Explainable AI uncovers how neural networks learn to regionalize in simulations of turbulent heat fluxes at FluxNet sites

Enter authors here: Andrew Bennett[1], Bart Nijssen[1]

[1]Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, USA

Corresponding author: Andrew Bennett (andrbenn@uw.edu)

Key Points:

- Explainable AI methods neural network behaviors learned to extract information from input data in a physically plausible way.
- The neural networks learned different behaviors at arid and non-arid sites, without aridity information in the training data.
- Linear decompositions of the neural networks uncovered how such models learn to regionalize.

Abstract

Machine learning (ML) based models have demonstrated very strong predictive capabilities for hydrologic modeling, but are often criticized for being black-boxes. In this paper we use a technique from the field of explainable AI (XAI), called layerwise relevance propagation (LRP) to "open the black box". Specifically we train a deep neural network on data from a set of hydroclimatically diverse FluxNet sites to predict turbulent heat fluxes, and then use the LRP technique to analyze what it learned. We show that the neural network learns physically plausible relationships, including different ways of partitioning the turbulent heat fluxes according to moisture or energy limiting characteristics of the sites. That is, the neural network learns different behaviors at arid and non-arid sites. We also develop and demonstrate a novel technique that uses the output of the LRP analysis to explore how the neural network learned to regionalize between sites. We find that the neural network primarily learned behaviors that differed between evergreen forested sites and all other vegetation classes. Our analysis shows that even simple neural networks can extract physically-plausible relationships and that by using XAI methods we can learn new information from the ML-based methods.

Plain Language Summary

Machine learning (ML) techniques have been shown to make very good predictions for hydrology, but it is difficult to understand why they make good predictions, when they might fail, or what they have learned. A new field of techniques known as explainable artificial intelligence (or XAI) attempts to make ML models more understandable and tractable. We use these techniques to analyze an ML model of evaporation and conductive heat transfer. We find that the ML model learns relationships which agree with physical understanding. Further, we show that they are able to distinguish between arid and non-arid sites, even though they are not provided with this classification up front. Finally, we show how to use XAI to examine how the ML model learned intersite behavior. In doing so, we find that the ML model learns different behaviors at evergreen forest sites than all other site types.

1 Introduction

The hydrologic sciences have a long history of using 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 "predictive performance" versus "explainability" (Lipton, 2017). With the recent surge in interest in using ML methods for hydrologic modeling as well as continuing advances in both process-based and data-driven models this debate continues. In this paper we hint that data-driven models may be used to refine theoretical underpinnings and improve hydrologic understanding. Specifically, we focus on a class of ML based models from the field of deep learning (DL), which generally are considered models with multiple hidden layers. We build on previous work that showed that DL parameterizations can be used directly in process-based models to represent individual processes, and improve their predictions. In this study, we show how our DL parameterizations identify physically relevant predictor variables in a way that coincides with physical understanding and 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 DL in geosciences to focus exclusively on model output. 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 increasingly clear that DL techniques can be used as tools for interpretation as well as 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 the use of XAI methods 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) provide an overview and taxonomy of these methods. They distinguish 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" ML approach and focused on predictive performance to train a DL parameterization for the prediction of turbulent heat fluxes. We then embedded this DL parameterization directly into a process-based hydrologic model (PBHM). We demonstrated that DL-based models that are trained out-of-sample are able to outperform locally-calibrated 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 providing the DL parameterization with updated soil moisture information from the PBHM 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 DL parameterization. Our experiments hinted that the improvements in performance were due to the DL model's ability to find physical relationships between input and output that had not been encoded explicitly in the physics-based models and 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 understanding of how turbulent heat fluxes are generated. We show that the network also learned 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, nor was any information about the long term aridity of each site.

We also show how the LRP method can be used to understand what the network has learned between sites. Transfer of hydrologic understanding between sites, whether in the context of prediction at ungauged sites or parameter regionalization, remains one of the fundamental problems in hydrology (Blöschl et al., 2019; Hrachowitz et al., 2013). 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 whether this is the case in our model we also explore how our DL parameterization 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 DL parameterizations 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 show how the LRP technique can be extended by using it to develop linear approximations of the neural network at each site. We use these linear approximations to analyze how the neural network generalized between locations. We find that the neural network primarily learns different behaviors at evergreen forested sites than at all other site types. Based on our analysis we believe that this new technique that uses LRP decompositions is a very promising analysis tool for understanding how to extract understanding from DL models.

2 Materials and Methods

2.1 Data and study sites

As in Bennett & Nijssen (2020), we analyzed 60 FluxNet sites (Pastorello et al., 2020) where data quality was robust enough and with a sufficient record length for a PBHM to be run. We required at least 3 years of half hourly data with at most 15% of the entire record missing. Missing data was gap-filled by the FluxNet teams with ERA-interim data that has been bias-corrected and downscaled. This resulted in 509 site-years worth of half hourly data. Figure 1 shows the locations and vegetation types of each of the sites.

