also like to thank the Bureau of Reclamation, National Aeronautics and Space Administration, and National Oceanic and Atmospheric Administration for project funding not directly related to the work presented here which made my PhD studies possible.

Chapter 1. INTRODUCTION

1.1 Background and Motivation

This dissertation is about how data-driven methods can be used in a variety of ways in the hydrologic sciences. Many of the problems in modern hydrology have not been solvable by empirical models or from theories which are derived from first principles (Blöschl et al., 2019). Advances in data-driven methods for science and engineering as well as increases in the amount of data, computational power, and data analysis tools are causing shifts in the ways in which science and engineering applications are approached (Brunton & Kutz, 2019). In hydrology this new perspective has begun to be widely acknowledged (Nearing et al., 2020; Peters-Lidard et al., 2018; Sivapalan & Blöschl, 2017). This shift towards data-intensive computing has been referred to as a fourth paradigm for science, alongside those of theory, experimentation, and simulation (Hey et al., 2009). That is not to say that the adoption of data-driven methods has been abrupt or unanticipated. Tukey (1962) laid out a convincing and continuously relevant treatise that appears very similar to the modern field of data science. In hydrology this was anticipated by Dooge (1988).

I present three studies from the data-driven fields of information theory and machine learning. Both of these fields have long histories in hydrology but have shown a recent surge in use due to the advances described above. Information theory is a branch of mathematics which deals with communication and spans a wide range of useful applications (Cover & Thomas, 2006). In hydrology and Earth sciences, information theory has been used for uncertainty quantification (Gong et al., 2013), the construction of parsimonious model structures (Singh & Guo, 1995), and understanding process interaction (Goodwell & Kumar, 2017). I use the last

concept to develop a model intercomparison framework that allows for insights into how total model implementation affects outputs of key hydrologic model outputs.

Similarly, machine learning (and particularly deep learning) has seen a huge increase in successful applications in a large number of fields, including hydrology (Shen, 2018). Some applications of machine learning to hydrology include replacing entire rainfall-runoff models with deep-learned variants (Kratzert et al., 2018), developing novel approaches for assimilating alternative sources of data to improve forecasts (Feng et al., 2020), optimizing hydropower generation by improving forecasts (Ahmad & Hossain, 2019), and providing high-resolution land cover classifications (Geng et al., 2015). I show that machine learning can be merged with more traditional process-based modeling approaches in Chapter 3, and then show that there are methods we can use to provide explanations for their operation.

1.1 Overview of Research Objectives

In this dissertation I will explore three data-driven methods for simulating and evaluating hydrologic processes. Specifically, I will pursue three areas of inquiry:

1. How can we meaningfully compare entire macroscale hydrologic models, which contain a variety of representations of hydrologic processes?
2. How can machine learning (ML) models be incorporated into process based hydrologic models (PBHMs)?
3. Do ML models of hydrologic processes learn physical behavior, and how can we learn to generalize from them?

In Chapter 2 I present a model intercomparison experiment that examined how model structures differ in a physically meaningful way rather than which model provides the best

predictions. I used three hydrologic models to simulate the hydrologic cycle in the Columbia river basin (CRB) in the Pacific Northwestern United States. These models each had different structures, process parameterizations, parameter values, but shared input meteorological forcing data. By examining the components of the water balance, consisting of precipitation, runoff, evapotranspiration, soil moisture, and snowpack, as represented by these different models, I demonstrate that an information theoretic measure called transfer entropy can compute a model "fingerprint" that can highlight processes interactions. Transfer entropy provides a way to quantify how knowledge of one process relates to the predictability of another, and is increasingly being used in hydrology to disentangle complex systems (Goodwell & Kumar, 2017; Ombadi et al., 2020; Ruddell et al., 2019). I show how we can use transfer entropy for model diagnostics and to quantify the process connections between each of the water balance components. Using this methodology, I analyzed four different hydroclimates in the CRB and compared process connectivity in the three different model configurations.

In Chapter 3 I incorporated ML-based process parameterizations directly into PBHMs, and showed that they provide not only more performant predictions at sub-diurnal timescales, but that they also better represent long term constraints. Specifically, I added an option for the turbulent heat flux parameterization that consists of a deep-learned neural network trained on observations from FluxNet sites. I developed two versions of this parameterization, both of which were directly coupled into the SUMMA hydrologic model. The first is a one-way coupling, which uses only SUMMA input forcings and parameters to predict latent and sensible heat fluxes. The second is a two-way coupling, which combines SUMMA derived fluxes and states with the inputs used for the one-way coupling. I show that this coupling provided not only accurate predictions, but was able to reproduce physical signatures that the DL models were not

explicitly trained to reproduce. Further, I show that when compared to PBHM simulations that were calibrated in-sample, the DL based parameterizations, which were trained out of sample, were routinely able to out-perform the PBHM. This further demonstrates that data-driven models can learn representations of complex phenomena in a way that exceeds the abilities of empirical or theoretical methods.

In Chapter 4 I explore how the types of models explored in Chapter 3 were able to learn such performant behaviors. I used a technique called Layerwise Relevance Propagation (LRP) to determine which inputs to the neural networks contributed most their predictions and how these contributions varied over time. I show that interpretable methods can connect data-driven methods to both empirical and theory-driven modeling applications and that the neural network was largely able to learn physically realistic transformations of the input data to generate estimates of turbulent heat fluxes. I also developed a novel method that allows for deeper understanding of what data-driven models have learned between sites. I show that general hydrologic principles are common between sites even though site-specific behaviors form a dominant control on model transferability.

In each chapter of this dissertation I used large amounts of data, either generated via simulation or measurement, to find insights into the hydrologic cycle. In doing so I highlight how data-driven methods can be used in all facets of hydrology. As data availability increases and methods improve, I expect the use of data-driven methods in Earth science to become the norm rather than the exception.

Chapter 2. QUANTIFYING PROCESS CONNECTIVITY WITH TRANSFER ENTROPY IN HYDROLOGIC MODELS

This chapter is published in the journal *Water Resources Research*. The supplemental material for this chapter is provided in Appendix A.

Bennett, A., Nijssen, B., Ou, G., Clark, M., & Nearing, G. (2019). Quantifying process connectivity with transfer entropy in hydrologic models. *Water Resources Research*, 55, 4613–4629. https://doi.org/10.1029/2018WR024555

2.1 INTRODUCTION

Approaches to hydrologic modeling show great diversity, reflecting the community's varying philosophical and practical viewpoints on the role of modeling in hydrology (Beven, 2002; Beven et al., 2012; Sivapalan et al., 2003; Wood et al., 2012). One framework for breaking down the diversity of hydrologic models describes the complexity introduced at each formal step of the model development process (Gupta & Nearing, 2014). This framework provides an informal methodology for classifying the complexity of model structures. In aggregate, models may differ at varying levels of this hierarchy by representing different processes, using different functional forms for the process parameterizations, or utilize different numerical methods of solutions to these equations. A more in-depth classification scheme described in (Kampf & Burges, 2007a) further reinforces the idea that hydrologic modeling involves a large number of subjective decisions.

The particular set of decisions used in constructing different models may produce different results even given similar input data (Best et al., 2015; Sellers et al., 1993). It is necessarily true that only one correct representation of any given system exists, however it is unlikely that we will be able to find it. Instead, model diversity is often used to characterize some range of behavior, which can be thought to represent possible outcomes. In this framework, hydrologic models may

be more appropriately thought of as hypotheses which can be tested (Beven et al., 2012; Clark et al., 2011a) or in a probabilistic sense which can be used to represent our uncertainty (Koutsoyiannis, 2005; Nearing et al., 2016; Weijs et al., 2010). Thus, we recognize that each of the steps in our modeling hierarchy is associated with some uncertainty which dictates the utility of the model for a particular application.

Model intercomparison and benchmarking experiments aim to understand the diversity of model behavior (Breuer et al., 2009; Sellers et al., 1991; Smith et al., 2013). While these experiments are able to characterize the differences in land surface models, the reasons for these differences are difficult to understand in part due to the lack of standardization of model structures (Koster & Milly, 1997; Nijssen et al., 2003). Other complications such as differences in parameters and numerical solver implementations can also contribute to this difficulty (Kavetski & Clark, 2010; Nearing et al., 2016a). This paper concerns itself with determining the overall effects that the totality of model implementation differences has on the generated outputs, including all of the factors described above.

Model intercomparison experiments often rely on metrics that are descriptive, focusing on aggregate differences of timeseries or spatial distributions (Clark et al., 2011b). Common choices for difference metrics include the root mean square error, skewness, or kurtosis. Each of these metrics highlights a different facet of model performance, so care must be taken in choosing the correct metric or weighting of metrics for ranking models depending on the goals of the analysis (Ritter & Muñoz-Carpena, 2013).

Additional metrics specifically derived for hydrologic systems such as the runoff ratio and aridity index are also commonly used to understand the behavior of models. Just as with performance metrics, these hydrologically motivated quantities highlight specific aspects of model

behavior and must be chosen in accordance with a specific goal.

Neither set of metrics accounts for nonlinearities in process representation which are ubiquitous in hydrologic models (Weijs et al., 2010). This omission may obscure features of model behavior. Model intercomparison experiments conducted this way ask, "What are the differences in model performance?", but do not necessarily provide insight into why or how models differ. To reason effectively about model improvement we must also ask why the differences between the models exist and develop tools and metrics that help us answer that question.

The ability to answer this question requires more advanced techniques which can account for the full range of model behavior (including nonlinearities, feedbacks, and emergent behavior), as well as being able to decompose individual processes. Information theory is one approach that can be used for this purpose. The use of information theory-based methods in the hydrologic sciences has a long history of diverse applications. For an overview of the development of these methods as well as a history of hydrologic applications see Singh (1997). Weijs et al. (2010) determined that a divergence score based on information theory had properties which lent themselves to the probabilistic view of modeling discussed previously. Following a similar argument, Gong et al. (2013) laid out a methodology for quantifying the amount of random uncertainty (that is, not due to a lack of correctness in process representations) and the total model structure uncertainty. Nearing, et al. 2016b) expanded on this work to quantify the uncertainties in boundary conditions and model parameters. Building on this general technique we will use similar methods to quantify the interaction between pairs of individual processes.

Our approach uses the time asymmetric quantity known as transfer entropy (Schreiber, 2000) to quantify how much information is transferred between mass flux terms of the water balance. We evaluate model output over a range of hydrologic regimes using these methods from which we

construct process networks that allow us to reason how model structure ultimately affects model output. Characterization of dynamics as process networks has been used in the hydrologic sciences mostly to analyze and understand observational data (Goodwell & Kumar, 2017; Ruddell & Kumar, 2009a, 2009b; Sendrowski & Passalacqua, 2017). Additionally, this technique has proven popular in other disciplines such as climate dynamics and biological engineering (Lee et al., 2012; Paluš, 2014; Runge et al., 2015; Sun et al., 2014).

We demonstrate these techniques as an evaluation tool for estimating process connectivity within hydrologic models to provide insight into how they operate and how they differ. We show that process level connectivity is not necessarily linked to the similarity of output timeseries behavior. By extension we show how, in some cases, mediating variables can have impacts on model output, thereby providing a level of understanding that would not be found using commonly-used error metrics.

2.2 METHODS

2.2.1 **Information metrics**

The basis of information theory was developed by Shannon (1948), who defined the quantity of information entropy for a continuous random variable, X, as

$$H(X) = -\int_X p(x) \log\big(p(x)\big)\, dx \tag{2.1}$$

where $p(x)$ is the associated probability density function. The multivariate case, or joint entropy, is given by

$$H(X,Y) = -\int_X \int_Y p(x,y) \log\big(p(x,y)\big)\, dy\, dx \tag{2.2}$$

where Y is another continuous random variable. We can relate the amount of shared information

between two continuous variables, X and Y, by the mutual information (Cover & Thomas, 2005)

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \quad 2.3$$

The mutual information can be intuitively thought of as the knowledge we gain about Y from measuring X, or vice-versa. From these definitions we can set out definitions for conditional forms of these quantities. We write the conditional entropy as

$$H(X|Y) = H(X) - I(X;Y) \quad 2.4$$

and the conditional mutual information (given a third random variable Z) is defined as

$$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z) \quad 2.5$$

For a more comprehensive overview of these quantities see Cover & Thomas (2005).

These quantities are all symmetric (except when the conditioning variable is changed), which means that they do not tell us anything about the amount each variable contributes individually to shared information. A method for accounting for information transfer from one variable to another was developed by Schreiber (2000) and has become a popular tool for estimating causal effects, timescales, and coupling strengths (Frenzel & Pompe, 2007; Hlinka et al., 2013; Ruddell & Kumar, 2009a). The quantity, known as transfer entropy, can be written as a conditional mutual information (Hlaváčková-Schindler et al., 2007)

$$T_{X\to Y} = I(Y; X^{-}|Y^{-}) \quad 2.6$$

where $T_{X\to Y}$ is the transfer entropy from X to Y, and X^{-} and Y^{-} denote the (potentially infinite) history of the variables X and Y, respectively. Using the full timeseries causes these calculations to become very high-dimensional and thus computationally intractable. It is often practical to designate some parameters to limit the history window (Ruddell & Kumar, 2009a). Generally, four parameters are used. The parameters τ and ω represent the time lags for the dependent and target variables, while k and l represent the dependent and target variable history window sizes. Then,

we can write the transfer entropy as

$$T_{X\to Y}(\tau,\omega,k,l) = I(Y_t; X_{t-\tau-k:t-\tau}|Y_{t-\omega-l:t-\omega}) \quad 2.7$$

where the ranges $a{:}\,b$ are inclusive only on the right (that is, encloses the range (a, b]) and denote vectors of the random variables over these ranges. When choosing these parameter values there is a tradeoff between computational complexity and estimation accuracy and stability (Schreiber, 2000). Choosing $\omega = 1$ is a natural choice, which only conditions on the immediately preceding history of the target variable. It is common to choose either $k = l$ or $l = 0$. If k and/or l are chosen to be large, the reliability of the estimate of transfer entropy is decreased (Hlinka et al., 2013). This decrease in reliability is a fundamental issue with estimating high dimensional probability distributions due to the curse of dimensionality (Weijs et al., 2018). To minimize these effects, we set $\tau = 1, \omega = 1, k = 0,$ and $l = 0$, and use the temporal resolution of the timeseries as the method of choosing a timescale. This simplification effectively encodes the assumption that the system under investigation possesses the Markov property, which says that the current state of the system depends only on the directly previous timestep. Our parameterization of transfer entropy is then given by

$$T_{X\to Y} = I(Y_t; X_{t-1}|Y_{t-1}) \quad 2.8$$

All quantities described thus far have a common unit which is dependent on the base of the logarithm used. We will use the natural logarithm, which gives information quantities in units of nats. This common unit allows cross variable comparison.

As a further complication, the probability distributions required to calculate these information and entropy measures are not generally known, so they must be estimated to calculate approximate quantities. There are many ways to estimate these distributions with varying degrees of complexity, bias, and stability (Gong et al., 2014; Hlaváčková-Schindler et al., 2007; Paninski,

2003). Estimation of these distributions is usually split into parametric and non-parametric techniques. Parametric methods assume some class of underlying distribution, which is then fit to the data based on some parameters, whereas non-parametric techniques do not assume any specific underlying distribution. Usage of either technique is context dependent. Previously, in hydrologic applications the classical histogram method has been popular. This is a non-parametric method, meaning it is applicable universally. However, entropy-based quantities computed with these techniques are known to have positive bias and can be unstable in high-dimensional calculations (Hlaváčková-Schindler et al., 2007). Goodwell & Kumar (2017) made use of kernel density estimators, a parametric approach, to avoid these challenges.

Our analysis will use k-nearest neighbor estimators, a non-parametric approach which has some advantages over the histogram method for large and high-dimensional datasets. It also has an advantage over kernel density estimators in that it does not assume any particular form of the density distribution, though it does require a larger amount of data for convergence. Nearest-neighbor estimators have become popular for large datasets because of their general applicability and scalable nature. We have chosen the estimators and parameters to minimize the average amount of bias for each computed information transfer. The simplified functional form of these estimators is dependent on the distance norm used to perform the neighbor search; we use the L^{∞} (i.e. the maximum or Chebyshev) norm, which for two vectors is the maximum difference along any coordinate dimension. Intuitively, these estimators provide an estimate of the local density parameterized by the distance to the k^{th} nearest neighbor. The further that distance, the less dense the probability at that value.

To describe our estimators, we begin with some notational conventions. We denote the distance to the k^{th} nearest-neighbor as ρ_k, the dimension of the space as d, the number of data points as N.

The digamma function is signified by ψ. The mean will be written as $\langle\cdot\rangle$.Then, n_x is the number of data points within $\langle\rho_k\rangle$ when projected into the subspace spanned by the domain of X (and similarly for Y). The volume of a d-dimensional unit hyper-ball is written as C_d. Estimator-based quantities are denoted with a hat. In all cases we fix the estimator to use a value of $k = 10$. In our application this value strikes a balance between bias and noise and is in line with values chosen in the literature (Kraskov et al., 2004; Runge et al., 2017).

Following these definitions the entropy estimator (Goria et al., 2005), the mutual information estimator (Kraskov et al., 2004) and the conditional mutual information estimator (Vlachos & Kugiumtzis, 2010) are given by

$$\hat{H}(X) = d\ln(\langle\rho_k\rangle) + \psi(N) - \psi(k) + \ln(C_d) \tag{2.10}$$

$$\hat{I}(X;Y) = \psi(k) - \frac{1}{k} - \langle\psi(n_x) + \psi(n_y)\rangle + \psi(N) \tag{2.11}$$

$$\hat{I}(X;Y|Z) = \psi(k) - \langle\psi(n_{xz}) + \psi(n_{yz}) - \psi(n_z)\rangle \tag{2.12}$$

Our transfer entropy estimator is then given by

$$\hat{T}_{X\to Y} = \hat{I}(Y_t; X_{t-1}|Y_{t-1}) \tag{2.13}$$

We compute the pairwise connections between variables in the model output and visualize the resulting information transfer network as a chord diagram. A simple template along with a description of its interpretation are shown in Figure . Chord diagrams display relative influences and provide a high-level understanding of the structure of the information flow networks enabling the identification of couplings. The total size of the outer arc lengths should not be interpreted as a total, unique depiction of information transferred to a process. Information flows that contribute to this length may contain redundant or synergistic components (Goodwell & Kumar, 2017; Weijs et al., 2018).

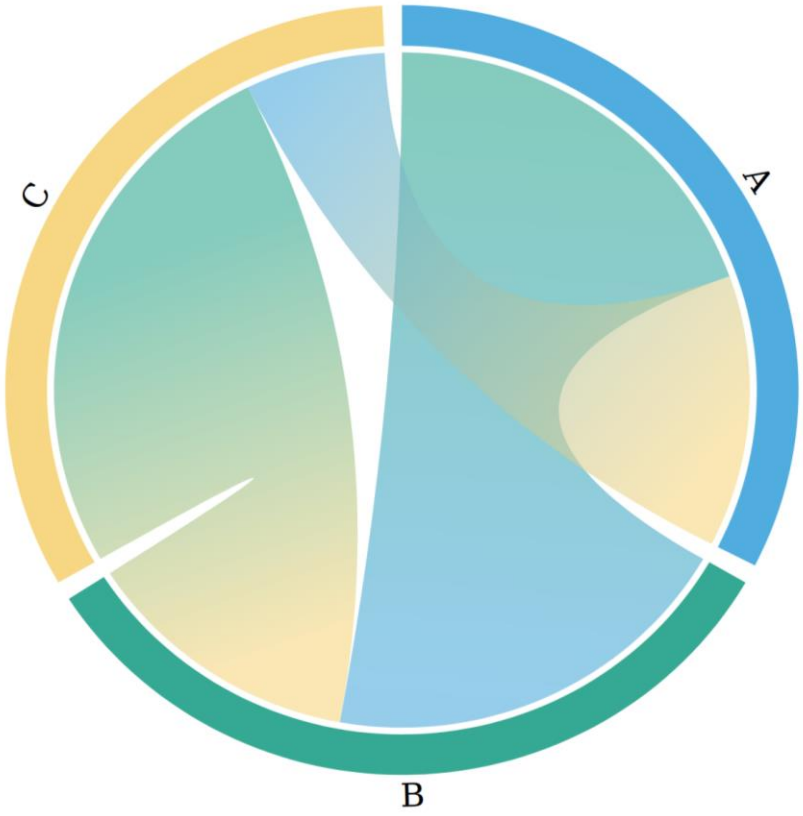

Figure 2.1 Template illustration of an information transfer network chord diagram. The outer circle is comprised of arcs whose relative lengths correspond to the sum of information received from other sources. The inner sections are comprised of chords, or ribbons, which indicate the direction and magnitude of information transfer. Note that the chords are asymmetric, with the coloration at the end of each chord indicating the source variable of the transfer. To illustrate this, consider the arc labeled A. Each of the other variables, B and C, transfer information to A at variable rates, indicated by the green and yellow chord ends that terminate at A, respectively. Similarly, B and C receive information at variable rates.

2.2.2 **Study domain**

Our study domain consists of the Columbia River Basin (CRB) and its adjacent coastal drainage areas located in the Pacific Northwest region of North America as shown in Figure 2.2. The total domain covers an area of 810,000 km^2, including southeastern British Columbia, Canada and most of the U.S. states of Idaho, Oregon, and Washington, the western part of Montana, as well as small portions of California, Nevada, Utah, and Wyoming. The hydroclimate in this region is highly diverse, making it ideal for comparing process representations. In the west, it is dominated by

moist, maritime conditions. The eastern portions are dominated by the high elevation Rocky Mountains. Laying between the Rocky Mountains and the Cascade Mountains is the Columbia Plateau which has a semi-arid climate due to the rain shadow cast by the Cascade Mountains.

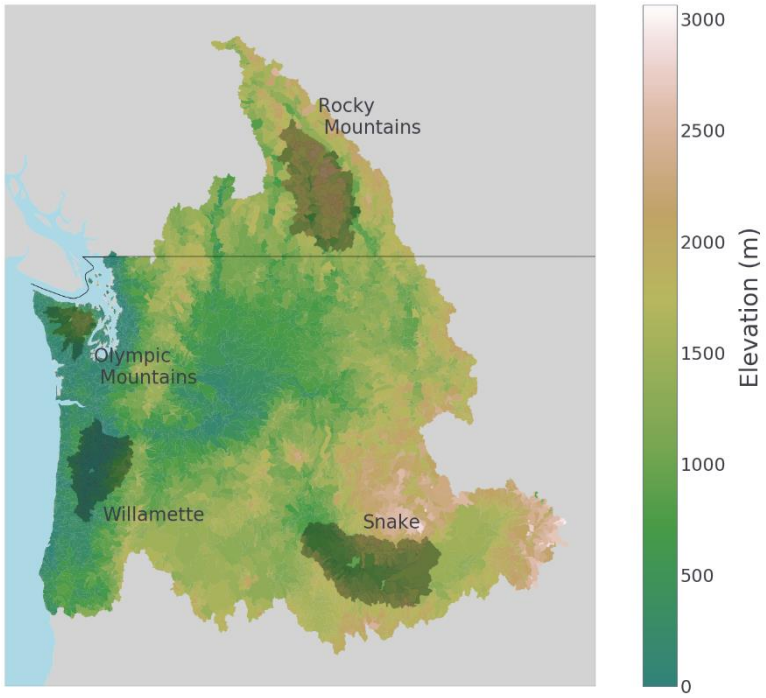

Figure 2.2. Simulation domain map (showing the SUMMA spatial discretization). Background coloration shows elevation, grey highlights are the analysis regions.

Within this domain we further analyze four distinct hydrologic regions whose basic characterizations are shown below:

- Snake River – Arid, warm
- Canadian Rocky Mountains – High-elevation, snow-dominant
- Olympic Mountains – Wet, high-elevation, large seasonal cycle
- Willamette River – Wet, warm

2.2.3 **Modeling setup**

The CRB domain was simulated for the period of 1950-2011 at sub-daily timestep with three different distributed hydrologic model setups. We used the Variable Infiltration Capacity (VIC) (Liang, 1994), Precipitation Runoff Modeling System (PRMS) (Leavesley et al., 1983), and a reference implementation of SUMMA (Clark, et al., 2015b; Clark, et al., 2015a). Both the VIC and PRMS setups were run at $1/16^{th}$ degree spatial resolution (consisting of 23,929 grid cells) and 3 hourly timestep, while the SUMMA instance was run using hydrologic response units (HRU) as its spatial discretization and an hourly timestep. The spatial discretization for SUMMA was derived from the United States Geologic Survey Geospatial Fabric (Viger & Bock, 2014), resulting in 11,723 HRUs. A description of the modeling decisions used in the SUMMA instance can be found in Table A1 of the supplementary materials.

All three model setups were forced using the dataset developed by (Livneh et al., 2013). The daily forcing data was disaggregated to the appropriate timestep using the Mountain Micro Climate Simulator (MTCLIM) (Bohn et al., 2013; Thornton & Running, 1999). MTCLIM was also used to provide forcing estimates for air pressure, specific humidity, shortwave radiation, and longwave radiation. The gridded data was interpolated onto the SUMMA HRU discretization using area-weighted averaging. The VIC and PRMS model setups were calibrated on runoff (RMJOC, 2018; Chegwidden et al., 2018), while the SUMMA implementation remained uncalibrated. SUMMA parameters were specified based on a combination of default values as well as some tuning for similar values to VIC and PRMS with respect to partitioning of precipitation between snow and rain as well as snow albedo values. Both PRMS and SUMMA were initialized with 10 year spin up periods. VIC, which in this application included a simple glacier model (Chegwidden et al., 2018), was initialized with a 210 year spin up period to account for the longer memory of the

glacier. For all models, the analysis period was 1960-2010. We expect that changes in the parameters and/or structures of any of these modeling setups will have impacts on the resulting information transfer networks because of changes in the marginal probabilities of any pair of processes. However, our main intent here is to describe how information metrics such as transfer entropy can be used to provide insight into differences in process parameterizations rather than examine differences in model calibrations.

2.2.4 **Experimental details**

For this experiment we restricted our analysis to the water balance. None of these model setups contained any lateral flow between elements or a regional groundwater aquifer, so we do not include a specific groundwater component in the water balance equation. We first aggregated the model output into daily values of soil moisture (SM), evapotranspiration (ET), precipitation (P), snow water equivalent (SWE), and runoff (R). Then, the water balance equation is

$$0 = \mathrm{P} - \mathrm{ET} - \mathrm{R} - \Delta\mathrm{SWE} - \Delta\mathrm{SM}$$

where $\Delta\mathrm{SWE}$ and $\Delta\mathrm{SM}$ are the change in SWE and SM over the course of the day for which the remaining terms are averages. From this breakdown we compare the seasonal water balance for the sub-regions described above. We define the seasons so that summer is the months of June, July, and August (JJA); fall is the months of September, October, and November (SON); winter is the months December, January, and February (DJF); and spring is the months of March, April, and May (MAM).

We then calculate both the transfer entropy for each pair of variables at a daily time scale as well as a monthly timescale. Both calculations use the same formulation of lag 1 transfer entropy, with lag 1 representing a single day for the daily time scale and a single month for the monthly time scale. For the monthly time scale, we first aggregated the time series to monthly values. To

ensure that the calculated information transfer is not a numerical artifact we also implement a randomized shuffled surrogate test where the order of the source and conditioning variables is shuffled and the information transfer is computed with these randomized timeseries (Ruddell & Kumar, 2009a). All results are reported for exceedance of a p-value of 0.01. Values which are not statistically significant are reverted to the null hypothesis that no information transfer occurs between variables.

The calculation on the water balance variables of the model output gives us a method for quantifying model structure. This is an important first step to show that this methodology is able to adequately characterize difference in model structures. By conducting a synthetic study on model output we are more clearly able to investigate the utility of viewing information transfers as a proxy for process connectivity since the terms are forced to balance, and we have clear definitions of what each signal represents. We find that these information transfer networks provide a good indication of the hydrologic state at multiple time scales, effectively quantifying the model structure in a concise manner.

2.3 RESULTS

The mean seasonal water balance for each of the sites is shown Figure 2.3. This illustrates that all three models conserve mass, which may not be apparent in the average yearly timeseries plots that will be shown. We can see that each of the four regions have significantly different seasonal fluxes (note changes in vertical scale in Figure 2.3 between the sites). The inter-model variability for each region gives us an initial view into aggregate model behavior. For instance, PRMS tends to have the smallest summer fluxes of the three models. The largest aggregate differences between models occur for evapotranspiration and changes in soil moisture. There is a small but noticeable discrepancy in the amounts of precipitation in the SUMMA instance due to the difference in

gridding and domain selection; particularly the Willamette (panel b) shows this difference. This region is especially sensitive due to orographic precipitation occurring on the eastern border of the domain caused by the Cascade Mountain range.

