Integrating EOFs into Machine Learning Algorithms to Emulate Climate Land Model

By

Kachinga Silwimba (kachingasilwimba@u.boisestate.edu)
Boise State University (BSU), Idaho, USA

Synthesis Article

April 2022

COMPREHENSIVE EXAMINATION
Advisor: Dr. Alejandro Flores
Committee Members: Dr. Jodi Mead, Dr. Anna Bergstrom, Dr. Catherine Olschanowsky

1. Introduction

1.1 Soil and Patterns

Land ecosystems significantly control the Earth's climate system through biophysical and biogeochemical feedback mechanisms. The mechanisms are primarily affected by changes in land surface albedo, sensible heat and latent heat exchanges with the atmosphere, and net land surface fluxes of greenhouse gases (GHG) (Sargsyan et al., 2014). Soil is the essential component of land that exerts control on the climate system by partitioning water, energy, and carbon exchanges between the land and atmosphere. Additionally, soil is the realm in which the atmospheric and hydrologic processes are connected with the biosphere and, therefore, play a vital role in sustaining life on Earth. Additionally, soil serves as a solid waste landfill, a wastewater filter, and the foundation for our cities, communities, and agricultural ecology. Soil has various properties that differ from place to place (Webster, 1985). The variations in the soil type depends on the two main climate factors precipitation and temperature. The amount of precipitation controls how much water permeates the soil, and salts and minerals dissolve in the water and move with it. Climate and temperature also influence the types of plants and other organisms that reside in the soil. Moreover, the water contained in the soil is called soil moisture is a crucial state variable in hydrology due to its various applications in the field and its influence on runoff production (Perry and Niemann, 2007). According to Western et al. (1999), insufficient representations of soil moisture patterns in hydrological models result in significant errors in the predicted runoff.

The spatio-temporal patterns of soil moisture considerably influences the responsiveness of stream discharge and rainfall events. Han et al. (2014) emphasized soil moisture content and temperature status significantly impact the hydrological and thermodynamic processes involving water and energy movement between the land surface and the atmosphere. These state variable's space-time dynamics depend on the hydraulic and thermal conductivity of the soil as well as other material characteristics. The presence of moist soil conditions is a crucial factor in the occurrence of floods. Plant growth and evapotranspiration are also influenced by soil moisture in areas where local evapotranspiration accounts for a significant portion of precipitation (i.e., where precipitation is high), soil moisture conditions can also contribute to the persistence of dry and wet spells (Kitanidis and Bras 1980; Eltahir and Bras 1994). Areas with high soil moisture that are well connected to channels have seen more discharge and erosion (Ntelekos et al., 2006). Despite this, Jaynes et al. (2003) showed that the geographical patterns of plant development and agricultural yield also reflect the spatiotemporal fluctuation of soil moisture. Additionally, soil moisture patterns appear different at different resolutions, and it is of great interest to study both temporal and spatial aspects of these patterns.

To gain a better understanding of soil moisture, both Wang et al. (2017) and Perry and Niemann (2007) analyzed soil moisture in a way that accounts for spatial variation and temporal evolution of distinct underlying patterns that has been difficult in the past by using Empirical Orthogonal Functions (EOF), which is a data-based method. The output of this analysis is a small number of spatial structures (EOF) that account for a substantial portion of the dataset's variation and temporal variable coefficients, which alter the influence of these spatial structures across time. Yoo and Kim (2004) used EOF to characterize soil moisture for agriculture sites while Jawson and Niemann (2007) characterized soil moisture at a large scale and their dependence on soil,

land-use, and topographic properties. Nevertheless, EOF decomposition has proved to be helpful in analyzing climate data. We will discuss further in section 4 how they can be significant as input in the machine learning (ML) model to emulate the climate model, which is of great interest in predicting soil features and other climate features.

1.2 Climate Model Emulation

According to Dagon et al. (2020) modeling plays a vital role in understanding, projecting, and predicting the land surface dynamics in the Earth system. The representation of soil in Earth System Models, like the Community Earth System Model (CESM), is an essential facet in modeling the response of the Earth System to climate change. Since their inception, land models have grown to represent critical processes like carbon cycling, ecosystem dynamics, terrestrial hydrology, and agriculture and serve as a lower boundary condition for atmospheric general circulation models. With increasing process representation they are computationally expensive. Additionally, several parameterization schemes are used by hydrologists and modelers to describe the water and energy balance (Han et al. 2012; Liang et al. 1994; Oleson et al. 2010; Sellers et al. 1996). Even though parameterization approaches are required since there are not enough computational resources to compute all length and time scales, they also cause significant bias and uncertainties in the data simulations (e.g., Bechtold et al. 2008; Farneti and Gent 2011; Wilcox and Donner 2007). Due to the uncertainties associated with model parameters, both Reichle et al. (2004) and Kato et al. (2007) were able to observe land models to be strongly limited in their ability to reproduce the observed soil moisture dataset.

The combination of EOF decomposition and ML approaches offer a potentially powerful way to reduce the dimensionality of datasets to a key subset of spatiotemporal modes in order to more efficiently ascertain model parameters, reducing significantly the number of full model runs that need to be performed. Climate models, such as the aforementioned CESM, are very computationally expensive to execute and frequently only performed in coordinated multinational studies designed to investigate specific scientific problems due to the wide range of spatial and temporal dimensions and a vast number of processes being modeled. Moreover, they generate enormous amounts of data, making it challenging to analyze and interpret using conventional tools and techniques. However, the use of ML to lower the computational cost of producing this data to extract more value from it after it has been created is of great interest in geoscientific research (Huntingford et al. 2019; Reichstein et al. 2019). To that effect, Costabile et al. (2012) stated that building and calibrating physically based models is challenging since they often require enormous amounts of data and physical parameters to represent the hydrological processes.

The development of new parameterization schemes based on ML is a potential step forward that can be achieved by fitting a statistical model using the output of relatively expensive physical models that better depict the subgrid dynamics (O'Gorman and Dwyer, 2018). To better understand the land model uncertainty due to parameter variation choices, Dagon et al. (2020) emulated a subset of Community Land Model, version 5 (CLM5) output using a series of artificial feed-forward neural networks to obtain optimal parameter values. The obtained optimal parameters were then tested on the global scale with CLM5 to assess the predictive skills of the model. As a result, the framework demonstrated a way to increase the computational efficiency

and reduce uncertainties in the land model originating from the parameters. Watson-Parris et al. (2021) designed a workflow shown in Figure 1.1 for the model calibration and emulation of models and outputs using the ESEm version 1.0.0 a scalable earth system emulator to validate it so that it can be used for the inference, while utilizing Community Intercomparison Suite (CIS) which is a user-friendly command line program that has been created to make it simple to compare data from models, in situ data, and remote sensing.

Deep neural networks (DNN) and ML are becoming increasingly important in scientific research. Reichstein et al. (2019) stated that using computationally expensive, physics-based models might be replaced in some circumstances by ML and DNN models. On the other hand, the authors suggested implementing a hybrid modeling approach, which couples physical process models with the versatility of data-driven ML. In this work, we synthesize ongoing work and suggest new approaches to the emulation, prediction, and analysis of the Land Model soil output using ML and Empirical Orthogonal Functions (EOF).