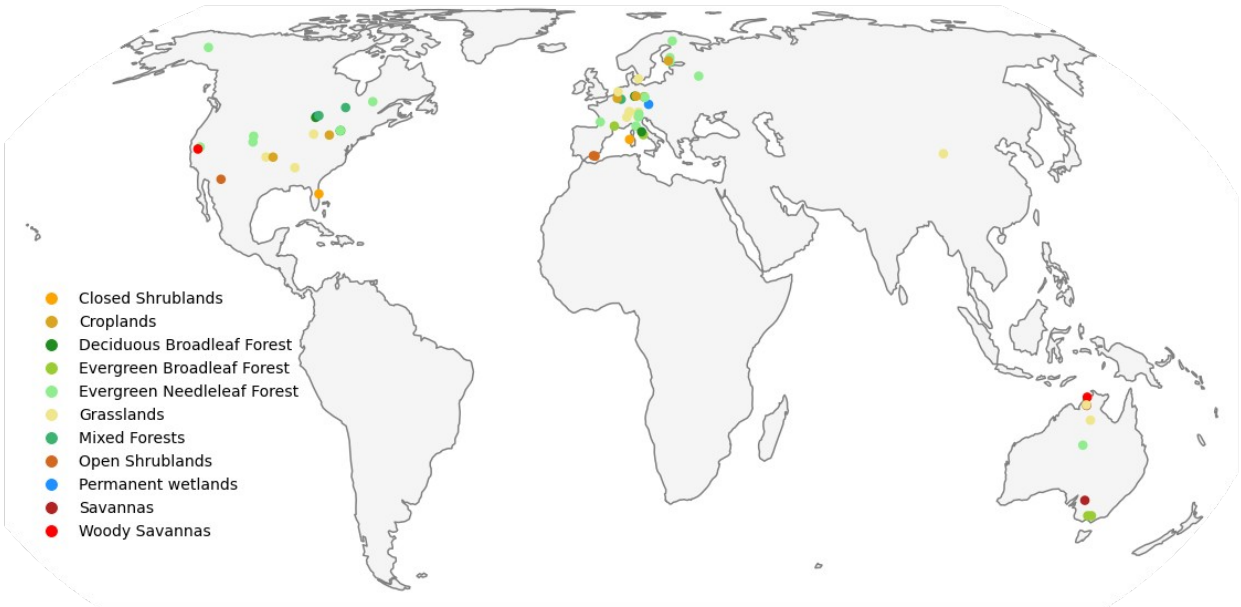


Figure 1 A map of the FluxNet sites used in the analysis, coded by IGBP land cover classification.

2.2 Coupled deep learning parameterization

To predict turbulent heat fluxes we use a deep dense neural network, also known as a multilayer perceptron architecture or a deep feedforward network (Goodfellow et al., 2016). 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 allowed us to use model-derived states as input to the neural network, both during training and during execution of the coupled model. The use of SUMMA as our PBHM allowed us to maintain the mass and energy balances while exploiting the flexibility and predictive capabilities of neural networks. Coupling the DL parameterization into SUMMA 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 be executed by 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 can more easily disentangle the impact of the input variables on the predicted turbulent heat fluxes.

We trained a 2 layer neural network with each layer consisting of 28 nodes with hyperbolic tangent activations. At each layer we incorporate dropout regularization (with dropout rate of 0.1). We used the mean squared error between predicted and observed heat fluxes at a half-hourly interval as our loss function. The neural network was optimized using the Adam method, which automatically tunes the learning rate and has been shown to work well in many settings (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.

The neural network we trained takes air temperature, relative humidity, shortwave radiation, soil moisture content, leaf area index (LAI) multiplied by the height of the vegetation canopy, and International Geosphere-Biosphere Programme (IGBP) land cover class (Loveland et al., 2000) 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 soil moisture to facilitate interpretation. Each input represents a single timestep at the half-hourly timescale and does not include any other temporal information. We refer to this neural network configuration as NNLRP throughout the remainder of this paper.

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). Bach et al. (2015) and Montavon et al. (2017) explain the original method in greater detail. For clarity we provide a high level description of the LRP algorithm.

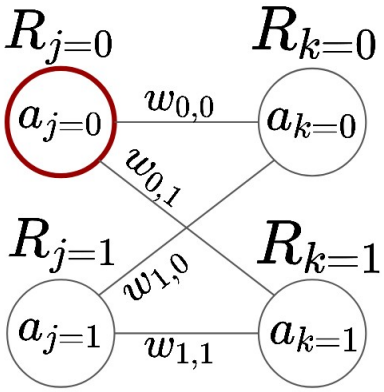
Intuitively, LRP works by taking advantage of the ability to backpropagate information from the outputs to the inputs of a neural network. Following training, neural networks can be used to make predictions using the forward pass. LRP uses the predictions made during the forward pass, along with a "rule"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

A number of functional relationships, referred to as rules, can be used to compute and backpropagate relevance, each with different purposes, interpretations, and theoretical properties. For a review of some of the most commonly used rules see Samek et al. (2019). Mamalakis et al. (2021) compared several of these rules and found that the "Z rule" for propagating relevance was best suited for applications in the geosciences. In this study we use the Epsilon rule (Equation 1), which is the same as the Z rule, but is more numerically robust when the denominator inside of the sum is small. The Epsilon rule propagates relevance according to the rule:

$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_j a_j w_{jk}} R_k \tag{1}$$

where the j, k subscripts 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. ϵ is a tunable parameter which is introduced to "absorb" some of the relevance when the sum of all of the contributions of the weights to the relevance in the denominator is small ($\sum_j a_j w_{jk}$in the denominator of Equation 1). For all relevance scores reported in this study we use $\epsilon = 0.001$. A schematic and example of how the relevance scores are calculated is shown in figure 2.

a) Simplified network architecture



b) Example relevance score calculation

$$R_{j=0} = \left[\frac{a_0 w_{0,0} R_{k=0}}{\epsilon + (a_0 w_{0,0} + a_1 w_{0,1})}\right] + \left[\frac{a_0 w_{0,1} R_{k=1}}{\epsilon + (a_0 w_{0,1} + a_1 w_{1,1})}\right]$$

Figure 2. Schematic of the relevance score calculation. A simplified network architecture with 2 layers, each of 2 nodes, is shown in panel a, with the node where j=0 outlined in red. Panel b shows the calculation of the relevance score the j=0 node in panel a.

The relevance score is approximately proportional to the derivative of the flux with respect to an input variable. We demonstrate this in figure S2 of the supporting information. Considering this interpretation of the relevance score as a sensitivity shows that a variable can be considered a "producer" of the flux when the relevance score is positive, and an "inhibitor" of the flux when the score is negative.

2.4 Using LRP to disentangle site similarity

One of the surprising findings of Bennett & Nijssen (2020) was that the DL based approaches outperformed the process-based model at sites where the DL models were not trained. This indicated that the neural network learned inter-site generalizations that were not encoded in the PBHM. We extended our use of LRP to better understand of how the NN learned to generalize between sites. We did 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. We used this decomposition into a set of local sensitivities of the inputs and flux responses of the outputs to build a linear model for each site.