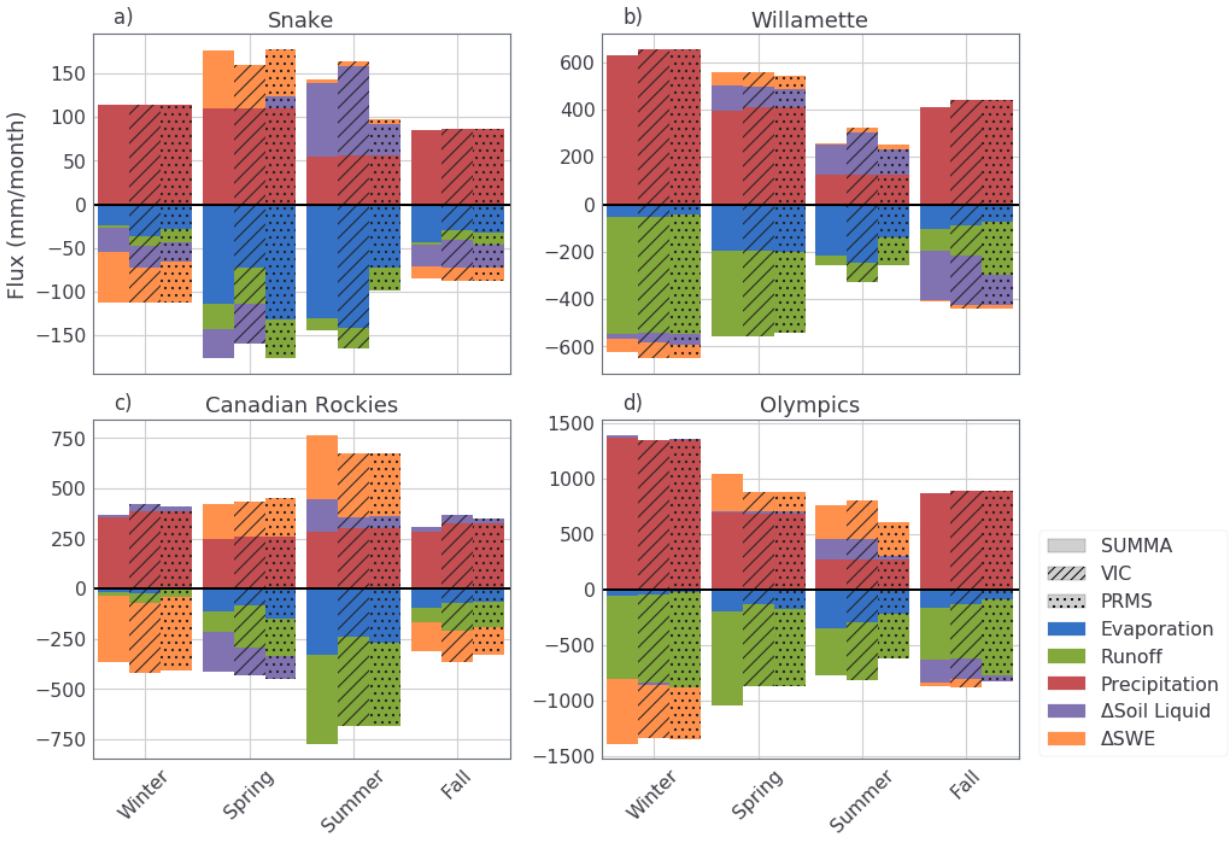

Figure 2.3. Seasonal water balance across the four selected regions for each of the three models.

We also compute the runoff ratio (R/P) for each combination of location and model (Table 1). The SUMMA instance shows the lowest values across all sites which may result from being uncalibrated (while VIC and PRMS were). SUMMA's runoff ratios in the Snake and Rocky Mountain regions are noticeably lower than VIC and PRMS. We will examine how this affects the information transferred to and from runoff.

	Snake	Willamette	Rocky Mountains	Olympic Mountains
SUMMA	0.136	0.643	0.534	0.768
VIC	0.234	0.653	0.670	0.797
PRMS	0.272	0.729	0.602	0.825

Table 2.1. Runoff ratio (R/P) for each region and model setup.

To explore how the runoff ratios relate to model behavior we compute the transfer entropy between each pair of variables in the water balance equation from the period of 1960-2009. The information transfer networks are displayed as chord diagrams. When the average behavior of the models computed via information transfer networks requires deeper inspection, we examine how these connections vary over the course of the water year.

2.3.1 **Snake River region**

The timeseries showing the daily median value for each day of the year along with the middle 50th interquartile range for the Snake region (Figure 2.4) displays several features of interest. The timeseries for ET (panel a) shows that PRMS behaves much differently than either SUMMA or VIC. PRMS peaks in ET earlier in the water year and has a much lower peak. VIC shows abrupt increases in ET at the start of May and June, dictated by a monthly varying leaf area index (LAI), which is used in its ET calculations. Soil moisture (panel b) features similar shapes for SUMMA and VIC with a predominant peak in the spring months. PRMS shows a much lower soil moisture dynamic range than the other models. SUMMA is most different than the other models in SWE (panel c) and runoff (panel d). SUMMA shows much larger snow accumulation, although the length of the snow season is not noticeably longer. The runoff in SUMMA generally features a lower baseline as well as less high frequency variability through most of the year and a large and abrupt spike in the runoff in the spring months as a result of snowmelt.

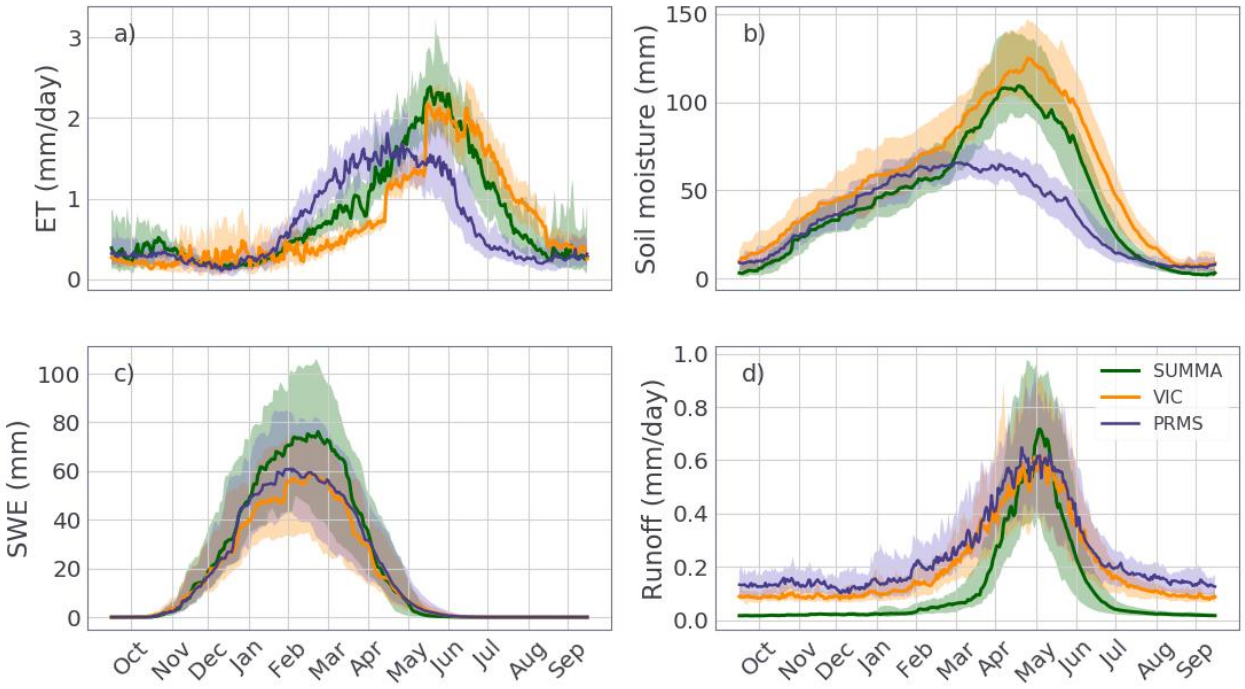

Figure 2.4. Median and 50% interquartile bounds for the four output variables of interest in the Snake region over the analysis period 1960-2010

Figure 2.5 shows a summary of the lag 1-day transfer entropy networks for each model and allows us to see how these differences in timeseries can affect other variables. These calculations are performed on the entire 50-year record, and thus show the average information transfer network for the Snake region. The largest differences in information transfers between the models is in runoff. Both VIC (panel b) and PRMS (panel c) show multiple sources contributing to runoff, with changes in soil moisture and precipitation dominating the information received. On the other hand, SUMMA does not receive information from any term except for changes in SWE (indicated by orange in the transfer entropy networks). This confirms that the spring spike in runoff can be attributed to snowmelt. Further, this single source of received information explains SUMMA's lower runoff from summer to late winter, showing that processes such as precipitation which can occur throughout the year do not seem to contribute directly to runoff at a 1-day lag in this

SUMMA implementation, as they do in VIC and PRMS. We note that if we were to only have the chord diagrams we would not be able to diagnose this behavior, but that we are able to build on our understanding by examining the timeseries and the chord diagrams simultaneously.

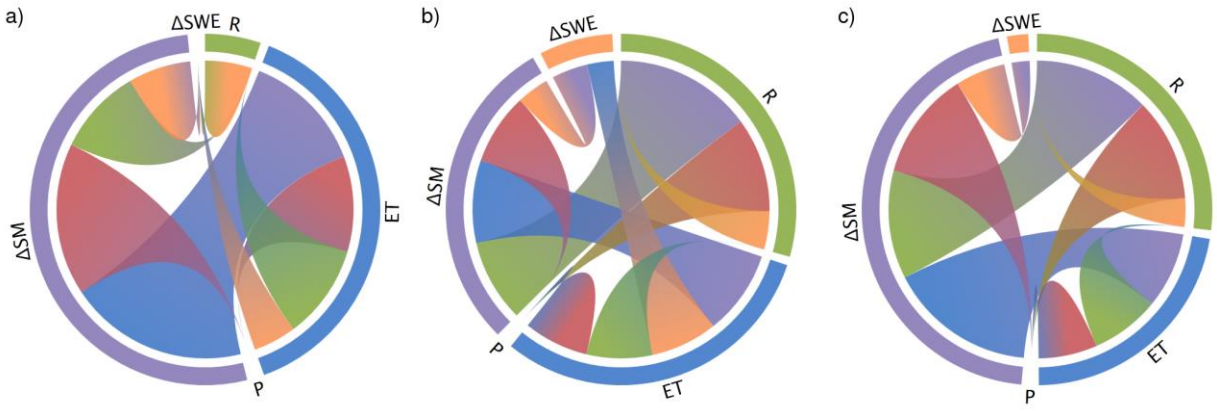

Figure 2.5. Lag 1 transfer entropies in the Snake region for SUMMA (panel a), VIC (panel b), and PRMS (panel c).

Another feature in Figure 2.4 and Figure 2.5 is that SUMMA and VIC are more similar in the soil moisture time series, even though SUMMA and PRMS are more similar for the information received by changes in soil moisture. Both SUMMA and VIC show an influence from ΔSWE on ET, while PRMS does not. This influence from ΔSWE on ET in SUMMA and VIC is communicated in turn to ΔSM and results in time series for soil moisture that show a distinct peak during the spring period.

2.3.2 **Olympic Mountains**

The annual median timeseries (Figure 2.6) for the Olympic Mountains show differences between the models. Notably, PRMS has a much smaller dynamic range in soil moisture, although the general shape is similar to that of VIC and SUMMA. As in the Snake, all three models have different ET, especially during the spring months. PRMS has a spikier runoff signature, while both

VIC and SUMMA have smoother shapes. SUMMA's runoff is different than VIC's, however, and features two peaks, one around November and December and one in April.

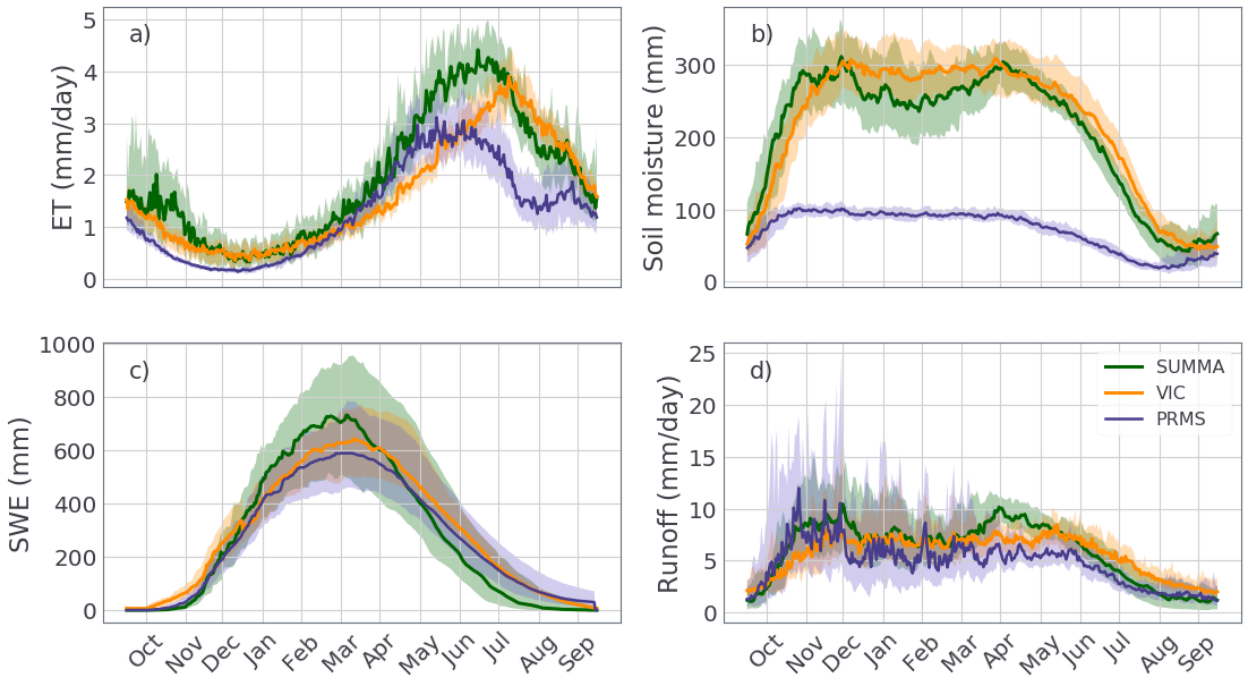

Figure 2.6. Median and 50% interquartile bounds for the four output variables of interest in the Olympic mountain region over the analysis period of 1960-2010

The chord diagrams (Figure 2.7) reflect some of these differences. Soil moisture changes are represented similarly by VIC and SUMMA and are noticeably different in PRMS. PRMS shows a much higher proportion of the information transfer going to soil moisture fluxes, with a large emphasis on runoff and precipitation as contributing factors. The smaller dynamic range of soil moisture in PRMS is the reason for this higher proportion of information transfer, as similarly sized runoff or precipitation events have a larger effective impact on PRMS than either VIC or SUMMA. Despite the larger proportion of information transfer towards soil moisture change, the relative contributions to the information transfer from the source variables to soil moisture change are similar for all three models. That is, despite the difference in dynamic range, the contributing

factors to the behavior are similar.

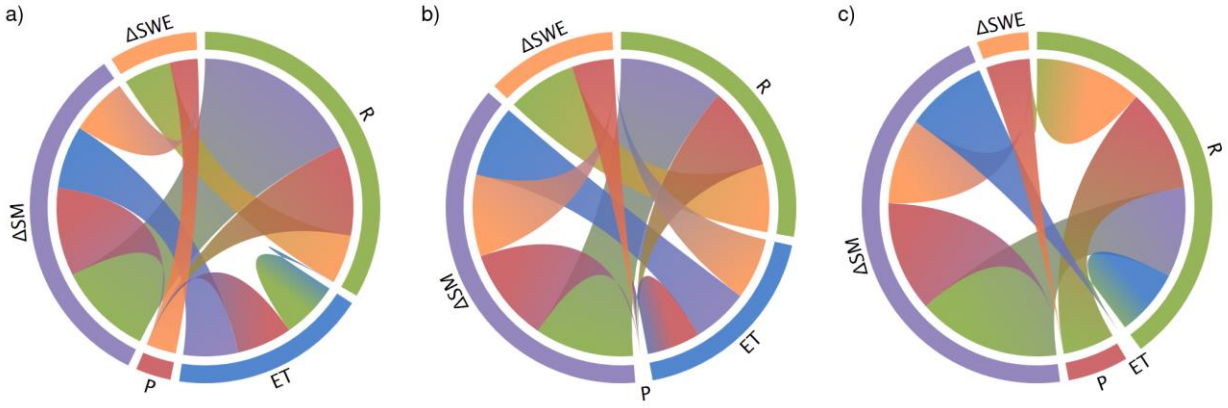

Figure 2.7. Lag 1 transfer entropies in the Olympic Mountain region for SUMMA (panel a), VIC (panel b), and PRMS (panel c).

Another large difference in the chord diagram for PRMS is the lack of any information transferred to ET, while both SUMMA and VIC's ET receives information from all other water balance terms. Part of the lack of information transfer to ET in PRMS is likely due to the diminished amount of high frequency variability in the winter months. Finally, we see that SUMMA and PRMS show some influence on precipitation. Given that precipitation is specified as a forcing variable in all models this is a surprising result. We will explore the reasons for this connection in the discussion.

2.3.3 **Canadian Rockies**

Figure 2.8 shows the annual median timeseries for the Canadian Rockies region. Differences are most noticeable between models in soil moisture and runoff, though all of the timeseries are more similar than we found in the Snake River region. Both VIC and PRMS are very similar in runoff as was also true in the Snake and which reflects the calibration to streamflow that was performed for these two models. The information transfer networks shown in Figure 2.9 show both similarities in process connectivity as well as differences.

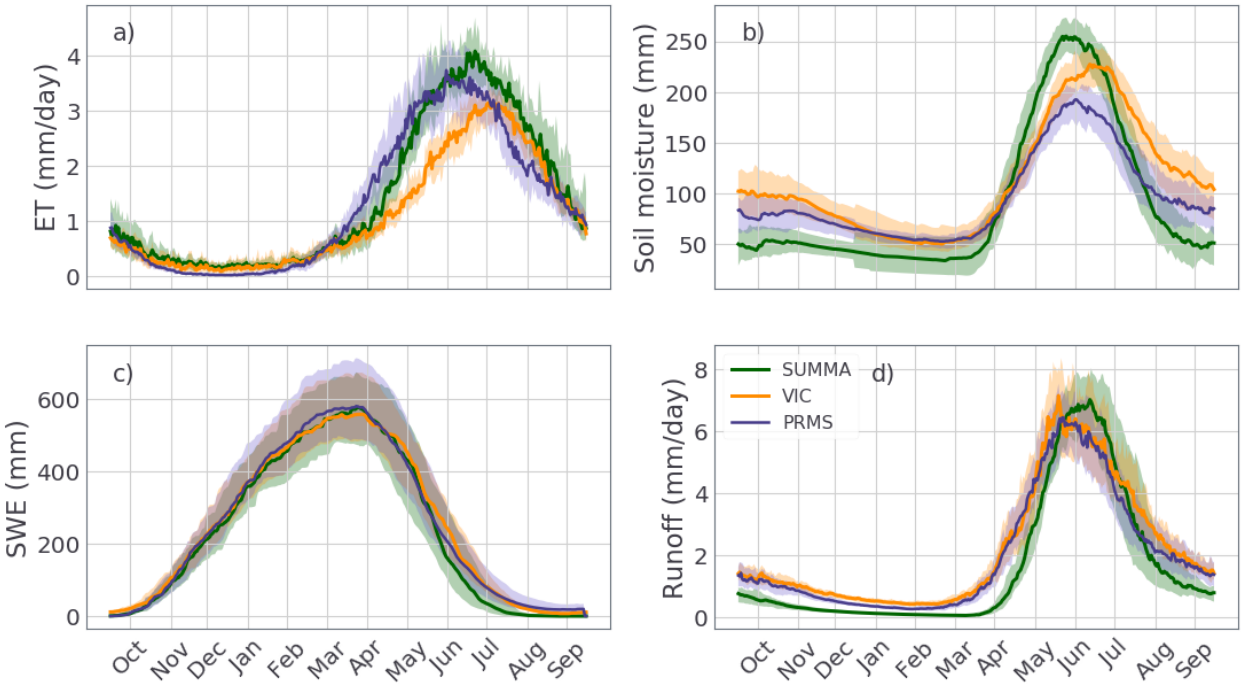

Figure 2.8. Median and 50% interquartile bounds for the four output variables of interest in the Canadian Rockies over the analysis period 1960-2010

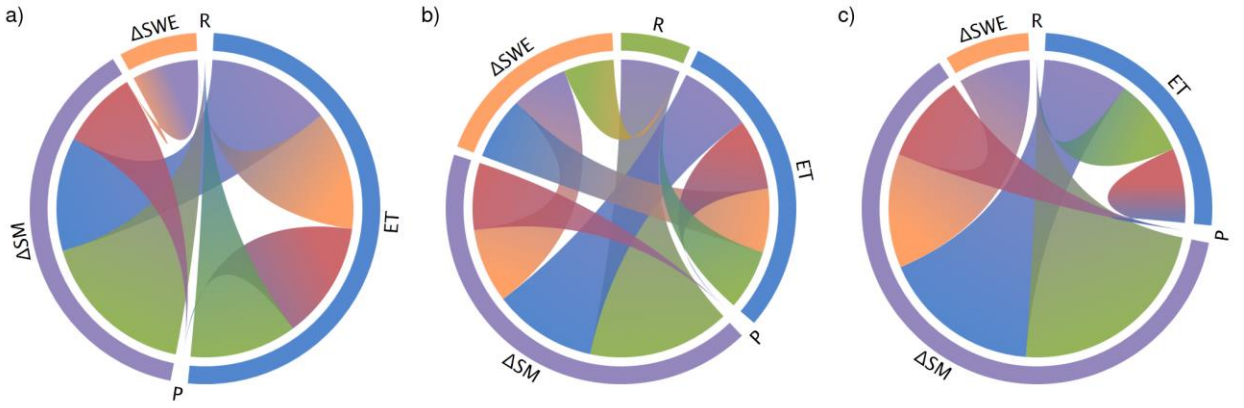

Figure 2.9. Lag 1 transfer entropies in the Canadian Rockies region for SUMMA (panel a), VIC (panel b), and PRMS (panel c).

None of the models showed changes in SWE as having an influence on runoff at a 1-day lag. However, all three models show a strong influence of runoff on soil moisture changes. VIC and PRMS show a less prominent influence in the reverse direction. SUMMA shows a higher proportion of information transferred to ET, though all three models show roughly equal

contributions from all other water balance variables.

The runoff ratio for SUMMA in this region was much lower than VIC or PRMS (Table 2.1) though the average information transfer was unable to give a clear picture as to what contributing factors differ between the models. To further investigate the lower runoff ratio, we also compute the information transfer to runoff as well as the correlation with runoff at a monthly timescale. That is, we aggregated the daily timeseries to monthly and calculated the information transfer at a 1-month lag (Figure 2.10). By calculating the transfer entropy to runoff at a monthly timescale we notice seasonal differences in information transfers. In this monthly information transfer to runoff we see the effect of snowmelt (along with an uptick in all other water balance terms) during the spring months in all three models. Runoff in VIC and PRMS receives more information outside of the spring months than it does in SUMMA. Runoff in SUMMA receives no information from precipitation and ET in the fall and winter. This shows that, as in the Snake, SUMMA does not respond directly to precipitation events until the soil moisture shows a large increase due to snowmelt.

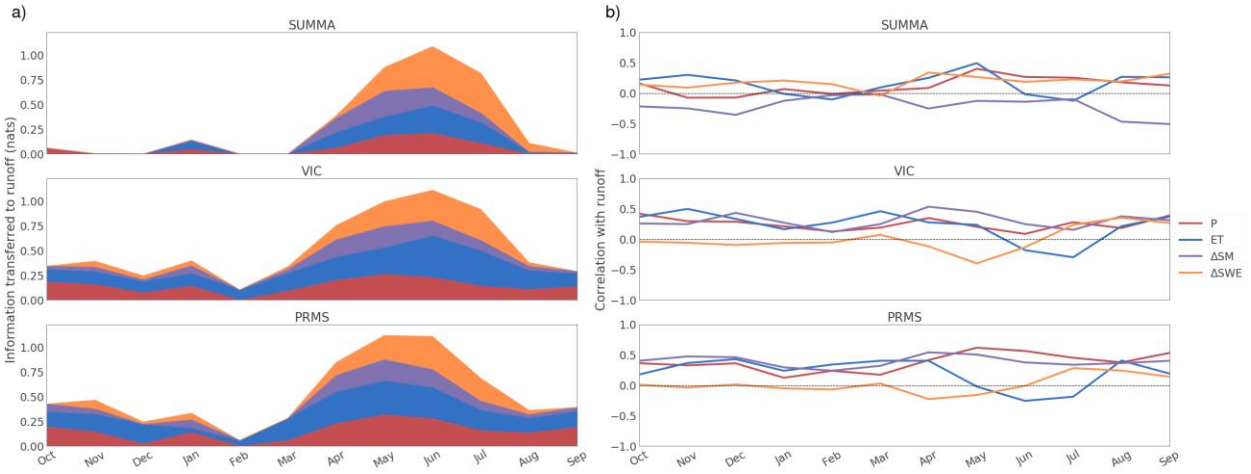

Figure 2.10. Monthly information transferred to runoff (panel a) and monthly correlation with runoff (panel b) in the Canadian Rockies.

In panel b of Figure 2.10 we calculate the correlation on the monthly values for the 50-year

analysis duration. The correlations with runoff do not show such an easily interpretable signal during the peak runoff during the spring. All models show no correlation with ΔSWE until April through September, but we do not see this synchronized influence from all other variables as in the information transfer.

2.3.4 **Willamette**

The annual median timeseries for the Willamette region (Figure 2.11) show that the three models behave more similarly there than in the Snake and Olympic Mountain regions. PRMS is most different in ET, soil moisture and runoff, while SUMMA appears most different in SWE. PRMS is the most unique of the three models in the chord diagrams (Figure 2.12), mostly due to the relationships between water balance variables and changes in SWE. Because these differences in timeseries are relatively small it is difficult to find attribution in the chord diagrams, which only display averages in information transfer.