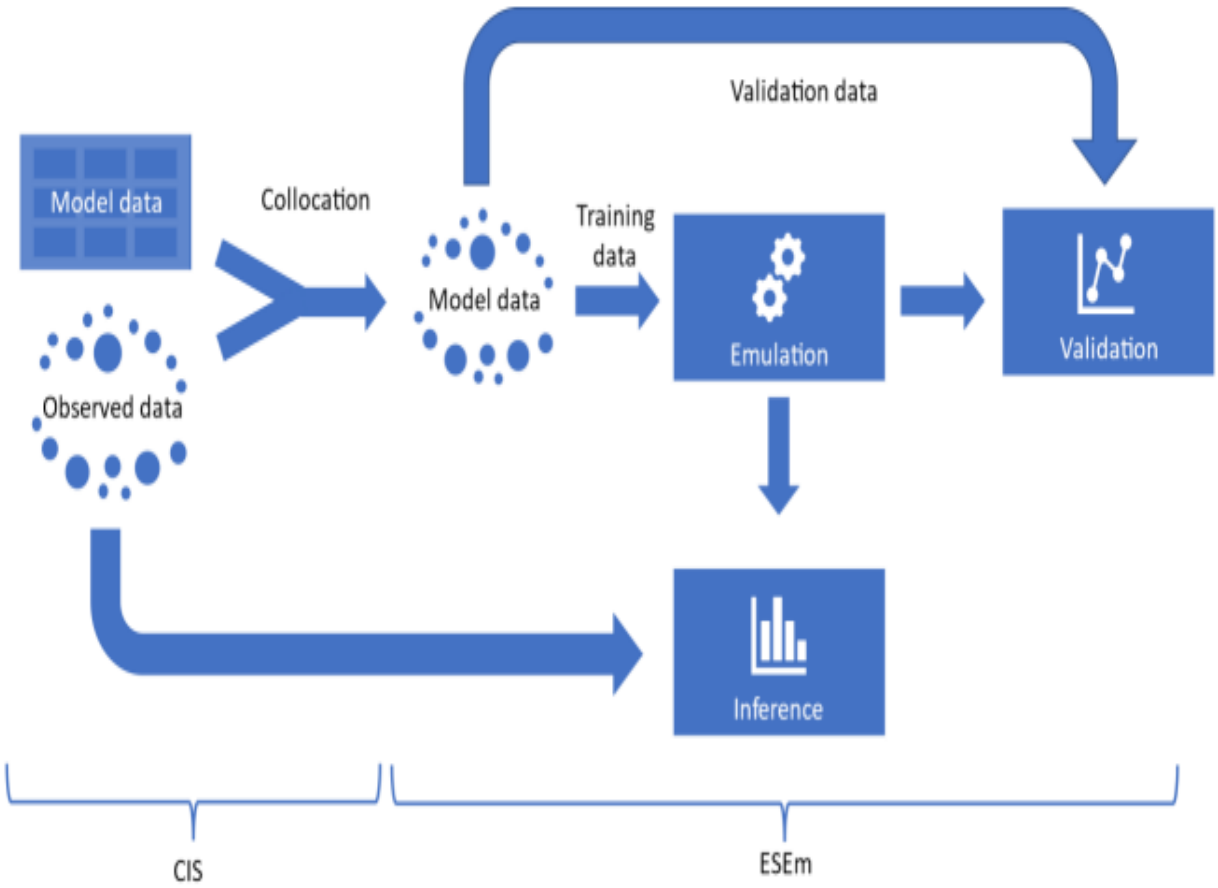

Figure 1.1: A schematic of a typical workflow using Community Intercomparison Suite (CIS) and ESEm version 1.0.0 to perform model emulation and calibration (Watson-Parris et al., 2021).

2. Climate Model Emulation Applications

An aim of emulator or surrogate models in hydrological simulations is to imitate climate models, parameter estimates more efficiently to reduce errors, and other physical properties in the geoscientific research.

2.1 Machine Learning in Geoscientific Research

The 1990s saw the publication of the first studies utilizing neural networks for meteorological and air quality applications (Schizas et al. 1991; Comrie 1997). These studies analyzed and predicted time series at specific station locations using multi-layer perceptron designs, typically with three layers. Furthermore, various simple semantic network algorithms were employed for post-processing and prediction optimization of numerical weather prediction output (Krasnopolsky and Lin 2012; Rasp and Lerch 2018), and as surrogate models for various parameterization techniques in climate models (Krasnopolsky and Fox-Rabinovitz 2006; Krasnopolsky et al. 2002).

Surrogate models statistically relate the input and output dataset computed through execution of the complex system simulation (Davis et al., 2017). The aim of developing surrogate models in climate science is to emulate computationally expensive Earth System Models (ESM) (e.g., Climate Land Models) simulations using ML and artificial intelligence algorithms. Various authors have used the ML algorithms for parameterization to obtain better parameters for General Circulation models to reduce uncertainties in the climate model (e.g., Dagon et al. 2020; O'Gorman and Dwyer 2018; Ricciuto et al. 2018). Another promising aspect of using ML algorithms is that they can be used to gain insight from the large multidimensional dataset into underlying physical processes (O'Gorman and Dwyer, 2018).

2.2 Biophysical Parameter Estimation

In order to examine how differences in parameter values affect model prediction, parametric uncertainty in land models has historically been investigated by experimentation with various parameter values (e.g., Gao et al. 2021; Hawkins et al. 2019; Huo et al. 2019; Dagon et al. 2020; Ricciuto et al. 2018). When the spatial domain is broad or global, hand tuning parameter values can be computationally inefficient, requiring numerous model simulations and a lot of processing time. However, it takes less computational time to simulate a single point, but the simulations' results are potentially not transferable to other regions. The solution to this problem is to use ML approaches to build a surrogate model that emulates the climate model behavior. This approach allows parameters to be optimized and tested quickly without executing the full model to analyze the parameter space more effectively (Sanderson et al., 2008). In this section, we explore how an emulator model can be used to emulate the climate models. To explain this, Dagon et al. (2020) research paper and other supporting works of literature that have employed emulator models for the parameterization of global circulation models will be synthesized.

To understand the land model uncertainty, Dagon et al. (2020) used artificial feed-forward neural networks (FFNN), a type of ML algorithm that learns the relationship between the input and

output data to explore the impact of the parameter selections on overall model uncertainty by exploring the CLM5 biophysical parameter space and identifying sensitive parameters. A FFNN was trained on the perturbed parameter ensemble (PPE) to emulate the subset of the CLM5 output. Moreover, the supervised neural networks were trained using known model parameter values. Also, the output of the simulation from the perturbed parameter ensemble (PPE) was generated using the Latin hypercube (LHC) (McKay et al., 2000) sampling technique to simulate 100 different parameters sets for each of the six CLM5 parameters listed in Table 2.1. These parameters were selected by conducting a parameter sensitivity simulation to determine and narrow down key parameters that can be utilized in the emulation and parameterization step. Two methods, parameter effect (PE) and pattern correlation calculation (PC), were employed in selecting valuable parameters. The parameter variations characterize the plant functional type (PFTs), which is the plant physiology in CLM partitioned in 15 batches plus bare ground. Using the LHC-generated parameters, they performed 100 run simulations of CLM5. The resulting output from the model, along with the scaling values for the parameter, were used to develop and train a collection of neural networks that mimic the land model. Speaking of emulation, Ricciuto et al. (2018) constructed the Polynomial Chaos surrogate employing the Bayesian Compressive Sensing algorithm to conduct a global sensitivity analysis (GSA) in a high dimensional land surface model to determine the efficient global parameter. Among the set of parameters tested, 20 were identified as sensitive, and all depended on the PFTs.

Table 2.1: CLM5 candidate parameters selected based on sensitivity tests (Dagon et al., 2020).

Parameter name	Description	Varies with PFT?
medlynslope	Slope of stomatal conductance–photosynthesis relationship	Yes
dleaf	Characteristic dimension of leaves in the direction of wind flow	Yes
kmax	Plant segment maximum conductance	Yes
ffff	Decay factor for fractional saturated area	No
dint	Fraction of saturated soil for moisture value at which dry surface layer initiates	No
baseflow_scalar	Scalar multiplier for base flow rate	No

The FFNN were trained to replicate the output of the perturbed parameter ensemble (PPE simulations) outlined above. Figure 2.1 shows a series of networks used with completely interconnected layers, each layer comprising several nodes. The six normalized parameter scaling values (p_i) needed to create the CLM PPE are the input values for the networks in the first layer, known as the input layer. The final layer, or the CLM output to forecast, is the output layer. According to Mitchell and Mitchell (1997) the hidden layers between the input and output layers contain different numbers of nodes. Through the specific activation functions, input values were transformed into values at each node in the hidden layer (n_j^1 and n_k^2), and the output layer (z_m).