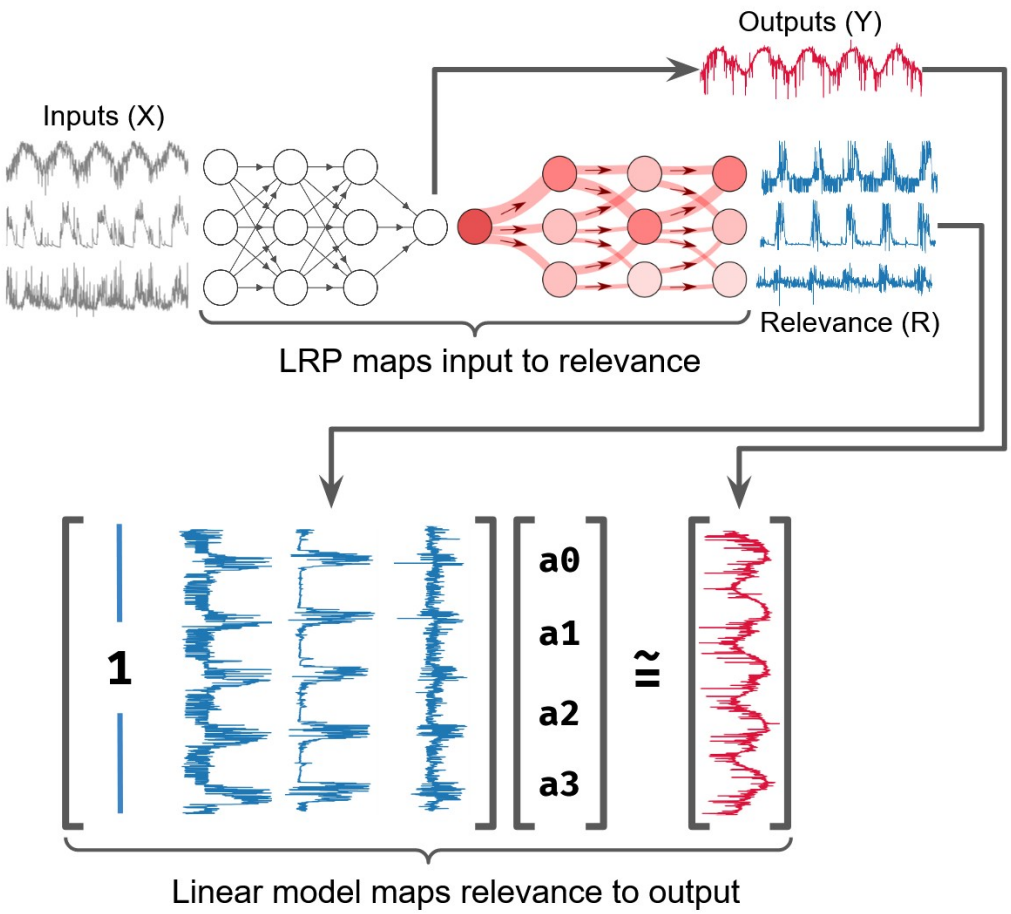


Figure 3. A schematic of how we build site-specific linear decompositions of NNLRP from the relevance timeseries. We first use LRP to produce timeseries of relevance scores for each of the input variables to NNLRP. These timeseries are then fit via linear regression against the turbulent heat fluxes that NNLRP produces as output. The resulting weights of the linear regression, shown in the schematic as a0 through a3 are site specific regression coefficients, though the system is larger in reality. See section 2.2 for descriptions of the inputs to NNLRP.

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 differ slightly from 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 independent variables are derived from the trained neural network.

We built these linear models by performing a multivariate linear regression at each site where the predictor variables are the set of relevance scores for each of the neural network inputs and the target variable is a turbulent heat flux (Figure 3). We found that this linearized model can almost exactly reproduce the relationship between the relevance scores and heat fluxes. This is shown in figure S3 of the supporting information.

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 a regression at one site and applying it to another we quantified the inter-site learning by the neural network. This allowed us to build graphs of site interactions which yield insight into the nature of variability of turbulent heat fluxes across sites.

3 Results

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 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 half-hourly model interval based on both SUMMA inputs as well as dynamically-updated SUMMA soil moisture. Bennett & Nijssen (2020) demonstrated good performance for NN2W coupled into SUMMA. 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 to aid interpretability. The network was reduced from approximately 13,000 parameters (NN2W) to roughly 1000 parameters (NNLRP). We also reduced the number of input features. However, it is promising that NNLRP obtained performance which continued to exceed that of SA. NNLRP performance relative to NN2W showed a 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 adding more neurons or layers and thus increasing model capacity. 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.

