


A Review of Machine Learning Applications in Land Surface Modeling

Sujan Pal *  and Prateek Sharma

Department of Atmospheric Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA;
prateek7@illinois.edu

* Correspondence: sujanp2@illinois.edu

Abstract: Machine learning (ML), as an artificial intelligence tool, has acquired significant progress in data-driven research in Earth sciences. Land Surface Models (LSMs) are important components of the climate models, which help to capture the water, energy, and momentum exchange between the land surface and the atmosphere, providing lower boundary conditions to the atmospheric models. The objectives of this review paper are to highlight the areas of improvement in land modeling using ML and discuss the crucial ML techniques in detail. Literature searches were conducted using the relevant key words to obtain an extensive list of articles. The bibliographic lists of these articles were also considered. To date, ML-based techniques have been able to upgrade the performance of LSMs and reduce uncertainties by improving evapotranspiration and heat fluxes estimation, parameter optimization, better crop yield prediction, and model benchmarking. Widely used ML techniques used for these purposes include Artificial Neural Networks and Random Forests. We conclude that further improvements in land modeling are possible in terms of high-resolution data preparation, parameter calibration, uncertainty reduction, efficient model performance, and data assimilation using ML. In addition to the traditional techniques, convolutional neural networks, long short-term memory, and other deep learning methods can be implemented.

Keywords: machine learning; land surface; land-atmosphere interactions; parameterizations; model uncertainty



Citation: Pal, S.; Sharma, P. A Review of Machine Learning Applications in Land Surface Modeling. *Earth* **2021**, *2*, 174–190.
<https://doi.org/10.3390/earth2010011>

Received: 6 February 2021

Accepted: 18 March 2021

Published: 20 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Machine learning (ML) and artificial intelligence (AI) progressively impact society, supported by significant enhancement of big data, computational efficiency, easily available data storage, and uninterrupted connectivity. ML algorithms are being increasingly applied in Earth and Environmental modeling studies as a result of growing resources of extensive data sets, easy computation, and upgraded ML algorithms. Some well-explored areas with ML applications in Earth Science are climate modeling [1,2], hydrologic modeling [3,4], remote sensing [5,6], etc. Land Surface models (LSMs) are the components of climate models, which simulate land surface processes, such as the partitioning and consumption of energy, moisture, momentum, and carbon. The understanding of land surface feedback on the meteorological and climatological features have gained attention in recent years [7–9]. The land provides forcing from the surface to the atmosphere via energy balance, surface heating (supplying buoyancy for convection), and surface roughness (generating eddies that modulate atmospheric boundary layer). Land-surface interactions are communicated through vegetation and soil moisture. LSMs are crucial to understand and predict the dynamics of the land surface and its involvement within the Earth system. The complex processes that interconnect different components of the terrestrial system, and the depth of convolution present in all of those processes, make the LSMs less amenable. With the advancement of land surface observation systems, such as the Global Observing System of World Meteorological Organization (WMO) and FLUXNET around the world,

improved estimates of land surface radiation, ecology, hydrology, biology are now possible. However, representing the land-surface interactions properly in the model still remains a challenge. Specifically, in the areas of complex terrain and strong surface-atmosphere coupling. Major components of an LSM include (1) atmospheric forcing, (2) physical, chemical, and biological processes, (3) parameters, and (4) outputs.

The adoption of ML in land surface modeling has so far been gradual, but the related fields are now highly emerging. ML-based methods can create viable, complementary avenues toward knowledge discovery in land modeling. ML is often categorized into traditional methods and deep learning (DL) based methods. While traditional methods include random forests (RF), support vector machines (SVM), classification and regression trees (CART); DL techniques include Deep Learning Artificial Neural Network (DL-ANN), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN) etc. Overall, ML models have repeatedly outperformed simpler statistical models and provided improved generalized solutions to unforeseen circumstances in many applications of Earth Sciences. The recent growth of big hydrologic data through remote sensing and data compilation has fostered the development.