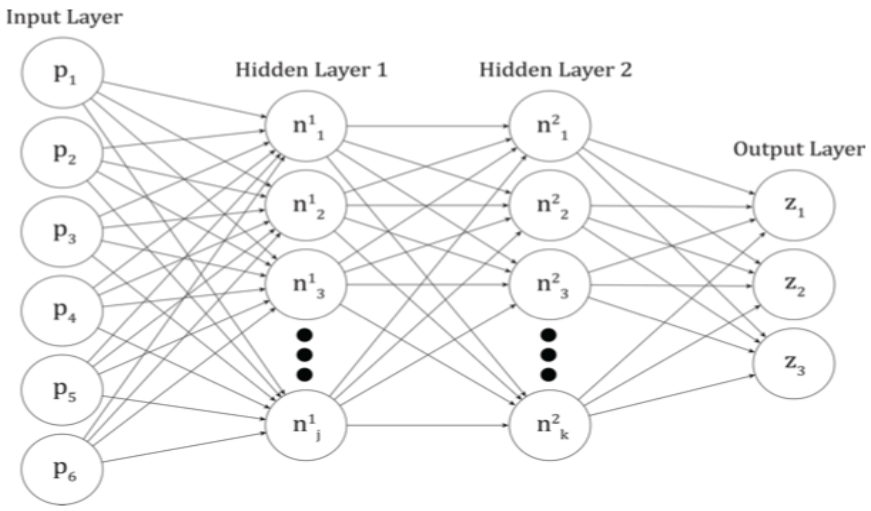

Figure 2.1: FFNN framework where p_i is the inputs as the land model parameter, z_m are the three outputs, and also two hidden layers with j and k nodes respectively (Dagon et al., 2020).

The trained neural network in Figure 2.1 is supervised learning, and the output values used to train the networks were the global five-year mean maps derived from the PPE simulations. To reduce the dimensionality of the output variables gross primary production (GPP) and latent heat flux (LHF) while conserving the information in the dataset, EOF and PCs (Lorenz 1970; Jollife 2002) were calculated. The decomposition was performed using the singular value decomposition (SVD) procedure discussed in Chapter 3. Similarly, the global gridded observation dataset FLUXNET-MTE product (Jung et al., 2011) was used as the target variable was gridded to match the CLM output resolution. FLUXNET-MTE data set provides global, gridded estimates of surface-to-atmosphere fluxes with monthly resolution and a spatial resolution of $0.5^\circ \times 0.5^\circ$ longitude and latitude (Swenson and Lawrence, 2014). FLUXNET-MTE estimates surface-atmosphere fluxes from a set of explanatory variables representing meteorological and land surface conditions using a machine learning technique (Model Tree Ensemble, or MTE) (Jung et al., 2011). The observed anomalies $\mathbf{X}_{\text{obs}}$ were calculated and projected into the EOF space $\mathbf{\Gamma}$ and $\mathbf{U}^*$ computed from the SVD in section 3 to produce observational estimates ($\mathbf{U}_{\text{obs}}$) with similar metrics as the PPE simulations used to train the neural networks. The variables considered were GPP and LHF, and observational estimates ($\mathbf{U}_{\text{obs}}$) were considered as a target defined in Eq. (2.2.1). However, the importance of decomposition of the data to be used as input in machine learning is discussed in section 3 of this synthesis paper.

$$\mathbf{U}_{\text{obs}} = \mathbf{X}_{\text{obs}} \times (\mathbf{\Gamma} \times \mathbf{U}^*)^{-1} \tag{2.2.1}$$

Moreover, considering two output variables (v) (GPP and LHF) and three modes (m) of variability, they optimized parameter values that minimize the error concerning the observational estimates. Six targets were obtained for the parameter calibration. The cost function $J(p)$ used (defined in Eq. (2.2.2)) determines the predictive model skills, which is based on the three modes of the weighted sum of the first three modes of the squared differences between emulator predictions and observations, where $\hat{U}_{v,m}$ is the prediction from the neural network emulator for a given parameter set while U_{obs} is the observational targets computed in Eq. (2.2.1) and $\sigma(U_{\text{obs}^*,v,m})$

is the standard deviation that accounts for the natural variability across all observational years. Each variable's total is weighted according to the percentage of variation that each EOF mode $(\lambda_{v,m})$ explains.

$$J(p) = \sum_{v=1}^{2} \left[\sum_{m=1}^{3} \lambda_{v,m} \left(\frac{\hat{U}_{v,m}(p) - U_{\text{obs},v,m}}{\sigma(U_{\text{obs}^*,v,m})} \right)^2 \right] \tag{2.2.2}$$

The CLM test case was performed using optimized parameter values and the same model setup as before. Dagon et al. (2020) concluded with Figure 2.2 a summary of the findings for the first three modes of GPP and LHF. The plots display the performance comparison of the models with the FLUXNET observational data. The findings for PC1, especially for PC1 GPP, demonstrated that the emulator forecasts (blue bars) and CLM test case results (green bars) were quite similar. Furthermore, due to the higher mode's noise-prone characteristics, the CLM test case did not capture the optimized predictions precisely. The findings confirm what Brenowitz and Bretherton (2018) and Rasp et al. (2018) elaborated that the parameterization might alternatively be entirely replaced by ML, which could learn directly from high-resolution simulations, while Reichstein et al. (2019) argued that data-driven ML methods in geoscientific research would not replace physical modeling. However, they will significantly enhance and complement them.

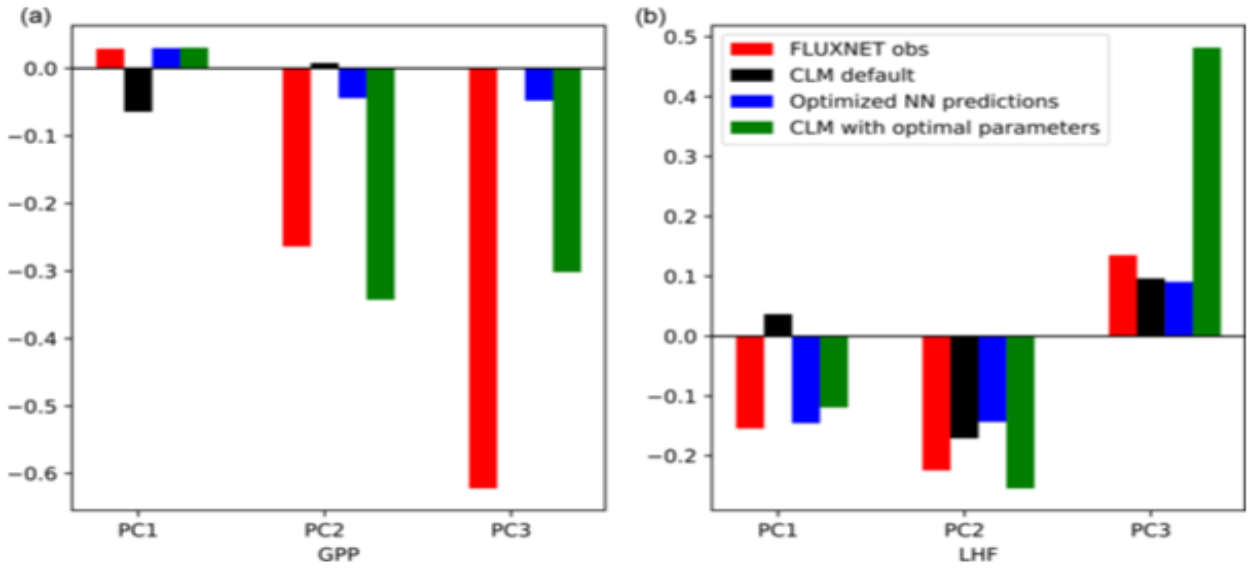

Figure 2.2: PC1, PC2, and PC3 for GPP (a) and LHF (b) comparing observations, model simulations, and emulator predictions. Observational estimates from FLUXNET shown in red bars, CLM default values shown in black bars, the optimized NN predictions shown in blue bars, and the results of the CLM test case with optimal parameters in green bars. (Dagon et al., 2020).

2.3 Predicting Hydrological Drivers using Emulators

In this section, we discuss how machine learning can be used to predict hydrological drivers. Due to their many benefits, ML models have been widely used in numerous hydrological simulations, including groundwater simulations (Mohanty et al., 2013), soil moisture estimation (Karthikeyan

and Mishra, 2021), stream flow level forecasts, and water quality modeling (Liu and Lu, 2014). ML models often require fewer hydrological parameters than physically based models because data-driven ML models can immediately bridge the mappings between hydrological drivers (such as precipitation) and responses (such as streamflow) without explicitly describing the hydrological processes (Tigkas et al., 2016). Most ML models have been used in predicting hydrological drivers in geological science. Wu et al. (2022) built a rainfall-runoff model to simulate streamflow using support vector regression(SVR), long short-term memory (LSTM) multilayer perceptron (MLP), and gradient boosting regression tree (GBRT). The designed framework of the models was incorporated with the EOF analysis to extract insight from a gridded climate dataset and used as predictands of the ML models as discussed in section 4. Moreover, the authors compared the performance of the SVR, MLP, LSTM, and GBRT in predicting streamflow.

