machine learning & knowledge extraction

MDPI

Article

A Hybrid Artificial Neural Network to Estimate Soil Moisture Using SWAT+ and SMAP Data

Katherine H. Breen [1,2,†], Scott C. James [1,3,*,†], Joseph D. White [4], Peter M. Allen [1] and Jeffery G. Arnold [5]

1 Department of Geosciences, Baylor University, Waco, TX 76798, USA; kathy_breen@baylor.edu (K.H.B.); peter_allen@baylor.edu (P.M.A.)
2 Goddard Space Flight Center, NASA, Greenbelt, MD 20771, USA
3 Department of Mechanical Engineering, Baylor University, Waco, TX 76798, USA
4 Department of Biology, Baylor University, Waco, TX 76798, USA; joseph_d_white@baylor.edu
5 USDA-Agricultural Research Service, Temple, TX 76502, USA; jeff.arnold@ars.usda.gov
* Correspondence: SC_James@Baylor.edu
† These authors contributed equally to this work.

Received: 29 June 2020; Accepted: 12 August 2020; Published: 21 August 2020

Abstract: In this work, we developed a data-driven framework to predict near-surface (0–5 cm) soil moisture (SM) by mapping inputs from the Soil & Water Assessment Tool to SM time series from NASA's Soil Moisture Active Passive (SMAP) satellite for the period 1 January 2016–31 December 2018. We developed a hybrid artificial neural network (ANN) combining long short-term memory and multilayer perceptron networks that were used to simultaneously incorporate dynamic weather and static spatial data into the training algorithm, respectively. We evaluated the generalizability of the hybrid ANN using training datasets comprising several watersheds with different environmental conditions, examined the effects of standard and physics-guided loss functions, and experimented with feature augmentation. Our model could estimate SM on par with the accuracy of SMAP. We demonstrated that the most critical learning of the physical processes governing SM variability was learned from meteorological time series, and that additional physical context supported model performance when test data were not fully encapsulated by the variability of the training data. Additionally, we found that when forecasting SM based on trends learned during the earlier training period, the models appreciated seasonal trends.

Keywords: physics-guided machine learning; hybrid architecture; LSTM; data-driven science

1. Introduction

In the geosciences, it is often necessary to simulate physical processes that are impractical or impossible to directly observe such as sub- or near-surface water flux. The classic approach to simulation and prediction for many years has been development, calibration, and validation of physically based models; however, these techniques have notable limitations including introduction of user and model bias, not to mention the computational burden. Over the past few decades, applications of artificial intelligence, most notably artificial neural networks (ANNs), have matured due to advances in computer hardware in conjunction with the availability of large geospatial datasets. In particular, ANNs developed to make predictions in the physical sciences informed by domain knowledge have grown in popularity with the rise of interest in explainable machine learning [1–4]. We define explainability as a clarification of how interactions between model design and domain knowledge (i.e., informed physical context) affects the outcome (i.e., physically plausible model output) [1].

Soil moisture (SM) is a key variable in hydrologic flux with a well-documented need for near-real-time predictions for risk assessments related to, for example, flooding and crop viability [5–9]. Many well-known physically based hydrologic models simulate and predict SM; however, the simulated metric is often a model-specific index of SM state dependent on quantities such as hydraulic conductivity and porosity. These are calculated within the model instead of explicitly supplied by the user, as the number of required input parameters can easily scale to a degree that the requirements for implementation outweigh the availability of measurements [10–12]. Model calibration is successful if there are many input parameters to which the prediction of interest is sensitive, which can be adjusted or, more accurately, *calibrated* to reduce residuals between model predictions and observed data. Additionally, calibration extends computational time because the required number of model runs scales with the number of parameters. For hydrologic models, traditional calibration techniques such as Bayesian inversion are only appropriate for select primary metrics (e.g., runoff or streamflows), and forecasting alternate quantities of interest (e.g., SM, evapotranspiration) using these physically based models is challenging [13]. Given the availability of computational resources and the wealth of remotely sensed, spatially continuous data, there is ample opportunity to apply ANNs to the geosciences [14]. Remotely sensed (RS) data are commonly used with physically based models to provide calibration data [15–18], particularly useful in areas where monitored networks may not be available. Using RS data in concert with ANNs provides the added benefit of leveraging the strength of ANN pattern recognition to identify signatures in RS data that may be difficult to detect manually.

In physical processes, there is often notable latency between application of the driver (cause) and the onset of the change (effect). In machine learning, the "cause" is represented by samples (tensors) of features (individual values or series) to which the target predictions (labels), or "effects," are sensitive. In many cases, the process of interest is affected by conditions over a period of time reflecting antecedent environmental conditions. For example, infiltration rates depend not only on the physical characteristics of the land surface (vegetation, porosity and grain size, topographic slope), but also the level of saturation. SM, and unsaturated flow in general, is a strong function of both the physical properties of the soil (e.g., pore size and permeability) and the atmospheric conditions (e.g., precipitation, relative humidity, and evapotranspiration).

Machine learning models depend on a large volume of cause and effect examples (samples and target labels) to develop a mapping that mimics the complex variability of a system. A sufficiently trained model has learned a mathematical representation of a physical process such that it is able to make accurate predictions on new data not used during training, i.e., the model is capable of generalization. When using spatiotemporal data, it is important that the model be evaluated for prediction accuracy in both time and space dimensions. When predicting SM, we are concerned with the static environmental conditions such as topography and land use as well as seasonality. To properly reflect transient processes, ANN architectures should consider both long and short timescales when predicting time series data while also taking into account static environmental features that do not vary in time (i.e., soil type, topography). Additionally, a spatiotemporal model is especially useful if it can make accurate predictions not only during the temporal window of the training period, but for the future as well.

Building known physical laws and constraints into machine learning (ML) models of physical systems can significantly improve the predictive accuracy and statistical efficiency of the model by limiting the optimization search space to the set of physically plausible possibilities [2,14,19,20]. Physical constraints are often enforced using data or feature augmentation, where additional samples and features are provided to enhance the variability of the problem presented to the model for training. Additionally, penalizing the optimization process when physically inconsistent predictions are made will steer the optimization algorithm to make choices that do not just reduce error, but do so in a physically meaningful way. The goal is to provide the model with a well-stated objective that does not violate known physics and is constrained by it.

In this work, an ensemble of ANNs was trained with eight experimental configurations to assess the strengths and weaknesses of the hybrid model architecture and the ability of a trained ANN to estimate and predict SM in environments and time periods outside of the training dataset. A long short-term memory (LSTM) network was coupled with a multilayer perceptron (MLP) network to map hydrologic inputs from the Soil & Water Assessment Tool (SWAT; features) to near-surface (0–5 cm) SM estimates from NASA's Soil Moisture Active Passive satellite (SMAP; labels) to within $\pm 0.04\,cm^3\,cm^{-3}$. LSTMs learn to predict patterns in time series data, while MLP networks are particularly adept at developing many-to-many mappings of discrete samples. As a result of the correlations of SM to both weather time series and soil/site physical properties, the strengths of both LSTMs and MLPs were leveraged through development of a hybrid ANN.

Two model designs were trained using data from (1) a single and (2) multiple watersheds, then evaluated on watersheds with different environmental conditions. Of course, the trained models were evaluated using data withheld from ANN training (test set). To assess the ability of the trained models to estimate (current) and forecast (future) SM, test data were used from the temporal window of the training period as well as a subsequent interval. Experiments were performed to assess the effectiveness of using augmented input data that set flags for unusually wet or dry days as well as antecedent conditions. Additionally, we evaluated whether adding static spatial data had significant effects on model accuracy, or if the physical information encapsulated by weather timeseries was sufficient to predict SM over a range of environmental conditions. Each model design was trained by optimizing two different loss functions: (1) a standard mean-squared-error (MSE) and (2) MSE amended with a component to provide physical context. Performance was evaluated by calculating the RMSE between hybrid model SM estimates and SMAP SM on the test set.

2. Materials and Methods

2.1. Site Descriptions

The Middle-Tennessee Elk river watershed (ELKR) was selected for hybrid-ANN development because it has a high density of SM monitoring stations from the Soil & Climate Analysis Network (SCAN) within the spatial extent of passive satellite SM pixels ($\approx 36\,km^2$), is topographically flat over the westernmost 200 km, and has relatively homogeneous land-use designations (Figure 1). In many cases, uncertainty associated with RS SM data decreases as the physical characteristics of the land surface become increasingly homogeneous. In addition, the ELKR site was selected because it has few developed areas. Here, SCAN data from the top 5 cm of the soil profile were used to replace missing data in SMAP time series, whenever possible. Previously, a "goodness-of-fit" assessment was performed wherein RS SM time series were compared with observed data over a nearly three-year period for the ELKR site [21].

Three additional watersheds were selected for this study based on a variety of criteria including proximity to the ELKR site, similarities and differences in environmental conditions, and the availability of in situ SM stations; the Yazoo (YAZO), Escalante-Sevier Lake (ESCL), and Washita (WASH) basins (Figure 1). The YAZO watershed borders the ELKR site and has similar environmental conditions. Although the YAZO site shares many of the characteristics of an "ideal" watershed for SMAP SM retrievals. As a subbasin of the Mississippi River, it is a complex hydrologic system with a lot of development—shipping routes, dams—that have significantly altered its hydrology. The ESCL site is in the southwest U.S. where environmental conditions are typically arid in contrast to the humid characteristics of the ELKR site. This basin was previously evaluated along with the ELKR site in a goodness-of-fit assessment using RS and in situ SM data [21]. The WASH site is a temperate area dominated by grasslands and was previously used as a core validation site during SMAP SM calibration and validation [22]; it is also a common study site for hydrologic modeling with SWAT [23,24].