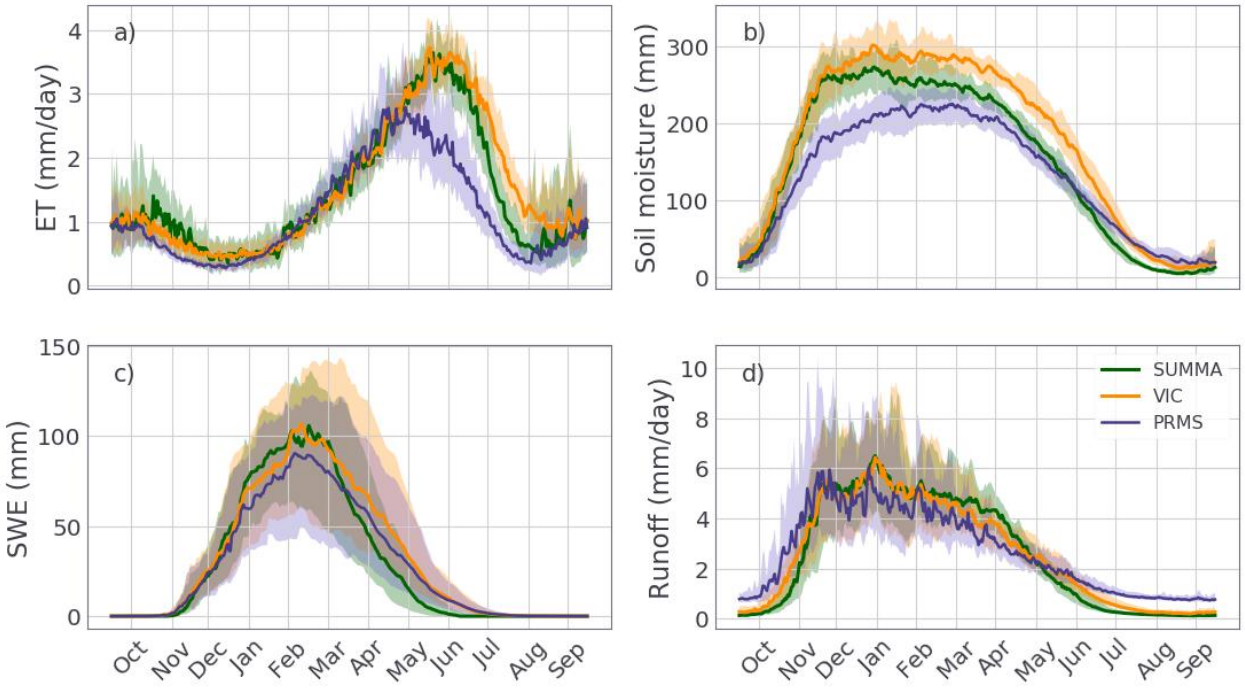

Figure 2.11. Median and 50% interquartile bounds for the four output variables of interest in the Willamette region over the analysis period 1960-2010

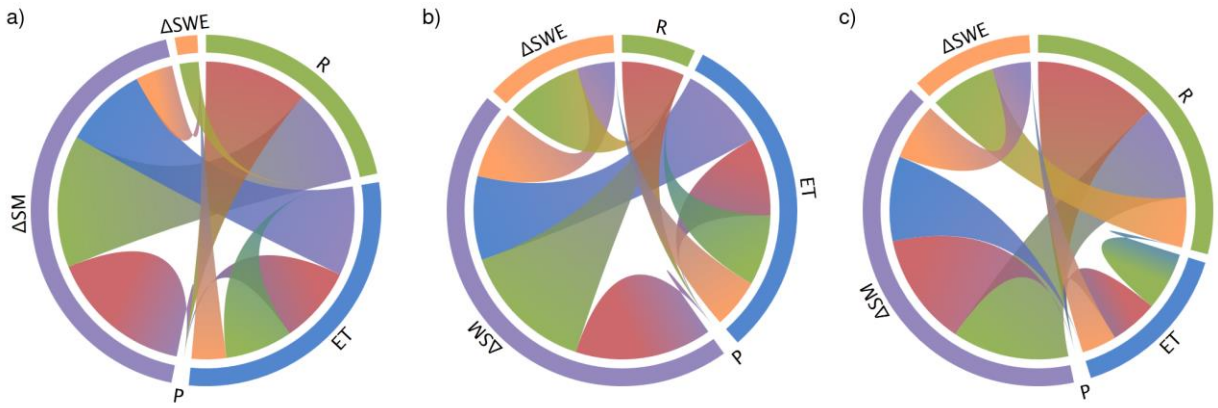

Figure 2.12. Lag 1 transfer entropies in the Willamette region for SUMMA (panel a), VIC (panel b), and PRMS (panel c)

As with the Canadian Rockies, we aggregated the model output to a monthly timestep, and then computed the information transferred to runoff for each month of the year (Figure 2.13). In all three models we see that the runoff is primarily influenced by precipitation, although a large component is also from ET. This new introduction of information transfer from ET to runoff is

due to the change in timescales which smooths out some of the high frequency variability and accounts for seasonal trends rather than daily variability. During the winter and spring months soil moisture fluxes also contribute to the information content of runoff. Most notably we see that runoff receives little to no information in VIC and SUMMA during the summer months, which coincide with near-zero runoff. On the other hand, we can see some small influence from both precipitation and ET during this time in PRMS, which maintains a higher amount of runoff than VIC and SUMMA during this time.

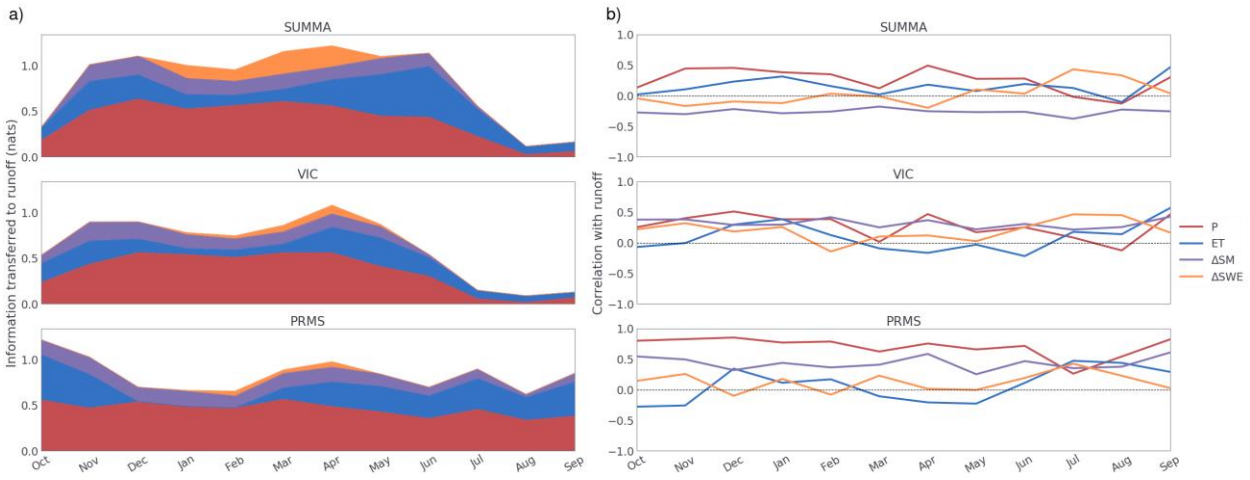

Figure 2.13. Monthly information transfer to runoff (panel a) and correlation with runoff (panel b) in the Willamette region.

Contrasting this with the correlations (Figure 2.13 panel b) we find that the information transfer maps more directly to the timeseries. We find that the correlation between precipitation and runoff in PRMS is quite high throughout the year, which is also reflected in the information transfer. However, one difference is that the correlation in all three models show a spike from ΔSWE in July and August. The cause of this increase in correlation can be seen in Figure 2.11 where both runoff and ΔSWE flatten out. This increased correlation may be misleading, because it corresponds to a period when there is no snow and precipitation and runoff are generally low. The information transfer to runoff in both SUMMA and VIC show this directly by the sharp decrease in information

transferred from precipitation.

2.4 Discussion

The deconstruction of hydrologic model outputs into information flow process networks reveals how model structure relates to model outputs. The calculation of transfer entropy shows the asymmetric and time dependent ways in which processes interact. By conditioning on the target variable's history we remove some effects of autocorrelation, which is not captured by quantities such as mutual information (including time-lagged variants). This allows us to look at active information transfers within hydrologic model output.

In the Snake River region we see that SUMMA's runoff is primarily driven by snowmelt and that precipitation exerts no direct influence on runoff at a 1-day lag because all rainfall infiltrates in the SUMMA implementation. Then, due to the aridity of the region, this rainfall contributes to ET rather than runoff, which is represented in the information transfer by a stronger linkage between precipitation and evaporation for SUMMA than for VIC and PRMS.

Also, in the Snake, we see that although VIC and SUMMA have more similar timeseries of soil moisture than PRMS, the process connectivity is actually more similar between SUMMA and PRMS. The similarity of the timeseries is a result of the mediated effect of ΔSWE through ET. This reveals how a network-based analysis can be used to reason about model behavior in a comprehensive manner and lead to a better understanding of why such behavior occurs. This is an example of how we can use an analysis of information network to understand differences in model behavior.

The analysis in the Olympic Mountains also shows how information transfer analyses can provide complementary insights to traditional methods. The timeseries (Figure 2.7) for soil moisture shows that VIC and SUMMA have a much larger dynamic range than PRMS. Despite

this large difference in dynamic range, the daily information transfer to soil moisture fluxes is similar between all three models, because transfer entropy is based on transition probabilities and does not directly account for magnitude.

Figure 2.7 shows that the daily information transfer networks for the Olympic Mountains include information flows from the model output data to precipitation. These linkages from SWE fluxes and runoff in SUMMA and PRMS, respectively, are spurious since precipitation is a prescribed forcing variable and the model simulations are uncoupled. (Smirnov, 2013) demonstrated that statistically significant non-zero transfer entropies can arise between unrelated variables when latent variables are present, when the analysis is conducted at too coarse a temporal resolution, or in the presence of high measurement noise.

One explanation for these spurious linkages is that on average the Olympic Mountains receive over 300 days of rain. The 1-day lag time period of the analysis is unable to clearly identify the causality of the process connections in this situation. It is possible that analysis on smaller timescales would be useful for regions which have very consistent driving from forcing variables. Further, we did not account for all possible variables in our analysis in order to keep the results tractable, but this could also be a source of this spurious linkage. These possible sources of error motivate further studies at finer time scales and with more variables, which should give further insight into model behavior.

In our analysis of the Canadian Rockies, we found that all three models performed similarly in both the average timeseries as well as the information transfer networks. To gain a more complete understanding of the information transferred to runoff in the Canadian Rockies we aggregated the data to a monthly time step, and then computed the lag 1-month transfer entropy to runoff. Additionally, we computed the monthly Pearson correlation coefficient with runoff (Figure 2.10).

Using a monthly timescale, we find that VIC and PRMS show larger amounts of transfer entropy to runoff in the fall and winter months than SUMMA. We see that although all three models have peak information transferred to runoff during May and June, SUMMA shows the most abrupt peak along with the highest contribution due to snowmelt. Information transfer for PRMS peaks in May, a month earlier than for both SUMMA and VIC. This shift earlier is mostly due to information received from ET, which also has an earlier peak in the timeseries shown in Figure 2.8. In all three models we see information transfer from soil moisture changes in early spring, giving way to transfers from changes in SWE. This shows that soil processes provide better predictive capability of the runoff in the early melt season, and that SWE fluxes themselves provide better predictive capability of runoff later in the melt season.

The monthly correlation with runoff shows more inter-model variability in the Canadian Rockies. However, these correlations do not clearly identify the driving variables at the monthly timescale, while transfer entropies clearly show how runoff is driven by other water fluxes. This highlights how the information transfer network approach can provide insights that are complementary to the understanding gained from traditional metrics used for model evaluation.

We also show the monthly information transfer to runoff and correlation with runoff for the Willamette River region (Figure 2.13), and again find that the transfer entropy and correlation coefficient tell different stories for the three models. The transfer entropy to runoff in PRMS shows that runoff is receiving information from precipitation and evapotranspiration throughout the year, while SUMMA and VIC show very little transfer during August, September, and October. The correlation coefficients with runoff for these two variables do not provide the same insights.

Our results show a new context in which information theory can be used to understand hydrologic systems. Previous work has shown how information theory provides a robust model

evaluation framework (Gong et al., 2014; Nearing, et al., 2016a; Weijs et al., 2010b) and how the process network approach can help us understand the structure of dynamics in observed data (Goodwell & Kumar, 2015; Kumar & Ruddell, 2010). We have bridged this gap and shown how process networks computed using transfer entropy provide a useful way to quantify model structure. Our results show that computing process networks with transfer entropy on the water balance components is a useful way to quantify the entire hydrologic system and that they can be a useful tool in the model evaluation and intercomparison toolbox.

2.5 Conclusions

Computing information transfer networks for a variety of hydrologic models can highlight process level differences which determine model output. Using these techniques we analyzed three hydrologic models (SUMMA, VIC, and PRMS) in a variety of hydroclimatic regimes to find both similarities and differences in the process connectivity.

We compute transfer entropy between pairwise combinations of variables in the water balance equation using a nearest-neighbor estimation technique. Based on the analysis of 50 years of daily output data we were able to quantify the strengths of relationships between variables. This allowed us to identify connections between water balance components that traditional error metrics are unable to discover.

In all models, precipitation is a driving variable, whereas runoff, soil moisture fluxes, and evapotranspiration all show complex interdependencies. Generally, the average behavior of these interdependencies varies more across sites than across models, though differences between models at a specific site also show large differences at times. However, some interdependencies are fairly stable overall. Across most models and sites, bidirectional information transfers can be seen between changes in soil moisture and runoff as well as between changes in soil moisture and

evapotranspiration. We also found specific process level differences. For example, in the Snake, SUMMA showed less fast response runoff due to increased infiltration and subsequent evaporation from the soil. By comparing transfer entropy to Pearson correlations at a monthly timescale we showed that transfer entropy tends to be a better proxy for when a process is active.

Our simple lag-1 information transfer networks provide a first order estimate of the information transferred between variables, but also indicate opportunities for further study. Looking at finer timescales, including more variables, adding further conditioning, and using multivariate approaches to distinguish between drivers, synergies, and feedbacks are some avenues for further analysis. Longer windows would allow further probing of interaction timescales and could provide further insights into the ways that emergent behavior develops within hydrologic models. Initial tests of our estimators show that we are able to robustly estimate 4-dimensional information measures, which open up some of these avenues of research. We believe that further refinements of the nearest neighbor estimators will allow us to tackle some of these problems by allowing faster convergence in higher dimensions. However, this type of analysis will always require large amounts of data for reliable estimation. Even when these data constraints are not an issue, considerable computational challenges in the estimation of these quantities remain, most notably the large computational cost compared to more traditional metrics.

Using a network-based approach can facilitate a deeper understanding of model behavior that would not be clear by computing classical difference methods. Building on these methods is a step towards holistic approaches to model evaluation that can be used to understand the role of individual processes as well as emergent properties. We hope that one day these techniques can also be used as a tool for guiding model selection and development by allowing us to better match process connectivity of observed quantities.

Chapter 3. DEEP LEARNED PROCESS PARAMETERIZATIONS PROVIDE BETTER REPRESENTATIONS OF TURBULENT HEAT FLUXES IN HYDROLOGIC MODELS

This chapter is in review for publication in the journal *Water Resources Research.* © American Geophysical Union. Used with permission. The supplemental material for this chapter is provided in Appendix B. It is also available as a preprint on the Earth and Space Science Open Archive.

Bennett, A. and Nijssen, B. (2021). Deep learned process parameterizations provide better representations of turbulent heat fluxes in hydrologic models. *Water Resources Research*, in review.
Bennett, A., & Nijssen, B. (2020). Deep learned process parameterizations provide better representations of turbulent heat fluxes in hydrologic models [preprint]. https://doi.org/10.1002/essoar.10505081.1

3.1 INTRODUCTION

The debates amongst the hydrologic modeling community about the use and utility of machine learning (ML) to simulate hydrologic processes indicate that much work remains to be done to understand the role and potential of machine learning in hydrologic modeling (Nearing, et al., 2020; Shen, 2018). While it is true that deep learning (DL) models have shown great promise and superior performance in many cases it is yet unclear how to make models that are both composable and transferable for scientific studies. In this paper we outline an approach for coupling DL parameterizations of individual process representations into existing hydrologic modeling frameworks. This coupling approach allows us to represent individual physical processes within a larger model using ML methods. The ability to couple model components will

address these composability and transferability questions, as well as allow use of these types of machine-learned models in areas which do not have readily available training data.

There are several reasons for the rapid advancement of ML-based approaches in hydrology (and other fields), including a greater abundance of publicly available data, increased computational resources, and better frameworks for selecting, fitting, and applying models. Along with this increase in interest, the community has also begun to think about how to incorporate aspects of physical theory into these data-driven models. This desire for physics-based machine learning is enticing for a number of reasons. As scientists we hope that the use of models which are based in, or constrained by, physical properties will allow us to learn about the underlying processes of the systems we are modeling. Not only that, we hope that such approaches will be able to efficiently extract information from a variety of datasets, from in situ observations to satellite remote sensing data, or be able to represent complex phenomena in a more efficient way.

While inclusion of empirical or statistical relationships of individual process representations into hydrologic models is common, this is not yet the case for parameterizations based on ML methods. One reason for this is that it is not clear how to combine ML models in the same way in which we have been able to include processes for which we have parsimonious descriptions and parameterizations which represent physical relationships between processes. In part, this is not surprising since machine learning is good at resolving relationships which we have not been able to decompose into easily describable parts. This “whole-system” or “black box” approach is conceptually appealing due to its simplicity, and is exemplified by rainfall-runoff modeling, which deep learning has proven to be very good at (Hu et al., 2018; Kratzert et al., 2018; Moshe

et al., 2020). However, by taking a more granular approach, we will show that DL models can be successfully incorporated as process modules into existing models.

In this paper, we look at turbulent heat fluxes, for which high-quality, long-term, local observations are available across a range of hydroclimates. While machine learning has been used for modeling of turbulent heat fluxes and evaporation (Jung et al., 2009; Tramontana et al., 2016) there have not yet been model intercomparisons with land surface models, much less integrations into land surface models. However, Best et al. (2015) showed that even simple statistical models are often able to outperform state of the art land surface models in simulation of latent and sensible heat fluxes. The authors postulated that the statistical models were better able to use the information in the meteorological forcing data than the physics-based approaches. This indicates there is strong motivation for incorporating data-driven techniques into complex land surface and hydrologic models. We believe that if these types of approaches are able to provide better performance than the physically motivated relationships we should work to understand how and why this performance is better and use them where appropriate and applicable.

Despite the statistical benchmarks' superior ability for predicting turbulent heat fluxes in Best et al. (2015), land surface models remain more suitable for a wide range of applications, because they represent a wider range of hydrologic processes and may be better suited for studies of environmental change. Such studies include drought prediction (Li et al., 2012), snow melt predictions under climate change (Musselman et al., 2017), and predicting volatile organic compound emissions (Lathière et al., 2006). That is not to say that ML models cannot be used in this way or incorporated into larger frameworks. Both Kratzert et al. (2018) and Jiang et al. (2020) make qualitative comparisons of internal ML model states to snowpack, but do not later

use the models for prediction of snowpack. We believe that it is likely that ML models will be used for such purposes in the near future.

Because the hydrology community is still learning the best ways to build and use ML models, there remains considerable room for incorporation of machine learning into more conventional process-based hydrologic models (PBHMs), which have the flexibility needed for general purpose modeling. This approach has been adopted recently by Brenowitz & Bretherton (2018) as well as Rasp et al. (2018) for parameterizing sub-gridcell scale processes, such as cloud convection, in atmospheric circulation models. Similarly, in oceanography, neural networks have been used to parameterize the turbulent vertical mixing in the ocean surface (Ramadhan et al., 2020).

In this study, we demonstrate how coupling ML models into a hydrologic model can yield better performance at estimating turbulent heat fluxes without sacrificing mass and energy balance closure or the ability to represent other processes such as runoff or snowpack. We have developed two ML models which are coupled into a PBHM. Our first model was only allowed to learn from the same meteorological data that is used to force the hydrologic model, while our second ML model is additionally trained with the inclusion of states derived from the hydrologic model. We show that both ML models are able to outperform the routines for simulating turbulent heat fluxes at subdaily timescales. We also show that the configuration which was trained using model states is better able to reproduce the long-term water balance. Our results indicate that approaches to coupling machine learning with PBHMs offer a promising avenue, which has only begun to be explored.

3.2 Materials and Methods

3.2.1 **Data and study sites**

We used data from 60 FluxNet sites (Pastorello et al., 2020) to run our experiments. These sites cover a large variety of vegetation and climate classifications. Our site selection process considered several criteria. We first filtered the full FluxNet dataset to make sure we only included sites which had energy balance corrected measurements of both sensible and latent heat fluxes, which will be discussed later. We then made sure that these sites had the necessary variables to force our models, which include precipitation, air temperature, incoming shortwave radiation, incoming longwave radiation, specific humidity, air pressure, and wind speed. We then removed sites which had either fewer than three years of contiguous data or more than 20% missing observations during the longest continuous period with observations. For the remaining sites, we used gap-filled data provided as part of the FluxNet dataset. Gap-filling was based on ERA-Interim (ERAI) (Dee et al., 2011) and includes downscaling and postprocessing explicitly for the purpose of model forcing. Time steps flagged as gap-filled were excluded from our performance analysis to ensure that we did not simply measure the ability of our simulations to model ERAI data. However, the gap-filled data is included when analyzing the water balance.

We also limited our analysis to sites which had an observed ET/P ratio of less than 1.1, calculated using the mean FluxNet-reported values of ET and P over the simulation period. This was done to accommodate our model structure, which enforces mass and energy balances on a point (or lumped) scale. Larger observed ET/P ratios likely occur at sites which have strong spatial gradients and flow convergence, so that moisture available for ET is not just the result of local precipitation. Our filtering process resulted in 60 sites with 508 site-years of data. A

breakdown of the site names, data periods, locations and site characteristics are given in Table 3.1. Likewise, Figure 3.1 shows the locations and vegetation classes for these same sites.

Site name	Latitude	Longitude	Vegetation Type	Start Time	End Time
BE-Vie	50.3	6	Mixed Forests	1-1996	12-2014
RU-Fyo	56.5	32.9	Evergreen Needleleaf Forest	1-1998	12-2014
CA-Qfo	49.7	-74.3	Evergreen Needleleaf Forest	1-2003	12-2010
BE-Lon	50.6	4.7	Croplands	4-2004	10-2013
US-Prr	65.1	-147.5	Evergreen Needleleaf Forest	11-2010	12-2014
NL-Hor	52.2	5.1	Grasslands	7-2004	4-2009
IT-MBo	46	11	Grasslands	1-2003	12-2013
IT-Tor	45.8	7.6	Grasslands	4-2008	12-2014
IT-SRo	43.7	10.3	Evergreen Needleleaf Forest	6-2000	4-2009
AU-Cpr	-34	140.6	Savannas	1-2010	12-2014
AT-Neu	47.1	11.3	Grasslands	1-2002	12-2012
ES-LJu	36.9	-2.8	Open Shrublands	1-2004	12-2013
US-NR1	40	-105.5	Evergreen Needleleaf Forest	1-2004	12-2008
US-Var	38.4	-121	Grasslands	11-2000	12-2011
US-Los	46.1	-90	Permanent wetlands	9-2000	2-2009
FI-Hyy	61.8	24.3	Evergreen Needleleaf Forest	10-2004	8-2012
CA-TP3	42.7	-80.3	Evergreen Needleleaf Forest	1-2002	12-2014
DE-Hai	51.1	10.5	Deciduous Broadleaf Forest	1-2000	8-2011
DE-Gri	51	13.5	Grasslands	1-2004	12-2014
FI-Let	60.6	24	Evergreen Needleleaf Forest	7-2009	12-2012
CZ-wet	49	14.8	Permanent wetlands	3-2009	12-2014
DK-Eng	55.7	12.2	Grasslands	6-2005	10-2008
DE-Tha	51	13.6	Evergreen Needleleaf Forest	1-1996	12-2014
US-Whs	31.7	-110.1	Open Shrublands	1-2007	12-2014
CA-TPD	42.6	-80.6	Deciduous Broadleaf Forest	1-2012	12-2014
IT-Lav	46	11.3	Evergreen Needleleaf Forest	1-2003	12-2014
FR-LBr	44.7	-0.8	Evergreen Needleleaf Forest	1-1996	12-2008
US-KS2	28.6	-80.7	Closed Shrublands	5-2003	12-2006
US-Goo	34.3	-89.9	Grasslands	5-2002	12-2006
US-WCr	45.8	-90.1	Deciduous Broadleaf Forest	8-2010	12-2014
US-IB2	41.8	-88.2	Grasslands	1-2004	12-2011
CA-Gro	48.2	-82.2	Mixed Forests	1-2003	12-2014
IT-Noe	40.6	8.2	Closed Shrublands	2-2004	12-2014
US-Blo	38.9	-120.6	Evergreen Needleleaf Forest	5-1998	12-2007
AU-Wac	-37.4	145.2	Evergreen Broadleaf Forest	5-2005	12-2008
AU-Wom	-37.4	144.1	Evergreen Broadleaf Forest	1-2010	12-2014
CH-Cha	47.2	8.4	Grasslands	1-2006	3-2014
AU-ASM	-22.3	133.2	Evergreen Needleleaf Forest	1-2010	12-2014
DE-Kli	50.9	13.5	Croplands	5-2006	12-2014
US-Ton	38.4	-121	Woody Savannas	1-2001	12-2014
FI-Sod	67.4	26.6	Evergreen Needleleaf Forest	4-2002	4-2005
CA-TP1	42.7	-80.6	Evergreen Needleleaf Forest	1-2002	12-2014

DE-Obe	50.8	13.7	Evergreen Needleleaf Forest	1-2008	12-2014
US-CRT	41.6	-83.3	Croplands	1-2011	12-2013
AU-DaS	-14.2	131.4	Savannas	1-2008	12-2014
IT-Cpz	41.7	12.4	Evergreen Broadleaf Forest	4-2000	1-2009
US-Syv	46.2	-89.3	Mixed Forests	9-2001	1-2008
IT-Ro2	42.4	11.9	Deciduous Broadleaf Forest	1-2002	2-2007
FR-Pue	43.7	3.6	Evergreen Broadleaf Forest	7-2004	3-2013
DE-Geb	51.1	10.9	Croplands	1-2001	12-2014
US-AR2	36.6	-99.6	Grasslands	5-2009	12-2012
AU-How	-12.5	131.2	Woody Savannas	4-2009	12-2014
US-GLE	41.4	-106.2	Evergreen Needleleaf Forest	9-2004	12-2014
AU-Stp	-17.2	133.4	Grasslands	4-2008	12-2014
IT-Ren	46.6	11.4	Evergreen Needleleaf Forest	8-2003	12-2013
ES-Amo	36.8	-2.3	Open Shrublands	6-2007	12-2012
CH-Fru	47.1	8.5	Grasslands	1-2006	2-2014
FI-Jok	60.9	23.5	Croplands	2-2000	11-2003
CN-HaM	37.4	101.2	Grasslands	1-2002	12-2004
US-ARM	36.6	-97.5	Croplands	1-2003	12-2012

Table 3.1. A listing of the sites, locations, IGBP vegetation types, and dates of simulation

As noted, we chose to use the FluxNet-provided energy balance corrected turbulent heat fluxes.The energy balance gap in eddy-covariance measurements is an extensively studied topic (Foken, 2008; Kidston et al., 2010; Wilson et al., 2002), though no strong consensus has been reached on how to account for gaps in the observed energy balance (or even whether one should). However, because we will be using models and methods that enforce energy conservation, we chose to use the corrected fluxes provided by the FluxNet data providers (Pastorello et al., 2020).

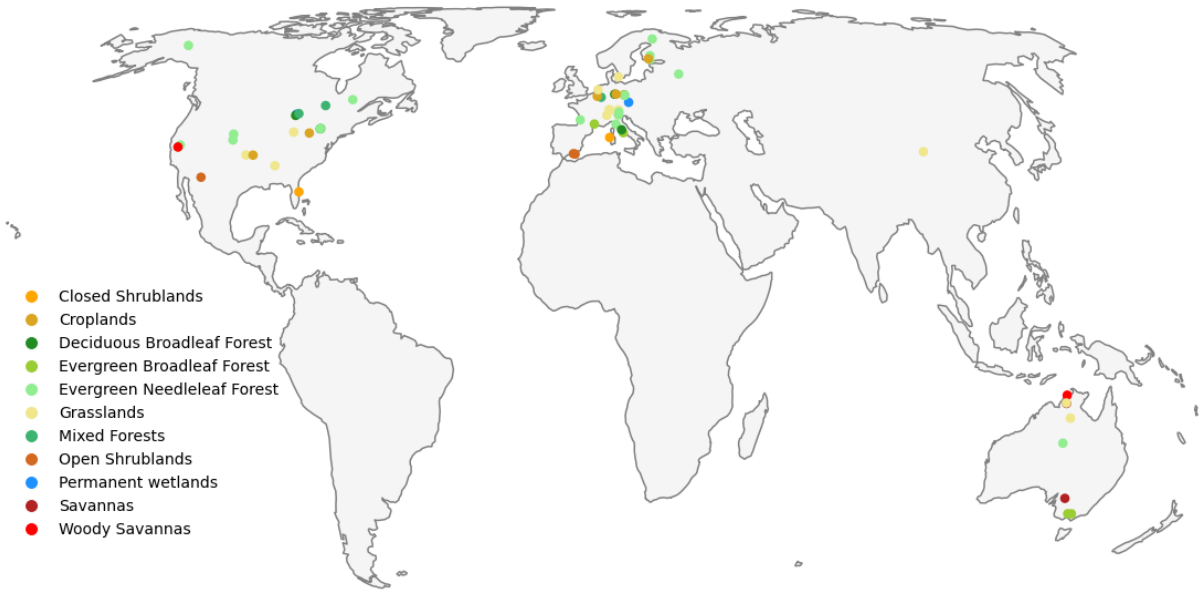

Figure 3.1. A map of the FluxNet sites used in the analysis, coded by the IGBP vegetation type.

3.2.2 SUMMA standalone simulations

We used the Structure for Unifying Multiple Modeling Alternatives (SUMMA) to simulate the hydrologic cycle (Clark et al., 2015a) including the resulting turbulent heat fluxes. SUMMA is a hydrologic modeling framework that allows users to select between different model configurations and process parameterizations. The clean separation between the numerical solver and flux parameterizations made it easier to couple our DL parameterizations into SUMMA. The core numerical solver in SUMMA enforces closure of the mass and energy balance and is used in all of our simulations.