Karthikeyan and Mishra (2021) trained the XGBoost models region-wise and layer-wise to estimate the multilayer soil moisture at five different depths of soil layers at a 1km spatial resolution. The algorithm was deployed to homogeneous areas to capture the complex relationship between relevant predictor factors and in-situ soil moisture at various layers over the Contaguous United States (CONUS). Ham et al. (2019) achieved skillful seasonal forecasts for El Niño and the Southern Oscillation (ENSO) events through applying the convolutional neural networks (CNN) model to the historical climate model simulations. Additionally, cluster analysis was implemented by Gibson et al. (2021) to focus on the larger, more predictable spatial precipitation patterns to increase the observed seasonal and identify the physical process contributing to the prediction skills. Moreover, they utilized the EOF analysis to decompose the gridded dataset. Despite that, Prakash et al. (2018) used multiple linear regression (MLR), recurrent neural network (RNN), and SVR to predict the soil moisture in advance, and they observed good performance in the model prediction skill for 1 and 2 days ahead and emphasized on improving the models by using other techniques. This could be helpful to farmers in adjusting their farming strategies without the need to run a climate model.

2.4 Performance Measurements of Emulators

Climate model emulation adds another source of uncertainty to projections, and this requires a robust quantification for the prediction to be relevant (Watson-Parris, 2021). The performance and forecasting reliability of the machine learning models can be evaluated using different metrics. According to Doycheva et al. (2017) and Patel and Ramachandran (2015), the commonly recommended metrics for validating the hydrological models are Percent bias (Pbias) Eq. (2.4.1) assesses how often simulated data has higher or smaller magnitudes than the observed data. Positive Pbias numbers signify overestimation, whereas negative values signify underestimating. Zero is the ideal value. Nash-Sutcliffe efficiency (NSE) Eq. (2.4.2) is frequently used to evaluate the accuracy of hydrological models. However, it can also be used to measure and describe the accuracy of model outputs. NSE has a range of negative infinity to 1. Therefore, when the observed mean is a better predictor than the model output, the efficiency is less than 0. conversely, an efficiency of 1 indicates that the model and the data are perfectly matched. The Mean Absolute Error (MAE) Eq. (2.4.3) determines the absolute difference between the actual and predicted values, aggregates these differences, and divides the total by the number of observations. A

model is perfect if its MAE value is 0. This suggests a symmetry between the actual and the predicted. The Root Mean Square Error (RMSE) Eq. (2.4.4) is the square root of the mean square of all the error measures for numerical predictions. It is also the standard deviation of the residuals (prediction errors) and measures how spread out these residuals are. The residual measures how far the data points are from the regression line. Where y_i^{obs} and y_i^{sims} are observed and predicted values respectively while $\overline{y_i^{\text{obs}}}$ is the mean of the observed values. Dagon et al. (2020) used the mean squared error (MSE) for the validation of the model used in the training of the emulator. However, Mandeville et al. (1970) stated that one of the best hydrological measures for evaluating the effectiveness of a hydrology model's fitting is the NSE.

$$\text{Pbias} = \frac{\sum_{i=1}^{N}\left(y_i^{\text{sim}} - y_i^{\text{obs}}\right)}{\sum_{i=1}^{N} y_i^{\text{obs}}} \times 100\% \tag{2.4.1}$$

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{N}\left(y_i^{\text{obs}} - y_i^{\text{sim}}\right)^2}{\sum_{i=1}^{N}\left(y_i^{\text{obs}} - \overline{y_i^{\text{obs}}}\right)^2} \tag{2.4.2}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\left(y_i^{\text{obs}} - y_i^{\text{sim}}\right)\right| \tag{2.4.3}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i^{\text{obs}} - y_i^{\text{sim}}\right)^2} \tag{2.4.4}$$

Wu et al. (2022) compared the performance of the ML model in predicting streamflow in Yingluoxia, Hongshuiba River, Fengle River, and Taolai River watershed using the NSE metric shown in Figure 2.3. The results showed higher NSE values when the EOF was used as the predictor of the streamflow, while lower NSE values were observed for the analysis without the EOF in all the four watersheds considered.

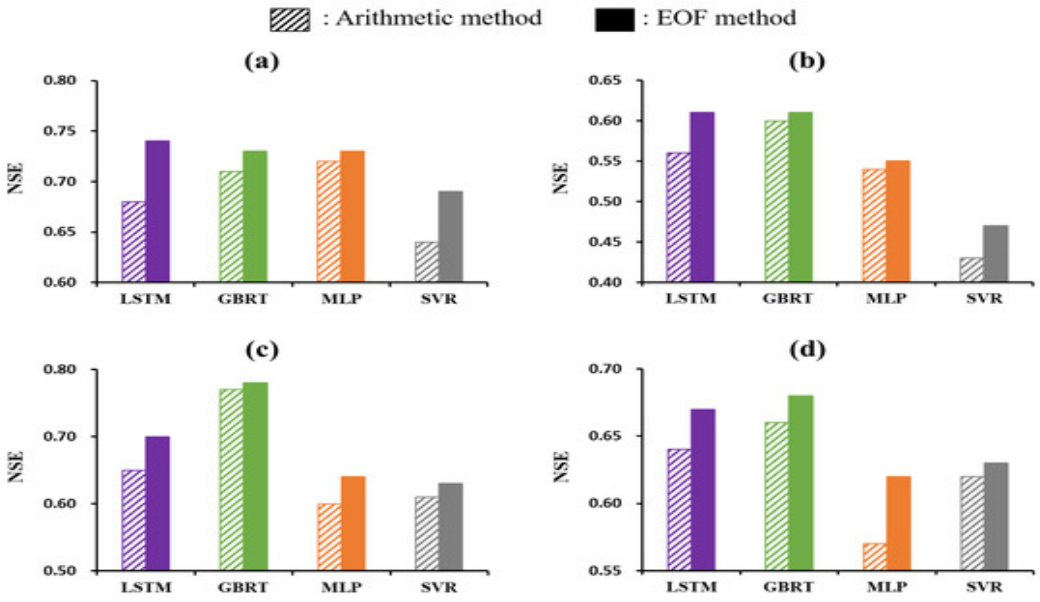

Figure 2.3: Model performance evaluation of prediction models (Wu et al., 2022)

3. Empirical Orthogonal Functions Analysis

3.1 A Brief Historical Background

Empirical Orthogonal Functions (EOF) analysis is a statistical technique that is frequently used to analyze large multidimensional datasets (Jollife, 2002). The EOF analysis is sometimes referred to as the principal component analysis in other fields of study. EOF break down the observed variability of a dataset into a collection of orthogonal spatial patterns (EOF) and a collection of time series referred to as expansion coefficients (ECs) or principal components (PCs). The history of application of EOF in the climate science dates back in the 1940s where Obukhov (1947), Lorenz (1970), and Kutzbach (1967) used the method. EOF were originally designed to decompose a continuous space-time field $X(t, \mathbf{s})$, where t and $\mathbf{s}$ represents time and spatial positions respectively. M represents the number of modes in the field applying the space $u_k(\mathbf{s})$ of the optimal set of basis functions and also the expansion functions of time $c_k(t)$.

$$\mathbf{X}(t, \mathbf{s}) = \sum_{k=1}^{M} c_k(t) u_k(\mathbf{s}) \tag{3.1.1}$$

In practice, EOF aim to identify new variables that represent the majority of the variance observed from the data through linear combinations of the original variables. Today, most meteorological centers regularly compare observations, reanalyses, and climate model simulations using EOF analysis (Hannachi et al. 2007; Wang et al. 2017). Moreover, the vast majority of the climate data covers enormous geographic areas and is either observed or modeled over long periods. Therefore, To investigate the variability of patterns in the large multidimensional climate model output, EOF analyses are often used (e.g, Dawson 2016; Wilks 2006). However, Wang et al. (2017) used the EOF as the diagnostic tool to evaluate soil and climate effects on regional soil moisture spatial variability. The authors identified the dominant spatial pattern in the soil for the three regions with different climate regimes across the continental U.S. used in the study.