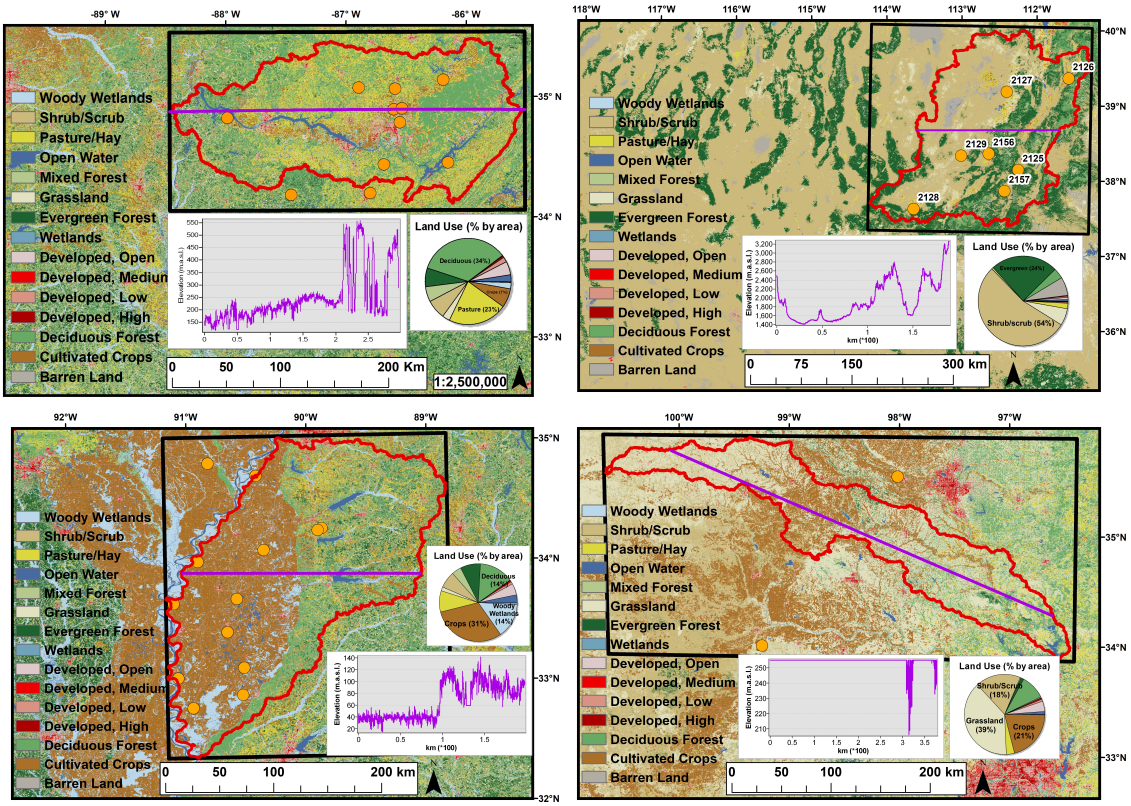

Figure 1. Study site land use (NLCD 2011), Soil & Climate Analysis Network (SCAN) stations (orange), and elevation profile (purple) for (clockwise from top left) Elk river watershed (ELKR), Yazoo (YAZO), Washita (WASH), and Escalante-Sevier Lake (ESCL). The red curves outline each basin.

2.2. Neural Network Architecture

Hybrid architectures leverage the strengths of member networks to uniquely process data. For example, Ref. [25] successfully used a hybrid convolutional LSTM neural network to analyze long blocks of text to identify the overall sentiment of the writer Natural language processing is a common application of LSTMS, where words and characters are analyzed as a series, and convolutional neural networks have proven effective in image and document classification. Hybrid architectures have also been employed with great success to predict changes in land surface characteristics [26], accelerate incompressible flow solvers [27], and locate the source of contaminants detected in time series [28].

The hybrid architecture (Figure 2) was selected to simultaneously interrogate both static and dynamic input data types as demonstrated by [29]. Initially, an MLP was developed in an attempt to map SWAT inputs and weather time series to SMAP SM time series, but this was met with limited success. Not only were unseasonably wet and dry days not replicated by the MLP, but error for estimates on the test set was not within acceptable thresholds defined in accordance with SMAP mission objectives [30] (RMSE $< 0.04\,\mathrm{cm}^3\,\mathrm{cm}^{-3}$). Although MLPs have been successfully used to develop surrogates of physically based models [31–33], the degree of difficulty increases when the data depend on processes occurring across both spatial and temporal scales. Time series data are inherently ordered whereas static data are not, and machine learning algorithms developed to recognize patterns in unordered data have limitations when applied to time series [34]. For this application, inputting the static and dynamic features to an MLP network had an accuracy no better than $\pm 0.04\,\mathrm{cm}^3\,\mathrm{cm}^{-3}$ when compared to SMAP SM retrievals. This was because the input feature vectors for each sample representing all static and dynamic data for a discrete location lost all temporal interrelationships. Conversely, if the input feature vector represented static data and only a single timestep, then the static data were repeated for each timestep, and the ANN was "over-informed" about static features.

Additionally, MLPs do not have the capacity to learn information relevant to the current sample from neighboring cells or previous time steps because all spatial or temporal coherence is lost.

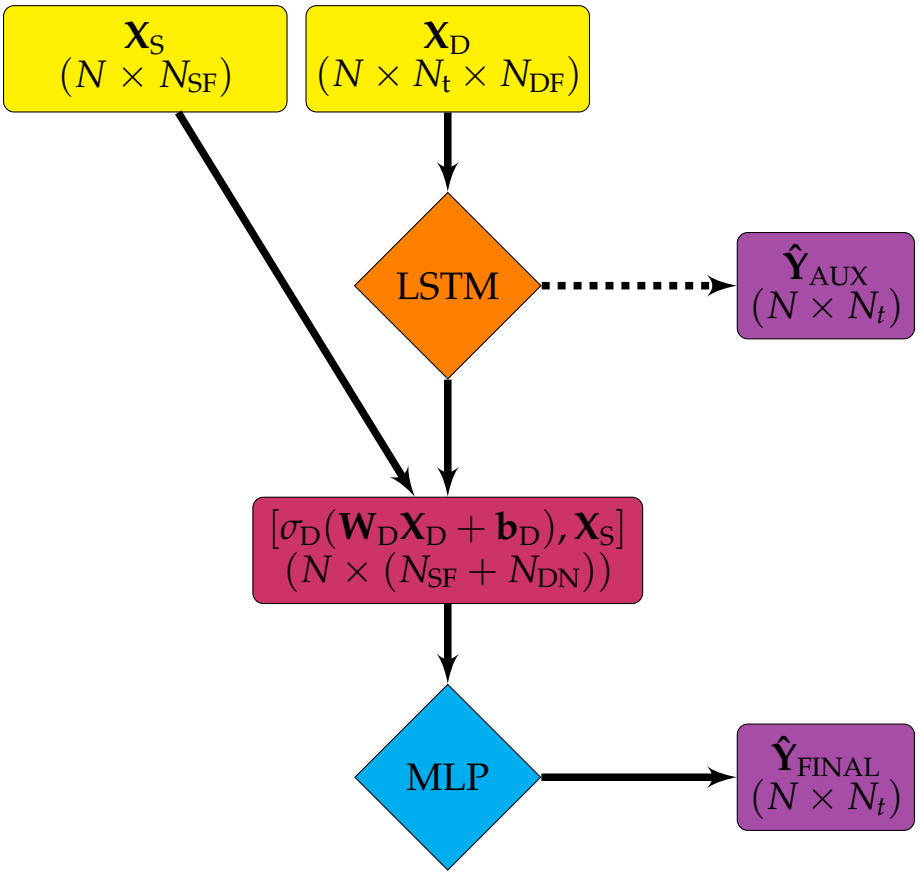

Figure 2. Hybrid architecture where $\mathbf{X}_D$ is the three-dimensional array of dynamic (time series) inputs, $\mathbf{X}_S$ is the two-dimensional array of static inputs, N is the number of samples, N_t is the number of time steps, N_{DF} is the number of dynamic features, N_{SF} is the number of static features, and N_{DN} is the number of nodes in the long short-term memory (LSTM) layers.

The strength of the LSTM architecture is its ability to "remember" characteristic events and sequences of events, such as mean-state conditions and seasonal behavior in time series data, to better replicate them in the future [35–38]. One difficulty is that while the LSTM architecture is good at replicating *commonly occurring* sequences of events, events that fall outside of short- or long-term mean state conditions are significantly more challenging not only because the frequency of above- or below-average events is small compared to the total number of events, but because there is momentum in the prediction and "discontinuities" are difficult to replicate, especially because they can be mistaken for noise. As recommended by [38], a value of 1 was supplied as the bias of the LSTM forget gate at initialization and all other LSTM biases were initialized as a vector of zeros, and LSTM weights were initialized as arrays of ones. MLP weights were randomly initialized.

During the training process, each ANN, composed of densely interconnected information-processing nodes organized into layers, identifies the nonlinear function, g, by mapping input features, $\mathbf{X}$, to the corresponding output labels, $\hat{\mathbf{Y}}$ such that the ANN yields the estimation $\hat{\mathbf{Y}}$

$$\hat{\mathbf{Y}} = g\left(\mathbf{X}; (\mathbf{W}, \mathbf{b}, \phi)\right), \tag{1}$$

where $\mathbf{W}$ and $\mathbf{b}$ are weight matrix and bias vector applied to $\mathbf{X}$, which form the mapping matrix $\boldsymbol{\Theta}$ and ϕ is the activation function that determines the contribution of each node in the current layer to the nodes in next layer.

2.3. Input Features

This study used input features corresponding to hydrologic inputs from a new version of SWAT, the Soil & Water Assessment Tool plus (SWAT+) [39]. The SWAT+ model is a physically based hydrologic model that predicts near-surface water budget in response to changes in climate, land use, and/or land-management practices over scales ranging from a single watershed to the continental U.S., which is run daily on a 10 km^2 grid, [39]. The SWAT+ model was selected because of the availability of its input datasets, which are amenable for machine learning studies applied to hydrologic modeling because the volume of data is large (≈500,000 nodes in the continental U.S.). Additionally, we wanted to develop a tool that could quickly and easily predict SM using common hydrologic model input data for assessments in regions without available calibration data. An initial attempt was made to train a surrogate model for SWAT+ by mapping SWAT+ inputs to SWAT+ SM predictions; however, results were unsatisfactory. The SWAT+ input dataset included only three parameters to which SM is sensitive to (soil depth, curve number, and percolation rate), and calibration efforts did not yield acceptable results. We instead decided to map the SWAT+ inputs to RS SM data.

The SWAT+ model uses static inputs for land use (National Land Cover Database; NLCD), soil type (State Soil Geographic dataset; STATSGO), topography (Advanced Spacebourne Thermal Emission and Reflection digital elevation map; ASTERM DEM), and weather data. Static parameters used in ANN development are listed in Table 1. The weather data were provided as measured or forecasted daily time series, or, when such data were not available or had missing values, these were replaced by the Agricultural Research Service (ARS) weather-generator model. The weather generator uses long-term monthly mean data for precipitation, temperature, solar radiation, relative humidity, and wind speed to project daily values on a 50 km^2 grid where each SWAT+ grid node is assigned the closest simulated weather station. Hourly meteorological data required for SWAT+ were acquired from the NOAA Integrated Surface Database (ISD; https://www.ncdc.noaa.gov/isd) and averaged into daily data (Table 2) with any missing data replaced by synthetic values from the SWAT+ weather generator module.

SWAT+ input data were partitioned into static and dynamic components. Static inputs (Table 1) define the initial conditions at each SWAT+ model grid cell, while dynamic inputs (weather) drive hydrologic flux at each grid cell. Static data shown in Table 1 are of numerical (measured or calculated) and categorical data types. Because the algorithms used only support numerical data, categorical data were converted using a technique called one-hot encoding. One-hot encoding means that categorical features are replaced by multiple binary features, one for each category, where the feature value is 1 if a given sample had that category as the original feature and zero otherwise. Static data were normalized using the Euclidean (L_2) norm.

Table 1. Static parameters from Soil & Water Assessment Tool plus (SWAT+) hydrologic input datasets used as features for the multilayer perceptron (MLP) network.

Parameter	Description	Type
CN2	Curve number	Numerical
SOILDEP	Depth of soil profile	Numerical
SLOPE	Percent slope	Numerical
SLOPELEN	Length of sloping surface (m)	Numerical
ITEXT	Soil textural characteristics	Categorical
IPLANT	Land-use type	Categorical
USLEK	Soil erodibility factor	Numerical
USLEC	Crop/vegetation management factor	Numerical

Dynamic inputs included all weather data used by SWAT+ (Table 2), primarily meteorological measurements, but weather-generator replacements for missing data (observations) and those estimated with the SWAT+ weather generator. As previously described, we purposefully used a combination of observed and generated weather times series because in a typical hydrologic modeling investigation, not all time series data required by the model may be available at a given location or over the period of time for the analysis. For this reason, we wanted to develop a tool that was developed with common data limitations in mind. Dynamic data were normalized using the L_{inf} norm, which encourages extreme values to not be treated as outliers.