SUMMA provides multiple flux parameterizations and process representations for many hydrologic processes. Because we were primarily interested in turbulent heat fluxes, we used a configuration for the other processes which would be suitable for general purpose hydrologic modeling, including runoff and snowpack simulations. For simulation of transpiration we used a

Ball-Berry approach for simulating stomatal conductance (Ball et al., 1987), an exponentially decaying root density profile, and soil moisture controls that mimic the Noah land surface model (Niu et al., 2011). Similarly, the radiative transfer parameterizations which are the primary controls on the sensible heat fluxes are also set up to mimic the Noah land surface model.

At each of the sites described in section 3.2.1 we independently calibrated a standalone SUMMA model using the dynamically dimensioned search algorithm (Tolson & Shoemaker, 2007) as implemented in the OSTRICH optimization package (Matott, 2017). The first year of available data was used for calibration. Because of the limited length of the data record at some sites, the calibration period was not excluded from subsequent analysis. The 10 parameters we chose to calibrate largely control water movement through the vegetation and soil domains. In the soil domain these include the residual and saturated moisture contents, field capacity, and controls on anisotropy of flows. In the vegetation domain these include controls on photosynthesis, rooting depth, wilting and transpiration water contents, amount of throughfall of precipitation through the canopy, and a generic scaling factor for the amount of vegetation. A summary of the calibration variables and test ranges is shown in the supplementary materials.

The calibrations were run to a maximum of 500 trial iterations, which provided good convergence across sites (see supplemental information for convergence plots). We used the mean square error at a half hourly timestep for both the latent and sensible heat as the objective function and saved the best set of parameters for each site to use as our comparison to the DL parameterizations. To provide good estimates of the initial soil moisture and temperature states we spun up the standalone SUMMA simulations for 10 years both before and after calibration (for a total of 20 spinup years). We will refer to the standalone calibrated SUMMA simulations as SA (StandAlone) for the remainder of the paper. To summarize, we independently calibrated a

set of parameters for each site, whose resulting best parameter set was used as an in-sample benchmark for comparison with our DL parameterizations.

3.2.3 **DL parameterization and simulations**

To produce each DL parameterization of turbulent heat fluxes we constructed our neural networks using the Keras python package (Chollet et al., 2015), using only dense layers. We chose a deep-dense architecture because it is the only network architecture that has robust implementation support for coupling to SUMMA. We will discuss the details of how we coupled the neural networks to SUMMA later in this section. After manual trial and error we settled on 6 layers each with 48 nodes. We used hyperbolic tangent (tanh) activations and stochastic gradient descent (SGD) with an exponential learning rate decay curve. We used the mean square error in the 30-min turbulent heat flux estimates as our loss function, similar to the objective function in our calibration of the standalone SUMMA simulations. Dropout was applied after the first layer and before the final layer with a retention rate of 0.9 to regularize.

When training the networks we performed a 5-fold cross validation. We used 48 sites to train each network and then applied it out of sample to each of the remaining 12 sites. The 48 sites used to train each network were randomly split into 80% training and 20% validation data. The validation data was used to define an early stopping criterion for the training procedure where training was stopped if the validation loss was not decreased for 10 training epochs. This procedure keeps the model from overfitting on the training data. The maximum number of training epochs was set to 500 epochs, with a batch size of 768 data points (or 14 days of data points). All data was shuffled before training to remove any temporal bias that the model could learn, which also reduces overfitting.

The first network we trained took only meteorological forcing data for the current timestep, as well as vegetation and soil types, and the calibrated SUMMA parameter values. We chose to include the calibration parameters to provide the same information to the neural networks as was provided to the calibrations, allowing for a more direct comparison and because the calibrated parameter values might be a proxy for site characteristics that can be associated with different responses among the sites. We denote this network NN1W, for Neural-Network-1-Way, because this configuration only takes meteorological forcing data and parameters, which cannot be changed by the rest of the SUMMA calculations. That is, the neural network provides information about turbulent heat fluxes to SUMMA, but SUMMA does not provide any internally-derived information to the neural network.

The second network we trained took all of the same data as the NN1W configuration, as well as a number of derived states that were taken from the output of the NN1W configuration. We included surface vapor pressure, leaf area index, surface soil layer volumetric water content, depth averaged transpirable water (as a volumetric fraction), surface soil layer temperature, depth averaged soil temperature, and a snow-presence indicator. These variables were chosen because they are used in the process-based SUMMA parameterizations for either latent or sensible heat, or affect the way in which the partitioning of the heat flux is distributed to the soil, vegetation, or snow domains. At runtime this network uses the additional variables as calculated internally by SUMMA, rather than the ones provided during training from NN1W. We denote this network NN2W, for Neural-Network-2-Way, because SUMMA internal states provide feedback to the ML model. That is, the neural network is provided inputs which are dependent on the state variables derived internally by SUMMA, which in turn depend on the turbulent heat fluxes that are predicted by the neural network.

After training each of these networks they were saved and translated into a format that could be loaded into Fortran via the Fortran Keras Bridge (FKB) package (Ott et al., 2020). The FKB package allows for translation of a subset of Keras model files (architecture, weights, biases, and activation functions) to be translated into a file format which can be loaded into the FKB Fortran library which implements several simple components for building and evaluating neural networks in Fortran, such as the deep-dense architecture used here.

We then extended SUMMA to allow for the use of these neural networks to simulate the turbulent heat fluxes. Normally SUMMA breaks the calculation of turbulent heat fluxes into several domains to delineate between heat exchanges in the vegetation and soil domains. Because we estimate these as bulk quantities we implemented this as only heat fluxes in the soil domain, and specified that the model should skip any computation of vegetation fluxes. We then specified that all ET computed by the neural network be taken from the soil domain as transpiration, according to SUMMA's internal routines. We chose this rather than taking all of the ET as soil evaporation because this allowed for a wider range of ET behaviors. In our simulations, the domain was split into nine soil layers, with a 0.01 m deep top layer. In SUMMA soil evaporation is only taken from the top soil layer and the shallow surface soil depth in our setup would not have allowed for sufficient storage to satisfy the predicted ET for many of the vegetated sites. Water removed as transpiration is weighted by the root density in each soil layer, which generally provides a large enough reservoir to satisfy the evaporative demand predicted by the neural networks. Another side-effect of our decision for taking all ET as transpiration is the removal of snow sublimation from the model entirely. As we will show in the results, the amount of snow sublimation in the SA simulations is negligible at most of our FluxNet sites, so we believe that this is an acceptable simplification for our initial demonstration. In cases where the

neural network predicts greater evaporation than is available in the soil SUMMA enforces the water balance and limits the evaporation to an amount it can satisfy.

3.3 RESULTS

We present our results in two categories. First, we compare the performance of the coupled neural network simulations to the standalone calibrated simulations (SA). We use two commonly used metrics for determining the performance of the simulated turbulent heat fluxes, the Nash-Sutcliffe efficiency (NSE) and Kling-Gupta efficiency (KGE) scores. Using two metrics in tandem allows us to be sure that our results are robust (Knoben et al., 2019). Then, we explore how the inclusion of NN-based parameterizations for turbulent heat fluxes affects the overall model dynamics. This analysis is crucial to ensure that the new parameterizations do not lead to unrealistic simulations of other processes

3.3.1 **Performance analysis**

Figure 3.2 shows the cumulative density functions of the performance metrics across all sites, evaluated on the half-hourly data for all non-gap-filled periods. For all cases we see that both NN1W and NN2W were able to outperform the SA simulations. NN1W showed a median increase in NSE of 0.07 for latent heat and 0.12 for sensible heat, while NN2W showed a median increase in NSE of 0.10 for latent heat and 0.14 for sensible heat. Likewise, for KGE these were 0.10 (latent) and 0.21 (sensible) for NN1W and 0.17 (latent) and 0.23 (sensible) for NN2W. Overall we see that the NN2W configuration slightly outperforms the NN1W configuration. However, it is possible that in both cases that there are additional performance gains to be made with better model architectures and/or training procedures. We will come back to this in the Discussion.

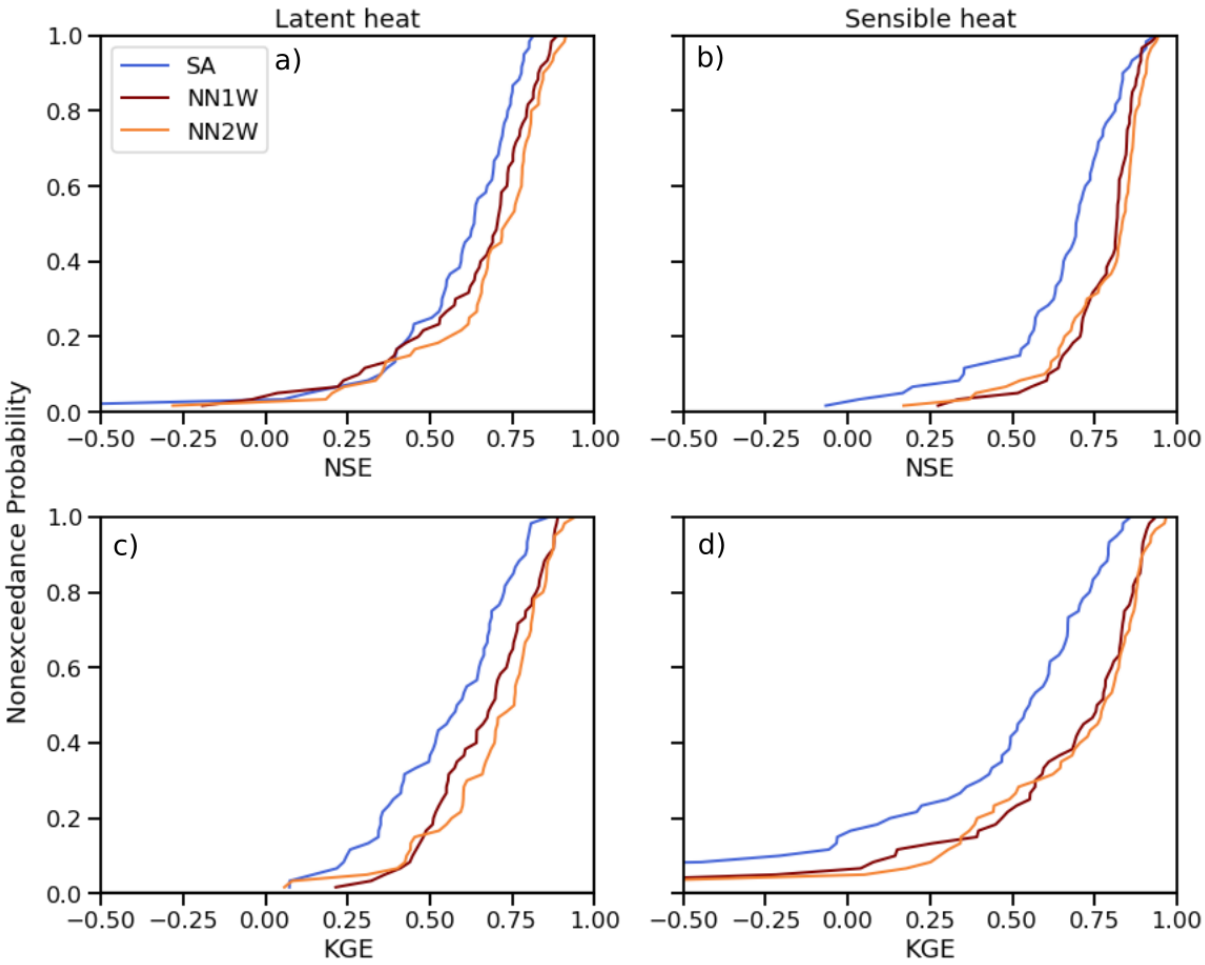

Figure 3.2 Empirical CDFs of performance measures for simulations across all sites. a) shows the NSE for latent heat, b) the NSE for sensible heat, c) the KGE for latent heat, and d) the KGE for sensible heat.

Even though the curves of the performance measures look quite similar between NN1W and NN2W, the performance differences from SA were not always perfectly correlated. Figure 3.3 shows the change in performance from SA for each site, ranked by SA performance. The maximum improvement that is possible is also shown to provide a reference to account for the fact that the range of both NSE and KGE is (-∞,1]. That is, there is more room for improvement for poorly performing sites than there is for well performing sites. For both performance measures and fluxes the general pattern of improvement follows the maximum improvement curve, with some added noise.

While on average the NN-based configurations performed better than the SA simulations, they performed worse at some locations. NN-based simulations generally had a higher NSE for sensible heat, but the KGE scores for sensible heat were more mixed, with SA outperforming the NN-based configurations at a number of sites. The NN-based configurations performed much worse at AT-Neu, DK-Eng, and CH-Cha (the outliers in the lowest 25th percentile of Figure 3.3d), where they failed in simulating large, upward, nighttime sensible heat fluxes. SA also performed poorly for these nighttime fluxes, but to a lesser extent. For latent heat, while some sites showed higher NSE and KGE values for SA results than for the NN-based simulations, more sites showed poor performance across all configurations when evaluated by NSE. Decreases in performance relative to SA mostly occurred where the NN-based configurations consistently overestimated latent heat during winter. For both conditions for which SA outperformed the NN-based configurations, we believe that the performance of the NN-based configurations can be improved if more training data or more sophisticated ML methods were used, since the number of outliers was small and the average performance improvement was large.

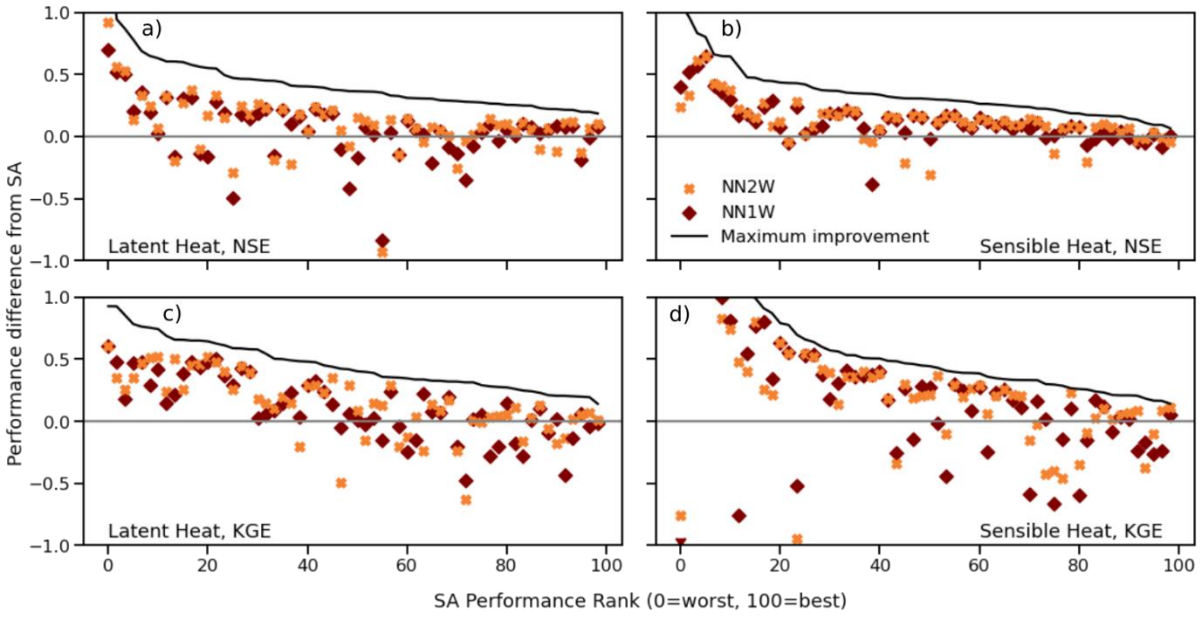

Figure 3.3 Scatter plots showing the performance of NN1W and NN2W against SA across all sites. Points above the grey zero line show configurations where the NN configuration improved performance over SA. The "Maximum improvement" line is based on the SA simulations.

We also compared the KGE for different periods of temporal aggregation to evaluate whether performance improvements of the NN configurations persisted across timescales (Figure 3.4). The KGE score was chosen here because it shows greater variability than the NSE score in Figure 3.3, though the results are similar for NSE. We see that the sub-daily aggregations, on average, showed better performance for both NN configurations, demonstrating that they were able to capture the diurnal cycle of turbulent heat fluxes. This is mostly due to the strong dependence of turbulent heat fluxes on solar radiation, which we will further explore in section 3.3.2. Both NN1W and NN2W were able to outperform SA across all timescales for sensible heat.

However, at daily and longer temporal aggregations differences between models were seen in latent heat performance. The NN1W configuration performed better at sub-daily timescales than

for daily or longer aggregations, for which performance was similar to SA. In contrast, the NN2W configuration performed better for latent heat than SA across all timescales.

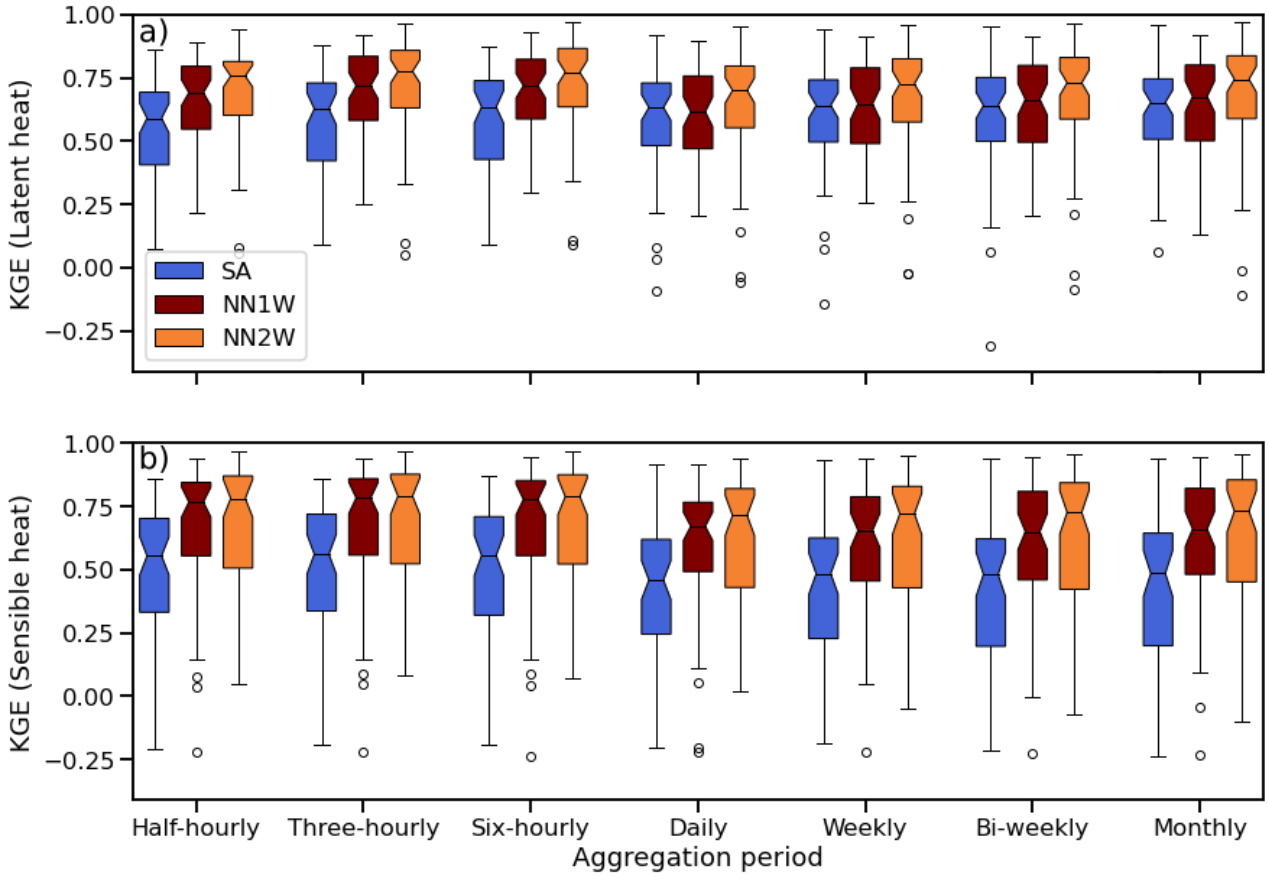

Figure 3.4 Performance of each model configuration for multiple temporal aggregations. Each box shows the interquartile range, with the median marked as the central line. A 95% confidence interval for the estimate of the median is represented by the notched portion. Outliers are shown as open circles.

3.3.2 **Diagnostic analysis**

In section 3.3.1 we demonstrated that the NN configurations were able to consistently outperform the SA configuration for both latent and sensible heat flux predictions at a half-hourly timestep. The range of performance differences shown in Figure 3.3 demonstrates that the NN-based simulations are significantly different from the physically-based representation in SA. Consequently, water and energy partitioning in the NN configurations is likely much different

than in SA. To explore the effect of the new NN-based parameterizations on the simulated water cycle we first compared the simulated evaporative fraction (ET/P) to the observed (Figure 3.5). In all three model configurations the KGE values tend to be higher for sites where the simulated evaporative fraction closely matches the observed value.

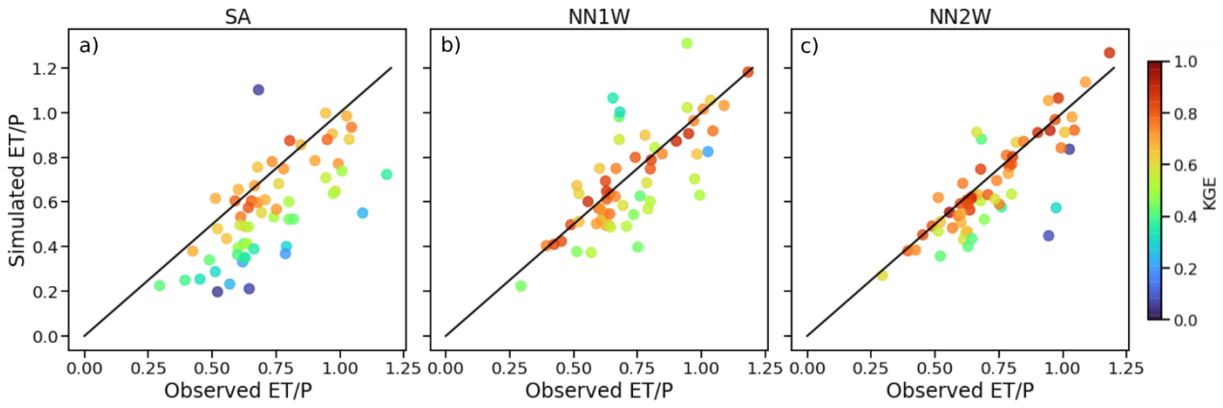

Figure 3.5 Comparison of evaporative fraction for each model configuration across all sites. The one-to-one line shows perfect correspondence with the observed values. Each point shows an individual site, averaged over the simulation period. Points are colored by their respective performance in terms of KGE of the latent heat at the half-hour timescale.

However, the SA configuration has a tendency to systematically underestimate total ET, while the NN configurations tend to match the observed evaporative fraction. The NN1W configuration shows more over-evaporation than NN2W, indicating that the introduction of soil states allows the model to perform better in moisture limiting conditions. This soil moisture feedback is the reason that the NN2W was able to perform better at daily and greater temporal aggregations for the prediction of latent heat.

The increased ET in the NN configurations affects the other water balance terms as shown in Figure 3.6. We first normalized each of the sites so that the water input (precipitation plus any storage drawdowns) summed to one, to facilitate comparison between sites. Generally, the

increased ET in the NN configurations corresponds to a decrease in runoff (R), rather than a drawdown in storage, indicating our simulations were sufficiently spun up.

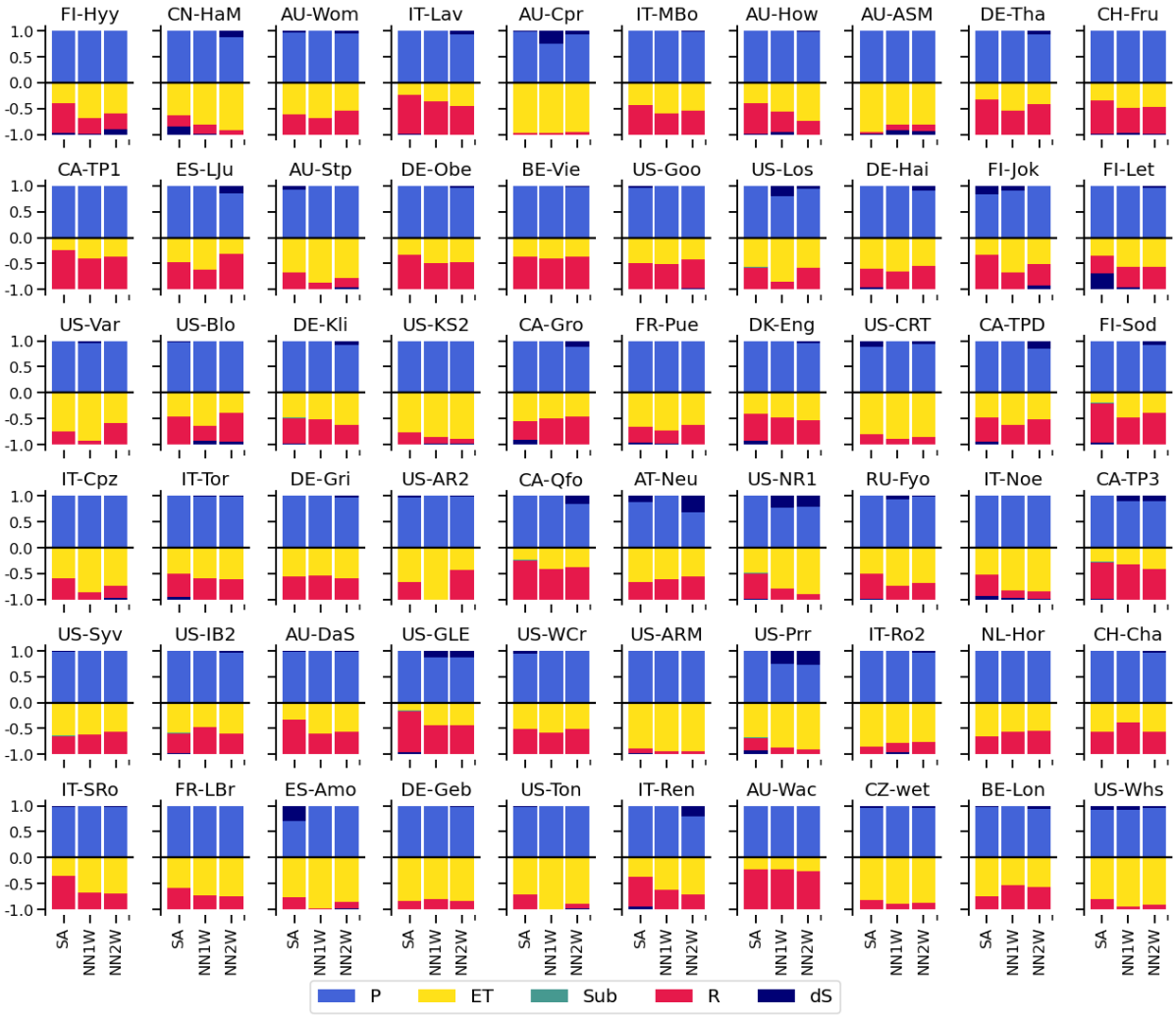

Figure 3.6 Breakdown of the water balance across configurations at each site, normalized so that inputs and outputs each sum to one on a per site-model basis. P is precipitation, ET is total evapotranspiration, Sub is sublimation, R is runoff, and dS is the change in moisture storage. Note that Sub only appears in SA and is a minor component that is present at only a few sites.

As noted when discussing Figure 3.4, we hypothesize that the NN-based simulations performed better at the sub-daily timescale because of their improved ability to model the diurnal cycle in the observations. We take the approach of Renner et al. (2019) by comparing the time

lag in the diurnal cycle between the turbulent heat fluxes and shortwave radiation. To compute this we fitted a regression equation of the form:

$$Q(t) = a_0 + a_1 SW(t) + a_2 \frac{dSW(t)}{dt} + \epsilon, \quad 3.1$$

where Q is the turbulent heat flux, SW is the shortwave radiation, a_i are the coefficients of the regression, and ϵ is the residual term (Camuffo & Bernardi, 1982). Then, the phase lag can be computed as

$$\phi = tan^{-1}\left(\frac{2\pi a_2}{a_1 n_d}\right), \quad 3.2$$

where n_d is the number of timesteps in a day (here, 48). We calculated this phase lag for each of the simulation configurations and the observations. Figure 3.7 shows how each of the simulations compare to the observed phase lag across all sites. For both latent and sensible heat we see that the NN-based configurations are better able to capture the diurnal phase lag seen in the observations, confirming our conclusion from Figure 3.4 that the improved sub-daily performance of the NN-based configurations is due to better representation of the diurnal cycle.