The difference between the process-based LSMs and ML models is that the model structure for LSMs is based on the underlying physical principles that we believe are valid, whereas ML model structure is completely data-driven. Despite this fundamental difference, they can complement each other. Two strategies popular in literature in this regard are: (1) integrating LSMs and ML, (2) introducing physical constraints on ML models. Strategy (1) is quite straightforward. One common way to implement it is to ensure the application of pure ML models only when the climatology is similar to the training data; otherwise, in case of rare events, the combined model relies on the LSM. In this way, modelers can avoid the uncertainty coming from ML models in the extreme cases while making the model less computationally expensive in general application. On the other hand, (2) acknowledges that data-driven algorithms often fail to satisfy the well-known physics law constraints as they are not generally enforced with those. They tend to violate constraints on individual samples while optimizing the overall performance. LSM outputs can be used to pose constraints on the ML models to overcome this. Recently, interrogative studies have also been invoked to interpret ML models to ask questions like (1) how trustworthy are ML models? or (2) how to interpret the processes that drive ML model systems?

The objectives of this review paper are (1) to feature the areas of improvement in land modeling using ML and (2) discuss the most important ML techniques in detail, and (3) suggest future directions for further improvements. To achieve these goals, we searched literature in google scholar database using a combination of the relevant key words: 'machine learning', 'land surface modeling', 'crop modeling', 'evapotranspiration estimation', 'parameters and uncertainty' etc. Moreover, the bibliographic lists of these articles were also considered to create an extensive list of articles. The most relevant 83 articles are cited and discussed in this paper.

This study demonstrates the potential of ML to propel LSMs and vice versa. Critically, this review paper provides information about (1) the importance of LSMs in different applications (2) the difficulties and limitations of traditional LSMs. (3) application of ML in LSMs in past literature. (4) concise and simple technical overview of ML for LSMs. (5) potential directions where ML can contribute to solving challenges in modeling land surface processes. The results from this study have implications for creating inexpensive, improved and tractable land surface models with less and quantifiable uncertainty. This in turn will help to construct upgraded coupled climate models to provide more realistic and refined future projections.

2. LSMs: Importance, Then and Now

As mentioned in the introduction, LSMs are numerical models that simulate land surface processes, such as absorption and partitioning of radiation, water, and carbon

between the land surface and atmosphere. Provided with meteorological forcing as inputs (from an atmospheric model either in ‘coupled’ mode or an ‘uncoupled’ mode), they estimate latent heat fluxes (LH), sensible heat fluxes (SH), carbon fluxes, surface runoff, deep drainage, reflected solar and emitted longwave radiation as output [10,11]. While LH and SH control the boundary layer properties and precipitation; net carbon flux influences the atmospheric CO₂ content. These estimates play a critical role in determining the effects of human-modified land surface and human emissions on changes in the climate. LSMs are perhaps the most efficient tools to predict how the continuously evolving earth surface will modify the hydroclimate in coming years and centuries. The extents of modeling activities with LSMs include multiple interlinked disciplines (such as atmospheric modeling, crop modeling, and hydrologic modeling) relevant to this overarching problem.

LSMs were originally developed by the atmospheric modeling community who needed physical boundary conditions consisting of energy and moisture partitioning, albedo, and surface roughness to indicate the impact of the surface on the atmospheric processes. Richardson [12], in 1922, first mentioned the importance of stomatal conductance on weather processes. Early studies, such as Charney et al. [13] used albedo as a proxy for vegetation and started investigating the effects of deforestation in terms of it. Starting from the 1980s, scientists started understanding the land surface-atmosphere interactions [14,15]. Garatt et al. [16] summarized the importance of land surface in climate modeling in a review paper. He discussed different boundary layer schemes and the results from global climate model (GCM) sensitivity studies using these schemes. He concluded that the regional and global climate is significantly influenced by albedo, surface moisture and roughness, and the inclusion of vegetation. However, till then, it was not clear how much spatial detail of the surface is sufficient to accurately represent the lower boundary conditions. For that decade, improvements of LSMs were driven by the need to understand the effects of deforestation in various parts of the world. In the 2000s, scientists started to visualize the importance of land in the context of sub-seasonal to seasonal forecasting. The land surface was identified as a slowly varying component of the earth system, which has a major role in modulating the atmospheric response at a longer timescale than weather prediction. Koster et al. papers [7,17], in connection with the Global Land-Atmosphere Coupling Experiment (GLACE), identified soil moisture as an important factor altering evaporation and precipitation. They also highlighted the regions where strong coupling between soil moisture and precipitation exists. For the first time, they introduced the concept of ‘coupling strength’ to quantify such coupling, which is still being widely used in land-atmosphere interaction studies. However, while modeling these interactions, there exists a huge variation among the global models, attributable to the uncertainties in terrestrial and atmospheric branches, and the models fail to represent the land-surface coupling accurately [18]. Specifically, they found systematic biases in near-surface temperature, humidity, and precipitation, which contribute to the uncertainty. Seneviratne et al. [19] summarized the findings related to soil moisture-precipitation relations in a review paper and concluded that the relationship between soil moisture and precipitation is evident in observations and models. However, significant uncertainty remains in quantifying those in terms of the strength of coupling, and persistence characteristics. These studies indicate the need for further improvement in land surface models. The need for LSMs to quantify such biogeophysical and biogeochemical feedbacks to the climate system has formed the basis of their recent development efforts.