Table 2. Weather-parameter feature vectors for the LSTM and their sources. Missing values from NOAA data were filled in using long-term monthly means and the SWAT+ weather generator.

Parameter	Source
Precipitation (mm)	NOAA ISD
Temperature (max, min, mean; °C)	NOAA ISD
Wind Speed (m/s)	NOAA ISD
Wind Direction (°)	NOAA ISD
Solar Radiation (mean, max; W m^{-2})	SWAT+ weather generator
Relative Humidity (kPa)	SWAT+ weather generator
Dewpoint (°C)	SWAT+ weather generator

Table 3 defines and enumerates input dimensions. Dynamic input data for the LSTM, $\mathbf{X}_{\mathrm{D}}$, had shape $(N \times N_t \times N_{\mathrm{DF}})$ where N was the number of samples, N_t was the number of daily time steps, and N_{DF} was the number of dynamic features per sample at each time step. Dynamic features were time series for weather parameters listed in Table 2, weather anomalies, and, for some experiments, flags for unseasonably wet or dry days and antecedent conditions (see Section 2.3). The output of the last LSTM layer (σ_{D}), of shape $(N \times N_{\mathrm{DN}})$, where N_{DN} was the number of LSTM nodes, represented the short- and long-term history remembered throughout the entire sequence of each sample. Static data ($\mathbf{X}_{\mathrm{S}}$) of shape $(N \times N_{\mathrm{SF}})$, where N_{SF} were the number of static features, were then concatenated with σ_{D} and input to the MLP. The error, or "loss" ($\mathcal{L}$), between ANN predictions and SMAP time series was calculated at the LSTM ("AUX") and hybrid LSTM-MLP ("FINAL") outputs. The AUX and FINAL losses were calculated to assess whether SMAP SM estimates were enhanced by processing additional static data or if the dynamic features alone contained sufficient information to make accurate SM estimates.

Table 3. Definitions of feature and label vector with dimensions for the LSTM (dynamic) input ($\mathbf{X}_{\mathrm{D}}$), MLP (static) input ($\mathbf{X}_{\mathrm{S}}$), Soil Moisture Active Passive (SMAP) SM timeseries ($\mathbf{Y}$), LSTM SM auxiliary estimate of $\mathbf{Y}$ ($\hat{\mathbf{Y}}_{\mathbf{AUX}}$), LSTM-MLP final SM estimate of $\mathbf{Y}$ ($\hat{\mathbf{Y}}_{\mathbf{FINAL}}$) for SNGLT and MULTT model designs.

Variable	Definition	Value
N_t	Number of time steps	1096
N_{DF}	Number of dynamic features	22 (without SAT flag) 44 (with SAT flags)
N_{SF}	Number of static features	56

Feature Augmentation

In the authors' opinion, it is critical to combine domain expertise for any machine learning exercise. Domain knowledge was provided to this physics-guided model by modifying model components (e.g., samples, labels, loss function, model architecture) to best reflect the physical system and enhance interpretability of the output [1,2].

The soil profile retains a "memory", i.e., a period of persistent moisture retention/deficit following an atypical event such as intense precipitation or dryness [40]. SM memory is not quantifiable using only meteorological variables (i.e., weather time series used in dynamic inputs for this study), therefore we felt it was necessary to augment dynamic samples with features that provided context specifically for atypical events. Additionally, we know that the amount of moisture present in the soil profile prior to an event, such as a heavy rainfall, will limit the amount of water that can infiltrate. By augmenting dynamic features, we sought to encourage the model to pay special attention to "important" time steps and predict SM of similar magnitude as labeled data as well as recognizing seasonal trends. This is important in risk analyses, where the magnitude of a flood, for example, determines how and when emergency response services are deployed.

Feature engineering is a common practice to incorporate physical knowledge into training data [1]. Use of first- and second-order features, where first-order features are actual data values and second-order features have been derived from actual data, is common in time series analyses [34,41,42]. First-order features were input to the ANN as both weather time series. Time series anomalies were calculated as second-order features using a moving average with a window of 90 days (approximately seasonal) and 50% window overlap. The window size and amount of overlap were chosen because a 90-day window approximated seasonal changes and the amount of overlap highlighted seasonal transitions. For some experiments (Table 4: MSEF, PGLF), additional second-order features were engineered to flag unseasonably wet or dry days (e.g., short-term weather anomalies or long-term floods and droughts). Hereafter, environmental conditions greater or less than one standard deviation from the seasonal moving average will be referred to as seasonally anomalous time steps (SATs). Here, second-order features were generated based on the statistical significance of first-order data at each time step. Flags for SATs were represented by an additional feature for each weather type where all time steps within one standard deviation of the moving average were assigned a value of 0 and SATs were given a value of 1. Antecedent conditions were appreciated by labeling the three preceding time steps to any SAT with a value of 0.75. The weights applied to SATs and the number of antecedent time steps were determined by training the five best-performing models with several different configurations of SAT flag weights and antecedent time steps.In this case, feature augmentation equated to creating additional binary categorical features that addressed if a condition at the given time step was an SAT or a precursor to an anomalous environmental state outside of mean-state conditions. The efficacy of using second-order features was assessed by interrogating the hybrid ANN network trained with and without SAT feature augmentation to determine if they improved overall predictions.

2.4. Output Labels

Near-surface (0–5 cm) SM data from NASA's SMAP satellite were downloaded from the NASA EarthView website (worldview.earthdata.nasa.gov). Data were retrieved as close to 6 AM as possible. Missing values in SMAP time series were replaced with in situ data from the nearest SCAN site (nrcs.usda.gov/scan/) to the SMAP grid cell center. Agreement between SMAP and SCAN for watersheds in the southern US was demonstrated by [21].

2.5. Physics-Guided Loss Function

As demonstrated by [43], the choice of loss function can significantly improve LSTM performance by as much as 50%. Ref. [3] showed that even for small datasets, incorporating prior knowledge into the network enhanced model performance and accuracy. For some experiments (Table 4), the traditional MSE loss function was amended with a physics-guided component that applied a penalty to the loss of an observed physical relationship between SM when the additional parameter was violated as suggested by [2]. Experiment PGLNF used a physics-guided loss function with no SAT flags in the dynamic inputs, and experiment PGLF used a physics-guided loss function with SAT input flags. The additional loss component assumed that on consecutive rain-free days, the relative differences between evapotranspiration (ET) on days i and $i-1$ should have the opposite sign as the relative

difference between SMs observed on the same days. Simply put, water flux in the near-surface environment (0–5 cm below ground surface), is balanced by infiltration from the surface (increase) and ET (decrease). If SM in the near-surface environment increased between day $i-1$ and i, ET should decrease given the absence of precipitation on either day. Loss was calculated as:

$$\mathcal{L} = \frac{1}{N}(\mathbf{Y}_{\mathrm{SMAP}} - \hat{\mathbf{Y}}_{\mathrm{SMAP}})^2 + \Delta_{\mathrm{sgn}}\zeta\sqrt{(\mathbf{Y}_{\mathrm{ET},i} - \mathbf{Y}_{\mathrm{ET},i-1})^2 + (\hat{\mathbf{Y}}_{\mathrm{SMAP},i} - \mathbf{Y}_{\mathrm{SMAP},i-1})^2}, \quad (2)$$

where $\mathbf{Y}_{\mathrm{SMAP}}$ are SMAP labels (target SM), $\hat{\mathbf{Y}}_{\mathrm{SMAP}}$ are ANN predictions of SMAP SMs (either AUX or FINAL output), ζ is a scaling factor, $\mathbf{Y}_{\mathrm{ET},i}$ are MODIS ET data on day i, $\mathbf{Y}_{\mathrm{ET},i-1}$ are MODIS ET values on day $i-1$, and

$$\Delta_{\mathrm{sgn}} = \begin{cases} 0 & \text{if } \mathrm{sgn}(\mathbf{Y}_{\mathrm{ET},i} - \mathbf{Y}_{\mathrm{ET},i-1}) == \mathrm{sgn}(\hat{\mathbf{Y}}_{\mathrm{SMAP},i} - \mathbf{Y}_{\mathrm{SMAP},i-1}) \\ 1 & \text{otherwise} \end{cases}. \quad (3)$$

This analysis used $\zeta = 0.1$, which was chosen by tuning the values {0.001,0.01,0.1,1.0} with the best-performing models (RMSE $< 0.04\,\mathrm{cm^3cm^{-3}}$) during hyperparameter optimization (see Section 2.7). The loss function was amended according to the physical relationship between ET and SM water balance flux components in the absence of rain to penalize $\mathcal{L}$ by the scaled RMSE of differences in ET and SM on subsequent days. It is acknowledged that this physical relationship is simplified; however, it was included to assess the efficiency of guiding the loss function. Because the relationship only holds on rain-free days, dry periods and droughts were preferentially "guided" by the augmented loss function, although feature augmentation in concert with LSTM memory should provide support for prediction of high SM due to intense precipitation events.

2.6. Experimental Design and Data Partitioning

Two model designs were developed for experimental analyses: (1) SNGLT (Figure 3), which trained a model on the ELKR site only and evaluated the trained model on test data for a range of environmental conditions, and (2) MULTT (Figure 4), which was trained on data from all sites (concatenated along samples axis), then evaluated on the test dataset for each site individually. Models were trained, validated, and tested on the sites shown in Figure 1, optimizing MSE and PGL loss functions, and using input feature datasets with and without SAT flagging as listed in Table 4. For all experiments, the RMSE of SM estimates at the AUX and FINAL outputs were calculated, and this metric was used to evaluate model performance of LSTM only vs. hybrid network SM prediction. The experiments were designed to interrogate the model architecture (LSTM only vs. hybrid), the use of SAT flagging, choice of loss function (MSE vs. PGL), and time period (TIME_E, TIME_P) to answer the following questions:

1. Was an LSTM network alone sufficient to estimate spatio-temporal data, or was the addition of static environmental data used in the hybrid architecture (soil and landscape characteristics) useful?
2. What were the effects of second-order features used to flag wetter/drier-than-normal time steps?
3. How did the two loss functions compare across the TIME_E (2016–2018) and TIME_P (2017–2019) intervals?