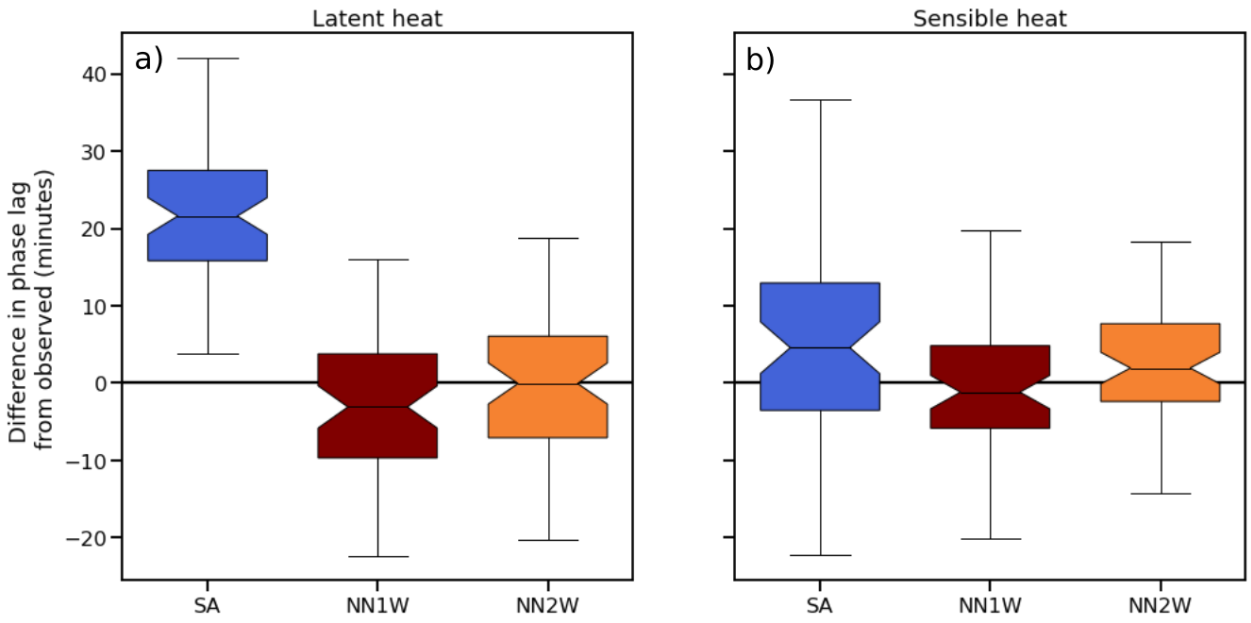

Figure 3.7 Difference in diurnal phase lag from observation. Positive values indicate that the simulated phase lag leads the observed phase lag.

3.4 Discussion

Our analysis shows that the DL parameterizations were able to outperform the standalone simulations for both latent and sensible heat fluxes. A large amount of the performance gains from the NN-based configurations was due to drastic improvements at sites where the SA configuration performed poorly. This is important to note, since our SA simulations were calibrated at site (and included the calibration period in the evaluation), while all NN-based simulations were trained out of sample in both time and space. This indicates that our NN-based configurations would likely be better able to represent turbulent heat fluxes in regions without measurements, implying that deep learning may be suitable for regionalization applications.

Both of the NN-based configurations represented the diurnal phase lag between shortwave radiation and turbulent heat fluxes better than SA. Renner et al. (2020) explored the ability of the land surface models used in the PLUMBER experiments (Best et al., 2015) to reproduce the

observed diurnal phase lag, finding similar deviations from the observed phase lag as our SA simulations. This indicates that the NN-based approach has been able to learn something that has not been codified in PBHMs, and could provide better insight into how turbulent heat fluxes are generated at the scales that FluxNet towers operate.

We also found that the NN2W configuration maintained higher performance than either NN1W or SA at longer than daily timescales, as well as more accurately reproduced the observed long-term evaporative fraction. This indicates that the synergy between the deep-learned parameterization and the soil-moisture state evolution in SUMMA was able to better capture the long-term dynamics than either a purely machine-learned or purely process-based approach. This lends credibility to our proposition that the synergy between data-driven and physics-based approaches will likely lead to better simulations than a rigid adherence to either one of the methods by themselves.

These performance gains came at the cost of drastically simplifying the way in which we represented evapotranspiration. The SA simulations partition the latent heat fluxes amongst the soil, snow, and vegetation domains separately, while the NN simulations were set up to only represent the latent heat as a bulk flux, whose withdrawals we set to be taken from each soil layer according to the root density in that layer. This leads to the SA simulations being able to represent a more diverse range of conditions. While this was not a problem for the NN simulations on average, we were able to identify two locations where our simplification to the way in which ET is taken from the soil led to poor performance. At US-WCr and US-AR2 both NN configurations underestimated ET, because the soil was too dry to meet evaporative demand for much of the time. At these two sites the NN simulations performed significantly worse than the SA simulations, indicating a clear failure mode of the neural network based approach. We

believe that this shortcoming can be addressed by developing strategies that better partition the latent heat fluxes amongst the soil, snow, and vegetation domains. This would also allow for adding snow sublimation back in, reducing the number of modifications which must be made to SUMMA in order to run with an embedded neural network.

Another area for development that we believe will result in further improvements to the predictions is the use of other neural network architectures. Many recent studies that used neural networks to predict hydrologic systems have shown that Long-Short-Term-Memory (LSTM) networks are superior at learning timeseries behaviors compared to the methods used here (Feng et al., 2020; Frame et al., 2020; Jiang et al., 2020; Kratzert et al., 2018). Likewise, convolutional neural networks (CNN) have been used extensively to learn from spatially distributed fields (Geng & Wang, 2020; Kreyenberg et al., 2019; Liu & Wu, 2016; Pan et al., 2019). To take advantage of these specialized architectures in existing PBHMs like SUMMA will require the investment in tools and workflows. As of the time of writing, the FKB library only supports densely connected layers, and a few simple activation functions. Implementing these layers in the FKB library, or some other framework that can be used to couple ML models with PBHMs, would open many possibilities for future research.

Alongside better tools for incorporating machine learning into process-based models, we believe that the development and identification of workflows to perform machine and deep learning tasks will be necessary for wider adoption in the field. For instance, we initially trained the NN2W networks using the SA soil states, which were drastically different from the spun up states in the NN configurations. This led to almost identical performance in the NN1W and NN2W simulations, since the soil state information from the SA simulations was very different from what the network saw during training. Only after realizing this and training the NN2W on

the states predicted by the NN1W simulations were we able to achieve better performance out of the NN2W simulations. Understanding whether there is a sort of iterative train-spinup-train workflow that balances overfitting and provides representative training data will be important for future studies.

Similarly, it is unclear whether there would be significant difficulties in trying to calibrate either of the NN-based models in new basins like we did for the SA simulations. Particularly, we do not know if the output of the neural networks is sensitive to the values of the calibration parameters. Our decision to include the calibrated parameter values in the training of the NN-based configurations was to provide the same types of information to both optimization procedures. In future studies it may be worthwhile to explore whether these parameters are necessary, or how regionalization of data-driven approaches should best be codified. It is also unclear whether our NN-based configurations are able to be calibrated efficiently for other processes such as streamflow.

Finally, model architectures that separate process parameterizations in as clean a way as possible will allow for more robust and rapid development of ML parameterizations of other processes. Building modular and general purpose ways to incorporate machine learning into process-based models will allow researchers to more efficiently evaluate different approaches. Exploring and answering these practical questions will likely lead to community accepted practices which can be adopted to accelerate research of other applications.

3.5 Conclusions

We have shown that coupling DL parameterizations for prediction of turbulent heat fluxes into a PBHM outperforms existing physically-based parameterizations while maintaining mass

and energy balance. We were able to couple our neural networks into SUMMA in two different ways, which both showed significant performance improvements when performed out of sample over the at-site calibrated standalone SUMMA simulations. The one-way coupling (NN1W), despite being conceptually simpler and not taking any model states as inputs, was able to improve simulations almost as much as the more complex two-way coupling (NN2W) at the sub-daily timescale. Both of the new parameterizations better represent the observed diurnal cycles and NN2W was better able to represent the long-term evaporative fraction as well as both turbulent heat fluxes at longer than daily timescales. We found that NN1W was also able to accurately predict sensible heat fluxes at greater than daily timescales, indicating that even "simple" DL parameterizations show great promise for coupling into PBHMs.

While we consider our new parameterizations a step forward in incorporating ML techniques into traditional process-based modeling, we have only scratched the surface on many of the different avenues which will surely be explored. We used the simplest possible network architecture, a deep-dense network. For spatial applications we suspect that CNN layers will prove invaluable. Likewise recurrent layers such as LSTMs have been dominant in the timeseries domain. More sophisticated architectures such as neural ordinary differential equations (Ramadhan et al., 2020) or those discovered through neural architecture search (Geng & Wang, 2020) are bound to be both more efficient and interpretable than our dense networks. The opportunities for incorporating and learning from ML-based models into the hydrologic sciences are virtually untapped. We believe that as the community builds tools and workflows around the existing ML ecosystems we will be able to unlock this potential.

Chapter 4. ON THE PHYSICAL INTERPRETATION OF A NEURAL NETWORK FOR SIMULATING TURBULENT HEAT FLUXES

This chapter is in preparation for publication in the journal *Water Resources Research.* © American Geophysical Union.

Bennett, A. and Nijssen, B. (2021). On the physical interpretation of a neural network for simulating turbulent heat fluxes. *Water Resources Research*, in preparation.

4.1 INTRODUCTION

The hydrologic sciences have a long history of making use of a wide variety of modeling philosophies (Baartman et al., 2020; Blöschl & Sivapalan, 1995; Kampf & Burges, 2007b). The framing of machine learning (ML) methods versus more process-based (PB) methods often pits "explainability" versus "predictive performance" (Lipton, 2017). With the recent uptick in interest in using machine learning (ML) methods for hydrologic modeling this debate continues. Advances in both process based and data-driven models continue that debate. In this paper we hint that data-driven models, specifically deep-learning (DL) based models, may offer ways to refine theoretical underpinnings and improve physics-driven modeling approaches. We build on previous work that showed that deep-learning parameterizations can be used directly in process-based models to represent individual processes, and improve the performance of their predictions. In this study we show how our deep learned process parameterizations identify physically relevant predictor variables in a way that coincides with physical intuition while maintaining better predictive capabilities than existing process-based models. Additionally, we show how we can use explainable artificial intelligence (XAI) techniques to gain process insights

that can guide the construction of robust and transferable models, and hint at important processes across a range of hydrometeorologic conditions.

Toms et al. (2020) pointed out that it is common for studies using deep learning (DL) in geosciences to focus exclusively on output of the network. Any interpretation of the models is done in an ad hoc fashion to ensure that the transformations from inputs to outputs are physically plausible. However, it is becoming clear that DL techniques can be used as tools for interpretation instead of primarily for predictive purposes (Barnes et al., 2020; Dobrescu et al., 2019; McGovern et al., 2019; Chen et al., 2020). This flipping of perspectives may allow for greater insight into what DL models are learning, and may allow for scientific understanding that will continue to advance hydrologic theory.

While using XAI methods, also referred to as interpretable machine learning, is relatively new in the geosciences, a large number of techniques have been developed with differing goals and domains of application. (Barredo Arrieta et al., 2020) provides an overview of these methods, and provides a taxonomy for classifying XAI methods. They note six modes of providing "post-hoc" explanations (that is, following training of the model) which are popular in the ML literature. These modes are visualization, local explanations, feature relevance ranking, explanations by example, text explanations, and model simplification. The technique we use here, Layerwise Relevance Propagation (LRP) (Bach et al., 2015), fits into several of these categories, namely "visualization", "local explanations", and "feature relevance". It has recently been shown that a large number of XAI techniques bridge these categories and have similar general properties. Particularly it has been shown that gradient and saliency maps (Simonyan et al., 2014), relevance/attribution based methods (such as LRP), local explanations (LIME, Ribeiro

et al., 2016) are all facets of the more general framework of Shapley Additive Explanations (Lundberg & Lee, 2017).

In Bennett & Nijssen (2020), we took the "traditional" approach and focused on predictive performance to train a deep learned parameterization for the prediction of turbulent heat fluxes. We demonstrated that DL-based models that are trained out-of-sample are able to outperform locally-calibrated process-based hydrologic models (PBHMs) at the half hourly timescale. We also showed that the DL parameterization was more accurate at representing the diurnal phase lag between shortwave radiation and latent heat. Further, we showed that coupling the DL parameterization to a process-based hydrologic model (PBHM), by providing it with updated soil moisture information on a per timestep basis enabled it to learn behavior that improved the long-term water balance compared to either the standalone PBHM or standalone NN. Our experiments hinted that the improvements in performance are due to the DL model's ability to find physical relationships between input and output that have not been encoded explicitly in the physics-based models. This also hinted that a synergy between PBHM and DL-based process parameterizations could provide ways to improve both modeling philosophies.

In this paper we take the perspective of Toms et al. (2020), by considering interpretability as our main objective. We continue to build on our methods of coupling physics-based and DL models for the simulation of turbulent heat fluxes. First, we explore whether the DL model learned physically plausible relationships and show that it was able to learn relationships which fit our physical intuition of how turbulent heat fluxes are generated. We will show that the network also learned that there was a connection between latent and sensible heat, particularly by learning different process relations between energy and moisture limited sites. The network

learned that soil moisture limitations can be used to predict the partitioning between latent and sensible heat, though this constraint was not encoded into the network a priori.

We will also show how the LRP method can be used to understand what the network has learned between sites. One of the fundamental problems in hydrologic modeling is being able to provide predictions for locations which have not been extensively observed (Blöschl et al., 2019; Hrachowitz et al., 2013). It has been previously shown that DL based models have been able to obtain state of the art performance in making predictions at sites where the model has not been trained, indicating that DL may offer a way forward in making predictions in ungauged basins (Kratzert et al., 2019). It has been suggested that data-driven models are more accurate out-of-sample because data-driven models (including DL) are able to extract more information from the given datasets than is currently extracted by PBHMs (Best et al., 2015; Loritz et al., 2018; Nearing, et al., 2020)**.** To explore if this is the case in our model we will also explore how the neural network learns to generalize across sites. In Bennett & Nijssen (2020), we found that the out-of-sample simulations from the DL models performed better than the in-sample, calibrated PBHM. This indicated that the neural networks were able to learn some generalized method of predicting turbulent heat fluxes that was not captured in the physics encoded by the PBHM and subsequent calibrations. We present a novel approach to comparing individual samples that make use of the DL model which can be used to explain what the model learned from one site when applied to another.

4.2 METHODS

4.2.1 **Data and study sites**

As in Bennett & Nijssen (2020) we analyze 60 FluxNet sites (Pastorello et al., 2020) where data quality is robust enough and with a sufficient record length for a PBHM to be run. We

required 3 years of half hourly data with at most 15% missing. Missing data was gap-filled by the FluxNet team with ERA-interim data that has been bias-corrected and downscaled. This resulted in 509 site-years worth of half hourly data. Figure 4.1 shows the locations and IGBP vegetation types of each of the sites.

Figure 4.1 A map of the FluxNet sites used in the analysis, coded by the IGBP vegetation type

4.2.2 Coupled deep learning parameterization

To predict turbulent heat fluxes we use a deep dense neural network. We chose this network to be consistent with Bennett & Nijssen (2020). It was originally chosen so that we could embed the neural network into the SUMMA hydrologic model (Clark et al., 2015a). This coupling allows us to use model-derived states in the training process of the neural network. Using SUMMA as a PBHM with the neural network as a process parameterization for turbulent heat fluxes allows us to maintain mass and energy balance while exploiting the flexibility and predictive capabilities of neural networks. The coupling between the two modeling frameworks was facilitated by the Fortran-Keras-Bridge (FKB) (Ott et al., 2020), which allows neural

networks which are trained via the Keras python package (Chollet et al., 2015) to Fortran-based models (such as SUMMA). Currently FKB only allows for densely connected networks, which is the reason for our architectural choice. Future developments may allow for more complex network architectures, which may improve both predictive capabilities as well as interpretability. Compared to the network which was used in Bennett & Nijssen (2020), the network that we train here is much smaller. By reducing the size of the network we are more easily able to disentangle the impact of the input variables on the predicted turbulent heat fluxes.

The neural network that we trained is 2 layers deep with each layer consisting of 28 nodes with tanh activations. At each layer we incorporate dropout regularization (with dropout rate 0.1) to reduce the amount of mixing between inputs in the LRP decomposition (Samek et al., 2019). We use mean squared error between predicted and observed heat fluxes as our loss function. The Adam method as the optimizer, which automatically tunes the learning rate and has been shown to work well in many settings with little tuning (Kingma & Ba, 2017). Training is stopped when the loss on the validation data has not been reduced for at least 5 training epochs to further reduce the possibility of overfitting. We refer to this neural network configuration as NNLRP throughout the remainder of this paper.

The neural network we trained takes air temperature, relative humidity, shortwave radiation, soil moisture content, LAI multiplied by the height of the vegetation canopy, and IGBP vegetation class as inputs. The network predicts latent and sensible heat fluxes. The soil moisture content is computed as the depth-average soil moisture of the top four (out of a total of 8) soil layers as computed by SUMMA. It is scaled between the moisture content at wilting point (0) and the moisture content at saturation (1) before it is used as an input to the neural network. Both the saturation and wilting points are site-specific values whose values were determined as

described in Bennett & Nijssen (2020). We used only the top four soil layers because it represented a good compromise between the total transpirable water and the surface layer moisture, which were used in Bennett & Nijssen (2020). We decided to include only a single input related to the soil moisture to make the interpretation simpler. Each input represents a single timestep at the half hourly timescale and includes no other temporal information. We will refer to the new model as NNLRP to denote that it was designed primarily for analysis with the LRP method, rather than for purely predictive purposes.

4.2.3 **Layerwise relevance propagation**

We use the layerwise relevance propagation (LRP) technique to interpret the system learned by NNLRP. The use of LRP in the geosciences is relatively new, though a good overview of the method within that context can be found in Toms et al. (2020), with more detail about the original method provided by Bach et al. (2015) and Montavon et al. (2017). For clarity we provide a high level description of the LRP algorithm.

Intuitively, LRP works by taking advantage of the ability to backpropagate information from the outputs to the inputs of a neural network. Following training, neural networks can be used to make predictions using the forward pass. LRP uses the predictions made during the forward pass, along with a "rule" for partitioning relevance between neurons to backpropagate a relevance score from outputs to inputs on a local scale. Relevance scores are computed for each forward pass, meaning we obtain timeseries of relevances for each input variable with respect to both latent and sensible heat outputs.

A number of rules can be used, each with different purposes, interpretations, and theoretical properties. For a review of some of the most commonly used rules see Samek et al. (2019). In this study we use the Epsilon rule, which is generally the same as the original rule proposed in

Bach et al. (2015), but avoids numerical artifacts in the case of weak or contradictory explanatory power. It allows for regularizing small connections, promoting sparsity, and reducing noise in the total relevance scores. The Epsilon rule propagates relevance according to the rule:

$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_j a_j w_{jk}} R_k \quad 4.1$$

where the j, ksubscripts denote the index of the nodes in the network, $a_j$is the output of the j^{th}node from the forward (predictive) pass, $w_{jk}$is the weight of the connection between the $j^{th}$and k^{th}nodes, and $R_k$is the relevance computed for the k^{th}node. $\epsilon$is a tunable parameter which is introduced to "absorb" some of the relevance when the contributions of the weights to the relevance from $R_k$is small. For all relevance scores reported in this study we use $\epsilon = 0.001$.

4.2.4 **Using LRP to disentangle site similarity**

One of the surprising findings of Bennett & Nijssen (2020) was that both of the DL based approaches outperformed the process-based model at sites where the DL models were not trained. This indicated that the neural network was learned generalizations that could be applied to sites it had never seen. While LRP is useful for understanding generalities of what the network learned, we would also like to build a better understanding of how it learned to generalize between sites. We do this by shifting the perspective of what the relevance scores represent.

Relevance scores derived from LRP are proportional to local sensitivities from model inputs to outputs and the method can be grounded in the theory of Taylor expansions (Montavon et al., 2017). The set of all relevance scores for a particular site can be seen as a decomposition of what the neural network learned about that site. This decomposition into a set of local sensitivities of the inputs and flux responses of the outputs can be used to build a linear model of the neural

network for a particular site. To build this linear model we perform a multivariate linear regression where each of the predictor variables is the set of relevance scores for each of the neural network inputs and the target variable is a turbulent heat flux. We show that this linearized model can almost exactly reproduce the relationship between the relevance scores and heat fluxes.

This perspective is similar to the Sparse Identification of Nonlinear Dynamics (SINDy) method, which has proven successful in discovering the governing equations of dynamical systems from data (Brunton et al., 2016). However, the approach and goal of our regression analysis are slightly different than those of SINDy. In our approach we do not require the promotion of sparsity that SINDy uses, since we have already allowed the neural networks to determine feature importance. Additionally, we do not use this regression approach to build an explanatory model which can be used separate from the neural network, but rather to understand how the neural network learned from different sites. For clarity, this linear model is not usable without the neural network because the dependent variables are derived from the trained neural network.

Then, our key insight is that the relevance scores are conditional on the weights and biases of the trained neural network, which accounts for the entire training dataset across sites. By fitting this type of regression at one site and then applying it to another we are quantified how the neural network was learned inter-site relationships. This allows us to build graphs of site interactions which yield insight into the nature of variability of turbulent heat fluxes across sites.

4.3 RESULTS

4.3.1 **Performance of the NNLRP model**

Before determining *what* the neural network learned, it is important to ensure that the neural network performed adequately. We measured the performance of the new network against those used in Bennett & Nijssen (2020). Figure 4.2 shows the results of calculating the Kling-Gupta Efficiency (KGE) score for each site at the half-hourly timestep against the observations across the entire simulation record. The SA (or standalone) simulations are the benchmark simulations that use the process-equations for turbulent heat fluxes in SUMMA. The SA simulations were calibrated in-sample (i.e., using local observations of the turbulent heat fluxes). The NN2W (or neural-network-2-way) is the coupled model in Bennett & Nijssen (2020). NN2W is a neural network run directly in SUMMA that predicts turbulent heat fluxes for each 30-minute model interval based on both SUMMA inputs as well as dynamically-updated SUMMA soil moisture. NN2W coupled into SUMMA previously showed good performance (Bennett & Nijssen, 2020). It was trained out of sample, meaning that the performance metrics were calculated for sites which the network was not trained on. In contrast, NNLRP was trained on the entire dataset and was thus was evaluated in-sample. This choice was motivated because we are not interested in using NNLRP to make predictions, but rather, we want to understand what NNRLP has learned during training.

It is unsurprising that NNLRP did not match the performance of NN2W, because we reduced the network from approximately 13,000 parameters (NN2W) to roughly 1000 parameters (NNLRP) and also reduced the number of input features. However, it is promising that NNLRP obtained performance which continues to exceed that of SA.

NNLRP performance relative to NN2W showed a much greater decline for sensible heat than for latent heat. During our design of NNLRP we considered including additional variables that were included in the training of NN2W but we were unable to improve performance for sensible heat without increasing the model capacity through adding more neurons or layers. In the interest of maintaining a simple network that would allow for robust interpretations of the LRP method, we opted to trade model simplicity for loss in model performance for sensible heat.

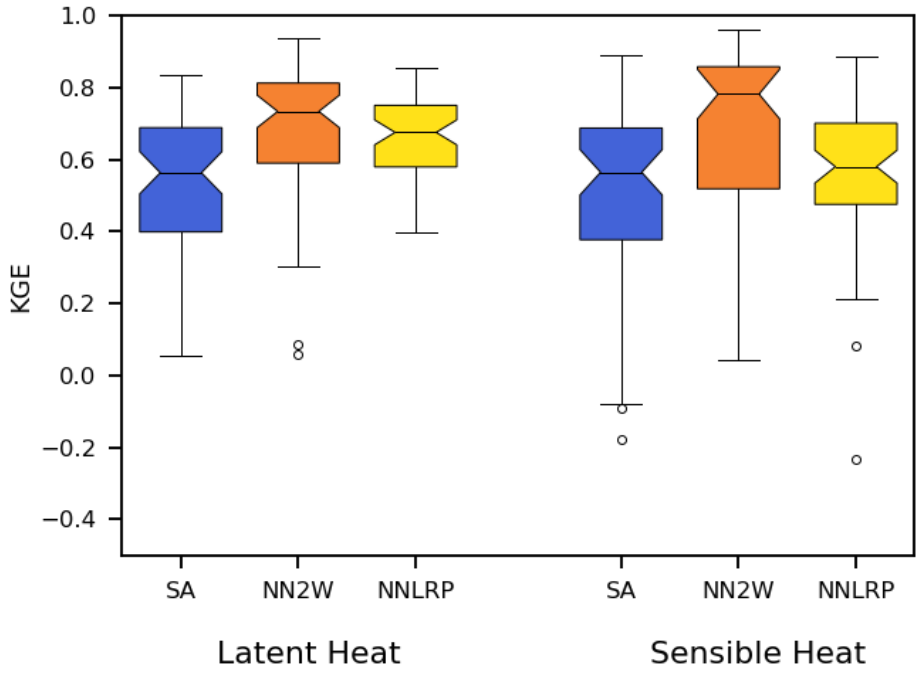

Figure 4.2. A comparison of the KGE performance of the neural network used in our analysis (NNLRP) against the SA and NN2W models reported in Bennett & Nijssen (2020). KGE scores were calculated based on observations of the turbulent heat fluxes at the FluxNet sites.

4.3.2 **Layerwise relevance propagation in the predictive model**

We computed the relevance of each of the input variables to the neural networks at each site. We computed timeseries of relevance scores for each of the input variables to gain an intuitive understanding of the relevance scores. Figures 4.3 and 4.4 show these timeseries for both an energy limited (CH-Fru, figure 4.3) and moisture limited (US-Whs, figure 4.4) site. CH-Fru is a grasslands site near the base of the Swiss Alps. US-Whs is a semi-arid shrubland located in the

Chihuahuan desert of the southwestern United States. To simplify the timeseries we show the average daily daytime values. We chose to illustrate the timeseries during the daytime because the turbulent heat fluxes are largest during this time. We omit the timeseries of LAI and vegetation relevance for simplicity.

At CH-Fru, in figure 4.3, we see large (in absolute value) relevance scores for latent heat from the air temperature and shortwave radiation. The importance of shortwave radiation and temperature is unsurprising and fits with physical understanding of the drivers of latent heat, namely available energy and atmospheric demand. Relative humidity also shows some importance in the prediction of latent heat, though less than air temperature or shortwave radiation. Soil moisture shows the smallest relevance scores for both latent and sensible heat, which is unsurprising since CH-Fru is not moisture limited. However, we do note the strong (negative) correlation between the latent heat relevance timeseries for humidity and soil moisture. We will investigate this behavior later in this section. Similarly, there appears to be a negative correlation between the temperature relevance timeseries for latent and sensible heat. These correlations hint that the network learned strategies for partitioning between heat fluxes, which is surprising, since the NNLRP network was not constrained to conserve energy, which means that it learned this partitioning directly from covariances in the training data. We will see that this partitioning behavior is also present in the relevance from soil moisture states.