At present, LSMs have expanded from their initial simple biophysical configurations [20] to include representations of stomatal functioning [21], scaling information from leaf to canopy [22], soil moisture dynamic and surface hydrology [23,24], crop processes [25,26], land surface heterogeneity [27], dynamic vegetation [28,29], urban environment [30], land cover management [31,32], plant demographic processes and plant hydraulics [33], groundwater dynamics [34], soil microbial dynamics [35], leaf mesophyll process, nitrogen, phosphorus, carbon cycling and their mutual interactions [36]. The incorporation of processes in LSMs is driven by the need for extensive user communities,

including ecologists, crop modelers, atmospheric scientists, biogeochemists, hydrologists, who explore interactions between different components of the system. Some widely used LSMs across the globe include Interaction Soil-Biosphere-Atmosphere (ISBA, [36]), The Community Land Model (CLM, [31]), JSBACH [37], Joint UK Land Environment Simulator (JULES, [38]), LPJ-GUESS [39], Noah-MP [40]. Along with the increasing capability of representing processes, LSMs are enhancing their spatial resolution as well, with the improvement in resolution of the atmospheric models. As the scope of LSMs broadens with the support of computational advancements, the questions of cognitive uncertainty and unresolved heterogeneity emerges as a challenge.

3. Complexity and Limitations of LSMs: Prospect of ML

The diversity of the interconnected processes in the terrestrial system, and the levels of entanglement present in these processes, pose a hurdle to build tractable land models. The propensity of scientists to focus on their own specific area of interest and the reality that the earth system is indeed complex are both responsible for this complex nature. Often, this reaches a point where no individuals are able to completely understand all aspects of any particular model, and the development teams strive to meet all the requirements placed on modern LSMs [11]. Even though, large uncertainty remains in our understanding and modeling of the interactions between atmospheric and terrestrial branches of the hydrologic cycle due to the non-trivial mechanisms at the land surface. Figure 1 illustrates the convoluted and connected processes in a typical LSM. The major parts, such as, atmosphere, hydrology, urban processes, agriculture; and plant physiology, soil biogeochemistry, soil physics related to each of those, are interlinked in an LSM. These major components are further segregated into smaller yet complicated processes. For example, agriculture includes fertilizer and pesticides usage, biomass burning, harvesting, irrigation, tillage, residual treatment etc. (Figure 1). The interactions are defined by the exchange of information between different parts of the model. However, some of the processes are still oversimplified in modeling. As such, most of the LSMs classify plant species into plant functional types (PFTs), within which the parameters are undifferentiated. Simulations consisting of a limited number of PFTs ignore biodiversity within a simulation grid. This may lead to uncertainty in the strength of climate responses when coupled to a climate model. Furthermore, understanding the combined effects of major greenhouse gases, such as Carbon dioxide, Methane, and Nitrogen dioxide on global warming are still at early stages due to constraints in the measurements of multiple gases. Limited models have the capacity to simulate such effects, which requires realistic carbon and nitrogen cycling processes.