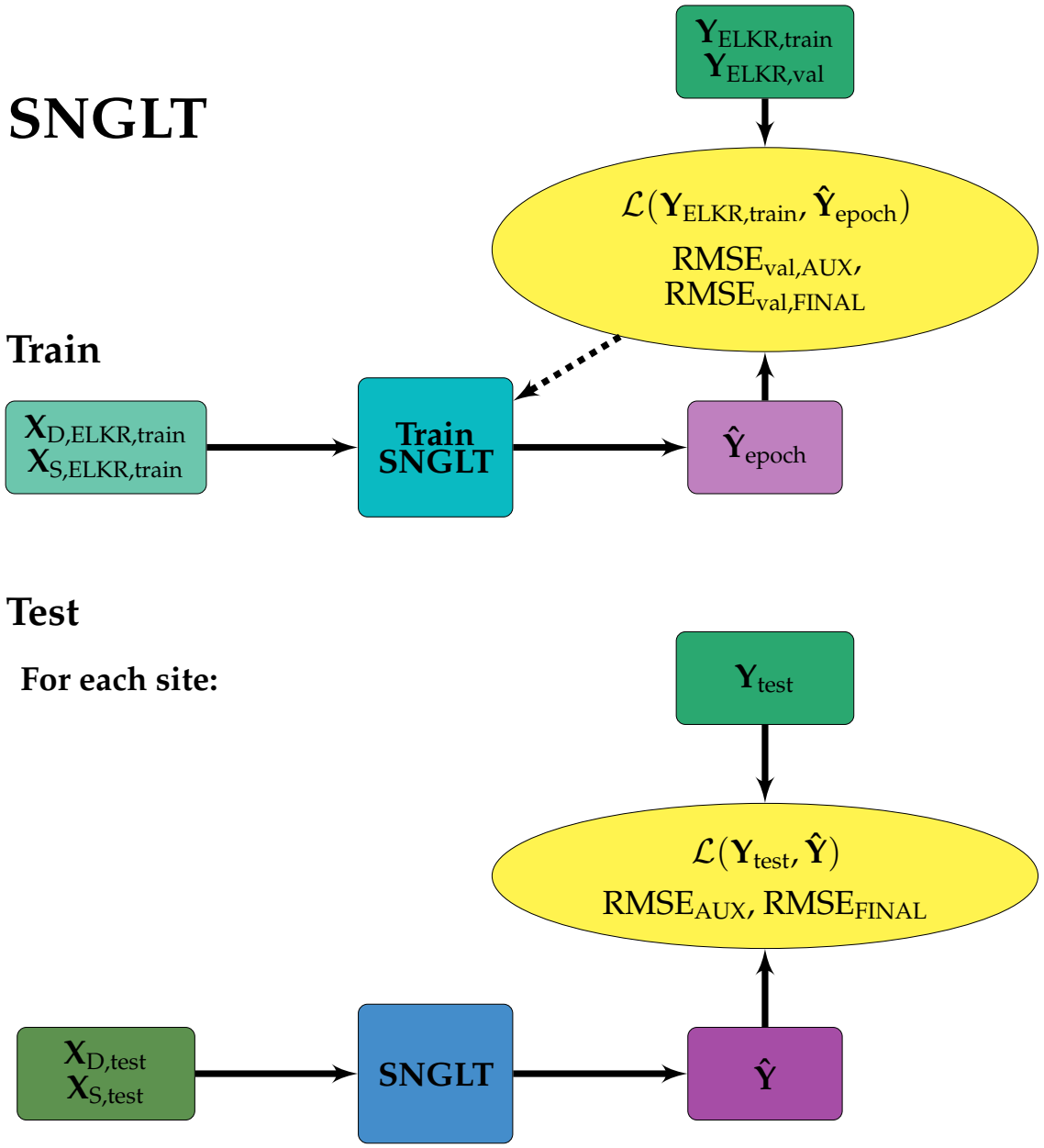

Figure 3. SNGLT model design. Solid arrows indicate the forward flow of information, dashed arrows indicate information used to make updates to model weights and/or inform early stopping. Training data were from the ELKR site only. The trained model was then applied to each site individually.

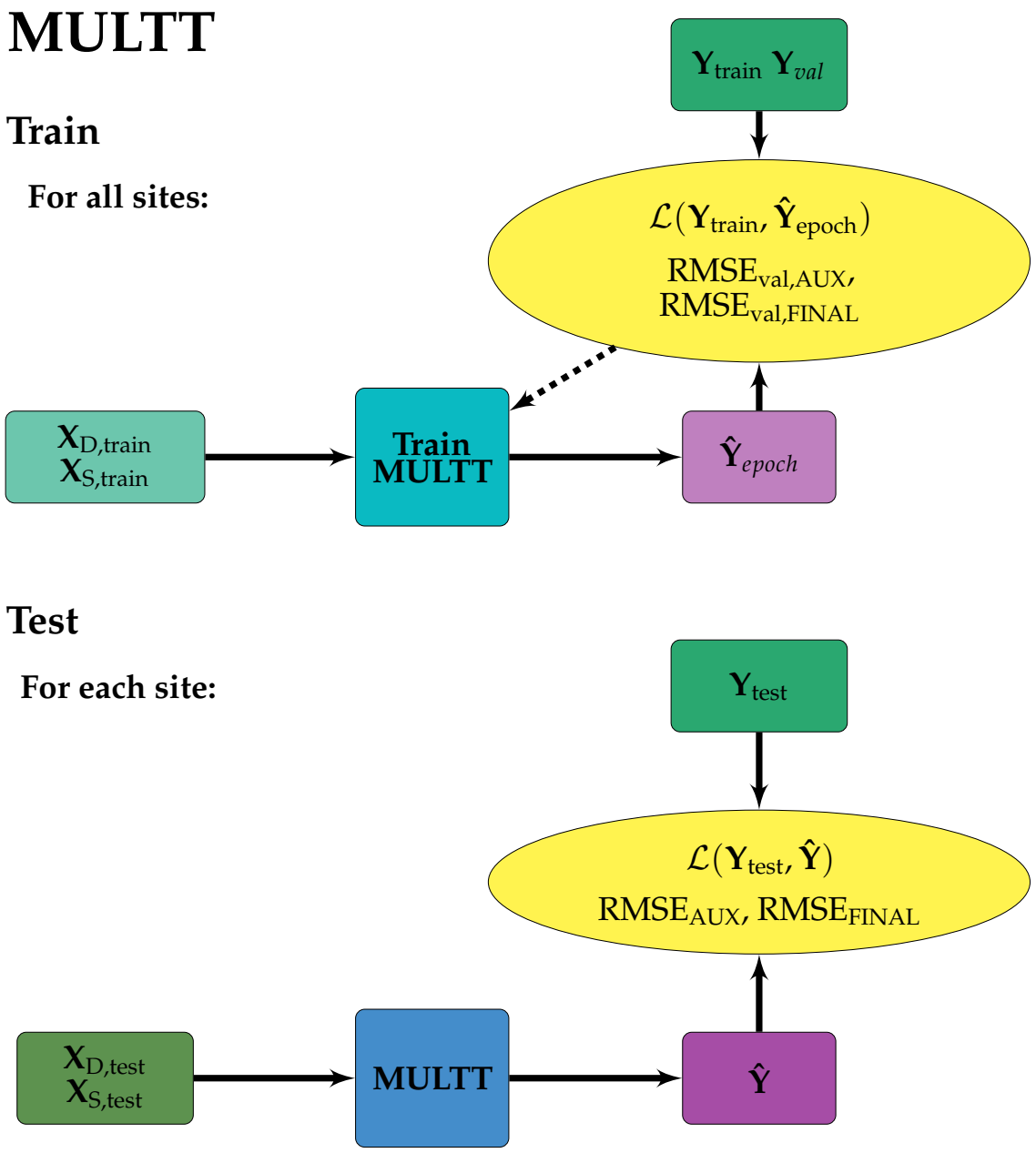

Figure 4. MULTT model design. Solid arrows indicate the forward flow of information, dashed arrows indicate information used to make updates to model weights and/or inform early stopping. Training data were from all sites, such that each tensor was a concatenation of training data from each site along the samples axis. The trained model was then applied to each site individually.

Table 4. Experimental design. Two model designs were trained and evaluated: (1) SNGLT was trained on the ELKR site only and tested on ELKR, YAZO, WASH, and ESCL sites individually and (2) MULTT was trained using data from all four sites at once, then evaluated on each site individually.

Model	Experiment	Training Sites	Loss Function	SAT Flags
SNGLT	SNGLT_MSENF	ELKR	MSE	No
SNGLT	SNGLT_MSEF	ELKR	MSE	Yes
SNGLT	SNGLT_PGLNF	ELKR	PGL	No
SNGLT	SNGLT_PGLF	ELKR	PGL	Yes
MULTT	MULTT_MSENF	ELKR, YAZO, WASH, ESCL	MSE	No
MULTT	MULTT_MSEF	ELKR, YAZO, WASH, ESCL	MSE	Yes
MULTT	MULTT_PGLNF	ELKR, YAZO, WASH, ESCL	PGL	No
MULTT	MULTT_PGLF	ELKR, YAZO, WASH, ESCL	PGL	Yes

Prior to model training, available data were partitioned into training, validation, and test datasets (Table 5). Training data, typically the majority (≈90%) of available input features and associated target labels, were used to train the network, i.e., develop the nonlinear transfer function of input features to output labels. Validation data were used during training to evaluate the performance of the current model state on data not used during training and validation. Often the number of optimization iterations (epochs) is determined according to the performance of the model on the validation dataset, where training is stopped when model performance on the validation set does not improve (early stopping). The test data were samples not used during training/validation, instead they were a "blind" dataset used to assess the prediction accuracy of the trained and validated model. All raw data (SMAP SM, SCAN SM, MODIS ET time series), preprocessed (normalized/one-hot encoded inputs) and partitioned datasets ($\mathbf{X}_{\mathrm{D}}$, $\mathbf{X}_{\mathrm{S}}$, $\mathbf{Y}_{\mathrm{SMAP}}$, $\mathbf{Y}_{\mathrm{ET}}$) are available as Supplementary Materials.

Table 5. Number of samples for training (N_{train}), validation (N_{val}), and test (N_{test}) sets for each site.

Site	N_{train}	N_{val}	N_{test}
ELKR	2871	319	10
YAZO	3235	360	10
WASH	4403	490	10
ESCL	4028	447	10

2.7. Hyperparameter Optimization

Hyperparameters were optimized through an exhaustive search of their combinations as listed in Table 6. The best set of hyperparameters was selected from those in Table 6 that had the lowest RMSE on the validation set. To reduce computational time, hyperparameter tuning over all configurations was performed on the training set for model experiments for the SNGLT design for PGL and MSE loss functions. Configurations that resulted in RMSE $< 0.04\,\mathrm{cm}^3\,\mathrm{cm}^{-3}$ at the FINAL output were then used for training model experiments for the MULTT design. Therefore, the optimal configurations chosen were the best-performing *overall* for both model designs.

The $\boldsymbol{\Theta}$ that yielded $\mathbf{Y} \approx \hat{\mathbf{Y}}$ was selected according to the optimal configuration of hyperparameters. Hyperparameters affect the architecture or specifics of network performance, both in terms of training speed and output performance (i.e., lowest validation loss), but are not inherently "learnable." The number of nodes and layers determine the density and depth of the network, respectively. Dropout rate, a form of regularization, specifies the random fraction of weights set to zero, which has the effect of minimizing overfitting by preventing the over-dependence of data flow through a select few

pathways through the neural network. During training, the dataset is divided into minibatches. For the hybrid architecture used in this work, minibatches are packets of the same sample indices from the LSTM and MLP inputs, such that dynamic and static data for the same samples are processed together. Smaller batch sizes tend to reach the minimized loss function faster because the network weights are updated more frequently per epoch (one complete exposure of the entire dataset to the neural network during training), but smaller batch sizes also increase training times because more time is spent passing smaller data blocks to the processor relative to the time necessary to complete forward and backward propagation [44]. Optimizers are algorithms that speed up the optimization of $\mathcal{L}$ by calculating the gradients of the loss function with respect to the weights and biases and appropriately setting the step length to establish the fastest path to the global minimum. Early stopping was implemented, which stopped training the model if the validation loss did not improve by a minimum value (Δ) within a given number of epochs (patience interval). Choice of loss function (PGL or MSE) was a hyperparameter to assess the effects of using the PGL.