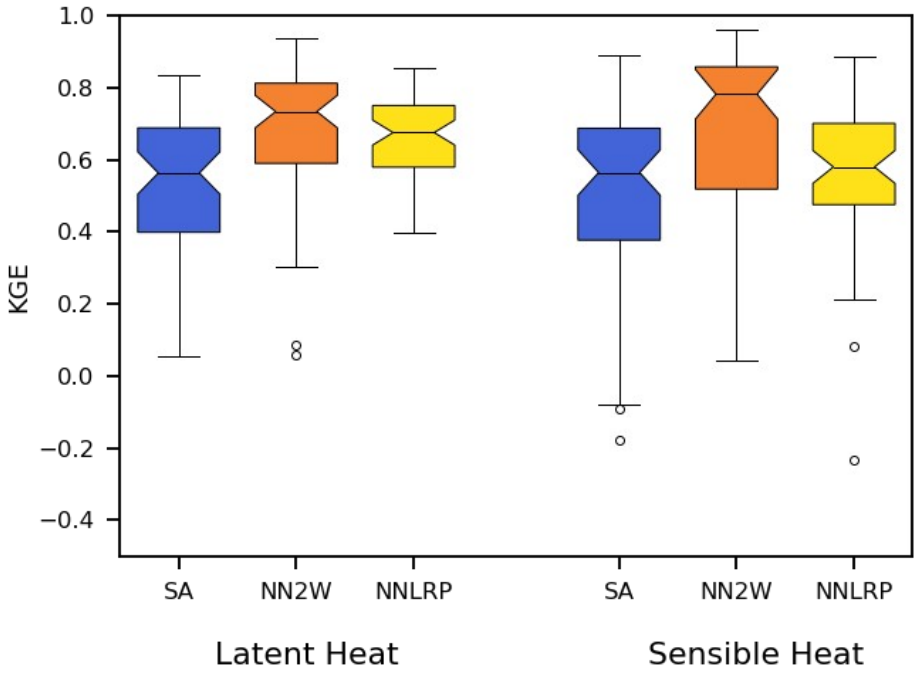


Figure 4. A comparison of the KGE performance of the neural network used in our analysis (NNLRP) against the SA and NN2W models reported in Bennett & Nijssen (2020). KGE scores were calculated based on observations of the half-hourly turbulent heat fluxes at the FluxNet sites.

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 5 and 6 show these timeseries for both an energy limited (CH-Fru, figure 5) and moisture limited (US-Whs, figure 6) 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 omitted the timeseries of LAI and vegetation relevance for simplicity.

At CH-Fru, we see large (in absolute value) relevance scores for latent heat from the air temperature and shortwave radiation (Figure 5). 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 negative correlations hint that the network learned strategies for partitioning between heat fluxes, which is surprising. 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.

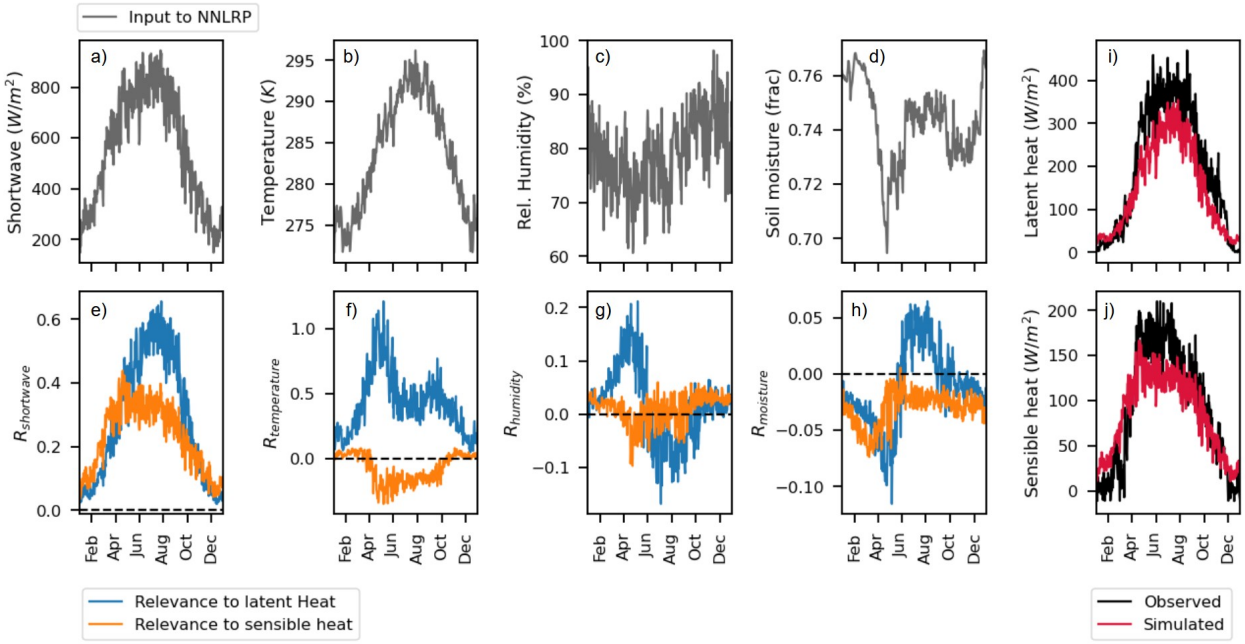


Figure 5 Timeseries for meteorological conditions and LRP-derived relevance values at CH-Fru. Subplots a-d show the observed forcings used as input to the neural network, while subplots e-h show the relevance timeseries for latent (blue) and sensible (orange) heat with respect to each of the input variables. Subplots i and j show the observed and simulated heat fluxes.

At US-Whs (Figure 6), we see some similar relationships. Air temperature is most relevant for latent heat, while 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 6d and indicates that the network learned when the site was moisture limited.