LSMs are often applied at large spatial scales aimed to simulate the interactions between climate and land surface. Nonetheless, validation data for these models are obtained from flux tower sites. This geographical gap usually limits the accuracy of the models. Microbes may play fundamental roles in altering biogeochemical cycling as ‘ecosystem engineers’ [41]. However, very few LSMs include the effects of such organisms in an explicit manner. This limits our ability to estimate the climatic impacts of changes in soil biological community composition and diversity. In addition, the unavailability of high-resolution land-surface data affects the LSMs in capturing the effects of spatial heterogeneity. The development of high-accuracy fine resolution data is important for the interpretation of observations and model simulations.

Some of the limitations of LSMs can be highly benefited from the enormous information currently available from satellite data. However, extracting useful knowledge from terabytes of data provided by observation and LMS simulations is challenging. In contrast, ML models are simple in nature in terms of structure and easy to simulate the output once trained properly. Figure 2 illustrates the general structure of a multilayer perceptron model, which is a commonly used feedforward ANN type ML model and uses supervised learning techniques for training. Compared to LSMs (Figure 1), the structure is much simpler and furthermore, the model parameters are data driven. ML techniques can help

and complement LSMs in several ways, including surrogate modeling, physics-guided machine learning, parameter estimation, and data assimilation to reduce the uncertainties and generate useful knowledge from large amounts of observational data. Some of the ML applications in land modeling are described in the next section.