Table 6. Parameters used during hyperparameter tuning. Optimal hyperparameters are shown in bold.

Hyperparameter	Values Interrogated
SAT antecedent day flag	0.25, 0.5, **0.75** *,**, 1
Number of antecedent time steps	2, **3** *,**, 5
ζ	0.001, **0.01** *,**, 0.1, 1.0
LSTM layers	**1** *,**, 2, 3, 4, 5
LSTM nodes	1, 2, 3, 4, **5** *,**, 10, 25, 50, 75
MLP layers	1, 2, 3, 5, **10** *,**, 20, 50, 100
MLP nodes	5, 10, 25, 50, 75, **100** *,**, 150, 200
Dropout rate	0.1, 0.2, **0.3** *, 0.5
MLP activation function	**ReLU** **, **ELU** *
LSTM activation function	sigmoid, **hyperbolic tangent** *,**
Optimizer	Adam, Adamax, Adagrad, Adadelta, **Nadam** *,**, SGD, RMSprop
Batch size	25, 50, **75** *, **100** **, 250, 500
Min delta	**0** **, 0.001, **0.0001** *, 0.00001
Patience	5, 10, **20** *,**, 50, 100, 250, 500

* Optimal parameters using PGL loss function. ** Optimal parameters using MSE loss function during hyperparameter tuning.

Custom-built software was developed to interrogate hyperparameter space and continually monitor the search progress. Initial software development allowed models to train without early stopping to compute the mean number of epochs (training iterations) required for the majority of model configurations to reach equilibrium, i.e., the rate of change of $\mathcal{L}$ had slowed such that fluctuations in sequential SM estimates from the FINAL output were negligible. More than 90% of trials indicated that models estimating SM to with RMSE $\leq 0.04\,\mathrm{cm}^3\,\mathrm{cm}^{-3}$ ran for 500 epochs or fewer. During hyperparameter optimization, each run was set to train for 500 epochs, although most trained for shorter periods upon implementation of early stopping.

3. Results and Discussion

Spatial generalizability was determined by applying each trained model design (SNGLT and MULTT) to the test datasets for all sites individually as shown in Figures 3 and 4. Performance metrics reported in Tables 7 and 8 and time series shown in Figures 5 and 6 show results for the average of samples in the test set for each site. The model design with the best performance (lowest RMSE, highest coefficient of determination (R^2)) for all study sites was able to generalize new data. To assess the ability

of the trained ANN to estimate and forecast SM, each model experiment was tested on data from the same period as the training data (TIME_E) as well as a future period (TIME_P). Accurate predictions for test data in TIME_E would indicate the model's ability to predict SM under meteorological conditions not explicitly included in the training data, and accurate future projections of SM in TIME_P would demonstrate the model's understanding of seasonality and intraseasonal variability. Error for SM estimates ($\hat{\mathbf{Y}}$) on the test set were calculated as the mean (μ) RMSE over all test samples as reported in Tables 7 and 8 for the TIME_E (1 January 2016–31 December 2018) and TIME_P (1 January 2017–31 December 2019) intervals, respectively. Error was calculated at both the auxiliary (AUX; LSTM only) and FINAL (hybrid LSTM-MLP) outputs to determine if the addition of static data significantly increased model performance, or if the variability of SM time series could be predicted using only meteorological data.

3.1. SAT Flag Effect

Overall, there did not appear to be any advantage or disadvantage to using SAT flags. Error at AUX and FINAL outputs was calculated over all time steps and for SATs only (not shown) to assess each models ability to replicate not only seasonality (all time steps) but unseasonably wet and dry days (SATs), which are important in risk analyses. Approximately 87% of the experiments with the minimum overall RMSE for either SNGLT or MULTT models also had the minimum RMSE for SATs. This indicated that the best-performing models were able to estimate not only mean trends and seasonality, but also unseasonably wet and dry days for sites not sampled during training (Tables 7 and 8). This showed that models that predicted SM with the highest accuracy did so because they were able to recognize subtle patterns in the time series data, and that our attempt at data augmentation did little to provide additional context. In hindsight, using both time-series anomalies and SAT flags as features to the LSTM was redundant because the anomalies already indicated unseasonably wet or dry days; the LSTM's ability to learn patterns on long and short timescales accounted for antecedent conditions. Hence, adding an additional feature that conveyed redundant information did not provide the model with any additional information.

3.2. SNGLT Model Performance

When assessed on the ELKR site for the interval TIME_E, SNGLT SM time series from the FINAL output compared well with SMAP SM time series ($R^2 \geq 0.96$; RMSE $\leq 0.04\,\mathrm{cm^3\,cm^{-3}}$) for all model experiments (Table 7). This result is unsurprising given that ELKR data were used to train SNGLT, and demonstrated that the hybrid model architecture successfully captured spatiotemporal variability within the ELKR site. When SNGLT was applied to the ELKR test data, error was reduced by $\geq 0.02\,\mathrm{cm^3\,cm^{-3}}$ between AUX and FINAL outputs for all experiments. The most significant improvement between AUX and FINAL outputs in model performance occurred for the SNGLT_MSENF experiment (RMSE $\approx 0.2\,\mathrm{cm^3\,cm^{-3}}$). Correlation between $\hat{\mathbf{Y}}_{\mathrm{ELKR,AUX}}$ and $\mathbf{Y}_{\mathrm{SMAP}}$ was high for most experiments ($R^2 \geq 0.96$), although low correlation was observed for the SNGLT_MSENF experiment (Table 7, Figure 7). Figure 7 shows that although non-zero SM estimates at the AUX output for SNGLT_MSENF correlate well with SMAP SM, 57% of timesteps were predicted with a SM of 0, resulting in overall poor correlation and high error ($R^2 = 0.021$, RMSE $= 0.239\,\mathrm{cm^3\,cm^{-3}}$). The SNGLT_MSENF experiment provided less physical context than the other experiments (no SAT flags or physics-guided loss). For this experiment, the physical processes governing SM were informed only by weather time series and anomalies as well as spatial data. Without additional information to inform the model of known physical laws, the SNGLT_MSENF experiment did not perform within acceptable error thresholds using meteorological data only. This showed that the additional physical context provided by spatial data enhanced SM estimation, in particular when little other physical context was available to the SNGLT model design during training

Table 7. Performance metrics for SNGLT and MULTT experiments performed on the training time period (TIME_E; 2016–2018) for all experiments: MSENF (MSE loss function, no SAT flagging), PGLNF (physics-guided loss, no SAT flagging), MSEF (MSE loss function, SAT flagging), and PGLF (physics-guided loss, SAT flagging). Performance metrics are shown for both auxiliary and final outputs, indicating whether the LSTM output (AUX) or hybrid output (FINAL) had better performance on the test set, i.e., if weather time series were sufficient to accurately predict SMAP SM or if the hybrid architecture yielded better accuracy. RMSEs and are listed above parenthetical R^2 statistics.

Exp	Output	ELKR		YAZO		WASH		ESCL	
		SNGLT	MULTT	SNGLT	MULTT	SNGLT	MULTT	SNGLT	MULTT
MSENF	Aux	0.239 (0.021)	0.058 (0.968)	0.228 (0.047)	0.071 (0.950)	0.149 (0.001)	0.053 (0.911)	0.153 (0.003)	0.048 (0.927)
	Final	0.035 (0.967)	0.032 (0.978)	87.217 (0.055)	0.054 (0.945)	60.581 (0.000)	0.030 (0.979)	0.237 (0.061)	0.036 (0.969)
MSEF	Aux	0.043 (0.963)	0.059 (0.967)	0.303 (0.069)	0.049 (0.909)	0.389 (0.079)	0.036 (0.965)	0.419 (0.018)	0.040 (0.915)
	Final	0.020 (0.991)	0.032 (0.982)	0.662 (0.004)	0.014 (0.995)	0.597 (0.000)	0.022 (0.981)	1.209 (0.001)	0.013 (0.927)
PGLNF	Aux	0.058 (0.967)	0.058 (0.967)	0.104 (0.563)	0.071 (0.950)	0.136 (0.013)	0.052 (0.923)	0.190 (0.109)	0.047 (0.919)
	Final	0.035 (0.990)	0.048 (0.759)	0.102 (0.564)	0.066 (0.844)	0.146 (0.012)	0.043 (0.487)	0.197 (0.115)	0.046 (0.528)
PGLF	Aux	0.050 (0.976)	0.061 (0.968)	0.144 (0.336)	0.059 (0.929)	0.202 (0.000)	0.041 (0.970)	0.219 (0.057)	0.046 (0.937)
	Final	0.031 (0.989)	0.075 (0.837)	0.107 (0.482)	0.062 (0.858)	0.144 (0.013)	0.070 (0.757)	0.197 (0.122)	0.055 (0.401)

When applied to the YAZO, WASH, and ESCL study sites, SNGLT model accuracy was poor, even for the neighboring YAZO watershed (Table 7). This suggested that the SNGLT training dataset did not "capture" the variability of environmental conditions in test data and was unable to estimate physically plausible SM time series outside of the training dataset within $0.04\,cm^3\,cm^{-3}$. Notably, RMSE between SMAP SM and SNGLT SM estimates increased between AUX and FINAL outputs for the SNGLT_MSENF experiment ($0.08 - 0.87\,cm^3\,cm^{-3}$). As previously mentioned, this experiment provided the least amount of physical context. The significant loss in accuracy between AUX and FINAL outputs for all experiments indicated that for study sites not included in the training dataset, the inclusion of static data harmed the generalizability of the SNGLT model. These results showed that critical learning of physical processes governing SM was performed by the SNGLT LSTM, which was able to generalize trends, if not absolute magnitude, in SM time series using meteorological features only. Additionally, our results show that the addition of static data not included in the training dataset did not enhance model performance without additional physical context provided during optimization.

Correlation at either AUX or FINAL outputs for SNGLT experiments was poor when applied to other study sites ($R^2 \leq 0.1$) with the exception of PGLNF and PGLF experiments on the YAZO test data ($0.48 \leq R^2 \leq 0.56$). Although error was high for the SNGLT model applied to YAZO test data (RMSE $\geq 0.1\ cm^3\ cm^{-3}$), correlation statistics for experiments optimizing the PGL function suggested that trend agreement was better for study sites with similar environmental characteristics (ELKR and YAZO) if the experimental configuration was informed of physical information at the loss function. This was likely due to a similarity in time series from ELKR and YAZO test data as opposed to true "learning" of SM variability and the ability to generalize.

The overall loss in performance when applying SNGLT to study sites not included in the training dataset was likely due to the small number of training samples used (Table 5) over a discrete set of

environmental conditions. As previously mentioned, an ANN requires a large volume of demonstrated cause and effect relationships to generalize to new data. Additionally, features specific to anthropogenic influence (other than "urban" and "developed" land use designations) were not parameterized in SWAT+ hydrologic inputs, and it makes sense that the environmental characteristics of the YAZO watershed could not be approximated by a model (SNGLT) trained on the nearby ELKR site.