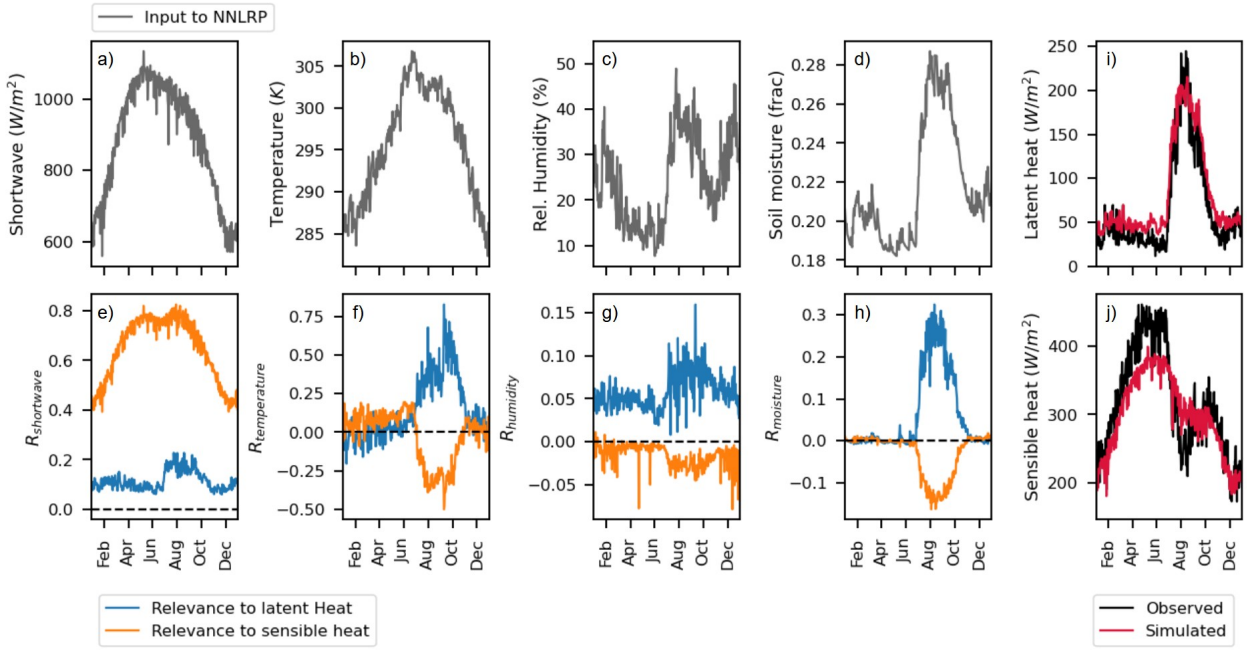


Figure 6. Timeseries for meteorological conditions and LRP-derived relevance values US-Whs.

We show the average (normalized) relevance scores of all of the model inputs in Figure 7, 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 of 1. The general ranking of relevance scores for both latent (Figure 7a) and sensible heat (Figure 7b) is stable across sites, particularly the primary importance of air temperature for latent heat and shortwave radiation for sensible heat.

It seems that NNLRP uses vegetation types to partition the latent and sensible heat fluxes differently in different ecosystems. The need to include vegetation types to maintain performance (as discussed in section 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 3.3.

Figure 7a 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 and relative humidity "inhibit" latent heat fluxes. Generally the positive relevance scores are much larger than the negative relevance scores, indicating that the network is more sensitive to changes that increase the predicted latent heat than changes that decrease it. On the other hand, only shortwave and relative humidity have consistently positive relevance scores for sensible heat fluxes (figure 7b). Air temperature, soil moisture, LAI, and vegetation type have consistently negative relevance scores.

The apportionment of relevance across sites for sensible heat (Figure 7b) shows more variation than that of latent heat (Figure 7a). This is largely due to the contributions of relative humidity and vegetation types. Because a vegetation type is site-specific and static through time, it is hard to disentangle it from the other variables which are temporally varying. We will

analyze the site-specific behavior further in section 3.3. An interesting feature of Figure 7 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 (energy-limited versus moisture-limited).

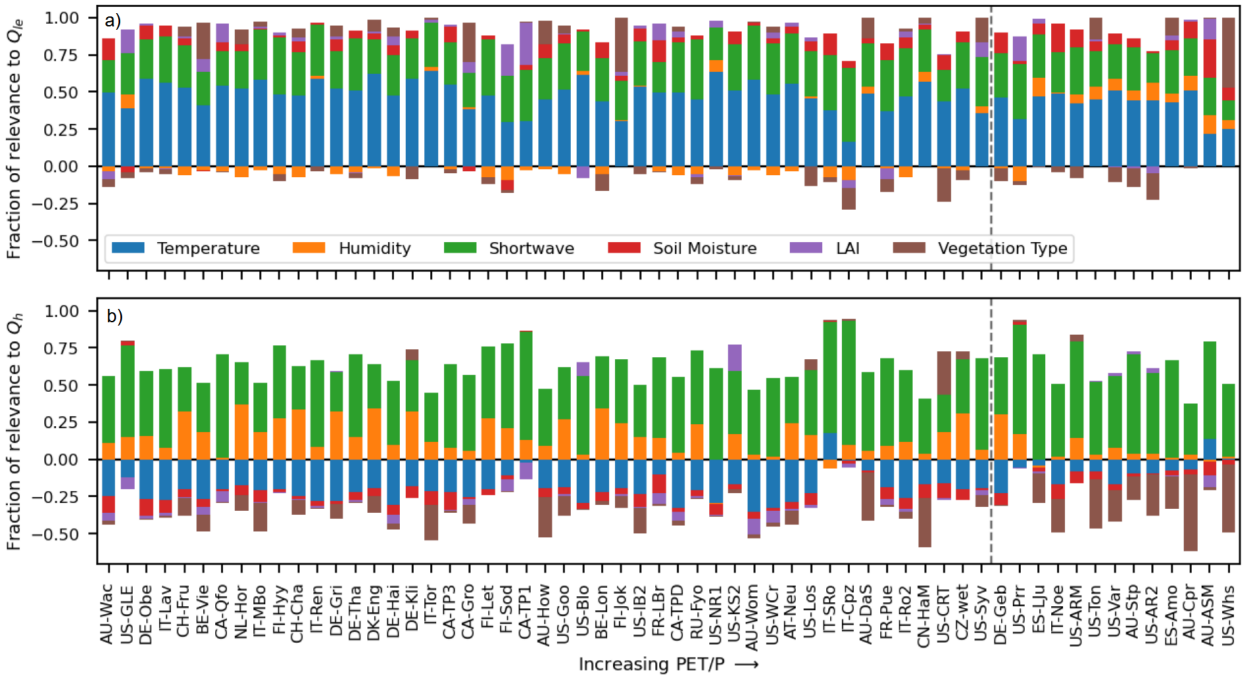


Figure 7. Average fraction of relevance by input variable. Panel a shows the relevance breakdown for latent heat, while panel b shows the breakdown for sensible heat. Sites are sorted by increasing PET/P. The dashed line shows the threshold of PET/P = 1, with energy-limited sites to the left and moisture-limited sites to the right.

Figure 6 shows breakdowns which are site specific, but we can also compare individual components across sites. For instance, 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 two examples of this in Figure 8. In Figure 8a we compute 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.