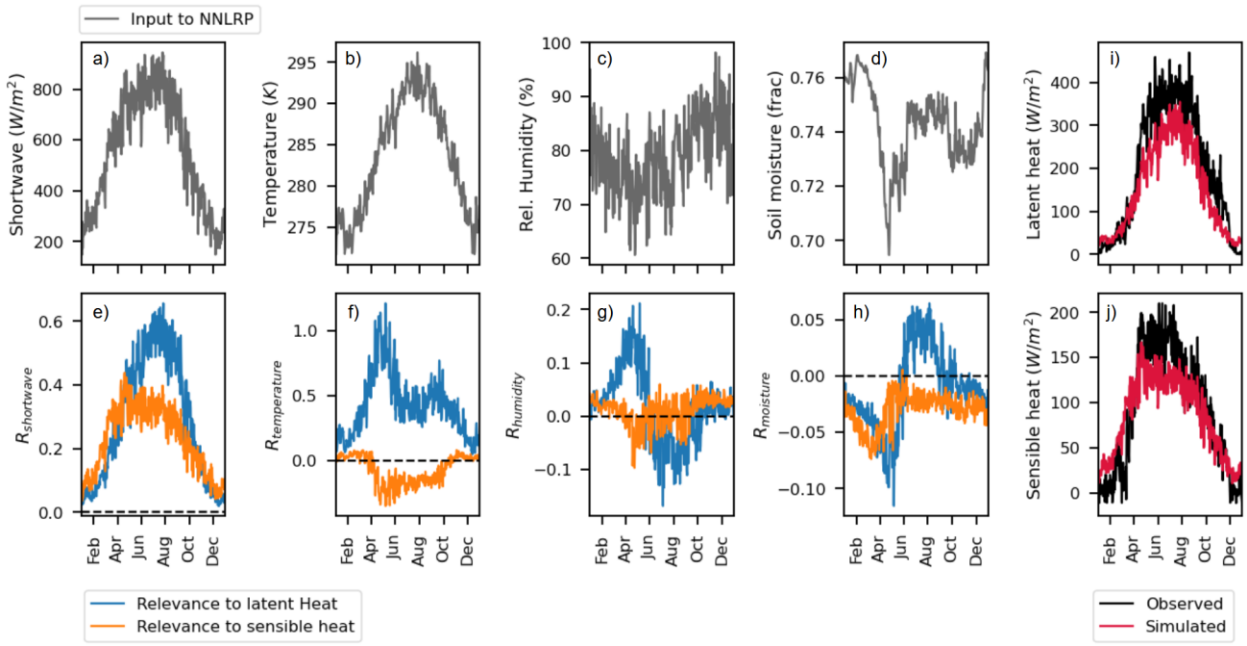

Figure 4.3 Timeseries for meteorological conditions and LRP-derived relevance values at CH-Fru. Subplots a-d show the observed forcings used as input to the neural network, while subplots e-h show the relevance timeseries for latent (blue) and sensible (orange) heat with respect to each of the input variables. Subplots i and j show the observed and simulated heat fluxes.

At US-Whs (figure 4.4), we see similar relationships. Air temperature is most relevant for latent heat and shortwave radiation is most relevant for sensible heat. We will show that these, and other relationships are quite stable across locations. Again, we see the strong negative correlation between latent and sensible heat relevances from temperature, indicating that the neural network uses temperature as a variable to partition energy between the heat fluxes. Unlike at CH-Fru, we see a large spike in the magnitudes of relevance from soil moisture to both latent and sensible heat. This spike in relevance corresponds to the soil moisture increase in figure 4.4d and indicates that the network learned when the site was moisture limited.

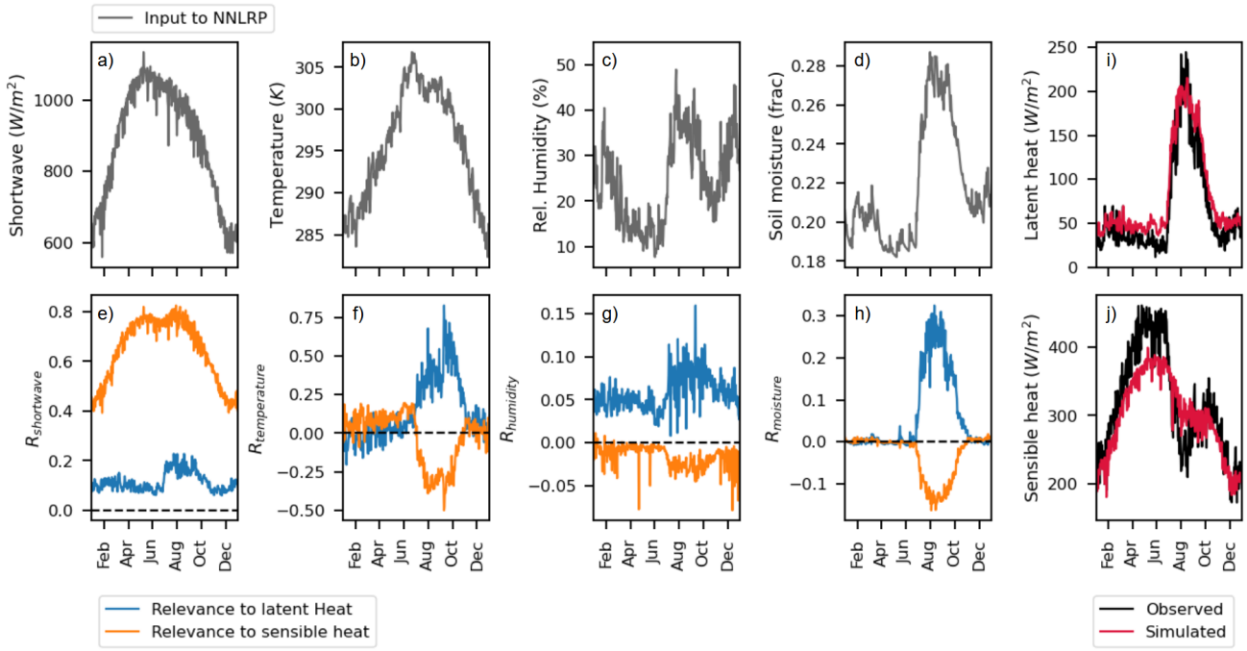

Figure 4.4. Timeseries for meteorological conditions and LRP-derived relevance values US-Whs.

We can see this regime shift by comparing the response of both relevance and heat flux to changes in soil states. To illustrate this, we selected the timestep where the relevance from air temperature to latent and sensible heats was largest, since it is a large driving factor in both heat fluxes. We then varied the soil moisture content near the surface from the wilting point to saturation, while keeping all other inputs constant, and computed latent heat, sensible heat, and the corresponding relevances for surface moisture (figure 4.5). Over the full range of soil moisture conditions, both CH-Fru and US-Whs show negative correlations between the fluxes and soil moisture relevance scores. Showing the relevance over a range of values with the other input variables fixed also shows an intuitive interpretation of the relevance score. It is approximately proportional to the derivative of the flux with respect to soil moisture. It indicates that when a relevance is positive with respect to a variable, that variable can be considered a "producer" of the flux, and when the relevance is negative it can be considered an "inhibitor" of the flux. Fixing all of the other input variables also shows potential limitations of our methods.

Both sites show peaks in latent heat at 75% saturation, after which increases in soil moisture lead to decreases in evaporation. It is unclear if this behavior is present in the observations or whether the NNLRP model extrapolates by reverting to the mean when it is provided with conditions that have not been observed. It is possible that this behavior is an artifact of soil saturation during precipitation events when measurement errors may be larger. However, it is also possible that this reflects observed behavior at vegetated sites where transpiration decreases when soils become fully saturated.

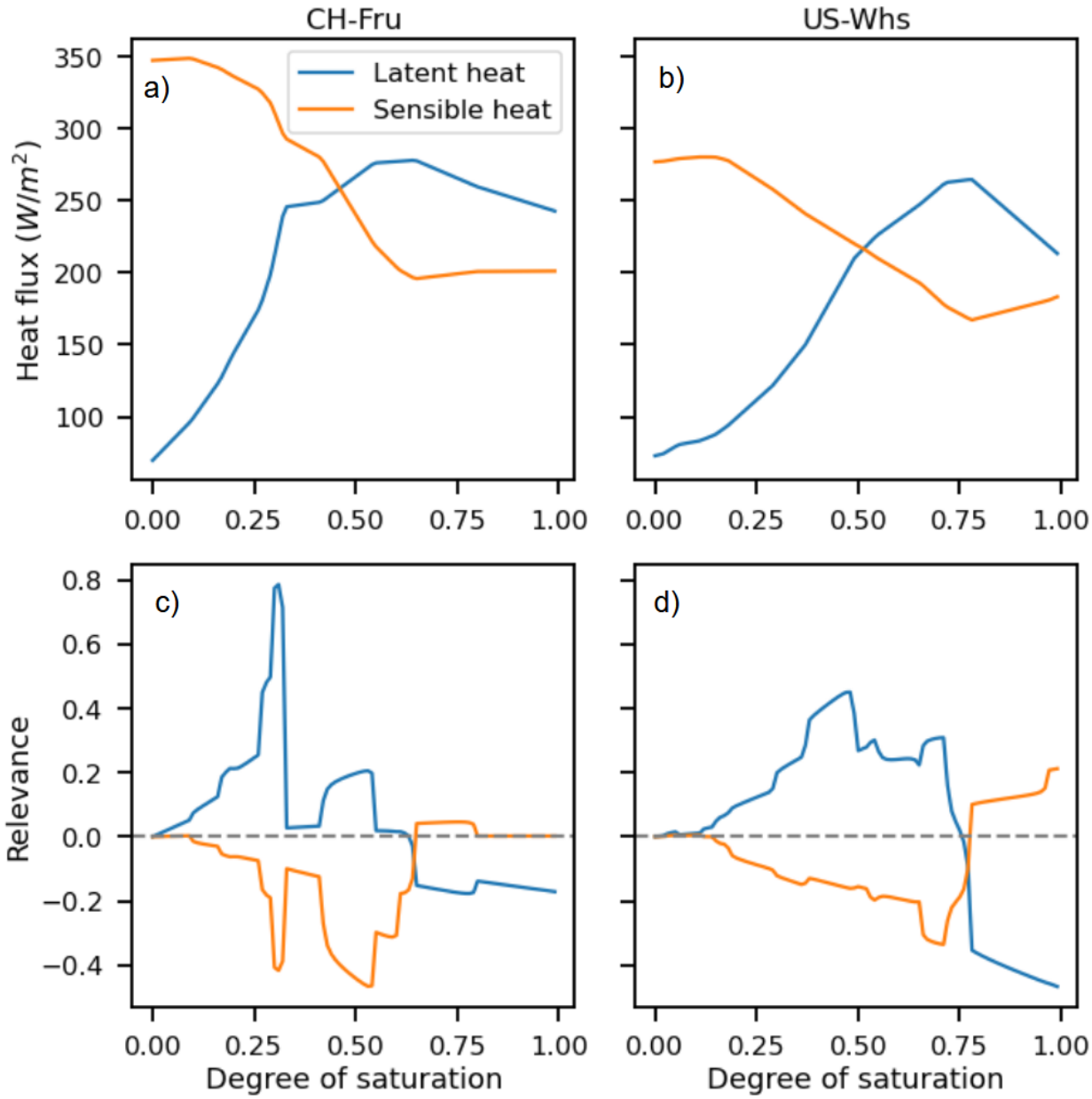

Figure 4.5. Sensitivity of heat fluxes and relevance scores over a range of saturation for CH-Fru and US-Whs

We show the average (normalized) relevance scores of all of the model inputs in figure 4.6, to provide a broader understanding of what the network finds important across sites. Sites were sorted in ascending order of aridity, defined as the long-term total potential evapotranspiration (PET) divided by the long-term total precipitation. PET is calculated according to the Hargreaves

formula (Hargreaves & Allen, 2003). The grey vertical dashed line shows the threshold for PET/P > 1. The general ranking of relevance scores for both latent and sensible heat is stable across sites, particularly the primary importance of air temperature for latent heat and shortwave radiation for sensible heat.

Figure 4.6a indicates that the network learned to use air temperature, relative humidity, shortwave radiation, and surface soil moisture to "produce" latent heat fluxes and vegetation type, soil type, and transpirable water to "inhibit" latent heat fluxes. On average the positive relevance scores are about two times as large as negative relevance scores, indicating that the network is more sensitive to changes that increase the predicted latent heat than changes that decrease it. This is particularly true when PET/P < 1 (energy limited sites), which shows that the network was learned that additional moisture was available for evapotranspiration.

On the other hand, only shortwave and relative humidity have consistently positive relevance scores (figure 4.6b). Air temperature, surface moisture, and vegetation type have consistently negative relevance scores. The consistently negative relevance of vegetation type is interesting, as it is a static input to the network. It seems that NNLRP uses vegetation type to partition the latent and sensible heat fluxes differently in different ecosystems. The need to include vegetation type to maintain performance (as discussed in section 4.3.1) indicates that the other inputs were not sufficient to distinguish between different vegetation types, and therefore site-specific behaviors of turbulent heat fluxes. The importance of vegetation type as a static feature shows that finding better input variables that are able to predict site-specific properties should improve the performance and generality of neural networks to predict turbulent heat fluxes. We will return to this in section 4.3.3.

The relevance breakdown across sites for sensible heat shows more variation than that of latent heat. This is largely due to the contributions of relative humidity and vegetation type. Because vegetation type is site-specific it is hard to disentangle it from the other variables which are temporally varying. We will analyze the site-specific behavior further in section 4.3.3. An interesting feature of figure 4.6 is that the relevance of relative humidity to latent heat tends to be negative for PET/P<1, and positive when PET/P>1. Similarly, the relevance of relative humidity to sensible heat is often a considerable fraction of the positive relevance when PET/P<1 and is greatly diminished when PET/P>1. This indicates that the network learned different relationships for these two regimes.

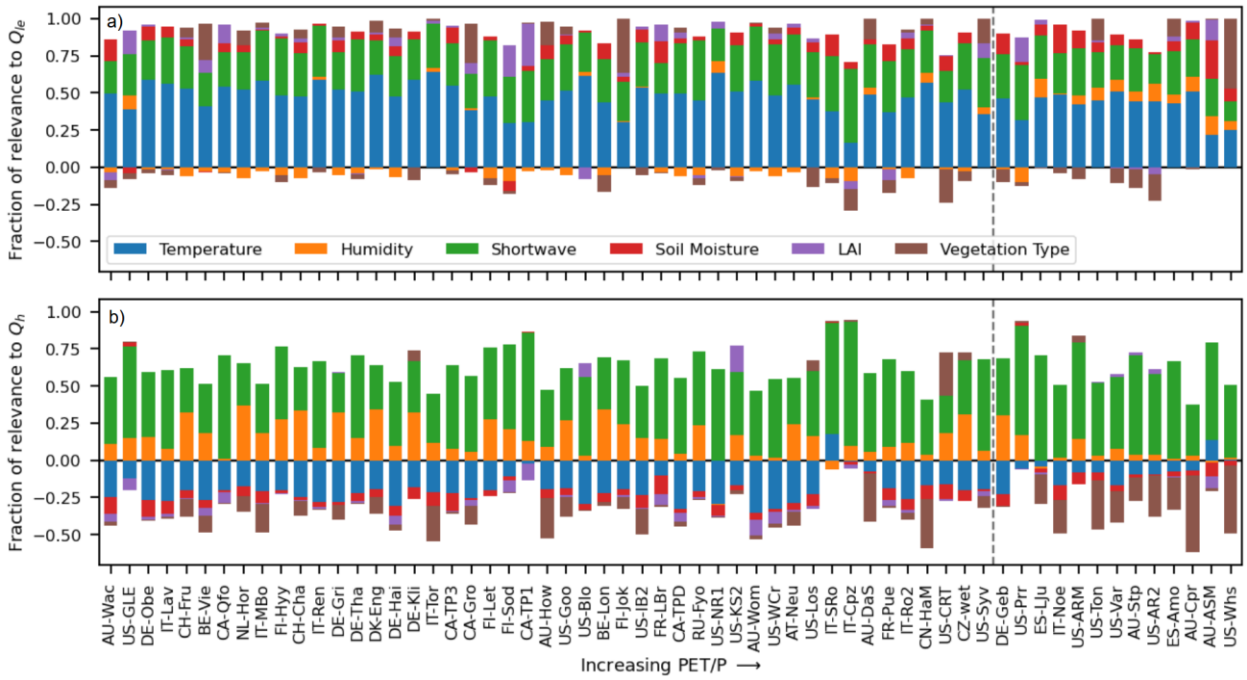

Figure 4.6. Average fraction of relevance by input variable. Sites are sorted by increasing PET/P. The dashed line shows the threshold of PET/P = 1, with energy-limited sites to the left and moisture-limited sites to the right.

The relevance curves shown in figure 4.5 are site specific, with the shape and magnitudes of the relevance varying largely between sites. However, the strength of the correspondence

between tradeoffs in relevance between latent and sensible heats is controlled by whether a site is energy limited. We show this in figure 4.7, where we computed the correlation between the soil moisture relevance timeseries to latent and sensible heat. For energy-limited sites (PET/P < 1), the correlation varies considerably. Moisture-limited sites (PET/P > 1) show consistently high negative correlations between the same soil moisture relevance timeseries. This high correlation indicates that the network identified when moisture contents are a primary control on the partitioning of energy between latent and sensible heat.

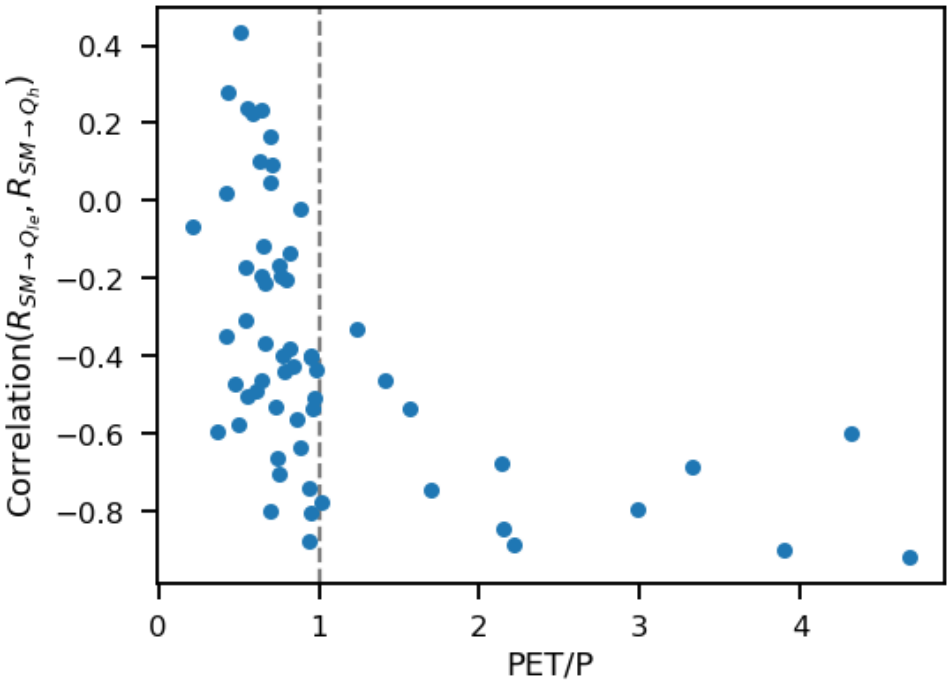

Figure 4.7. Correlation between the relevance between latent and sensible heat with respect to soil moisture.

Another tradeoff that the network learned was the relationship between soil moisture and relative humidity, as previously discussed. To show this, we performed a similar analysis as in figure 4.7, but instead computed the correlation between the relevance of soil moisture to latent heat and the relevance of relative humidity to latent heat. As PET/P increases this correlation

goes from strongly negative to moderately positive, indicating that the neural network learned specific behaviors based on the covariance of these two variables. When moisture is abundant (PET/P << 1), the relative humidity is likely to be high enough that evaporation should be limited by the atmospheric demand, while soil moisture is likely to be high meaning that there is plenty of moisture to be evaporated, but nowhere for it to go. On the other hand, at arid (PET/P >> 1) high levels of humidity are likely to be driven by evaporation, indicating that if humidity is high, the neural network tends to predict that more water can be evaporated.

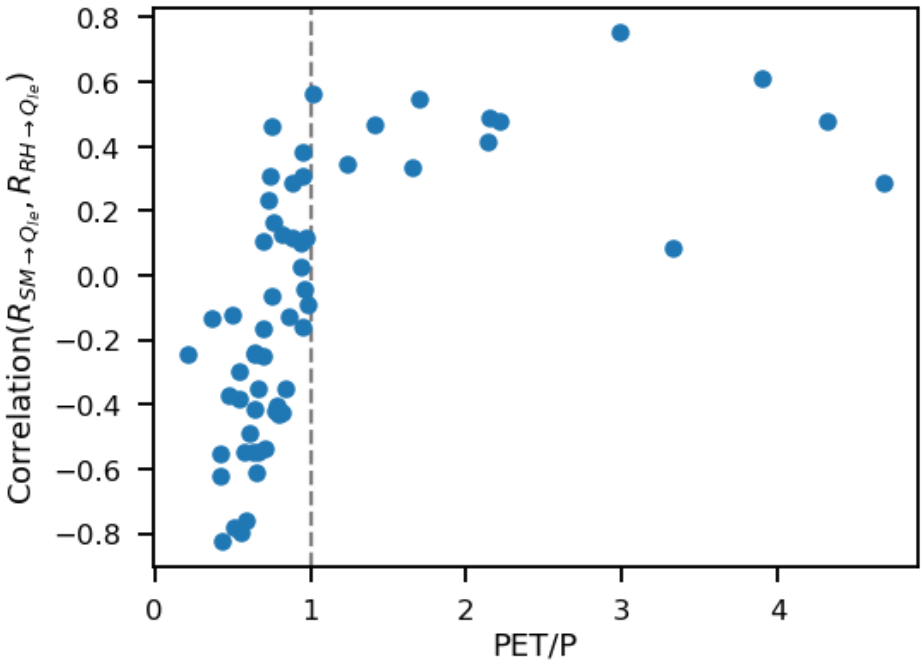

Figure 4.8. Correlation between the relevance of latent heat with respect to soil moisture and relative humidity.

4.3.3 Using LRP to decompose inter-site predictions

Thus far, we have only discussed general properties of NNLRP. As we outlined in section 4.2.4 we can use the relevance score to develop a linear model for each site. This linearized approximation reproduces the neural network output to a high degree of accuracy. We demonstrate this in figure 4.9, where we fit a linear model that uses the relevance scores as inputs to determine the turbulent heat fluxes at the 30-minute time scale. We then compare this

fit to the full timeseries of turbulent heat fluxes simulated by the neural network. We find that the linear models are able achieve KGE values >0.95 on average, confirming our hypothesis that the relevance decomposition provides good explanatory power of the time series of turbulent heat fluxes at each site.

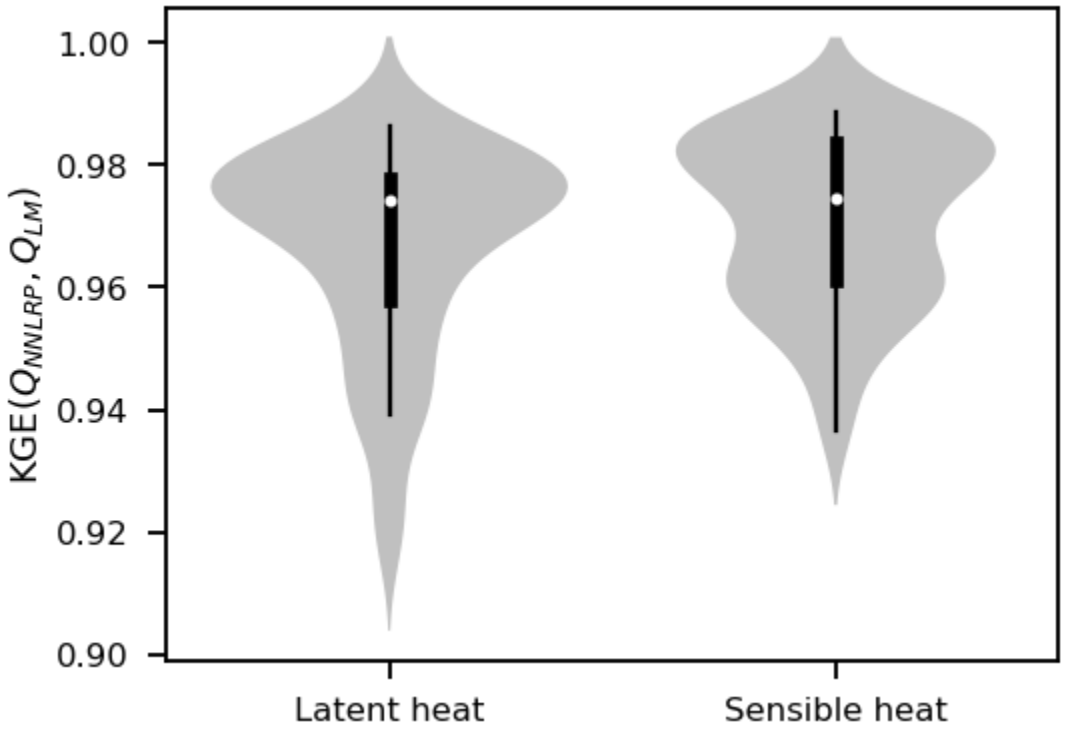

Figure 4.9. Violin plots showing the distribution of KGE between heat flux predicted by the linearized model (Q_{LM}) and NNLRP (Q_{NNLRP}) across all sites. The white dot represents the median, thick black box the interquartile range, and thin black line represents the 95% coverage range.

The success of the linear model that maps relevance to heat flux can then be used to investigate how well the neural network takes information from one site and applies it to another. To do so we fit a linear model at one site, then apply it to a "target" site. We then calculate the KGE between the output of the linear model and the output of the neural network at the target site and call this the inter-site "explainability score". We compute this explainability score for all site pairs, resulting in a matrix of scores, which can be thought of as a weighted-directed graph.

To introduce sparsity into our graph we prune connections so that each site only points to the sites for which it is the best predictor. We also prune connections that do not provide good predictions, with a lower bound for making predictions that achieve a KGE score of at least 90% of that which NNLRP scored. To ensure that record-length did not affect the scores, we used the same number of data points to compute each regression, equal to the number of timesteps at the site with the shortest record (at site CA-TPD, where n_T = 57552 30-minute timesteps).

To make further distinctions between sites we have also taken the coefficients of the multiple linear regression for each site and clustered them using K-means clustering to color the nodes in Figures 4.10 and 4.11. We set the number of clusters to 4 as a balance between visual simplicity and ability to resolve interesting relationships. This 4-category clustering divides the classification into non-Evergreen Needleleaf Forest, and three categories of Evergreen Needleleaf forest (with one outlier at AU-Wac, an Evergreen Broadleaf Forest).

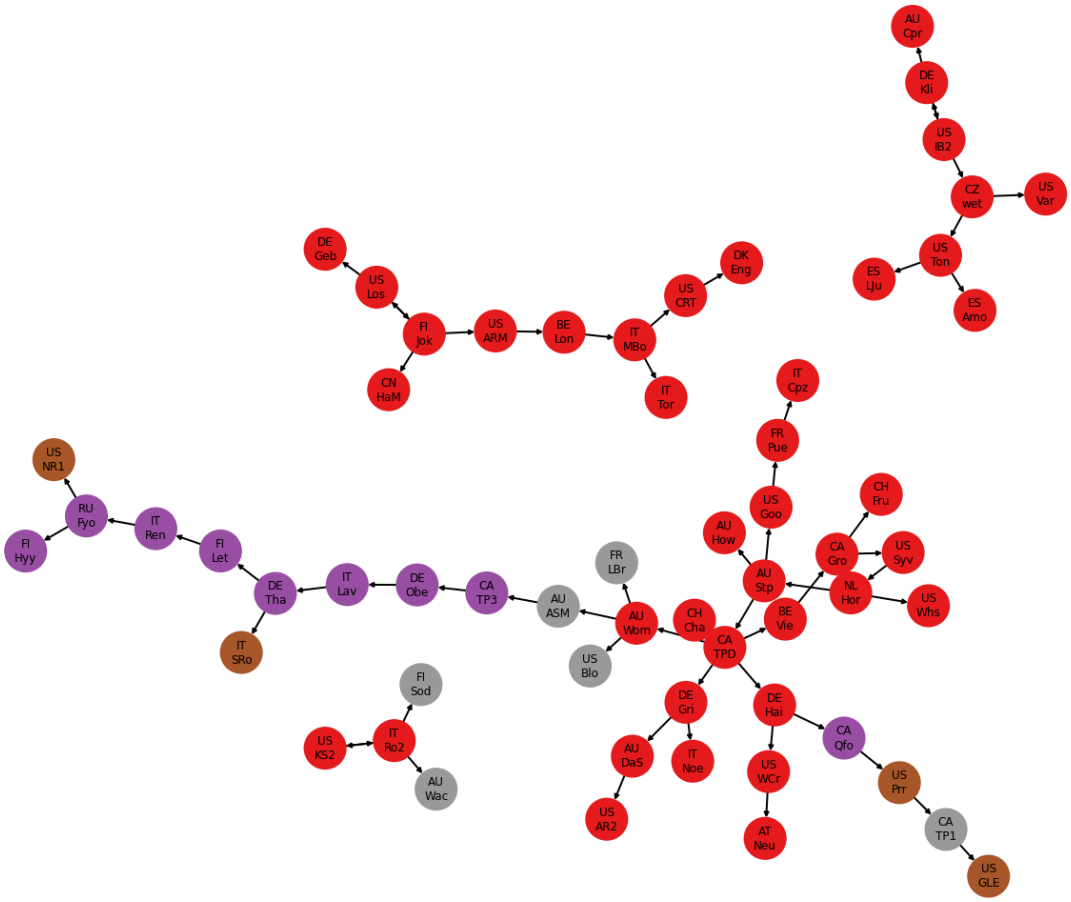

Figure 4.100. Site interaction graph determined by the inter-site explainability for latent heat. Site names are given as the center of each node. Nodes are colored by their k-means clustering. An arrow from a node to another indicates that it is the best predictor for the site being pointed to.