As shown in Figure 7 and Table 7, comparisons of $\hat{\mathbf{Y}}_{ELKR,FINAL}$ vs. $\mathbf{Y}_{SMAP}$ had higher correlation (red; $R^2 \geq 0.97$) than $\hat{\mathbf{Y}}_{\text{ELKR,AUX}}$ vs. $\mathbf{Y}_{SMAP}$ (blue; $R^2 \leq 0.98$), during the TIME_E period, although overall correlation was high at both output. Correlations for the TIME_P period were low ($R^2 < 0.2$) with a notable decrease in accuracy between the AUX and FINAL outputs particularly for MSENF and MSEF experiments. Figure 5 shows time series for the SNGLT_PGLNF experiment for the TIME_E (top) and TIME_P (bottom) periods applied to ELKR test data. The FINAL output time series for the TIME_P period (Figure 7 bottom) replicates the trends of the time series for the TIME_E period, indicating overfitting. In particular, the SNGLT model erroneously "predicted" a drought in the late summer/early fall of 2017 that was learned from the 2016 drought [21], apparent in time series for the PGLNF and PGLF (not shown) experiments (Figure 5). Additionally, Figure 7 and Table 8, suggest a decrease in performance was observed between AUX and FINAL outputs for SNGLT model experiments using the MSE loss function during the TIME_P period (Figure 7 and Table 8). These results showed that without the PGL term, the static data did not improve forecasting of SM trends. These results showed that critical learning of SM variability was performed by the LSTM in the hybrid model designs.

Table 8. Performance metrics for SNGLT and MULTT experiments performed on the prediction time period (TIME_P; 2017–2019) for all experiments: MSENF (MSE loss function, no SAT flagging), PGLNF (physics-guided loss, no SAT flagging), MSEF (MSE loss function, SAT flagging), and PGLF (physics-guided loss, SAT flagging). Performance metrics are shown for both auxiliary and final outputs, indicating whether the LSTM output (AUX) or hybrid output (FINAL) had better performance on the test set, i.e., if weather timeseries were sufficient to accurately predict SMAP SM or if the hybrid architecture yielded better accuracy. RMSEs and are listed above parenthetical R^2 statistics.

Exp	Output	ELKR		YAZO		WASH		ESCL	
		SNGLT	MULTT	SNGLT	MULTT	SNGLT	MULTT	SNGLT	MULTT
MSENF	Aux	0.208 (0.023)	0.154 (0.130)	0.272 (0.061)	0.154 (0.227)	0.120 (0.039)	0.172 (0.029)	0.101 (0.027)	0.245 (0.121)
	Final	84.92 (0.000)	0.131 (0.147)	87.20 (0.018)	0.157 (0.233)	83.81 (0.007)	0.177 (0.030)	80.25 (0.004)	0.242 (0.113)
MSEF	Aux	0.258 (0.017)	0.300 (0.070)	0.259 (0.109)	0.217 (0.170)	0.200 (0.065)	0.182 (0.017)	0.213 (0.119)	0.079 (0.175)
	Final	0.303 (0.009)	384.47 (0.005)	0.216 (0.080)	432.658 (0.188)	0.247 (0.001)	342.28 (0.028)	0.383 (0.010)	0.086 (0.203)
PGLNF	Aux	0.123 (0.153)	0.198 (0.151)	0.169 (0.287)	0.300 (0.236)	0.135 (0.059)	0.140 (0.035)	0.196 (0.199)	0.080 (0.072)
	Final	0.121 (0.144)	0.189 (0.143)	0.168 (0.279)	0.293 (0.248)	0.145 (0.057)	0.137 (0.050)	0.202 (0.200)	0.080 (0.173)
PGLF	Aux	0.171 (0.076)	0.213 (0.078)	0.175 (0.135)	0.191 (0.029)	0.134 (0.084)	0.210 (0.005)	0.174 (0.192)	0.238 (0.133)
	Final	0.169 (0.079)	0.172 (0.082)	0.144 (0.158)	0.103 (0.016)	0.154 (0.073)	0.154 (0.025)	0.221 (0.190)	0.177 (0.094)

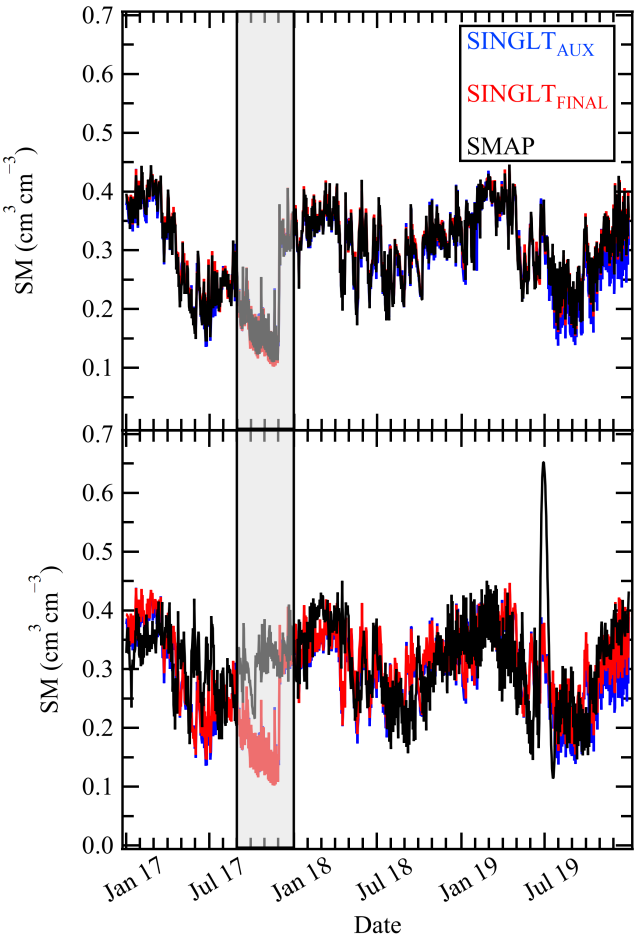

Figure 5. SNGLT_PGLNF model results for the ELKR site during the TIME_E (**top**) and TIME_P (**bottom**) periods. Results shown are for the SNGLT model evaluated on the ELKR test set. The gray box shows a severe drought observed in the late summer and fall of 2016 that was falsely predicted during the TIME_P period, indicating overfitting to the training data.

For SNGLT experiments, accurate results with respect to absolute magnitude were made only at the ELKR site, but that there was an overfitting to the seasonal variability in SM time series for test data not encapsulated the spatial or temporal training inputs. Seasonality was well-represented by the LSTM as shown by consistent RMSEs from the AUX output for all experiments, and performance of the hybrid model was enhanced when additional physical information was provided during optimization.

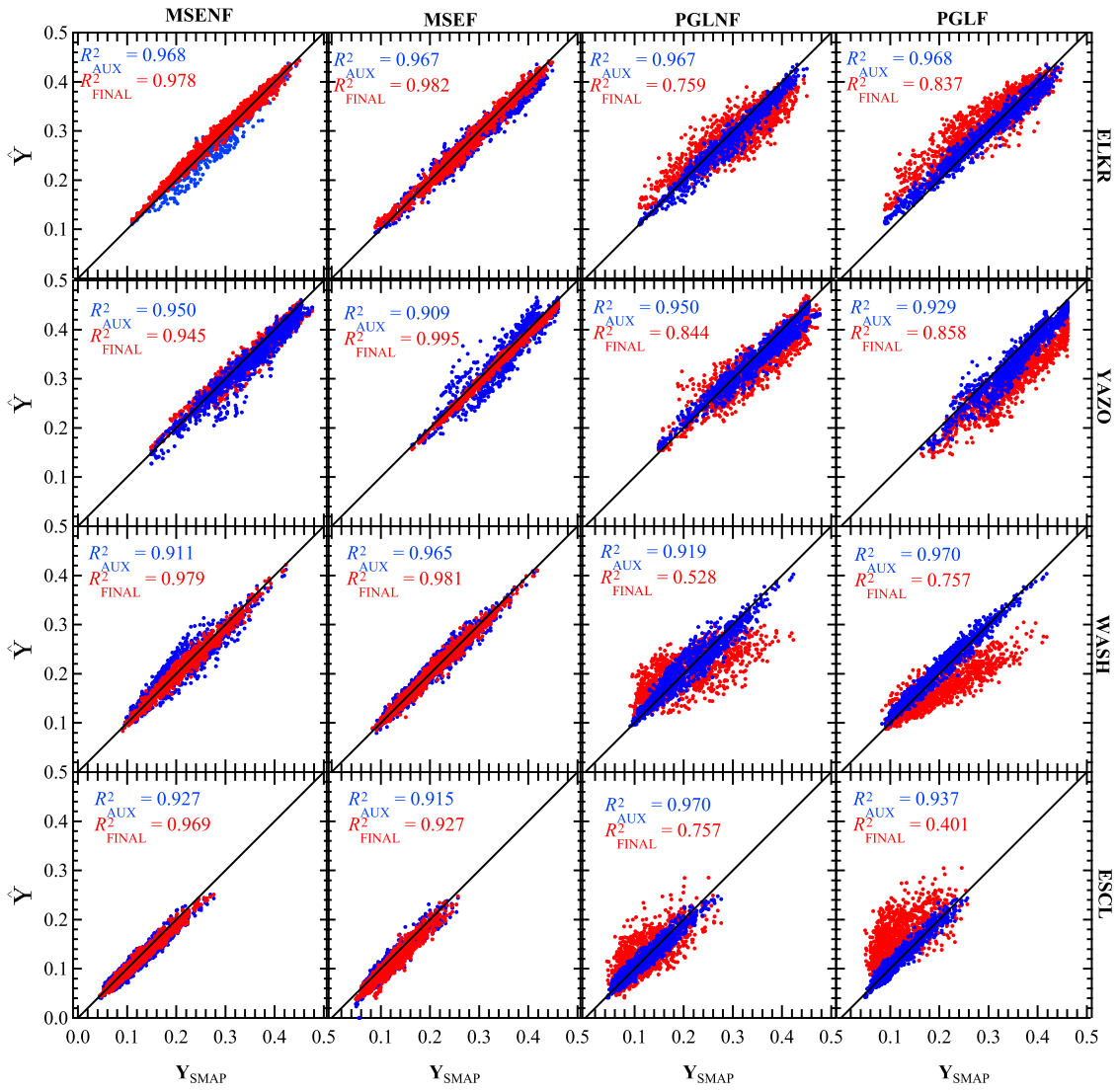

Figure 6. Results for MULTT experiments trained and tested on all four watersheds during the TIME_E period (1 January 2016–31 December 2018). The rows show results for each study site (from top): ELKR, YAZO, WASH, ESCL. The columns are (from left): MULTT_MSENF, MULTT_MSEF, MULTT_PGLNF, MULTT_PGLF. Results for $\hat{\mathbf{Y}}_{AUX}$ vs. $\mathbf{Y}_{SMAP}$ are shown in blue, and $\hat{\mathbf{Y}}_{FINAL}$ vs. $\mathbf{Y}_{SMAP}$ are shown in red. The output with the highest accuracy is shown as the top layer in each subplot, highlighting the effectiveness of each loss function.