3.2 Data Formatting

We consider a space-time field $\mathbf{X}(t, \mathbf{s})$ corresponding to a gridded climate data set where $\mathbf{X}$ represents the value of the field for example soil moisture, at time t and spatial position $\mathbf{s}$, respectively. The observed soil moisture gridded dataset can be represented, as

$$\mathbf{X} = \left[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\right]^T = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \tag{3.2.1}$$

where x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$ denotes the value of the field at discrete time t_i and grid point $\mathbf{s}_j$. The map or value of the field is represented by

$$\mathbf{x}_t = \left[x_{t1}, x_{t2}, \ldots, x_{tn}\right]^T, \quad \forall t \in \{1, \ldots, n\} \tag{3.2.2}$$

Let $\bar{x}_i$ represent the field's temporal average at the i spatial grid point. The soil moisture climatology of the field is defined by

$$\overline{X} = \left[\bar{x}_1, \ldots, \bar{x}_p\right], \qquad \text{where} \quad \bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{ki} \tag{3.2.3}$$

The seasonal cycles, e.g., monthly means from the soil moisture for a given observation time, are subtracted from all data collected to calculate spatial anomalies. Seasonal cycle refers to predictive and repetitive patterns over time. From the climatology in Eq. (3.2.3) defined at $(t, \mathbf{s}_k)$, $t = 1, \ldots, n$ and $k = 1, \ldots, p$ the anomaly field is represented in Eq. (3.2.4) and Eq. (3.2.5) whose columns have zero-mean values. The dash can be dropped for the simplification of further computations.

$$x_{tk}^{'} = x_{tk} - \bar{x}_k \tag{3.2.4}$$

we can also represent the data matrix in matrix notation, from here we use $\mathbf{X}$ for simplicity.

$$\mathbf{X}^{'} = \mathbf{X} - \mathbb{1}_n \bar{\mathbf{x}} \tag{3.2.5}$$

3.3 Formulation and Computation

To compute the EOF solution, the method based on singular value decomposition (SVD) is employed. Through SVD, the EOF method can be applied to enormous datasets because there is no need to build a covariance matrix that may be very large to compute (Dawson, 2016). Below we can see how the EOF can be obtained from the anomaly data matrix in Eq. (3.2.5). First, we can define the sample covariance matrix as follows:

$$\mathbf{S} = \frac{1}{n}\mathbf{X}^{'\mathbf{T}}\mathbf{X}^{'} \tag{3.3.1}$$

The covariances s_{ij}, $i, j = 1, \ldots, p$ in the sample covariance matrix between the field's time series at any pair of grid points $(\mathbf{s}_i, \mathbf{s}_j)$ can be expressed as

$$s_{ij} = [\mathbf{S}]_{ij} = \frac{1}{n}\sum_{t=1}^{n} x_{ti} x_{tj} \tag{3.3.2}$$

According to Van Loan and Golub (1996), computing the covariance matrix Eq. (3.3.1) and solving the eigenvalue problem in practice is unnecessary. The SVD is a powerful method from linear algebra mostly used for the analysis of multivariate data, since any $n \times p$ data matrix $\mathbf{X}$ can be decomposed. The equation of the SVD of $\mathbf{X}$ can be expressed as following:

$$\text{SVD}(\mathbf{X}^{'}) = \mathbf{A}\boldsymbol{\Gamma}\mathbf{U}^{*\mathbf{T}} \tag{3.3.3}$$

The dimensions of $\mathbf{A}$ and $\mathbf{U}^*$ are respectively $n \times r$ and $r \times p$ unitary matrices, i.e.

$$\mathbf{U}^{*T}\mathbf{U}^{*} = \mathbf{A}^T\mathbf{A} = \mathbf{I}_r, \qquad \text{where } r \leq \min(n, p) \tag{3.3.4}$$

There are numerous ways to express the SVD, but because Eq. (3.3.3) eliminates unnecessary zero singular values, it offers a compact form (Hannachi et al., 2007). The singular vectors are in the columns of $\mathbf{A}$ and $\mathbf{U}^*$, while the singular values are on the leading diagonal of $\mathbf{\Gamma}$ with dimensions $n \times n$, i.e.‘

$$\mathbf{\Gamma} = \begin{bmatrix} \lambda_{11} & 0 & \cdots & 0 \\ 0 & \lambda_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_r \end{bmatrix} \tag{3.3.5}$$

The singular vectors in $\mathbf{U}$ are the standardized PCs or ECs, which is an advantage of utilizing the SVD approach. The EOF analysis generates ($p =$ sampling locations) EOF/EC pairs, but only $\min(n, p)$ eigenvalues, where ($n =$ sampling times) are positive, and only a subset of these are positive. The eigenvalues of the EOF and ECs generally cause them to be rearranged in descending order, with the first EOF (EOF1) having the largest eigenvalue. To determine the percentage of variance explained (EV) by each EOF, divide the corresponding eigenvalue by the total eigenvalues. (Korres et al., 2010).

$$EV_i = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{3.3.6}$$

3.4 Selection and Reconstruction of Significant Modes

The EOF and ECs can be used to rebuild the entire variability of the dataset after decomposition by choosing all EOF/EC pairs. However, only the first few EOF and EC pairs that account for the most significant proportion of variation are often chosen to estimate and compress a dataset. A reduction in the dataset's dimensionality is the method's outcome. A better version of the soil moisture dataset is constructed by truncating the EOF because random noise present in the higher order EOF is removed by selecting EOF with the highest explained variance (Halldor and Venegas 1997; Preisendorfer and Mobley 1988). Moreover, EOF differ from each other significantly. To this effect, Korres et al. (2010) stated that knowing the difference is one of the prerequisites for the physical interpretation of EOF. Two sets of rules are applied to estimate the number of significant patterns: measuring the uncertainty for the eigenvalues, Eq. (3.4.1) is utilized. This is summarized by the rule of thumb defining the typical error $\Delta(\lambda)$ of eigenvalues, and this approach was used by (North et al., 1982) to select statistically significant EOF. The number of independent samples is represented by s. In this method, an EOF is statistically significant if its lower confidence limit exceeds the upper confidence limit of its most significant eigenvalue, for example, 95%. The methods depend on the number of samples (s). Another method worthy of mentioning utilizes the Monte Carlo simulations to calculate the uncertainty of the eigenvalues (Preisendorfer and Mobley, 1988).

$$\Delta(\lambda_i) \approx \lambda_i \left(1 \pm \sqrt{\frac{2}{s}}\right) \tag{3.4.1}$$

4. Integrating Modes into Emulators

Understanding the functionality and capability of integrating EOFs analysis into ML models for emulation and prediction, to make hydrological simulations using ML and make greater use of the readily available gridded climate data is very useful. Wu et al. (2022) compared the EOF preprocessing approach and straightforward arithmetic preprocessing method to ascertain the impact of using EOF analysis in streamflow prediction. They observed that the LSTM model coupled with EOF analysis displayed a better performance when compared to the other three models, SVR, MLP, and GBRT not coupled. As anticipated, EOF analysis handled the gridded climate data more effectively than the arithmetic technique. Dagon et al. (2020) also used EOF analysis to mimic the global fields' spatial variability without explicitly expressing simulated data at each grid point because it would need a more sophisticated neural network architecture. EOFs applied on the gridded data and used as input series into ML models give a novel insight (Wu et al., 2022). Moreover, We display a framework in Figure (4.1) integrating EOFs as input space into ML models designed by Wu et al. (2022) from our synthesis papers. The EOFs were calculated from four daily gridded hydrological data, namely precipitation (P), temperature (T), wind speed (W), and net solar radiation (R).