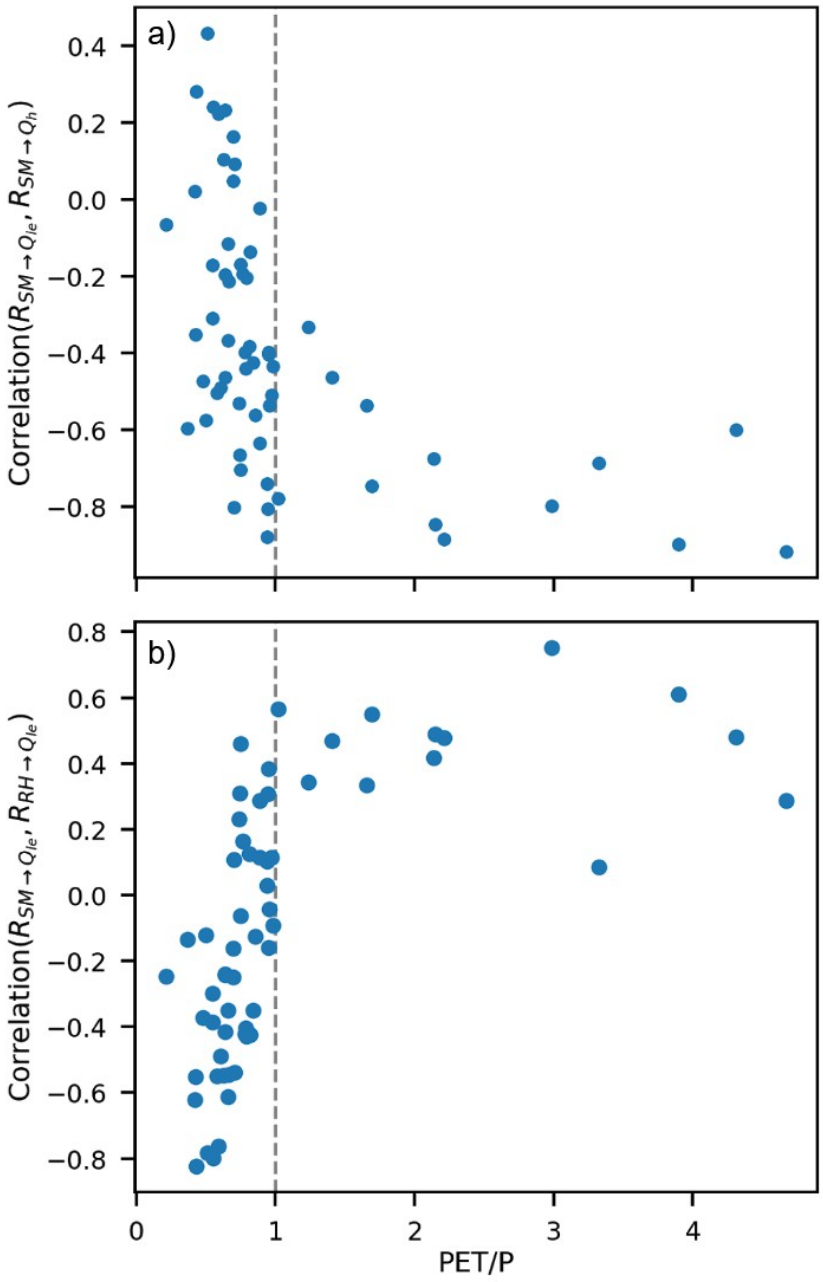


Figure 8. NNLRP exhibits different behavior in energy and moisture limited sites. Panel a shows the correlation between the relevance between latent and sensible heat with respect to soil moisture, indicating NNLRP learned to partition between turbulent heat fluxes based on soil moisture availability. Panel b shows the correlation between the relevance from soil moisture to latent heat and the relevance from relative humidity and latent heat, indicating that NNLRP learned different physical relationships at sites of varying aridity (PET/P).

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 8b, but instead computed the correlation between the relevance of soil moisture to latent heat and the relevance of relative humidity to latent heat as shown in Figure 8a. 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.

At energy limited sites (PET/P << 1) when humidity is a strong control on evaporation (high relevance from humidity to latent heat) the atmospheric demand for more moisture would be low. If atmospheric demand is low, then it does not matter how much moisture is in the soil, resulting in low relevance from soil moisture to latent heat. On the other hand at arid sites (PET/P >> 1) when there is enough soil moisture to evaporate, which coincides with higher amounts of relative humidity. The relationships between the relevance timeseries at different sites hint at how NNLRP was able to learn about long-term behaviors from the short-term input data based on the covariances presented in the data. This indicates that NNLRP was able to learn some physical relationships which we did not encode or provide as input.

3.3 Using LRP to decompose inter-site predictions

Thus far, we have only discussed general properties of NNLRP. As we outlined in section 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 by fitting a linear model that uses the relevance scores as inputs to determine the turbulent heat fluxes at the half-hourly time scale. We then compared this fit to the full timeseries of turbulent heat fluxes simulated by the neural network. We found that the linear models were able to achieve KGE values larger than 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. A figure showing the KGE values (evaluated against the output of NNLRP at the half-hourly interval) for each site is shown in Figure S2 in the supporting information.

We used these linear approximations to quantify the similarity of representations learned by NNRLP by clustering the regression coefficients using agglomorative clustering (Day & Edelsbrunner, 1984), resulting in the dendrogram shown in Figure 9. The hierarchical clustering in figure 9 shows two main groupings. The green cluster is comprised of all of the evergreen needleleaf forest sites, and a single evergreen broadleaf site (AU-Wac). The purple cluster contains all other sites. We examined the regression coefficients between the two clusters and found that the main difference between the two groups was the coefficient for the vegetation type. This clustering indicates that NNLRP learned specific behaviors for the evergreen forested sites and non-evergreen-forested sites. Further, based on the height of the green cluster in Figure 9 NNLRP learned a more diverse set of behaviors for the green cluster than for the purple cluster. This also indicates that the sites in the green cluster are more unique than those in the purple cluster.