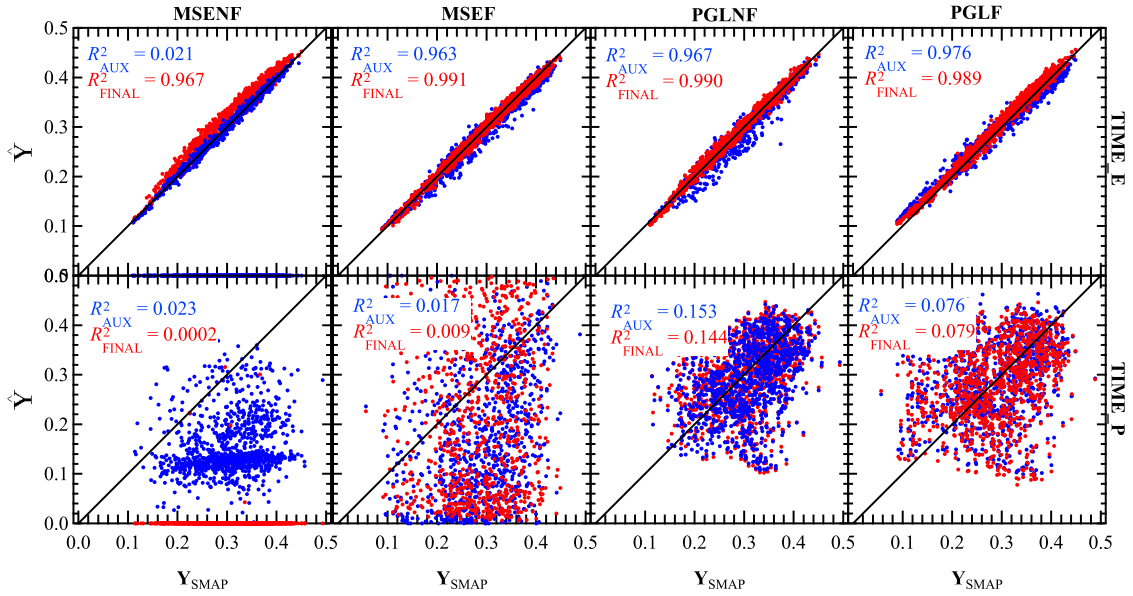

Figure 7. ELKR model results from the AUX (blue) and FINAL (red) outputs plotted vs. SMAP SM labels. Results are shown for the trained SNGLT model applied to ELKR test sets for the TIME_E (top; 1 January 2016–31 December 2018) and TIME_E (bottom; 1 January 2017–31 December 2019) time periods. Results are shown for all experiments (from left) SNGLT_MSENF, SNGLT_MSEF, SNGLT_PGLNF, and SNGLT_PGLF. The output with the highest accuracy is shown as the top layer in each subplot, highlighting the effectiveness of each loss function.

3.3. MULTT Model Performance

The MULTT model design estimated SM at the FINAL output for all study sites with RMSE $< 0.04\,\mathrm{cm^3\,cm^{-3}}$ during the TIME_E interval for at least one experiment per study site (Table 7). As listed in Table 5, the training datasets for the WASH and ESCL sites were considerably larger than ELKR and YAZO, having approximately 1000 more samples each. The uneven number of samples per site did not degrade model performance for any site in particular, as the best-performing MULTT experiments had RMSEs at the FINAL output ranging from $0.013 - 0.032\,\mathrm{cm^3\,cm^{-3}}$ for all sites. This indicated that the MULTT model design learned to generalize SM time series for all study sites using input data with more heterogeneous environmental conditions.

For test datasets in the TIME_E period, 21 of 28 (75%) of the minimum RMSEs were reported for MULTT experiments using the MSE loss function while R^2 for the cross-plots of $\hat{\mathbf{Y}}$ vs. $\mathbf{Y}_{SMAP}$, or R^2, were >0.9 (Table 7). This finding was contrary to the results from SNGLT, which required additional guidance from the PGL function during optimization to compensate for variability not accounted for in the training dataset. As previously noted, this suggested that robust knowledge of SM variability was largely provided by the diversity of samples in the training dataset, which included a sufficient number of cause and effect examples for the MULTT model to reliably generate SMAP-like SM time series. In fact, MULTT SM estimates at either the AUX or FINAL outputs were not within acceptable error thresholds for experiments using the PGL function (MULTT_PGLNF, MULTT_PGLF), and error increased between AUX and FINAL outputs for the MULTT_PGLF experiment. This suggested that for large training datasets, the physical context provided by the PGL function was less accurate than the processes learned from the training data.

Figure 6 shows that although R^2 were high for both AUX and FINAL outputs for the majority of PGL-function experiments ($R^2 > 0.7$), bias increased between AUX and FINAL outputs. Notable biases were observed for the MULTT_PGLF model applied to the WASH and ESCL test datasets, where SM estimates for the FINAL output were drier or wetter than their SMAP-observed equivalents by SMAP by 0.05–0.1 $\mathrm{cm^3\,cm^{-3}}$, respectively. On average, RMSE decreased by $\approx 0.02\,\mathrm{cm^3\,cm^{-3}}$ between AUX

and FINAL outputs for the MULTT model design when assessed on data in the temporal window of the training data (TIME_E period). These results indicated that the inclusion of static (site) data enhanced the precision of SM seasonality learned by the LSTM, while the physical context provided by the PGL function was less robust than physical information in the static data. The gain in precision between AUX and FINAL outputs when using the MSE loss function could be useful in applications where the magnitude of SM is important, such as risk analyses. In addition, using a larger volume of static data accounting for more heterogeneous site conditions could be useful if there were significant changes in spatial characteristics locally, such and land use, that could significantly alter water balance for notionally equivalent meteorological conditions. Because it is computationally expensive to train an ANN yet relatively cheap to apply one, the heterogeneity of environmental conditions represented in inputs could afford a trained model a longer "lifetime" before retraining would be required due to significantly altered spatial characteristics of a subset of the study area. Our results suggested that for such applications, this hybrid architecture could be trained over a longer time period using forecasted weather data to make accurate predictions under new environmental conditions. New environmental conditions could be, for example, data from a different watershed or real/hypothetical changes in static input parameters for sensitivity studies.

The majority of minimum RMSEs for the TIME_P period were reported for experiments using the PGL function, and overall, seasonal trends were learned (Table 8). None of the experiments performed on the TIME_P period had RSME within error thresholds and overall correlation was low at either output ($R^2 \leq 0.203$). The minimum FINAL output RMSE for any model performed on the TIME_P period was 0.101 $cm^3\,cm^{-3}$, which, depending on the environmental conditions, could be equivalent to a discrepancy between mean seasonal conditions and a drought or flood. This revealed that addition of static features did not reliably improve SM prediction precision unless the loss function at the FINAL output provided physical context as shown in Equation (2). For testing on the TIME_P period, experiments using the MSE loss function estimated and predicted SMs that were sometimes as much as 1000 times greater than SM predictions at the AUX output. As previously noted, this indicated that when making future predictions, the physical information encapsulated in static data was less influential on SM than the physical context provided by the PGL function. Our results suggested that the most useful application of the hybrid architecture (trained with the PGL function) for forecasts would be applications wherein seasonal trends are of interest.

3.4. Explainability

In this work, a hybrid ANN architecture predicted SMAP-like SM time series using meteorologic and hydrologic input data (features). In addition to assessing the ability of each model design to predict SM over a range of environmental characteristics, we also sought to address explainability (i.e., reasons why interactions between the data and model design resulted in a particular outcome). Overall, we found that physical processes governing SM variability were learned most robustly when the larger training dataset of meteorologic time series was used. In summary, LSTM output was largely consistent across model (SNGLT vs. MULTT) and experimental design configurations (with and without PGL), demonstrating that regardless of spatial variability, generalizability of the physical problem was achievable using only time series input data. Importantly, this revealed that SM variability, including daily SM and antecedent conditions, was learned independently of soil characteristics such as porosity, runoff potential, and hydraulic conductivity. These parameters are domain- and grid-specific in physics-based software like SWAT+, and often lead to calculations of SM that are model-specific indices of near-surface conditions at any given time step [12]. It is important to note that our data-driven design learned SM processes in a context that was reflective of "true" physics.

Daily SM predictions were often improved upon supplying the hybrid ANN with the static spatial data, as reflected in the decreased error from AUX to FINAL outputs. The exception to this finding occurred when using the PGL function during optimization with the MULTT model design, indicating that the SM estimates learned from the data were more robust than those informed

in conjunction with the PGL function. For the SNGLT model, application of the PGL function decreased error at the FINAL output, especially at the ELKR site, likely because the small training dataset did not include the variability necessary for the ANN to infer a site's physical processes. This result highlights a common weakness in physics-based modeling, where process specification and domain discretization are inescapably unable to capture the true complexity of the physical system. Physics-based model parameters are also simplified for the sake of computational tractability (e.g., zones of piecewise constancy). Using a data-driven framework, we predicted SM that was physically plausible and computationally inexpensive without numerically specifying physical processes.

4. Conclusions

This work has demonstrated that accurate SM predictions can be made using the same data input to SWAT+ (weather time series and hydrologic data), despite the fact that these data would not lead to accurate predictions from this physically based model because calibration of SWAT+ for SM was unsuccessful. This indicated that the neural network approach was better able to reproduce the complexity and variability of this dynamic system than even a computationally expensive physically based model.

Our results showed that critical knowledge of physical processes governing SM variability was learned by the LSTM network from meteorological time series and anomalies regardless of spatial characteristics. Model outputs from the hybrid network, which included both dynamic and static data types, were more sensitive to the variability and volume of input data. The model trained on a single watershed, SNGLT, was not able to generalize to new data within acceptable error thresholds, although errors associated with output from the LSTM were consistent for all study sites. For small volumes of data (<10,000 samples), we found that the additional physical context provided by the PGL function enhanced model performance for the FINAL output of the hybrid model. Future prediction of SM using the trained SNGLT model indicated overfitting to trends in the training data, most notably a drought in the late summer and fall of 2016. The MULTT model design predicted accurate SM for all watersheds when using the MSE loss function. Because the MULTT model was trained using data from all watersheds considered in the study, not only was the training dataset larger, it also comprised a wider variety of environmental conditions. For large volumes of data, the physical domain knowledge encapsulated in the PGL function provided weaker physical constraints than patterns learned from the dataset itself; however; for future predictions, the physical context provided by the PGL function was essential. Static data are most useful for applications in which the precision of SM predictions on a daily time step is important, such as risk analyses. We recommend that for such an application, the training dataset use weather-forecast time series so that the model can make informed forecasts with maximized accuracy.

Although SMAP provides daily predictions, it is advantageous to have near-real-time predictions within the 24-h latency of the SMAP SM retrieval algorithm. Risk assessments for droughts/floods, crop viability, and soil health depend on accurate SM quantification. Additionally, SMAP is a highly sophisticated machine, and like all such devices, will eventually fail or go offline. In the event of device failure, accurate, if provisional, SM predictions will be needed quickly and easily.