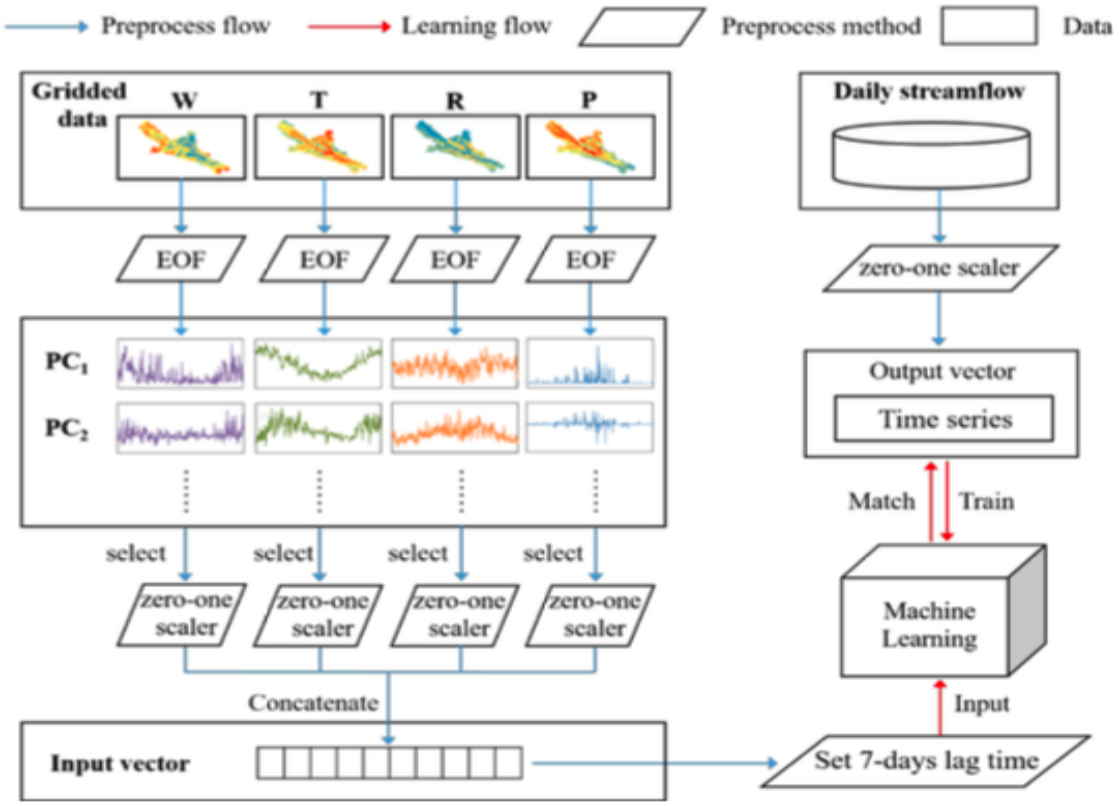

Figure 4.1: The framework of integrating EOF Analysis into ML model (Wu et al., 2022)

The four ML models described above were used in rainfall-runoff modeling. The four types of ML models discussed above were trained to construct the best simulation model on their own in rainfall-runoff modeling. According to Wu et al. (2022) the architecture in Figure 4.1 made it

possible to predict streamflow by only utilizing climatic data as driving forces. The framework can therefore be used to forecast future streamflows under various climate scenarios and generalized to study other hydrological fields and environmental spatio-temporal events (e.g., soil moisture). The GBRT and SVR libraries were individually trained to create the ideal simulation model. Both the SVR and GBRT libraries were Scikit-learn modules. TensorFlow and Keras were the MLP and LSTM libraries, respectively. The Scikit-learn module served as the EOF processing and normalization operations platform (Pedregosa et al., 2011).

Amato et al. (2020) decomposed the output space (target variable) into fixed temporal modes and their corresponding spatial modes using EOFs. As already discussed, the output space was considered the target variable for solving the spatial regression problem using any ML techniques. They modeled the coefficients with a deep feed-forward fully connected neural network (Goodfellow et al., 2016) to show the effectiveness of the proposed scheme and used the spatial covariates as inputs into the neural networks depicted in Figure 4.2. The suggested deep learning (DL) technique by Amato et al. (2020) has two fundamental benefits. First, many traditional ML regression techniques cannot accommodate multiple outputs; as a result, one would have to construct unique models for each mode without being able to benefit from the similarities between the tasks. Second, the reconstructed spatio-temporal field of interest, the final prediction objective of the suggested DL technique, directly minimizes the loss.

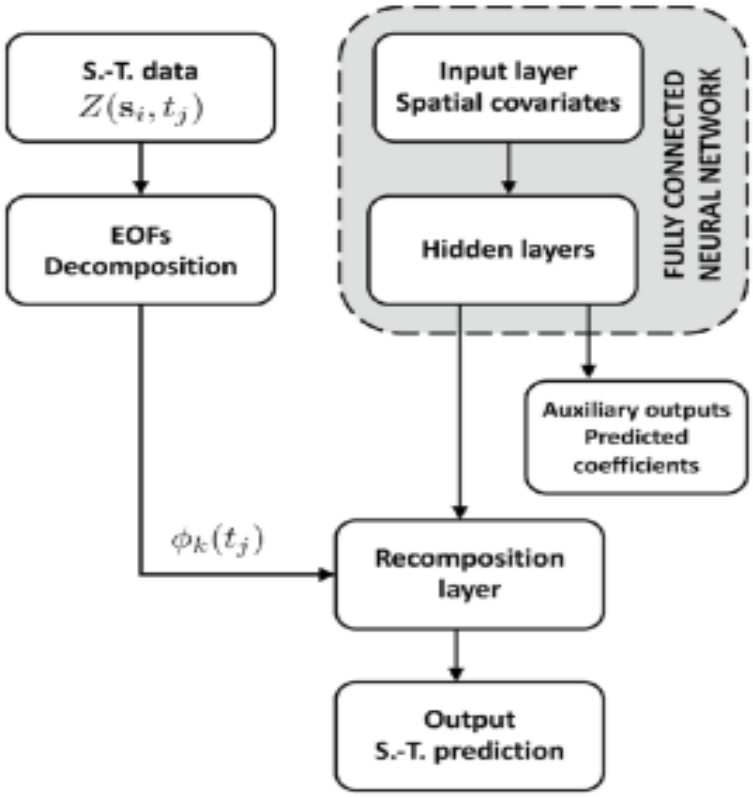

Figure 4.2: The architecture of the proposed. EOFs are used to decompose the spatio-temporal signal and obtain the temporal bases. Then, the corresponding spatial coefficients are learned using a fully connected neural network with reconstructed spatio-temporal (Amato et al., 2020).

5. Gaps and Future Research

The prospective applications described above all share ML and EOFs analysis approaches that have recently become accessible to emulate the physical climate model. However, according to the research articles reviewed, the applications of ML models and EOFs analysis to geoscientific research have various gaps, limitations, and challenges that will further stimulate the development of methodologies. We will explain some of them below and outline future research ideas. Additionally, our analysis is limited to the number of research articles reviewed.

5.1 Gaps, Challenges and Limitations

Watson-Parris (2021) stated that to ensure that the model does not aim to predict outside of the distribution of the training dataset, the training of climate emulators requires a baseline training dataset that spans all feasible outcomes. According to Sexton et al. (2019) and (McCollum et al., 2020), it is essential to consider this carefully when creating ensembles, and it is worth considering when designing future multi-model experiments. This will ensure that the emulators are not extrapolating past training points but rather interpolating between them. Besides that, choosing the ML algorithm to use during the model development stage in the analysis is also one of the significant concerns for scientists because it has a massive impact on the results and how they interpret them (e.g.,Lapuschkin et al. 2019).

Another challenge is the inclusion of biases in the training data and ML models. For example, Wang et al. (2017) observed underestimation of the impact of vegetation on the soil moisture spatial variability study regions when using EOF analysis. This bias was attributed to the dominance of natural grasses in the selected sites. Moreover, Mason and Patil (2015) noted that Earth science data are widely dispersed and large (big data challenges in the geoscientific context), so to use ML in Earth system science, users must first identify the essential datasets and then acquire, store and clean the data. This process takes up around 80% of a user's time. Therefore, the limitations have made it difficult to develop reliable ML models for prediction. However, we can reliably predict how the weather will change on a timescale of days, not months.

Seasonal weather forecasts, predicting catastrophic events like fires or floods, and long-term climate projections remain challenging tasks (Reichstein et al., 2019). Speaking of EOFs, Dommenget and Latif (2002) highlighted the challenges of interpreting the EOF modes as potential physical modes because they have problems identifying the dominant centers of action. Finding techniques to decrease the system's dimensionality of EOFs into a few modes is difficult. Another systematic challenge of the EOFs analysis is linking the modes obtained to the physical dynamics or physics of the system (Hannachi et al., 2007).