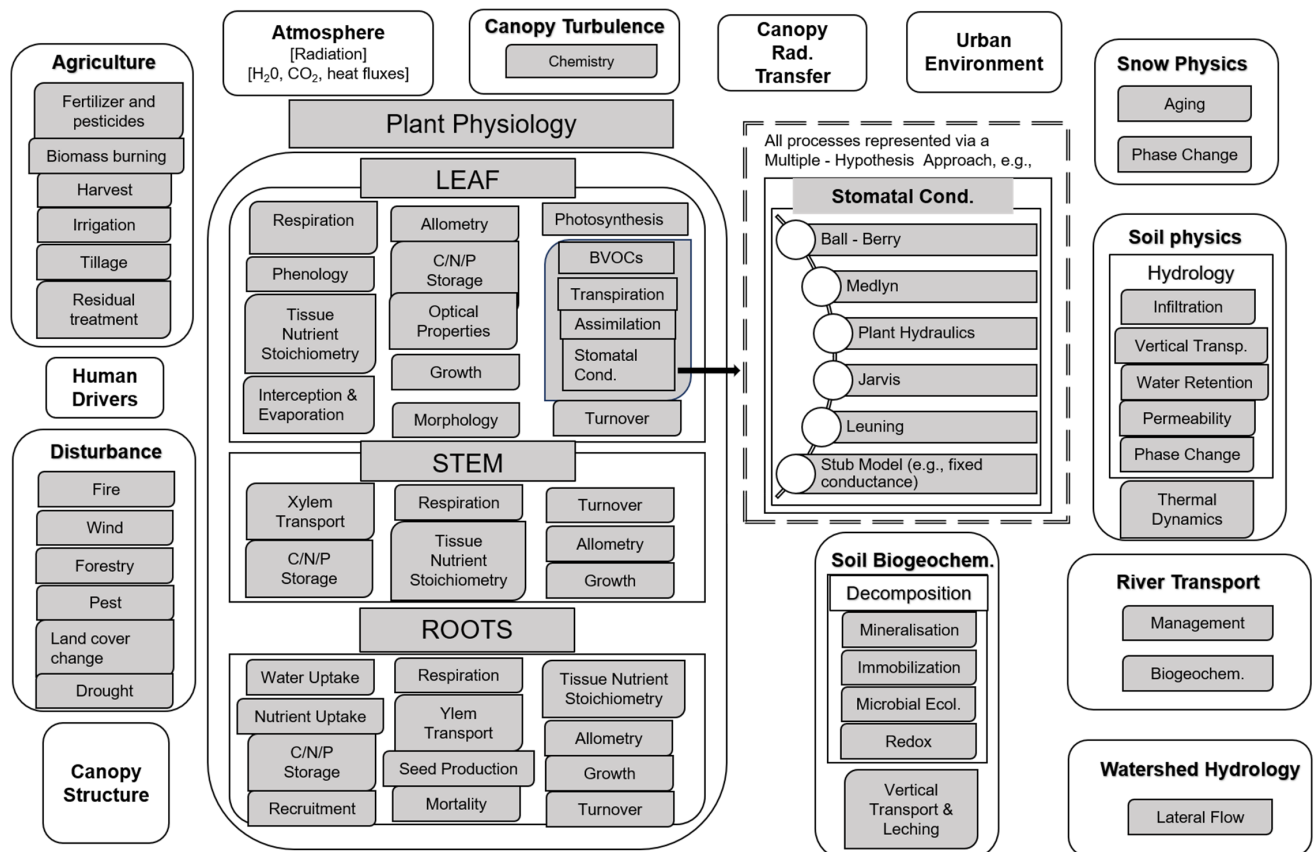


Figure 1. Interconnected complex processes included in a typical LSM. Adapted and modified from Fisher et al. [11].

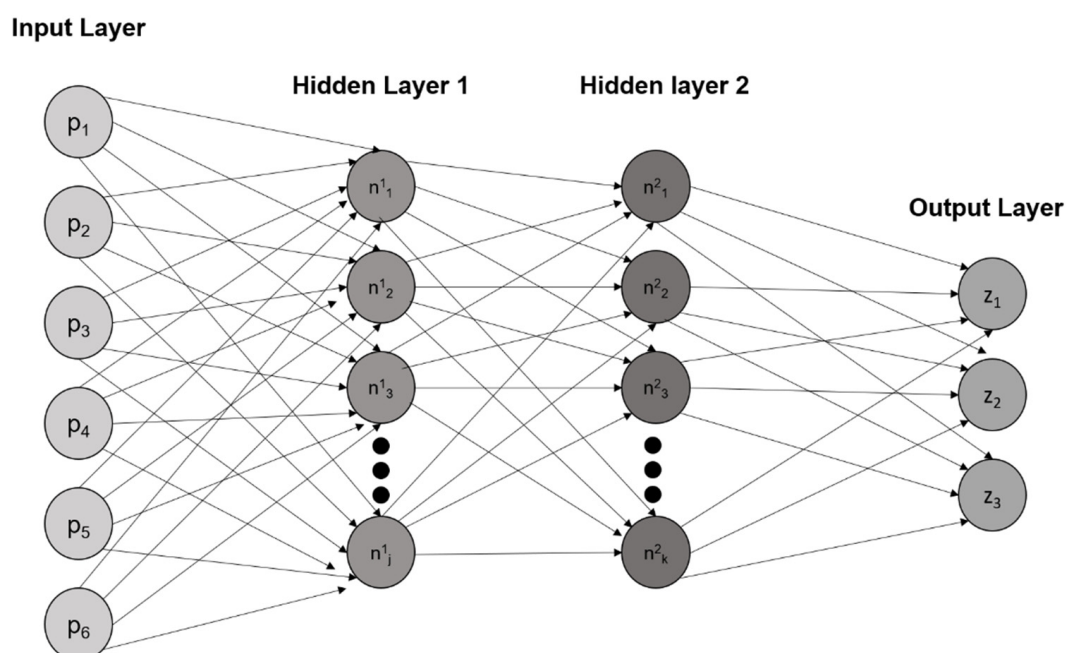


Figure 2. General structure of a multilayer perceptron model.

4. Major Applications of Machine Learning in Land Surface Modeling

4.1. Estimation of Evapotranspiration

Evapotranspiration (ET) is perhaps the most discussed variable in the land surface modeling since it is a major part of both water and energy (in the form of LH) balance. It is also a critical factor in the carbon cycle, acting as a trade-off between photosynthesis and transpiration. Since direct measurement of global terrestrial ET is implausible, one of the central uses of LSMs is providing ET estimates using other comprehensively measured hydrometeorological variables. Some noteworthy applications regarding ET estimation are listed below.