We see a much higher degree of connectivity in the site interconnection graph for latent heat (figure 4.10) in the red nodes (which represent non-evergreen needleleaf sites) than in the brown, purple, and grey nodes (which all represent evergreen needleleaf sites, except for AU-Wac which is an evergreen broadleaf site). This higher interconnectivity of the red nodes reveals that non-evergreen needleleaf sites are more like each other than evergreen needleleaf sites. Further, there are very few red nodes connected to the other classifications, indicating that the clustering uncovered different, unique, descriptions of how turbulent heat fluxes are generated.

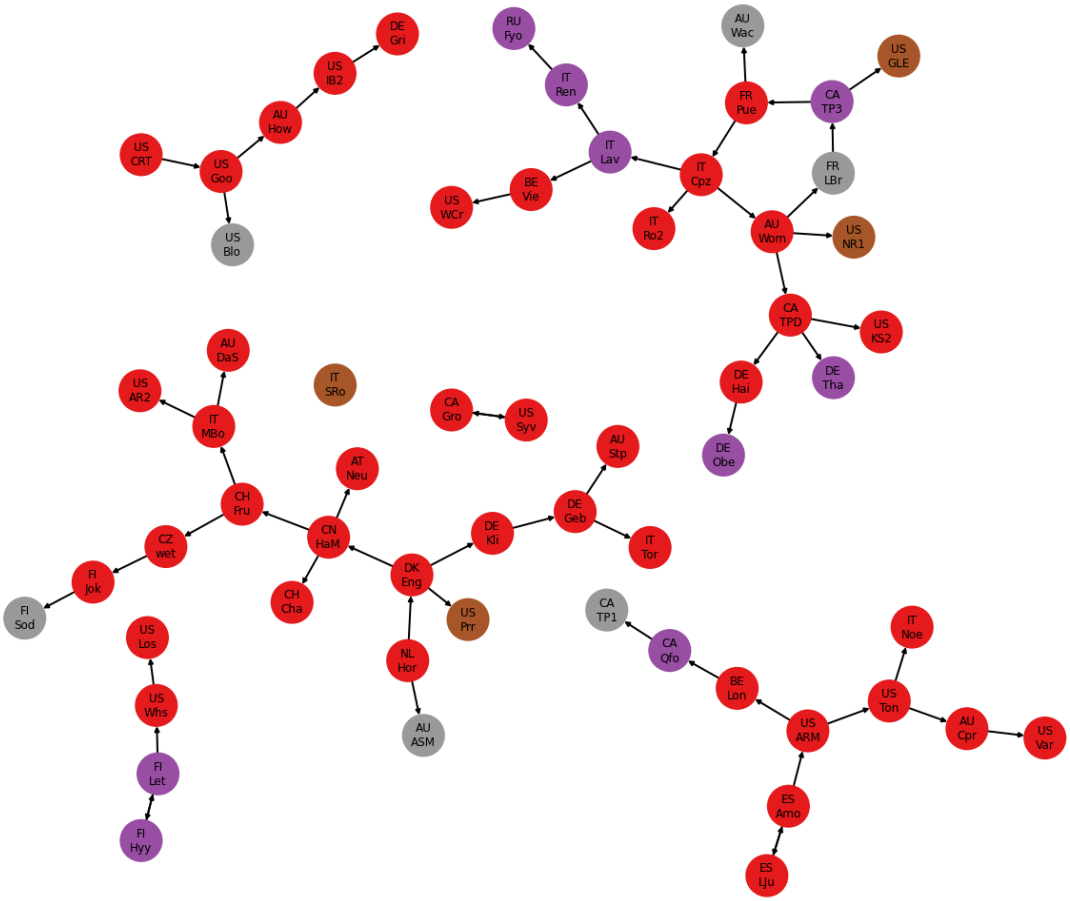

Figure 4.11. Site interaction graph determined by site interaction strength for sensible heat. Site names are given as the center of each node. Nodes are colored by their k-means clustering. An arrow from a node to another indicates that it is the best predictor for the site being pointed to.

The site interconnection graph for sensible heat (figure 4.11) shows similar behaviors. Again, we see that red nodes are more likely to have more connections, and non-red nodes tend to be closer to leaf nodes. However, we see much less clustering together of the non-red nodes. This suggests that the clustering was driven more by the regression coefficients for latent heat than those associated with sensible heat. This, in turn, means that NNLRP learned a more complex relation between site-specific details and latent heat than for sensible heat. It is possible that the simpler representation of sensible heat is related to NNLRP's poorer performance at predicting that flux, but also points to the processes that generate latent heat being more tied to vegetation characteristics than those which generate sensible heat.

In addition to pruning the site interaction graphs we can aggregate them to analyze which sites are difficult to predict and which sites are good predictors. For this analysis, we fit linear models at every site and use these models to predict all other sites. To analyze which sites are hard to predict we fix the target site and count the number of sites whose linear models made predictions at the target site with a KGE value below some threshold (figure 4.12; here, we choose KGE=0.25). These KGE scores are calculated between the linear model and the output from NNLRP as a measure of how well the linear model was able to reproduce NNLRP. Likewise, to analyze which sites are good predictors, we fix the source site and count the number of sites for which the source site model made predictions with a KGE value above some threshold (figure 4.13; here, we choose KGE=0.75).

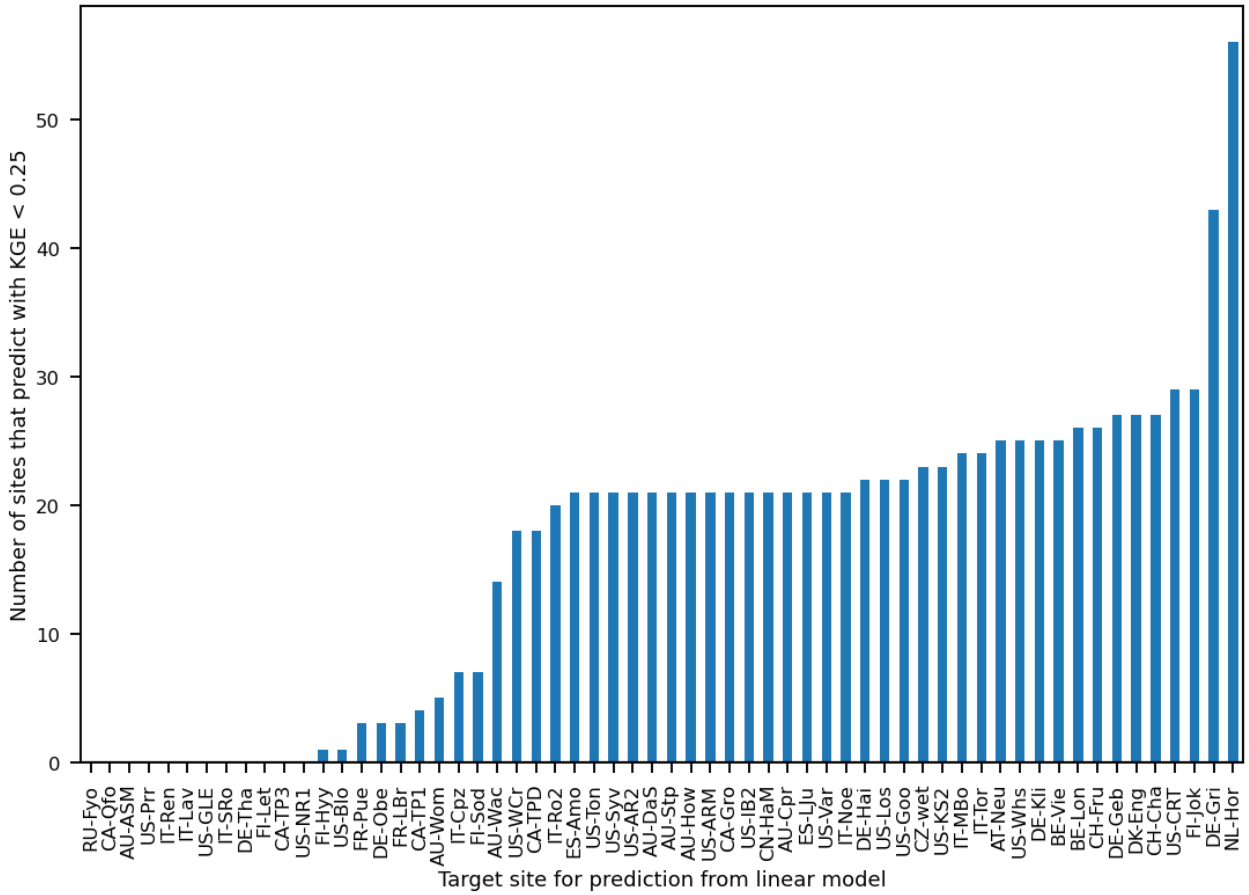

Figure 4.12. Quantification of hard to predict sites. Each target site was predicted by linear models fitted at all sites. The number of sites which made poorly performing predictions at the target site (KGE<0.25) are counted.

In figure 4.12 we see that there are three main site groupings. About one third of sites are almost never poorly predicted by the other sites. Two thirds of all sites are predicted poorly by 20-30 sites, and two sites are predicted poorly by more than 40 other sites. These sites, DE-Gri and NL-Hor, are both grassland sites. We hypothesize that the poor ability for other sites to be used to predict DE-Gri and NL-Hor might be related to data quality, site-specific characteristics, or human interventions. Overall, figure 4.12 indicates that the linear representations of turbulent heat fluxes were able to be transferred between sites without a complete loss of predictive capabilities, except in a few cases.

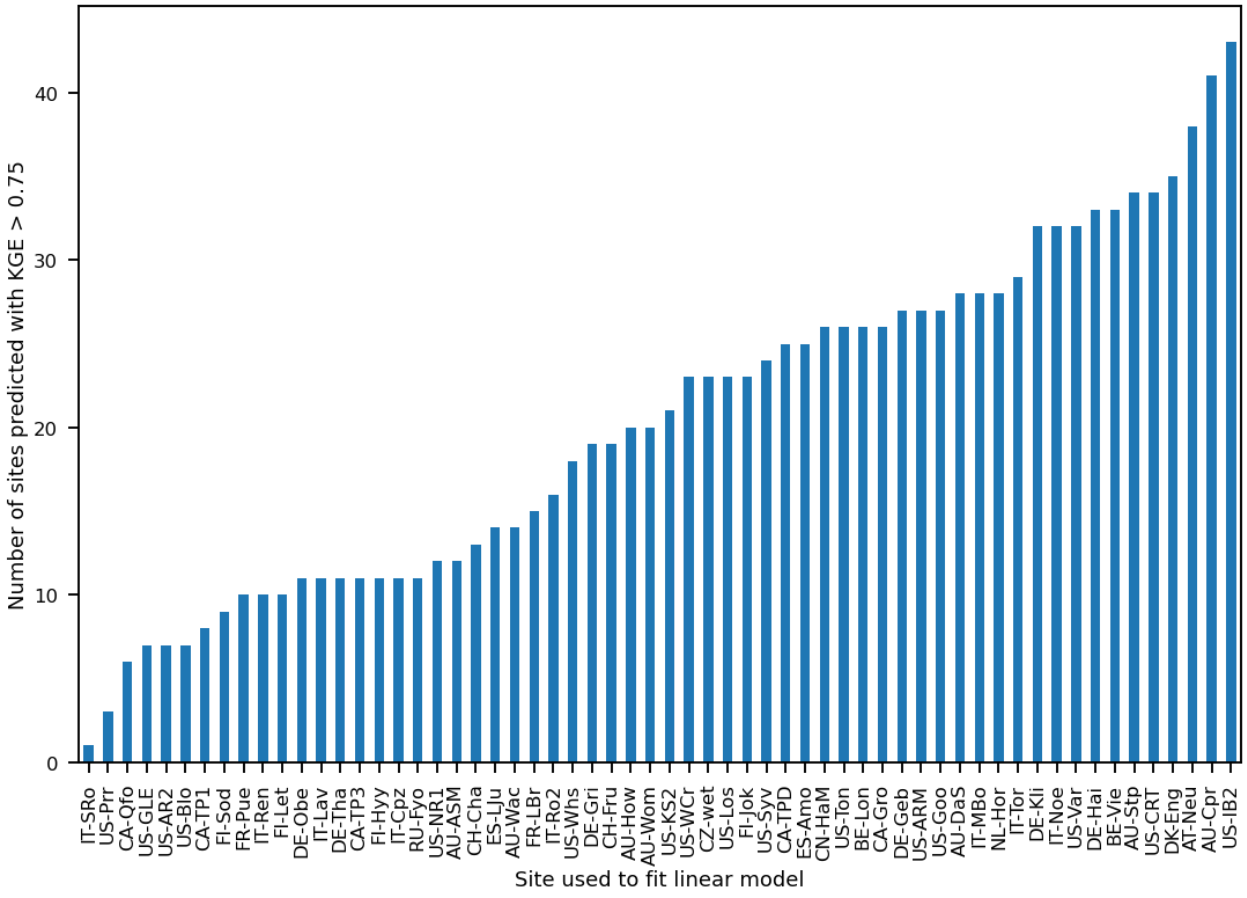

Figure 4.13. Quantification of sites which were good predictors. Each source site was used to

predict all sites. The number of target sites where the source site model provides good predictions (KGE>0.75) are counted.

Figure 4.13 reinforces that NNLRP is able to generalize well by learning common behaviors across sites (rather than learning many site-specific representations). This is demonstrated by the fact that almost all linear models fit at individual sites provide good (KGE>0.75) predictions at more than 10 other sites. Linear models from only two sites, IT-SRo and US-Prr, generalize well to fewer than 5 sites. Interestingly, both these sites are never predicted poorly in figure 4.12, indicating that there is likely a more generalizable model representation for these sites than the one that was fit by the linear model on their respective relevance scores. On the other hand, AU-Cpr and US-IB2 provide good predictions for more than 40 other sites. Sites which generalize well might be used as indicator sites to better inform future data collection and model development activities.

4.4 Discussion

Our LRP-based analysis of a neural network for simulating latent and sensible heat fluxes identified relationships between inputs and outputs that generally agree with physical intuitions and hydrologic theory. Further, we showed that the network was uncovered constraints and learn how to partition turbulent heat fluxes in a physically plausible way. For instance, NNLRP was predicted that at arid sites the importance of soil moisture to latent heat should be inversely proportional to the importance of soil moisture for sensible heat.

While LRP analysis does not provide us with (parsimonious) symbolic relationships between inputs and outputs, it does indicate that neural networks may be capable of learning physical behavior even when they are not specifically guided to do so. Building models which directly

encode constraints or promote known relationships may allow us to build networks that are more realistic.

Even though we say that the neural network learned physically plausible relationships, much work remains to be done to adequately constrain deep-learning based models of physical processes. For instance, in figure 4.5 we found that the relevance scores were quite jagged over the full range of degrees of soil saturation. While there may be soil moisture thresholds in real systems, we do not expect that the sensitivities produced by the neural network will be identical to those observed in the environment. Incorporating more observations into the training dataset across a wider range of environments should further constrain these sensitivities. Additionally, further refinement of neural network architectures and input features should result in better estimates of these sensitivities.

Sampling a full range of one variable while holding all other inputs constant is an easy way to screen for model sensitivity and can expose ways in which DL models fail (or make incorrect inferences) (Szegedy et al., 2014). Though catastrophic failure modes in DL models have been observed in other applications (Huang et al., 2017; Nguyen et al., 2015), the results from our analyses show that the NNLRP configuration does not “blow up” when pushed to the edges of the data distributions on which it was trained. We believe that this is because our dataset covers the phase space well and is generally well constrained. DL-based solutions to problems which incorporate much higher dimensional data with more inputs or with spatio-temporal awareness seem to be more likely to produce catastrophic failure modes.

To our knowledge, the use of LRP relevance decompositions to build linear models to compare inter-site relationships is a new technique. This approach allowed us to look at which sites the DL model was able to use for predictions of the other sites. Taking a graphical approach

we were able to see that certain sites were “indicator” sites, while others were not well predicted by any sites. It may be possible to relate this clustering technique to develop data-driven approaches to classifying catchment similarity (Wagener et al., 2007). We imagine that this type of approach might also be used to make recommendations for where future observations might be made or to better understand and categorize land-atmosphere interactions.

It is important to make the distinction that our results are based on the simplest neural network available, a densely connected network. Both convolutional and recurrent neural networks (CNNs and RNNs, respectively) have been used to great effect in hydrology and can aid interpretation when implemented carefully. For instance, the hidden states of RNNs can be viewed as proxies for stateful quantities such as snowpack (Hoedt et al., 2021; Jiang et al., 2020; Kratzert et al., 2018) while CNNs can distill spatial relationships (Castelluccio et al., 2015; Geng et al., 2015). LRP has been more successfully applied to CNNs than deep-dense networks, due to their reduced dimensionality and preservation of local structures (Samek et al., 2019). LRP can also be applied to RNNs, though the methodology is not as well-developed as for convolutional networks (Arras et al., 2017, 2019). Future applications of such methods in conjunction with more advanced XAI methods will likely be able to uncover physical relationships in higher fidelity than previous methods.

4.5 Conclusions

The use of XAI methods can help interpret how neural networks make their predictions. In this study we have shown how a particular technique, LRP, can be used to understand a neural network for predicting turbulent heat fluxes. LRP decomposes each individual prediction that the neural network makes into a set of relevance scores, which explain how important each input feature was to that prediction. This can be done for all predictions, producing timeseries of

relevance scores. We showed that the overall importance of variables to each latent and sensible heat follow physical intuition. For latent heat we found that air temperature and shortwave radiation were both drivers of latent heat production across sites. For sensible heat the shortwave radiation was the main driver, while air temperature was used to partition between latent and sensible heat. Further, at many sites the relative humidity was an important factor for predicting sensible heat.

We also showed that NNLRP was learned partitioning behaviors. At arid sites NNLRP learned to use soil moisture as a strong indicator for the partitioning between latent and sensible heat. NNLRP also learned different behaviors for using relative humidity at moisture and energy limited sites. This indicates that neural networks can automatically discover and encode information about physical processes that it has not been told about, purely from data. While we are far from being able to translate these discoveries into new theory, it does indicate the possibility that we may in the future. Improvements in XAI methods and improving the types of ML models which we use for scientific applications will further the goal of developing new theory from ML based models.

Alongside improvements to the XAI and ML methods, we also argue that it is important to continue to design experiments to address questions that cannot be investigated with straightforward applications of other methods. We used the LRP decomposition to compare what NNLRP learned between sites. LRP analysis provided a way to cluster the sites and identified sites that were unique, as well as “indicator” sites which provide good predictions for large numbers of other sites. XAI methods offer ways in which we can learn from the trained networks, rather than just being able to make predictions. Training networks with architectures

which promote interpretability and continuing to develop ways to extract information from them looks to be a promising way to learn from large datasets.

Chapter 5. CONCLUSIONS AND FUTURE WORK

5.1 CONCLUSIONS

In this dissertation I used data-driven techniques to analyze the hydrologic cycle. This encompassed studies disentangling process connectivity in hydrologic models with information theory, parameterizing hydrologic models with machine learning, and interpreting machine-learned parameterizations from a physical perspective. Throughout these studies I have developed and used methods that synthesize large amounts of data, either modeled or observed, to build a better understanding of how to build and apply hydrologic models.

In chapter 2 I showed that a holistic approach to model intercomparison and evaluation makes it possible to disentangle the complex interdependencies of macroscale hydrologic models. I used transfer entropy to compare three macroscale hydrologic model setups (SUMMA, VIC, and PRMS) in the Pacific northwestern United States and showed that model structure and parameterization can drastically alter hydrologic process interactions.

In chapter 3 I showed how individual hydrologic processes can be parameterized by deep-learned models, and then incorporated into existing process-based hydrologic modeling frameworks. I trained two coupled DL-PBHM configurations, NN1W and NN2W, to simulate turbulent heat fluxes across a large array of FluxNet sites. Both NN1W and NN2W outperformed the calibrated process-based formulation, SA, which was calibrated in-sample, while the DL based configurations were trained out of sample. NN2W also reproduced both short and long-term signatures much more closely than the process-based formulation, indicating that NN2W was able to learn relationships that approximate physical behaviors quite well.

To further explore this, in chapter 4 I analyzed a neural network that is similar to the NN2W parameterization from chapter 4 to determine whether the learned relationships are physically

realistic, and how the model transferred learned behaviors between sites. I used the LRP technique to map the importance of each input feature to the output of the network. I showed that even simple neural networks can learn behaviors that are physically plausible. Further, I showed that the network learned to partition between dominant processes in moisture and energy limited sites. To understand how the network learned to generalize from site to site I developed a novel method for linearizing the behavior predicted by the neural network. This methodology showed that the network learned site-specific behavior based primarily on vegetation type, but that it also identified indicator sites which are able to build models that are predictive at a large number of other sites.

5.2 Future work

The work in this dissertation is merely a series of vignettes of what is possible with data-driven methods in the hydrologic sciences. Many other applications have been previously successful, and many more will be in the future. To this end, I will speculate as to the avenues which might be most fruitful based on my experience. Many information theoretic measures require the explicit ability to estimate joint probability densities. For sites where there are many concurrent measurements, or where simulation is possible this is not an issue. However, the use of process networks as a multi-process and inter-process evaluation criteria may require more advanced methods for inferring joint distributions of processes which are not co-observed. There is promising work to this effect in developing more advanced copula-based techniques (which relate joint to univariate probability distributions).

Similarly, advances in coupling process-based hydrologic models to deep learning based models is, in many ways, a technological problem. The technology factor in the rise of deep-learning has been documented, and likely plays a strong role in the recent uptake in machine

learning in hydrology. Additionally, hydrology specific tools for normalizing data, providing meaningful aggregations (like signatures), enforcing constraints (mass and energy balance), or other sorts of structure will likely continue to make machine learning useful and interpretable.

To that end, further developments in interpretable methods will likely prove useful for ensuring that machine learning is providing robust predictions, but also that they can provide predictions which we can learn from. With advances in model structures, such as building implicit layers which can perform arbitrary computation, we should be able to build confidence in machine learned models for providing not only predictively accurate, but also physically consistent, representations of hydrologic processes.

BIBLIOGRAPHY

Ahmad, S. K., & Hossain, F. (2019). A generic data-driven technique for forecasting of reservoir inflow: Application for hydropower maximization. *Environmental Modelling & Software*, *119*, 147–165. https://doi.org/10.1016/j.envsoft.2019.06.008

Baartman, J. E. M., Melsen, L. A., Moore, D., & van der Ploeg, M. J. (2020). On the complexity of model complexity: Viewpoints across the geosciences. *CATENA*, *186*, 104261. https://doi.org/10.1016/j.catena.2019.104261

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Ball, J. T., Woodrow, I. E., & Berry, J. A. (1987). A Model Predicting Stomatal Conductance and its Contribution to the Control of Photosynthesis under Different Environmental Conditions. In J. Biggins (Ed.), *Progress in Photosynthesis Research: Volume 4 Proceedings of the VIIth International Congress on Photosynthesis Providence, Rhode Island, USA, August 10–15, 1986* (pp. 221–224). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-0519-6_48

Barnes, E. A., Toms, B., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2020). Indicator Patterns of Forced Change Learned by an Artificial Neural Network. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002195. https://doi.org/10.1029/2020MS002195

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bennett, A., & Nijssen, B. (2020, March 12). Deep learned process parameterizations provide better representations of turbulent heat fluxes in hydrologic models [preprint]. https://doi.org/10.1002/essoar.10505081.1

Best, M., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., et al. (2015). The Plumbing of Land Surface Models: Benchmarking Model Performance. *Journal of Hydrometeorology*. https://doi.org/10.1175/JHM-D-14-0158.1

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., et al. (2015). The Plumbing of Land Surface Models: Benchmarking Model Performance. *Journal of Hydrometeorology*, *16*(3), 1425–1442. https://doi.org/10.1175/JHM-D-14-0158.1

Beven, K. (2002). Towards an alternative blueprint for a physically based digitally simulated hydrologic response modelling system. *Hydrological Processes*, *16*(2), 189–206. https://doi.org/10.1002/hyp.343

Beven, K., Cloke, H., Wagener, T., Lees, M. J., & Wheater, H. S. (2012). Comment on "hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water" by Eric F. Wood et al. *Water Resources Publications*, *2*(1), 1–10. https://doi.org/10.1029/2010WR010090

Blöschl, G., & Sivapalan, M. (1995). Scale issues in hydrological modelling: A review. *Hydrological Processes*, *9*(3–4), 251–290. https://doi.org/10.1002/hyp.3360090305

Blöschl, Günter, Bierkens, M. F. P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., et al. (2019). Twenty-three unsolved problems in hydrology (UPH) – a community perspective. *Hydrological Sciences Journal*, *64*(10), 1141–1158. https://doi.org/10.1080/02626667.2019.1620507

Bohn, T. J., Livneh, B., Oyler, J. W., Running, S. W., Nijssen, B., & Lettenmaier, D. P. (2013). Global evaluation of MTCLIM and related algorithms for forcing of ecological and hydrological models. *Agricultural and Forest Meteorology*, *176*, 38–49. https://doi.org/10.1016/j.agrformet.2013.03.003

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic Validation of a Neural Network Unified Physics Parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298. https://doi.org/10.1029/2018GL078510

Breuer, L., Huisman, J. A., Willems, P., Bormann, H., Bronstert, A., Croke, B. F. W., et al. (2009). Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use. *Advances in Water Resources*, *32*(2), 129–146. https://doi.org/10.1016/j.advwatres.2008.10.003

Brunton, S. L., & Kutz, J. N. (2019). *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108380690

Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, *113*(15), 3932–3937. https://doi.org/10.1073/pnas.1517384113

Camuffo, D., & Bernardi, A. (1982). An observational study of heat fluxes and their relationships with net radiation. *Boundary-Layer Meteorology*, *23*(3), 359–368. https://doi.org/10.1007/BF00121121

Castelluccio, M., Poggi, G., Sansone, C., & Verdoliva, L. (2015). Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. *ArXiv:1508.00092 [Cs]*. Retrieved from http://arxiv.org/abs/1508.00092

Chegwidden, O. S., Nijssen, B., Rupp, D. E., Arnold, J. R., Clark, M. P., Hamman, J. J., et al. (2018). How do modeling decisions affect the spread amongh hydrologic climate change projections? *Earth's Future*, *In Review*. https://doi.org/10.1007/BF00702739.Sunda

Chollet, F. & others. (2015). Keras. Retrieved from https://github.com/fchollet/keras

Clark, M. P., Kavetski, D., & Fenicia, F. (2011a). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*(9), 1–16. https://doi.org/10.1029/2010WR009827

Clark, M. P., Kavetski, D., & Fenicia, F. (2011b). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*(9), 1–16. https://doi.org/10.1029/2010WR009827

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., et al. (2015a). A unified approach for process-based hydrologic modeling: 1. Modeling concept: A unified approach for process-based hydrologic modeling. *Water Resources Research*, *51*(4), 2498–2514. https://doi.org/10.1002/2015WR017198

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Gochis, D. J., et al. (2015). A unified approach for process-based hydrologic modeling: 2. Model implementation and case studies. *Water Resources Research*. https://doi.org/10.1002/2015WR017200

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. a, Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., et al. (2015b). A unified approach for process-based hydrologicmodeling: 1. Modeling concept. *Water Resources Research*, *51*(4), 1–17. https://doi.org/10.1002/2015WR017200.A

Cover, T. M., & Thomas, J. A. (2005). Elements of Information Theory Telecommunication Transmission Handbook, 3rd Edition. *Elements of Information Theory*. https://doi.org/10.1002/047174882X

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. New York, NY, USA: Wiley-Interscience.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., et al. (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *137*(656), 553–597. https://doi.org/10.1002/qj.828

Dobrescu, A., Giuffrida, M. V., & Tsaftaris, S. A. (2019). Understanding Deep Neural Networks for Regression in Leaf Counting. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 2600–2608). Long Beach, CA, USA: IEEE. https://doi.org/10.1109/CVPRW.2019.00316

Dooge, J., C. I. (1988). Hydrology in perspective. *Hydrological Sciences Journal*, *33*(1), 61–85. https://doi.org/10.1080/02626668809491223