We noticed gaps in applying EOFs and ML in geoscientific research through the surveyed research articles. To begin with, we observed that no studies have focused on comparing ML models with incorporated EOFs in predicting soil moisture by using a gridded climate dataset from a climate land model (e.g., CLM). Moreover, no studies have used EOFs on the CLM dataset simulated with different soil texture experiments to analyze spatial-temporal patterns across the continental USA. Furthermore, the usage of other hydrological variables, such as total evapotranspiration,

soil temperature, runoff flux, etc, as input into the ML to predict the soil moisture has not yet been utilized. These gaps, limitations, and challenges need to be addressed in future studies.

5.2 Future Research Ideas

In the previous section, we noticed some limitations, challenges, and gaps in applying EOFs and ML in geoscientific research. Therefore, further research in climate model emulation and the application of EOFs is needed to close the gaps in geoscientific research. These may focus on:

- Designing a multi-model ensemble techniques framework for both ML and climate models to reduce the uncertainties and errors in predicting soil moisture.
- Applying EOFs to the Community Land Model simulations and analyze soil moisture patterns across the Contiguous United States (CONUS) and evaluate different types of ML models to predict soil moisture and compare their performance using the performance metrics in section 2.4.
- Designing a framework to emulate climate land model and detect extreme events using deep learning.
- Applying EOF to analyze the climatology variability of soil moisture over the CONUS.
- Implementing a neural network across each land point on the CONUS using the EOF as the input to emulate soil moisture.

References

Amato, F., F. Guignard, S. Robert, and M. Kanevski (2020). A novel framework for spatio-temporal prediction of environmental data using deep learning. *Scientific reports 10*(1), 1–11.

Bechtold, P., M. Köhler, T. Jung, F. Doblas-Reyes, M. Leutbecher, M. J. Rodwell, F. Vitart, and G. Balsamo (2008). Advances in simulating atmospheric variability with the ecmwf model: From synoptic to decadal time-scales. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography 134*(634), 1337–1351.

Brenowitz, N. D. and C. S. Bretherton (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters 45*(12), 6289–6298.

Busch, F. A., J. D. Niemann, and M. Coleman (2012). Evaluation of an empirical orthogonal function–based method to downscale soil moisture patterns based on topographical attributes. *Hydrological Processes 26*(18), 2696–2709.

Comrie, A. C. (1997). Comparing neural networks and regression models for ozone forecasting. *Journal of the Air & Waste Management Association 47*(6), 653–663.

Costabile, P., C. Costanzo, F. Macchione, and P. Mercogliano (2012). Two-dimensional model for overland flow simulations: A case study. *European Water 38*, 13–23.

Dagon, K., B. M. Sanderson, R. A. Fisher, and D. M. Lawrence (2020). A machine learning approach to emulation and biophysical parameter estimation with the community land model, version 5. *Advances in Statistical Climatology, Meteorology and Oceanography 6*(2), 223–244.

Davis, S. E., S. Cremaschi, and M. R. Eden (2017). Efficient surrogate model development: optimum model form based on input function characteristics. In *Computer Aided Chemical Engineering*, Volume 40, pp. 457–462. Elsevier.

Dawson, A. (2016). eofs: A library for eof analysis of meteorological, oceanographic, and climate data. *Journal of Open Research Software 4*(1).

Dommenget, D. and M. Latif (2002). A cautionary note on the interpretation of eofs. *Journal of climate 15*(2), 216–225.

Doycheva, K., G. Horn, C. Koch, A. Schumann, and M. König (2017). Assessment and weighting of meteorological ensemble forecast members based on supervised machine learning with application to runoff simulations and flood warning. *Advanced Engineering Informatics 33*, 427–439.

Eltahir, E. A. and R. L. Bras (1994). Precipitation recycling in the amazon basin. *Quarterly Journal of the Royal Meteorological Society 120*(518), 861–880.

Farneti, R. and P. R. Gent (2011). The effects of the eddy-induced advection coefficient in a coarse-resolution coupled climate model. *Ocean Modelling 39*(1-2), 135–145.

Gao, X., A. Avramov, E. Saikawa, and C. A. Schlosser (2021). Emulation of community land model version 5 (clm5) to quantify sensitivity of soil moisture to uncertain parameters. *Journal of Hydrometeorology 22*(2), 259–278.

Gibson, P. B., W. E. Chapman, A. Altinok, L. Delle Monache, M. J. DeFlorio, and D. E. Waliser (2021). Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Communications Earth & Environment 2*(1), 1–13.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep learning*. MIT press.

Halldor, B. and S. A. Venegas (1997). A manual for eof and svd analyses of climate data. *McGill University, CCGCR Report 52*.

Ham, Y.-G., J.-H. Kim, and J.-J. Luo (2019). Deep learning for multi-year enso forecasts. *Nature 573*(7775), 568–572.

Han, X., H.-J. H. Franssen, C. Montzka, and H. Vereecken (2014). Soil moisture and soil properties estimation in the community land model with synthetic brightness temperature observations. *Water resources research 50*(7), 6081–6105.

Han, X., X. Li, H. Hendricks Franssen, H. Vereecken, and C. Montzka (2012). Spatial horizontal correlation characteristics in the land data assimilation of soil moisture. *Hydrology and Earth System Sciences 16*(5), 1349–1363.

Hannachi, A., I. T. Jolliffe, and D. B. Stephenson (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology: A Journal of the Royal Meteorological Society 27*(9), 1119–1152.

Hawkins, L. R., D. E. Rupp, D. J. McNeall, S. Li, R. A. Betts, P. W. Mote, S. N. Sparrow, and D. C. Wallom (2019). Parametric sensitivity of vegetation dynamics in the triffid model and the associated uncertainty in projected climate change impacts on western us forests. *Journal of Advances in Modeling Earth Systems 11*(8), 2787–2813.

Huntingford, C., E. S. Jeffers, M. B. Bonsall, H. M. Christensen, T. Lees, and H. Yang (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters 14*(12), 124007.

Huo, X., H. Gupta, G.-Y. Niu, W. Gong, and Q. Duan (2019). Parameter sensitivity analysis for computationally intensive spatially distributed dynamical environmental systems models. *Journal of Advances in Modeling Earth Systems 11*(9), 2896–2909.

Jawson, S. D. and J. D. Niemann (2007). Spatial patterns from eof analysis of soil moisture at a large scale and their dependence on soil, land-use, and topographic properties. *Advances in Water Resources 30*(3), 366–381.

Jaynes, D., T. Kaspar, T. Colvin, and D. James (2003). Cluster analysis of spatiotemporal corn yield patterns in an iowa field. *Agronomy Journal 95*(3), 574–586.

Jollife, I. (2002). Principal component analysis 2 edition springer. *New York*.

Jung, M., M. Reichstein, H. A. Margolis, A. Cescatti, A. D. Richardson, M. A. Arain, A. Arneth, C. Bernhofer, D. Bonal, J. Chen, et al. (2011). Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *Journal of Geophysical Research: Biogeosciences 116*(G3).

Karthikeyan, L. and A. K. Mishra (2021). Multi-layer high-resolution soil moisture estimation using machine learning over the united states. *Remote Sensing of Environment 266*, 112706.

Kato, H., M. Rodell, F. Beyrich, H. Cleugh, E. van GORSEL, H. Liu, and T. P. Meyers (2007). Sensitivity of land surface simulations to model physics, land characteristics, and forcings, at four ceop sites. *Journal of the Meteorological Society of Japan. Ser. II 85*, 187–204.

Kitanidis, P. K. and R. L. Bras (1980). Real-time forecasting with a conceptual hydrologic model: 2. applications and results. *Water Resources Research 16*(6), 1034–1044.

Korres, W., C. Koyama, P. Fiener, and K. Schneider (2010). Analysis of surface soil moisture patterns in agricultural landscapes using empirical orthogonal functions. *Hydrology and Earth System Sciences 14*(5), 751–764.

Krasnopolsky, V. M., D. V. Chalikov, and H. L. Tolman (2002). A neural network technique to improve computational efficiency of numerical oceanic models. *Ocean Modelling 4*(3-4), 363–383.