This work demonstrated the strengths and weaknesses of providing physical context during ML model training and explained under what circumstances the models performed best and why. Our approach incorporated understanding of the physical processes governing SM, most notably that SM is sensitive to current and antecedent meteorological conditions. Unlike SWAT+, the ANNs developed in this work predicted SM that was not dependent upon model-specific static quantities such as porosity and hydraulic conductivity. In some cases, the ANN SM predictions were enhanced using static data, although our results showed that the majority of physical learning was acquired from meteorological data (i.e., primary drivers for near-surface hydrologic processes). We are confident that

SM predicted using this hybrid model framework would be suitable for input to other models as a replacement for model-specific SM indices, although more work is required to verify this.

Supplementary Materials: All data and computer code are available online at: https://storage.googleapis.com/breen_make_supplementary/Breen_supplementary.7z. The volume of data is quite large. If interested in downloading Python scripts only, the file Breen_supplementary_scripts.7z may be downloaded directly from the MDPI webpage http://www.mdpi.com/2504-4990/2/3/16/s1.

Author Contributions: Conceptualization of methodology and application, K.H.B., S.C.J., J.D.W., and P.M.A.; data curation, software development, and formal analysis, K.H.B.; supervision and investigation, S.C.J. and J.D.W.; SWAT+ data acquisition, J.G.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a Graduate Teaching Fellowship from Baylor University, a summer 2019 NASA internship at Goddard Space Flight Center, and two fellowships at the Institute for Pure and Applied Mathematics at UCLA in the Fall of 2018 and 2019.

Acknowledgments: This research was supported by a Graduate Teaching Fellowship from Baylor University, a summer 2019 NASA internship at Goddard Space Flight Center, and two fellowships at the Institute for Pure and Applied Mathematics (IPAM) at UCLA in the Fall of 2018 and 2019. The authors would like to extend special thanks to the IPAM organizers, participants, and staff for fostering a multi-disciplinary environment that provided the authors with priceless learning opportunities.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Roscher, R.; Bohn, B.; Duarte, M.F.; Garcke, J. Explainable machine learning for scientific insights and discoveries. *IEEE Access* **2020**, *8*, 42200–42216. [CrossRef]
2. Karpatne, A.; Watkins, W.; Read, J.; Kumar, V. Physics-guided neural networks (PGNN): An application in lake temperature modeling. *arXiv* **2017**, arXiv:1710.11431.
3. Von Rueden, L.; Mayer, S.; Garcke, J.; Bauckhage, C.; Schuecker, J. Informed machine learning—Towards a taxonomy of explicit integration of knowledge into machine learning. *Learning* **2019**, *18*, 19–20.
4. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics informed deep learning (Part II): Data-driven discovery of nonlinear partial differential equations. *arXiv* **2017**, arXiv:1711.10566.
5. Mosavi, A.; Ozturk, P.; Chau, K.W. Flood prediction using machine learning models: Literature review. *Water* **2018**, *10*, 1536. [CrossRef]
6. Hunt, J. Floods in a changing climate: A review. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **2002**, *360*, 1531–1543. [CrossRef] [PubMed]
7. Kang, Y.; Khan, S.; Ma, X. Climate change impacts on crop yield, crop water productivity and food security—A review. *Prog. Nat. Sci.* **2009**, *19*, 1665–1674. [CrossRef]
8. Godfray, H.C.J.; Beddington, J.R.; Crute, I.R.; Haddad, L.; Lawrence, D.; Muir, J.F.; Pretty, J.; Robinson, S.; Thomas, S.M.; Toulmin, C. Food security: The challenge of feeding 9 billion people. *Science* **2010**, *327*, 812–818. [CrossRef]
9. Ramankutty, N.; Mehrabi, Z.; Waha, K.; Jarvis, L.; Kremen, C.; Herrero, M.; Rieseberg, L.H. Trends in global agricultural land use: Implications for environmental health and food security. *Annu. Rev. Plant Biol.* **2018**, *69*, 789–815. [CrossRef]
10. Neitsch, S.; Arnold, J.; Kiniry, J.; Williams, J. *Soil & Water Assessment Tool Theoretical Documentation. Version 2009*; TWRI Report. Technical Report, TR-406; Texas Water Resource Institute: College Station, TX, USA, 2011.
11. Beven, K.; Kirkby, M.; Schofield, N.; Tagg, A. Testing a physically-based flood forecasting model (TOPMODEL) for three UK catchments. *J. Hydrol.* **1984**, *69*, 119–143. [CrossRef]
12. Koster, R.D.; Guo, Z.; Yang, R.; Dirmeyer, P.A.; Mitchell, K.; Puma, M.J. On the nature of soil moisture in land surface models. *J. Clim.* **2009**, *22*, 4322–4335. [CrossRef]
13. Beven, K. Changing ideas in hydrology—The case of physically-based models. *J. Hydrol.* **1989**, *105*, 157–172. [CrossRef]
14. Bergen, K.J.; Johnson, P.A.; Maarten, V.; Beroza, G.C. Machine learning for data-driven discovery in solid Earth geoscience. *Science* **2019**, *363*, eaau0323. [CrossRef] [PubMed]
15. Nilawar, A.P.; Calderella, C.P.; Lakhankar, T.Y.; Waikar, M.L.; Munoz, J. Satellite soil moisture validation using hydrological SWAT model: A case study of Puerto Rico, USA. *Hydrology* **2017**, *4*, 45. [CrossRef]

16. Zhuo, L.; Han, D.; Dai, Q. Soil moisture deficit estimation using satellite multi-angle brightness temperature. *J. Hydrol.* **2016**, *539*, 392–405. [CrossRef]
17. Breaker, L.C.; Rao, D.B.; Kelley, J.G.; Rivin, I. Development of a real-time regional ocean forecast system with application to a domain off the U.S. East Coast. *Mar. Technol. Soc. J.* **2004**, *38*, 61–79. [CrossRef]
18. Fereidoon, M.; Koch, M.; Brocca, L. Predicting rainfall and runoff through satellite soil moisture data and SWAT modelling for a poorly gauged basin in Iran. *Water* **2019**, *11*, 594. [CrossRef]
19. Noé, F.; Olsson, S.; Köhler, J.; Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **2019**, *365*, eaaw1147. [CrossRef]
20. Riley, P. Three Pitfalls to Avoid in Machine Learning. 2019. Available online: https://www.nature.com/articles/d41586-019-02307-y (accessed on 20 November 2019).
21. Breen, K.; White, J.; James, S. Are extreme soil moisture deficits captured by remotely sensed data retrievals? *Int. J. Remote Sens. Lett.* **2020**, *11*, 767–776. [CrossRef]
22. Colliander, A.; Jackson, T.J.; Bindlish, R.; Chan, S.; Das, N.; Kim, S.; Cosh, M.; Dunbar, R.; Dang, L.; Pashaian, L.; et al. Validation of SMAP surface soil moisture products with core validation sites. *Remote Sens. Environ.* **2017**, *191*, 215–231. [CrossRef]
23. Sun, L.; Seidou, O.; Nistor, I.; Goïta, K.; Magagi, R. Simultaneous assimilation of in situ soil moisture and streamflow in the SWAT model using the Extended Kalman Filter. *J. Hydrol.* **2016**, *543*, 671–685. [CrossRef]
24. Tobin, K.J.; Bennett, M.E. Constraining SWAT calibration with remotely sensed evapotranspiration data. *JAWRA J. Am. Water Resour. Assoc.* **2017**, *53*, 593–604. [CrossRef]
25. Rhanoui, M.; Mikram, M.; Yousfi, S.; Barzali, S. A CNN-BiLSTM model for document-level sentiment analysis. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 48. [CrossRef]
26. Zhang, X.; Zhang, Q.; Zhang, G.; Nie, Z.; Gui, Z.; Que, H. A novel hybrid data-driven model for daily land surface temperature forecasting using long short-term memory neural network based on ensemble empirical mode decomposition. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1032. [CrossRef]
27. Rahman, S.; Rasheed, A.; San, O. A hybrid analytics paradigm combining physics-based modeling and data-driven modeling to accelerate incompressible flow solvers. *Fluids* **2018**, *3*, 50. [CrossRef]
28. Bilgera, C.; Yamamoto, A.; Sawano, M.; Matsukura, H.; Ishida, H. Application of convolutional long short-term memory neural networks to signals collected from a sensor network for autonomous gas source localization in outdoor environments. *Sensors* **2018**, *18*, 4484. [CrossRef]
29. Leontjeva, A.; Kuzovkin, I. Combining static and dynamic features for multivariate sequence classification. In Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 17–19 October 2016; pp. 21–30. [CrossRef]
30. Entekhabi, D.; Yueh, S.; O'Neill, P.; Kellogg, K.; Alen, A.; Bindlish, R.; Brown, M.; Chan, S.; Colliander, A.; Crow, W.T.; et al. *SMAP Handbook*; JPL Publication: Pasadena, CA, USA, 2014.
31. James, S.C.; Zhang, Y.; O'Donncha, F. A machine learning framework to forecast wave conditions. *Coast. Eng.* **2018**, *137*, 1–10. [CrossRef]
32. DeVries, P.M.; Thompson, T.B.; Meade, B.J. Enabling large-scale viscoelastic calculations via neural network acceleration. *Geophys. Res. Lett.* **2017**, *44*, 2662–2669. [CrossRef]
33. Ahmmed, B.; Mudunuru, M.; Karra, S.; James, S.; Vesselinov, V. A comparative study of machine learning models for predicting the state of reactive mixing. *arXiv* **2020**, arXiv:2002.11511.
34. Nakano, K.; Chakraborty, B. Effect of data representation for time series classification—A comparative study and a new proposal. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 62. [CrossRef]
35. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
36. Greff, K.; Srivastava, R.K.; Koutník, J.; Steunebrink, B.R.; Schmidhuber, J. LSTM: A search space odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *28*, 2222–2232. [CrossRef] [PubMed]
37. Gers, F.A.; Schmidhuber, J.; Cummins, F. Learning to forget: Continual prediction with LSTM. In Proceedings of the 1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470), Edinburgh, UK, 7–10 September 1999; Volume 2, pp. 850–855. [CrossRef]
38. Jozefowicz, R.; Zaremba, W.; Sutskever, I. An empirical exploration of recurrent network architectures. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2342–2350.

39. Bieger, K.; Arnold, J.G.; Rathjens, H.; White, M.J.; Bosch, D.D.; Allen, P.M.; Volk, M.; Srinivasan, R. Introduction to SWAT+, a completely restructured version of the soil and water assessment tool. *JAWRA J. Am. Water Resour. Assoc.* **2017**, *53*, 115–130. [CrossRef]
40. Koster, R.D.; Suarez, M.J. Soil moisture memory in climate models. *J. Hydrometeorol.* **2001**, *2*, 558–570. [CrossRef]
41. Nanopoulos, A.; Alcock, R.; Manolopoulos, Y. Feature-based classification of time-series data. *Int. J. Comput. Res.* **2001**, *10*, 49–61.
42. Esling, P.; Agon, C. Time-series data mining. *ACM Comput. Surv. (CSUR)* **2012**, *45*, 12. [CrossRef]
43. Aigner, S.; Körner, M. The importance of loss functions for increasing the generalization abilities of a deep learning-based next frame prediction model for traffic scenes. *Mach. Learn. Knowl. Extr.* **2020**, *2*, 6. [CrossRef]
44. Hoffer, E.; Hubara, I.; Soudry, D. Train longer, generalize better: Closing the generalization gap in large batch training of neural networks. *arXiv* **2017**, arXiv:1705.08741v2.