Feng, D., Fang, K., & Shen, C. (2020). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *ArXiv:1912.08949 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1912.08949

Foken, T. (2008). The Energy Balance Closure Problem: An Overview. *Ecological Applications*, *18*(6), 1351–1367. https://doi.org/10.1890/06-0922.1

Frame, J., Nearing, G., Kratzert, F., & Rahman, M. (2020). *Post processing the U.S. National Water Model with a Long Short-Term Memory network* (preprint). EarthArXiv. https://doi.org/10.31223/osf.io/4xhac

Frenzel, S., & Pompe, B. (2007). Partial mutual information for coupling analysis of multivariate time series. *Physical Review Letters*, *99*(20), 1–4. https://doi.org/10.1103/PhysRevLett.99.204101

Geng, J., Fan, J., Wang, H., Ma, X., Li, B., & Chen, F. (2015). High-Resolution SAR Image Classification via Deep Convolutional Autoencoders. *IEEE Geoscience and Remote Sensing Letters*, *12*(11), 2351–2355. https://doi.org/10.1109/LGRS.2015.2478256

Geng, Z., & Wang, Y. (2020). Automated design of a convolutional neural network with multi-scale filters for cost-efficient seismic data classification. *Nature Communications*, *11*(1), 3311. https://doi.org/10.1038/s41467-020-17123-6

Gong, W., Gupta, H. V., Yang, D., Sricharan, K., & Hero, A. O. (2013). Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach. *Water Resources Research*, *49*(4), 2253–2273. https://doi.org/10.1002/wrcr.20161

Gong, W., Yang, D., Gupta, H. V., & Nearing, G. (2014). Estimating information entropy for hydrological data: One-dimensional case. *Water Resources Research*. https://doi.org/10.1002/2014WR015874

Goodwell, A., & Kumar, P. (2015). Information theoretic measures to infer feedback dynamics in coupled logistic networks. *Entropy*, *17*(11), 7468–7492. https://doi.org/10.3390/e17117468

Goodwell, A., & Kumar, P. (2017). Process Network Approach To Infer Ecohydrologic Shifts, 1–21. https://doi.org/10.1002/2016WR020216.Temporal

Goodwell, A. E., & Kumar, P. (2017). Temporal information partitioning: Characterizing synergy, uniqueness, and redundancy in interacting environmental variables. *Water Resources Research*, *53*(7), 5920–5942. https://doi.org/10.1002/2016WR020216

Goria, M. N., Leonenko, N. N., Mergel, V. V., & Inverardi, P. L. N. (2005). A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Journal of Nonparametric Statistics*, *17*(3), 277–297. https://doi.org/10.1080/104852504200026815

Gupta, H. V, & Nearing, G. S. (2014). Debates—The future of hydrological sciences: A (common) path forward? Usingmodels and data to learn: A systems theoretic perspective on the future of hydrological science. *Water Resources Research*, *50*, 1–9. https://doi.org/10.1002/2013WR015096.Received

Hargreaves, G. H., & Allen, R. G. (2003). History and Evaluation of Hargreaves Evapotranspiration Equation. *Journal of Irrigation and Drainage Engineering*, *129*(1), 53–63. https://doi.org/10.1061/(ASCE)0733-9437(2003)129:1(53)

Hey, T., Tansley, S., & Tolle, K. (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Retrieved from https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/

Hlaváčková-Schindler, K., Paluš, M., Vejmelka, M., & Bhattacharya, J. (2007). Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, *441*(1), 1–46. https://doi.org/10.1016/j.physrep.2006.12.004

Hlinka, J., Hartman, D., Vejmelka, M., Runge, J., Marwan, N., Kurths, J., & Palŭs, M. (2013). Reliability of inference of directed climate networks using conditional mutual information. *Entropy*, *15*(6), 2023–2045. https://doi.org/10.3390/e15062023

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of Predictions in Ungauged Basins (PUB)—a review. *Hydrological Sciences Journal*, *58*(6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183

Hu, C., Wu, Q., Li, H., Jian, S., Li, N., & Lou, Z. (2018). Deep Learning with a Long Short-Term Memory Networks Approach for Rainfall-Runoff Simulation. *Water*, *10*(11), 1543. https://doi.org/10.3390/w10111543

Huang, S., Papernot, N., Goodfellow, I., Duan, Y., & Abbeel, P. (2017). Adversarial Attacks on Neural Network Policies. *ArXiv:1702.02284 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1702.02284

Jiang, S., Zheng, Y., & Solomatine, D. (2020). Improving AI System Awareness of Geoscience Knowledge: Symbiotic Integration of Physical Approaches and Deep Learning. *Geophysical Research Letters*, *47*(13), e2020GL088229. https://doi.org/10.1029/2020GL088229

Jung, M., Reichstein, M., & Bondeau, A. (2009). Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model, 13.

Kampf, S. K., & Burges, S. J. (2007a). A framework for classifying and comparing distributed hillslope and catchment hydrologic models. *Water Resources Research*, *43*(5). https://doi.org/10.1029/2006WR005370

Kampf, S. K., & Burges, S. J. (2007b). A framework for classifying and comparing distributed hillslope and catchment hydrologic models: DISTRIBUTED MODEL REVIEW. *Water Resources Research*, *43*(5). https://doi.org/10.1029/2006WR005370

Kavetski, D., & Clark, M. P. (2010). Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction. *Water Resources Research*, *46*(10), 1–27. https://doi.org/10.1029/2009WR008896

Kidston, J., Brümmer, C., Black, T. A., Morgenstern, K., Nesic, Z., McCaughey, J. H., & Barr, A. G. (2010). Energy Balance Closure Using Eddy Covariance Above Two Different Land Surfaces and Implications for CO2 Flux Measurements. *Boundary-Layer Meteorology*, *136*(2), 193–218. https://doi.org/10.1007/s10546-010-9507-y

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Koster, R. D., & Milly, P. C. D. (1997). The interplay between transpiration and Runoff formulations in land surface schemes used with atmospheric models. *Journal of Climate*, *10*(7), 1578–1591. https://doi.org/10.1175/1520-0442(1997)010<1578:TIBTAR>2.0.CO;2

Koutsoyiannis, D. (2005). Uncertainty, entropy, scaling and hydrological stochastics. 1. Marginal distributional properties of hydrological processes and state scaling. *Hydrological Sciences Journal*, *50*(3), 381–404. https://doi.org/10.1623/hysj.50.3.381.65031

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, *69*(6), 16. https://doi.org/10.1103/PhysRevE.69.066138

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall-Runoff modelling using Long-Short-Term-Memory (LSTM) networks. *Hydrology and Earth System Sciences Discussions*, 1–26. https://doi.org/10.5194/hess-2018-247

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. *Water Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019WR026065

Kreyenberg, P. J., Bauser, H. H., & Roth, K. (2019). Velocity Field Estimation on Density-Driven Solute Transport With a Convolutional Neural Network. *Water Resources Research*, *55*(8), 7275–7293. https://doi.org/10.1029/2019WR024833

Kumar, P., & Ruddell, B. L. (2010). Information driven ecohydrologic self-organization. *Entropy*, *12*(10), 2085–2096. https://doi.org/10.3390/e12102085

Lathière, J., Hauglustaine, D. A., & Friend, A. D. (2006). Impact of climate variability and land use changes on global biogenic volatile organic compound emissions. *Atmos. Chem. Phys.*, 19.

Lee, J., Nemati, S., Silva, I., Edwards, B. A., Butler, J. P., & Malhotra, A. (2012). Transfer Entropy Estimation and Directional Coupling Change Detection in Biomedical Time Series. *BioMedical Engineering Online*, *11*, 1–17. https://doi.org/10.1186/1475-925X-11-19

Li, L., Wang, Y.-P., Yu, Q., Pak, B., Eamus, D., Yan, J., et al. (2012). Improving the responses of the Australian community land surface model (CABLE) to seasonal drought. *Journal of Geophysical Research: Biogeosciences*, *117*(G4). https://doi.org/10.1029/2012JG002038

Lipton, Z. C. (2017). The Mythos of Model Interpretability. *ArXiv:1606.03490 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1606.03490

Liu, Y., & Wu, L. (2016). Geological Disaster Recognition on Optical Remote Sensing Images Using Deep Learning. *Procedia Computer Science*, *91*, 566–575. https://doi.org/10.1016/j.procs.2016.07.144

Livneh, B., Rosenberg, E. A., Lin, C., Nijssen, B., Mishra, V., Andreadis, K. M., et al. (2013). A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions. *Journal of Climate*, *26*(23), 9384–9392. https://doi.org/10.1175/JCLI-D-12-00508.1

Loritz, R., Gupta, H., Jackisch, C., Westhoff, M., Kleidon, A., Ehret, U., & Zehe, E. (2018). On the dynamic nature of hydrological similarity. *Hydrology and Earth System Sciences*, *22*(7), 3663–3684. https://doi.org/10.5194/hess-22-3663-2018

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *ArXiv:1705.07874 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1705.07874

Matott, L. S. (2017). OSTRICH: an Optimization Software Tool, Documentation and User's Guide, Version 17.12.19. University at Buffalo Center for Computational Research. Retrieved from www.eng.buffalo.edu/~lsmatott/Ostrich/OstrichMain.html

McGovern, A., Lagerquist, R., Gagne, D. J., Jergensen, G. E., Elmore, K., Homeyer, C., & Smith, T. (2019). Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning in: Bulletin of the American Meteorological Society Volume 100 Issue 11 (2019). *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199. https://doi.org/10.1175/BAMS-D-18-0195.1

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, *65*, 211–222. https://doi.org/10.1016/j.patcog.2016.11.008

Moshe, Z., Metzger, A., Elidan, G., Kratzert, F., Nevo, S., & El-Yaniv, R. (2020). HydroNets: Leveraging River Structure for Hydrologic Modeling. Retrieved from https://arxiv.org/abs/2007.00595v1

Musselman, K. N., Clark, M. P., Liu, C., Ikeda, K., & Rasmussen, R. (2017). Slower snowmelt in a warmer world. *Nature Climate Change*, *7*(3), 214–219. https://doi.org/10.1038/nclimate3225

Nearing, G. S., Tian, Y., Gupta, H. V., Clark, M. P., Harrison, K. W., & Weijs, S. V. (2016). A philosophical basis for hydrological uncertainty. *Hydrological Sciences Journal*, *61*(9), 1666–1678. https://doi.org/10.1080/02626667.2016.1183009

Nearing, G. S., Mocko, D. M., Peters-Lidard, C. D., Kumar, S. V., & Xia, Y. (2016a). Benchmarking NLDAS-2 Soil Moisture and Evapotranspiration to Separate Uncertainty Contributions. *Journal of Hydrometeorology*, *17*(3), 745–759. https://doi.org/10.1175/JHM-D-15-0063.1

Nearing, G. S., Mocko, D. M., Peters-Lidard, C. D., Kumar, S. V., & Xia, Y. (2016b). Benchmarking NLDAS-2 Soil Moisture and Evapotranspiration to Separate Uncertainty Contributions. *Journal of Hydrometeorology*, *17*(3), 745–759. https://doi.org/10.1175/JHM-D-15-0063.1

Nearing, G. S., Ruddell, B. L., Bennett, A. R., Prieto, C., & Gupta, H. V. (2020). Does Information Theory Provide a New Paradigm for Earth Science? Hypothesis Testing. *Water Resources Research*, *56*(2). https://doi.org/10.1029/2019WR024918

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., et al. (2020). What Role Does Hydrological Science Play in the Age of Machine Learning? *Water Resources Research*, *n/a*(n/a), e2020WR028091. https://doi.org/10.1029/2020WR028091

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 427–436). https://doi.org/10.1109/CVPR.2015.7298640

Nijssen, B., Bowling, L. C., Lettenmaier, D. P., Clark, D. B., El Maayar, M., Essery, R., et al. (2003). Simulation of high latitude hydrological processes in the Torne-Kalix basin: PILPS Phase 2(e) 2: Comparison of model results with observations. *Global and Planetary Change*, *38*(1–2), 31–53. https://doi.org/10.1016/S0921-8181(03)00004-3

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research: Atmospheres*, *116*(D12). https://doi.org/10.1029/2010JD015139

Ombadi, M., Nguyen, P., Sorooshian, S., & Hsu, K. (2020). Evaluation of Methods for Causal Discovery in Hydrometeorological Systems. *Water Resources Research*, *56*(7), e2020WR027251. https://doi.org/10.1029/2020WR027251

Ott, J., Pritchard, M., Best, N., Linstead, E., Curcic, M., & Baldi, P. (2020). A Fortran-Keras Deep Learning Bridge for Scientific Computing. *ArXiv:2004.10652 [Cs]*. Retrieved from http://arxiv.org/abs/2004.10652

Paluš, M. (2014). Cross-scale interactions and information transfer. *Entropy*, *16*(10), 5263–5289. https://doi.org/10.3390/e16105263

Pan, B., Hsu, K., AghaKouchak, A., & Sorooshian, S. (2019). Improving Precipitation Estimation Using Convolutional Neural Network. *Water Resources Research*, *55*(3), 2301–2321. https://doi.org/10.1029/2018WR024090

Paninski, L. (2003). Estimation of Entropy and Mutual Information. *Neural Computation*, *15*(6), 1191–1253. https://doi.org/10.1162/089976603321780272

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., et al. (2020). The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Scientific Data*, *7*(1), 225. https://doi.org/10.1038/s41597-020-0534-3

Peters-Lidard, C. D., Hossain, F., Leung, L. R., McDowell, N., Rodell, M., Tapiador, F. J., et al. (2018). 100 Years of Progress in Hydrology. *Meteorological Monographs*, *59*(1), 25.1-25.51. https://doi.org/10.1175/AMSMONOGRAPHS-D-18-0019.1

Ramadhan, A., Marshall, J., Souza, A., Wagner, G. L., Ponnapati, M., & Rackauckas, C. (2020). Capturing missing physics in climate model parameterizations using neural differential equations. *ArXiv:2010.12559 [Physics]*. Retrieved from http://arxiv.org/abs/2010.12559

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689. https://doi.org/10.1073/pnas.1810286115

Renner, M., Brenner, C., Mallick, K., Wizemann, H.-D., Conte, L., Trebs, I., et al. (2019). Using phase lags to evaluate model biases in simulating the diurnal cycle of evapotranspiration: a case study in Luxembourg. *Hydrology and Earth System Sciences*, *23*(1), 515–535. https://doi.org/10.5194/hess-23-515-2019

Renner, M., Kleidon, A., Clark, M., Nijssen, B., Heidkamp, M., Best, M., & Abramowitz, G. (n.d.). How well can land-surface models represent the diurnal cycle of turbulent heat fluxes? *Journal of Hydrometeorology*, 1–56. https://doi.org/10.1175/JHM-D-20-0034.1

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *ArXiv:1602.04938 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1602.04938

Ritter, A., & Muñoz-Carpena, R. (2013). Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *Journal of Hydrology*, *480*, 33–45. https://doi.org/10.1016/j.jhydrol.2012.12.004

Ruddell, B. L., & Kumar, P. (2009a). Ecohydrologic process networks: 1. Identification. *Water Resources Research*, *45*(3), 1–23. https://doi.org/10.1029/2008WR007279

Ruddell, B. L., & Kumar, P. (2009b). Ecohydrologic process networks: 2. Analysis and characterization. *Water Resources Research*, *45*(3), 1–14. https://doi.org/10.1029/2008WR007280

Ruddell, B. L., Drewry, D. T., & Nearing, G. S. (2019). Information Theory for Model Diagnostics: Structural Error is Indicated by Trade-Off Between Functional and Predictive Performance. *Water Resources Research*, *55*(8), 6534–6554. https://doi.org/10.1029/2018WR023692

Runge, J., Petoukhov, V., Donges, J. F., Hlinka, J., Jajcay, N., Vejmelka, M., et al. (2015). Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature Communications*, *6*, 1–10. https://doi.org/10.1038/ncomms9502

Runge, J., Sejdinovic, D., & Flaxman, S. (2017). Detecting causal associations in large nonlinear time series datasets. https://doi.org/arXiv:1702.07007

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.). (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6

Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters*, *85*(2), 461–464. https://doi.org/10.1103/PhysRevLett.85.461

Sellers, A., Yang, Z., & Dickinson, R. (1991). The Project for Intercomparison of Land-surface Parameterization Schemes, 1335–1350.

Sellers, A., Pitman, A., Love, P., Irannejad, P., & Chen, T. (1993). The Project for Intercomparison of Land Surface Parameterization Schemes (PILPS): Phases 2 and 3*, (1992), 489–504.

Sendrowski, A., & Passalacqua, P. (2017). Process connectivity in a naturally prograding river delta. *Water Resources Research*, *53*, 1–23. https://doi.org/10.1002/2016WR020339.Received

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(July 1928), 379–423. https://doi.org/10.1145/584091.584093

Shen, C. (2018). A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. *Water Resources Research*, *54*(11), 8558–8593. https://doi.org/10.1029/2018WR022643

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *ArXiv:1312.6034 [Cs]*. Retrieved from http://arxiv.org/abs/1312.6034

Singh, V. P., & Guo, H. (1995). Parameter estimation for 3-parameter generalized pareto distribution by the principle of maximum entropy (POME). *Hydrological Sciences Journal*, *40*(2), 165–181. https://doi.org/10.1080/02626669509491402

Sivapalan, M., & Blöschl, G. (2017). The Growth of Hydrological Understanding: Technologies, Ideas, and Societal Needs Shape the Field. *Water Resources Research*, *53*(10), 8137–8146. https://doi.org/10.1002/2017WR021396

Sivapalan, M., Blöschl, G., Zhang, L., & Vertessy, R. (2003). Downward approach to hydrological prediction. *Hydrological Processes*, *17*(11), 2101–2111. https://doi.org/10.1002/hyp.1425

Smirnov, D. A. (2013). Spurious causalities with transfer entropy. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *87*(4), 1–12. https://doi.org/10.1103/PhysRevE.87.042917

Smith, M., Koren, V., Zhang, Z., Moreda, F., Cui, Z., Cosgrove, B., et al. (2013). The distributed model intercomparison project - Phase 2: Experiment design and summary results of the western basin experiments. *Journal of Hydrology*, *507*, 300–329. https://doi.org/10.1016/j.jhydrol.2013.08.040

Sun, J., Cafaro, C., & Bollt, E. M. (2014). Identifying the coupling structure in complex systems through the optimal causation entropy principle. *Entropy*, *16*(6), 3416–3433. https://doi.org/10.3390/e16063416

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *ArXiv:1312.6199 [Cs]*. Retrieved from http://arxiv.org/abs/1312.6199

Thornton, P. E., & Running, S. W. (1999). An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity, and precipitation. *Agricultural and Forest Meteorology*, *93*(4), 211–228. https://doi.org/10.1016/S0168-1923(98)00126-9

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004723

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability. *Journal of Advances in Modeling Earth Systems*, *12*(9). https://doi.org/10.1029/2019MS002002

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., et al. (2016). Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences*, *13*(14), 4291–4313. https://doi.org/10.5194/bg-13-4291-2016

Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67.

Vlachos, I., & Kugiumtzis, D. (2010). Non-uniform state space reconstruction and coupling detection. https://doi.org/10.1103/PhysRevE.82.016207

Wagener, T., Sivapalan, M., Troch, P., & Woods, R. (2007). Catchment Classification and Hydrologic Similarity. *Geography Compass*, *1*(4), 901–931. https://doi.org/10.1111/j.1749-8198.2007.00039.x

Weijs, S. V., Schoups, G., & Van De Giesen, N. (2010). Why hydrological predictions should be evaluated using information theory. *Hydrology and Earth System Sciences*, *14*(12), 2545–2558. https://doi.org/10.5194/hess-14-2545-2010

Weijs, S. V., Foroozand, H., & Kumar, A. (2018). Dependency and Redundancy: How Information Theory Untangles Three Variable Interactions in Environmental Data. *Water Resources Research*, *54*(10), 7143–7148. https://doi.org/10.1029/2018WR022649

Weijs, Steven V., van Nooijen, R., & van de Giesen, N. (2010a). Kullback–Leibler Divergence as a Forecast Skill Score with Classic Reliability–Resolution–Uncertainty Decomposition. *Monthly Weather Review*, *138*(9), 3387–3399. https://doi.org/10.1175/2010MWR3229.1

Weijs, Steven V., van Nooijen, R., & van de Giesen, N. (2010b). Kullback–Leibler Divergence as a Forecast Skill Score with Classic Reliability–Resolution–Uncertainty Decomposition. *Monthly Weather Review*, *138*(9), 3387–3399. https://doi.org/10.1175/2010MWR3229.1

Wilson, K., Goldstein, A., Falge, E., Aubinet, M., Baldocchi, D., Berbigier, P., et al. (2002). Energy balance closure at FLUXNET sites. *Agricultural and Forest Meteorology*, *113*(1), 223–243. https://doi.org/10.1016/S0168-1923(02)00109-0

Wood, E. F., Roundy, J. K., Troy, T. J., van Beek, L., Bierkens, M. F. P., Blyth, E., et al. (2012). hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research*, *48*(1), 1–10. https://doi.org/10.1029/2010WR010090

Xingyuan Chen, Peishi Jiang, Justine E.C. Missik, Zhongming Gao, Brittany Verbeke, & Heping Liu. (2020). Opening the black box of LSTM models using XAI. Presented at the American Geophysical Union Fall Meeting, Virtual: American Geophysical Union.

APPENDIX A

Decision Name	Decision Value	Decision Description	Value Description
soilCatTbl	ROSETTA	Soil category dataset	Merged Rosetta table with STASRUC
vegeParTbl	MODIFIED_IGBP_MO	Vegetation category dataset	MODIS 20-category dataset
soilStress	NoahType	Soil moisture control on stomatal resistance	Threshold linear function of volumetric liquid water content
stomResist	BallBerry	Function for stomatal resistance	Ball-Berry (Ball et al., 1987)
num_method	itertive	Numerical method	Iterative solver
fDerivMeth	analytic	Method used to calculate flux derivatives	Analytic derivatives
LAI_method	specified	Method to determine LAI and SAI	LAI/SAI computed from green vegetation fraction
f_Richards	mixdform	Form of Richards' equation	Mixed form
groundwatr	qTopmodl	Groundwater parameterization	TOPMODEL parameterization (Beven & Freer, 2001)
hc_profile	pow_prof	Hydraulic conductivity profile	Power-law profile
bcUpprTdyn	nrg_flux	Upper boundary condition for thermodynamics	Energy flux
bcLowrTdyn	zeroFlux	Lower boundary condition for thermodynamics	Zero flux
bcUpprSoiH	liq_flux	Upper boundary condition for soil hydrology	Liquid water flux
bcLowrSoiH	zeroFlux	Lower boundary condition for soil hydrology	Zero flux

veg_traits	CM_QJRMS1988	Parameterization for vegetation roughness length and displacement heights	(Choudhury & Monteith, 1988)
canopyEmis	difTrans	Parameterization of canopy emissivity	Parameterized as a function of diffuse transmissivity
snowIncept	lightSnow	Parameterization for snow interception	Maximum interception capacity an inverse function of new snow density
windPrfile	logBelowCanopy	Wind profile through the canopy	Logarithmic profile below the vegetation canopy
astability	louisinv	Stability function	Inverse power function (Louis,
			1979)
canopySrad	BeersLaw	Canopy shortwave radiation	Beer's Law (as implemented in VIC)
alb_method	varDecay	Albedo representation	Variable decay rate (Dickinson et al., 1993)
compaction	anderson	Snow compaction algorithm	Semi-empirical method (Anderson, 1976)
snowLayers	CLM_2010	How to combine and divide snow layers	CLM type: rules depend on layer index
thCondSnow	jrdn1991	Type of thermal conductivity in snow	(Jordan, 1991)
thCondSoil	funcSoilWet	Type of thermal conductivity in soil	Function of soil wetness
spatial_gw	localColumn	Spatial representation of groundwater	Separate groundwater representation in each local soil column
subRouting	timeDlay	Method for sub-grid routing	Time-delay histogram

Table A1. Summary of modeling decisions used in the SUMMA instance.

APPENDIX B

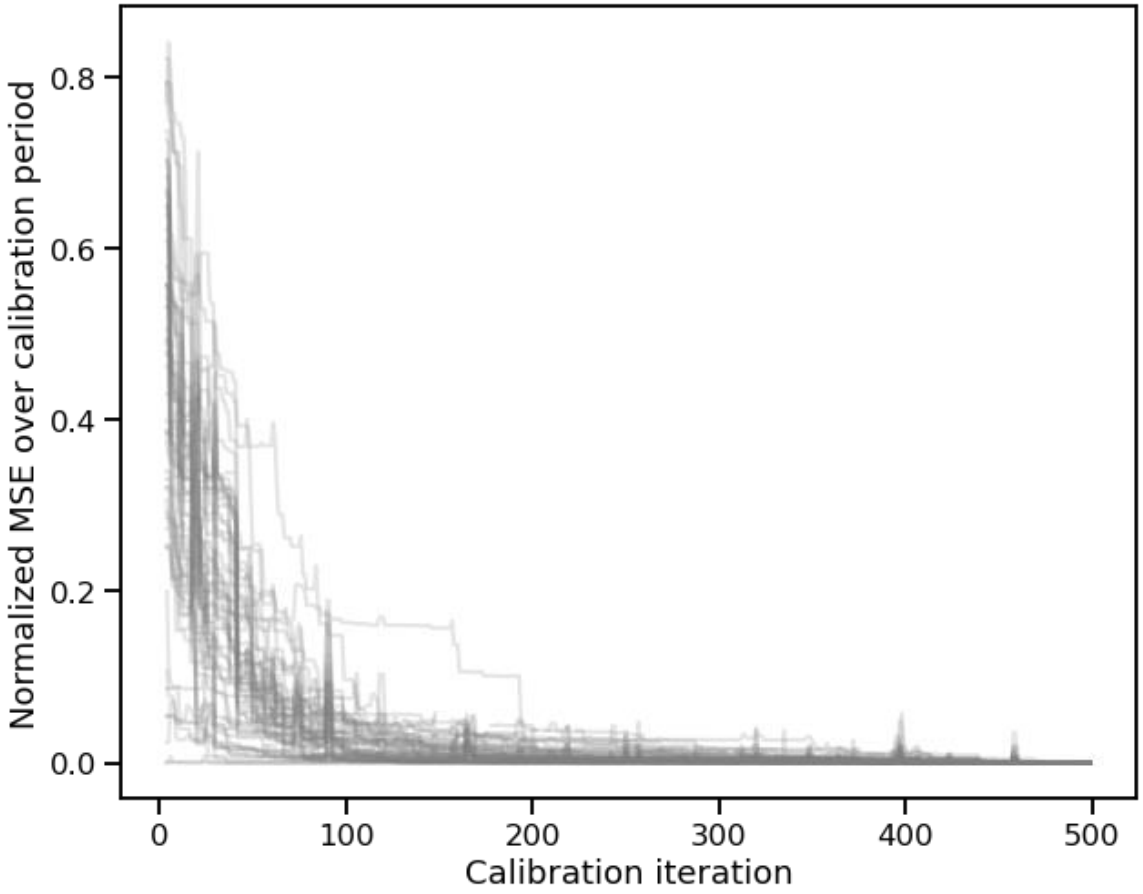

Figure B1. Normalized MSE over calibration iterations. Each curve represents reduction in MSE over the calibration period for a single site. Each site is normalized separately (such that 0.0 is the minimum MSE achieved for a site), and the 5-iteration rolling minimum is taken to reduce the noise between iterations.

Parameter Name	Description	Initial value	Lower bound	Upper bound
vcmax_Kn	Control on maximum carboxylation rate for photosynthesis	0.6	0.1	1.2
laiScaleParam	Scale parameter for LAI values, from IGBP tables	1.0	0.5	3.0
rootingDepth	Deepest depth of rooting zone	Determined by vegetation type (Zeng, 2001)	0.5 * initial	1.5 * initial
canopywettingfactor	Fraction of precipitation captured by vegetation	0.7	0.01	0.9
kAnisotropic	anisotropy factor for lateral hydraulic conductivity	1.0	0.5	5.0
theta_res	Residual soil moisture content	Determined by soil type (source)	0.001	0.2

theta_sat	Saturation soil moisture content	Determined by soil type (source)	0.3	0.7
fieldCapacity	Equilibrium soil moisture content after drainage and ET	(theta_sat + theta_res) / 2	theta_res	theta_sat
critSoilTranspire	Soil moisture content level at which transpiration becomes lmited	(theta_sat + theta_res) / 2	theta_res	theta_sat
critSoilWilting	Soil moisture content level at which transpiration stops	(theta_sat + critSoilTranspire) / 2	theta_res	critSoilTranspire

Table B1. Listing of parameters used for calibration of standalone simulations.