Krasnopolsky, V. M. and M. S. Fox-Rabinovitz (2006). Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. *Neural Networks 19*(2), 122–134.

Krasnopolsky, V. M. and Y. Lin (2012). Research article a neural network nonlinear multimodel ensemble to improve precipitation forecasts over continental us.

Kutzbach, J. E. (1967). Empirical eigenvectors of sea-level pressure, surface temperature and precipitation complexes over north america. *Journal of Applied Meteorology and Climatology 6*(5), 791–802.

Lapuschkin, S., S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K. Müller (2019). Unmasking clever hans predictors and assessing what machines really learn. nat. commun. 10 (1), 1–8 (2019).

Liang, X., D. P. Lettenmaier, E. F. Wood, and S. J. Burges (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research: Atmospheres 99*(D7), 14415–14428.

Liu, M. and J. Lu (2014). Support vector machine—an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river? *Environmental Science and Pollution Research 21*(18), 11036–11053.

Lorenz, E. N. (1970). Climatic change as a mathematical problem. *Journal of Applied Meteorology and Climatology 9*(3), 325–329.

Mandeville, A., P. O'connell, J. Sutcliffe, and J. Nash (1970). River flow forecasting through conceptual models part iii-the ray catchment at grendon underwood. *Journal of Hydrology 11*(2), 109–128.

Mason, H. and D. Patil (2015). Data driven. creating a data culture.

McCollum, D. L., A. Gambhir, J. Rogelj, and C. Wilson (2020). Energy modellers should explore extremes more systematically in scenarios. *Nature Energy 5*(2), 104–107.

McKay, M. D., R. J. Beckman, and W. J. Conover (2000). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics 42*(1), 55–61.

Mitchell, T. M. and T. M. Mitchell (1997). *Machine learning*, Volume 1. McGraw-hill New York.

Mohanty, S., M. K. Jha, A. Kumar, and D. Panda (2013). Comparative evaluation of numerical model and artificial neural network for simulating groundwater flow in kathajodi–surua interbasin of odisha, india. *Journal of Hydrology 495*, 38–51.

Niu, Z.-w., L. DAN, and Z.-z. Yang (2003). Porous membrane templated synthesis of polymer pillared layer. *Chinese journal of polymer science 21*(3), 381–384.

North, G. R., T. L. Bell, R. F. Cahalan, and F. J. Moeng (1982). Sampling errors in the estimation of empirical orthogonal functions. *Monthly weather review 110*(7), 699–706.

Ntelekos, A. A., K. P. Georgakakos, and W. F. Krajewski (2006). On the uncertainties of flash flood guidance: Toward probabilistic forecasting of flash floods. *Journal of Hydrometeorology 7*(5), 896–915.

Obukhov, A. M. (1947). Statistically homogeneous fields on a sphere. *Usp. Mat. Nauk 2*(2), 196–198.

O'Gorman, P. A. and J. G. Dwyer (2018). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems 10*(10), 2548–2563.

Oleson, K. W., D. M. Lawrence, B. Gordon, M. G. Flanner, E. Kluzek, J. Peter, S. Levis, S. C. Swenson, E. Thornton, J. Feddema, et al. (2010). Technical description of version 4.0 of the community land model (clm).

Patel, S. S. and P. Ramachandran (2015). A comparison of machine learning techniques for modeling river flow time series: the case of upper cauvery river basin. *Water resources management 29*(2), 589–602.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research 12*, 2825–2830.

Perry, M. A. and J. D. Niemann (2007). Analysis and estimation of soil moisture at the catchment scale using eofs. *Journal of Hydrology 334*(3-4), 388–404.

Prakash, S., A. Sharma, and S. S. Sahu (2018). Soil moisture prediction using machine learning. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 1–6. IEEE.

Preisendorfer, R. W. and C. D. Mobley (1988). Principal component analysis in meteorology and oceanography. *Developments in atmospheric science 17*.

Rasp, S. and S. Lerch (2018). Neural networks for postprocessing ensemble weather forecasts. *Monthly Weather Review 146*(11), 3885–3900.

Rasp, S., M. S. Pritchard, and P. Gentine (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences 115*(39), 9684–9689.

Reichle, R. H., R. D. Koster, J. Dong, and A. A. Berg (2004). Global soil moisture from satellite observations, land surface models, and ground data: Implications for data assimilation. *Journal of Hydrometeorology 5*(3), 430–442.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, et al. (2019). Deep learning and process understanding for data-driven earth system science. *Nature 566*(7743), 195–204.

Ricciuto, D., K. Sargsyan, and P. Thornton (2018). The impact of parametric uncertainties on biogeochemistry in the e3sm land model. *Journal of Advances in Modeling Earth Systems 10*(2), 297–319.

Sanderson, B. M., R. Knutti, T. Aina, C. Christensen, N. Faull, D. Frame, W. Ingram, C. Piani, D. A. Stainforth, D. Stone, et al. (2008). Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes. *Journal of Climate 21*(11), 2384–2400.

Sargsyan, K., C. Safta, H. N. Najm, B. J. Debusschere, D. Ricciuto, and P. Thornton (2014). Dimensionality reduction for complex models via bayesian compressive sensing. *International Journal for Uncertainty Quantification 4*(1).

Schizas, C. N., S. Michaelides, C. S. Pattichis, and R. Livesay (1991). Artificial neural networks in forecasting minimum temperature (weather). In *1991 second international conference on artificial neural networks*, pp. 112–114. IET.

Sellers, P., D. Randall, G. Collatz, J. Berry, C. Field, D. Dazlich, C. Zhang, G. Collelo, and L. Bounoua (1996). A revised land surface parameterization (sib2) for atmospheric gcms. part i: Model formulation. *Journal of climate 9*(4), 676–705.

Sexton, D., A. Karmalkar, J. Murphy, K. Williams, I. Boutle, C. Morcrette, A. Stirling, and S. Vosper (2019). Finding plausible and diverse variants of a climate model. part 1: establishing the relationship between errors at weather and climate time scales. *Climate Dynamics 53*(1), 989–1022.

Stevens, B. and S. Bony (2013). What are climate models missing? *Science 340*(6136), 1053–1054.

Swenson, S. and D. Lawrence (2014). Assessing a dry surface layer-based soil resistance parameterization for the community land model using grace and fluxnet-mte data. *Journal of Geophysical Research: Atmospheres 119*(17), 10–299.

Tigkas, D., V. Christelis, and G. Tsakiris (2016). Comparative study of evolutionary algorithms for the automatic calibration of the medbasin-d conceptual hydrological model. *Environmental Processes 3*(3), 629–644.

Van Loan, C. F. and G. Golub (1996). Matrix computations (johns hopkins studies in mathematical sciences). *Matrix Computations*.

Wang, T., T. E. Franz, R. Li, J. You, M. D. Shulski, and C. Ray (2017). Evaluating climate and soil effects on regional soil moisture spatial variability using eof s. *Water Resources Research 53*(5), 4022–4035.

Watson-Parris, D. (2021). Machine learning for weather and climate are worlds apart. *Philosophical Transactions of the Royal Society A 379*(2194), 20200098.

Watson-Parris, D., A. Williams, L. Deaconu, and P. Stier (2021). Model calibration using esem v1. 1.0–an open, scalable earth system emulator. *Geoscientific Model Development 14*(12), 7659–7672.

Webster, R. (1985). Quantitative spatial analysis of soil in the field. In *Advances in soil science*, pp. 1–70. Springer.

Western, A. W., R. B. Grayson, and T. R. Green (1999). The tarrawarra project: high resolution spatial measurement, modelling and analysis of soil moisture and hydrological response. *Hydrological processes 13*(5), 633–652.

Wilcox, E. M. and L. J. Donner (2007). The frequency of extreme rain events in satellite rain-rate estimates and an atmospheric general circulation model. *Journal of Climate 20*(1), 53–69.

Wilks, D. (2006). Statistical methods in the atmospheric sciences 2nd edn (new york: Academic).

Wu, Y., Y. Chen, and Y. Tian (2022). Incorporating empirical orthogonal function analysis into machine learning models for streamflow prediction. *Sustainability 14*(11), 6612.

Yoo, C. and S. Kim (2004). Eof analysis of surface soil moisture field variability. *Advances in Water Resources 27*(8), 831–842.