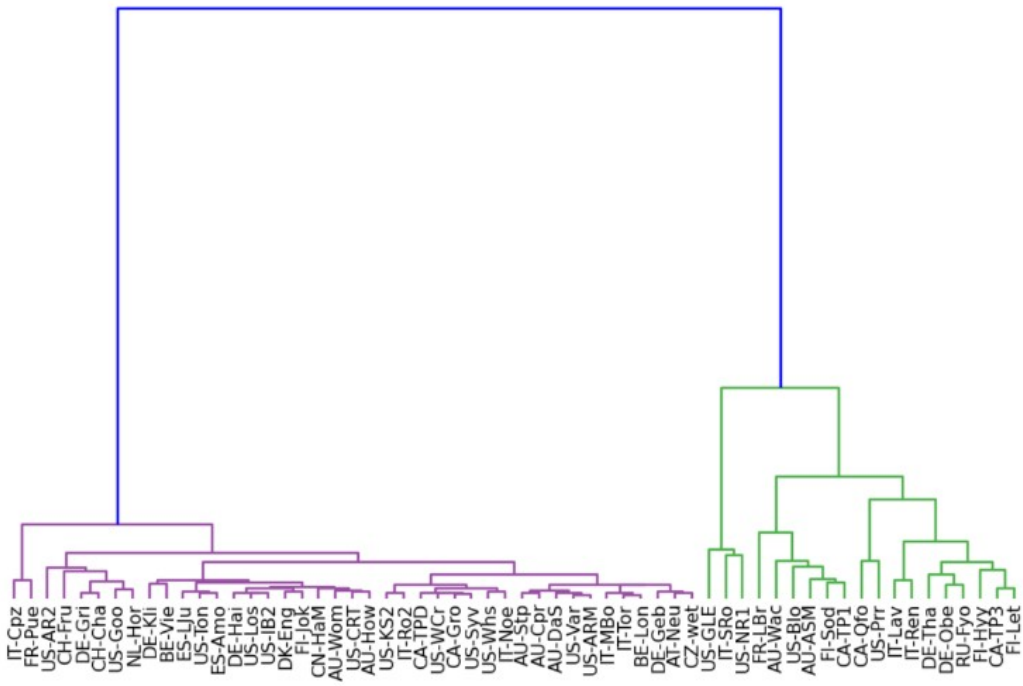


Figure 9. Dendrogram of site clusters based on the regression coefficients based on the methodology described in section 2.4. The height between each branch represents the distance between successive clusters. The purple and green clusters were selected because they were the "level" of clustering with the greatest distance between them.

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 a "source" site, then applied it to a "target" site. We then calculated 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 computed this explainability score for all site pairs, resulting in a matrix of scores, which can be thought of as a weighted-directed graph. We also pruned 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 (randomly sampled) at the site with the shortest record (at site CA-TPD, where n_T = 57552 half-hour timesteps). The full heatmaps of these inter-site explainability scores are shown in Figure S3 of the supporting information.

We aggregated these weighted-directed graphs to analyze which sites are difficult to predict and which sites are good predictors.We then performed two analyses. First, we examined for each site how well its locally-trained linear model performed at all other sites. Second, we examined for each site how well all remotely-trained models (models trained at other sites) performed locally. That is, in the first analysis we examined the performance when we used each site as a fixed "source" model for all other sites, and in the second analysis we examined the performance of each site as a "target". We used the modified KGE (denoted KGE_m) score, which is a normalized version of the KGE (Mathevet et al., 2006), and is calculated as

$$KGE_m = \frac{KGE}{2 - KGE}$$

We show the results of these analyses in Figure 10. The x-axis shows the site performance when its linear model was used as a predictor (source) and the y-axis shows the performance when a site is predicted using the linear models trained at other sites (target). The grey dashed lines at $KGE_m \cong -0.17$ represent the benchmark value when the model output is compared against the mean of the NNLRP model output (Knoben et al., 2019). This divides each panel of Figure 10 into four quadrants. Sites in the lower left quadrant were both bad predictors and were not predicted well by any other sites. None of our sites fell into this quadrant indicating that NNLRP always learned at least some generalizable behaviors. Sites which fell in the lower right quadrant were good predictors, but not able to be predicted. Only 2 sites fell within this quadrant in Figure 10a, NL-Hor and DE-Gri. We will speculate in section 4 as to why these sites fall in this quadrant. Sites in the upper left quadrant sites were bad at predicting other sites but able to be predicted well. This quadrant is entirely dominated by the evergreen needleaf and broadleaf forested sites, which form their own cluster that slightly crosses over to the upper right quadrant. These sites all showed much greater variability in their performance as a predictor, seen by the long right-ward tails in the interquartile range of Figure 10a. Finally, sites in the upper right quadrant of Figure 10a are sites which were both good at predicting other sites as well as at being predicted by other sites. These sites tended to have greater variability in their performance when being predicted by linear models from other sites, seen by the long downward tails in the interquartile range of Figure 10a. We see that, outside of the evergreen needle and broadleaf sites and the two outliers in the bottom right quadrant, the remainder of sites are tightly clustered into this quadrant. Based on this analysis we conclude that NNLRP learned a wide range of generalizable behaviors between sites, with some specific differences between evergreen needle and broadleaf sites and other land cover types.

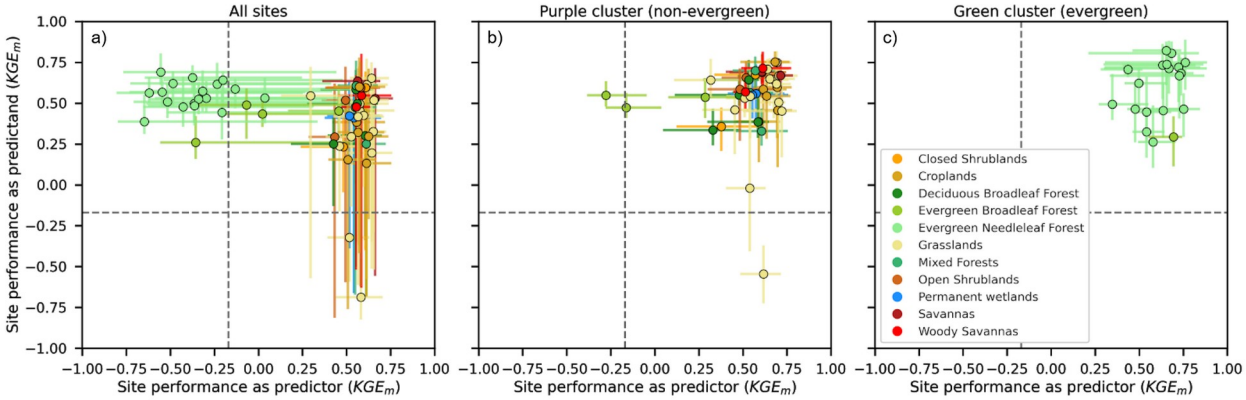


Figure 10. Scatter of site performance counts when using a linear model from one site applied at another. The x-axis shows the modified KGE when the linear model of a site is used to predict all other sties. The y-axis shows the modified KGE with linear models of all other sites are used to predict a single site. Dots show the median value across all sites, and the lines show the interquartile ranges. Dashed lines show the the threshold value of $KGE_m \cong -0.17$. Points above these thresholds indicate that they performed better than simply using the mean of the NNLRP output for that site. Panel a shows the results for the entire dataset. Panels b and c show the results when you restrict the analysis to only a single cluster from the clustering in figure 9.

In panels b and c of Figure 10 we separated the analysis into the two clusters found by the clustering in Figure 9. In both cases, separating the clusters reduces the long tails in the interquartile ranges. The majority of the sites in Figure 10b still fall into the upper-left quadrant, though the two outliers from before are still present, as well as two evergreen broadleaf sites (IT-Cpz and FR-Pue) which fall slightly into the upper left quadrant. In Figure 9 both of these sites are the most unique within the purple cluster, where they form the first branch of the purple cluster. In Figure 10c we see a large change in the performance of the green cluster from Figure 9. All sites fall entirely within the upper right quadrant, meaning they are all able to predict and be predicted reliably by each other. This, along with the clustering from Figure 9, shows that NNLRP learned two specific sets of behavior between the two clusters.

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 understanding and hydrologic theory. Further, we showed that the network uncovered constraints and learned how to partition turbulent heat fluxes in a physically plausible way. For instance, NNLRP 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. NNLRP was able to learn these partitionings only by looking at half-hourly data, with no explicit temporal memory, or estimates of PET or precipitation as inputs. This hints that neural networks are able to extract and learn about longer term site characteristics that are somehow implicitly encoded in the input data.

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 and/or easier to extract scientific knowledge from.

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. 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 (as shown in Figure S1 of the supporting information). 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. We found that NNLRP learned different behaviors between the evergreen needleleaf sites and sites with all other IGBP classifications. We showed how these linear decompositions of relevance demonstrated that these two clusters not only had different regressions, but were much better at within-cluster than between-cluster prediction, both as source and target sites. We found that the largest difference

between the regression coefficients of these two clusters was for the vegetation classification. We tried excluding vegetation type when we trained the NNLRP, but were unable to get good performance without it. It seems that NNLRP was able to learn a number of physically realistic behaviors, but still needed to use the simple static vegetation classification to learn site-specific behaviors. Inclusion of more physically meaningful quantities, such as stomatal or soil resistances, in the training data may allow neural networks to learn even more physically realistic behaviors, without the need for a static land use classifier. Inclusion of such terms might allow for better estimation and separation of transpiration and evaporation, which may improve the ability for neural networks to generalize even further.

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. For instance, we found that DE-Gri and NL-Hor, while good at predicting other sites, were difficult to predict given a linear model from another site. There may be several reasons for this behavior. If these sites exhibit a very diverse set of behaviors they would be able to be good predictors but not be easily predicted by sites which are less diverse. It may also be that these sites are subject to active water management, which may make them difficult to model.

It is important to make the distinction that our results are based on the simplest neural network available, a feedforward 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 to densely connected 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.

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 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

still lack methods to directly translate these discoveries into a new theory, it does indicate the possibility that we may be able to do so 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. Our new analysis based on these decompositions 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.

Acknowledgments, Samples, and Data

We would like to thank Yifan Cheng, Martyn Clark, Erkan Istanbulluoglu, and Grey Nearing for reading and commenting on early versions of this manuscript. Their comments improved the clarity and framing of our work. The code to process, configure, calibrate/train, run, and analyze the FluxNet data is available at https://doi.org/10.5281/zenodo.4706145. The SUMMA model configuration for NNLRP is available at https://doi.org/10.5281/zenodo.4706106. We would like to acknowledge high-performance computing support from Cheyenne (doi:10.5065/D6RX99HX) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation.

References

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

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. 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

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

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

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

Day, W., & Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, *1*(1), 7–24.

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

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

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)

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

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

Ian Goodfellow, Yoshua Bengio, & Aaron Courville. (2016). *Deep Learning*. MIT Press. Retrieved from https://mitpress.mit.edu/books/deep-learning

Kampf, S. K., & Burges, S. J. (2007). 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

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

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

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

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

Loveland, T. R., Reed, B. C., Brown, J. F., Ohlen, D. O., Zhu, Z., Yang, L., & Merchant, J. W. (2000). Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data. *International Journal of Remote Sensing*, *21*(6–7), 1303–1330. https://doi.org/10.1080/014311600210191

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

Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. A. (2021). Neural Network Attribution Methods for Problems in Geoscience: A Novel Synthetic Benchmark Dataset. *ArXiv:2103.10005 [Physics]*. Retrieved from http://arxiv.org/abs/2103.10005

Mathevet, T., Michel, C., Andréassian, V., & Perrin, C. (2006). A bounded version of the Nash-Sutcliffe criterion for better model assessment on large sets of basins. *IAHS Press*, 9.

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

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

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

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

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

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

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

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

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.