Alemohammad et al. [42] developed an ANN approach to estimate monthly LH, SH, and gross primary productivity (GPP) at a global scale for 8 years at $1^\circ \times 1^\circ$ resolution using remotely sensed solar-induced fluorescence (SIF) measurements in addition to conventional precipitation, temperature, soil moisture, snow cover and net radiation (R_n) as inputs. When compared to eddy covariance measurements from FLUXNET, their method 'WECANN' outperformed traditional surface fluxes products provided by Global Land Evaporation Amsterdam Model (GLEAM), Moderate Resolution Imaging Spectroradiometer (MODIS) and European Centre for Medium-Range Weather Forecasts (ECMWF). The ML retrievals also demonstrated capability to represent the extents of several extreme drought and heat events. They were also able to analyze the effects of extreme climatic events on surface turbulent fluxes and GPP. Prior to this study, neural networks were primarily applied to satellite observations to estimate LH [43]. Some studies also applied generalized neural networks and artificial intelligence models at a local-regional scale [44–47]. However, these techniques, when applied to the global scale, often failed to predict extremes and perform outside their calibration range [48].

A major limitation of traditional ML methods in Earth Science, as pointed out by Zhao et al. [49], was the issue of not conserving the surface energy budget, leading to unrealistic predictions of various surface fluxes causing problems in implementation with a coupled atmospheric model. Physics-based models, although complicated, tend to have superior interpretability [50]. To overcome this, [49] developed the first 'physics-constrained' ML model for ET estimation, which applied ML techniques while keeping the energy conservation equations valid, to provide a global ET estimate. Their hybrid model was able to learn the nonlinear relations from the data while obeying the physical laws. The study was significantly robust as they used 82 eddy covariance sites from FLUXNET with nine different PFTs and implemented a feedforward ANN escorted by that. This kind of modeling can enhance the capability to simulate ET during extreme climatic conditions like heatwaves and droughts.

Another study, Pan et al. [51], recently attempted to estimate global terrestrial ET (GTET) using two machine learning algorithms, RF, and Model Tree Ensemble and predicted its annual mean, interannual variability, and trends. They concluded that the utilization of satellite retrievals and deep-learning methods, and model-data fusion advances the predictive understanding of GTET. Critically, the ML methodology provided similar results to remote sensing-based products indicating a significant increasing trend in GTET.

4.2. Parameter Estimation and Uncertainty Assessment

LSMs are not properly equipped to accurately simulate the real world and lead to significant uncertainty when trying to capture the land-atmosphere interactions. One major reason behind this is the use of parameterizations for complex processes such as ET estimation, which includes an abundance of soil and vegetation parameters. For example, current global LSMs have over 20 parameters that are linked to just ET estimation [52]. Other sources of uncertainty in the LSMs include internal variability and forcing uncertainty.

Chaney et al. [52] provided global parameter estimates at 5 km spatial resolution for the Noah LSM using 85 eddy covariance sites in the global FLUXNET network and linked the most sensitive parameters to local environmental characteristics using an ML

algorithm—extremely randomized trees (Extra-trees). They concluded that FLUXNET could be potentially used to tune the parameters of LSM. The calibration led to a significant enhancement in model performance in terms of Kling-Gupta Efficiency increasing from 0.54 to 0.71. Furthermore, their leave-one-out cross validation method revealed the potential to relate calibrated model parameters to local environmental conditions.

Uncertainty quantification of LSMs is closely related to the parameter optimization problem. As such, it is far-fetched to obtain the accurate set of sensitive parameters for all observable variables, and different combinations of parameters may show similar skill to reproduce observations, also termed as ‘equifinality’ in hydrology [53]. An alternative method is to construct the probability density function of parameters, explaining the uncertainty in parameter estimation. Swada [54] applied ML techniques for parameter optimization and uncertainty assessment with global satellite observations. The technique was to create an ML model, which emulates the relationship between model parameters and LSM outputs and is computationally cheaper than LSM. They applied this technique, involving ML and Markov Chain Monte Carlo (MCMC), to EcoHydro-SiB (a modified version of [20]) model over a part of the Sahel region in Africa, and was successful in obtaining the nonparametric posterior distribution of four unknown parameters and enhanced the soil moisture and vegetation dynamics simulation skill of the LSM. The addition of ML techniques made optimization 50,000 times faster. They highlighted the importance of selecting suitable prior, consideration of error in meteorological forcing, and inclusion of satellite observations in the modeling chain.

In a recent study, Dagon et al. [55] investigated the biogeophysical parameter space of CLM5 and determined the sensitivity of parameters to get insight into the role of parameter choices on the overall model uncertainty. They implemented an ML approach to globally calibrate six parameters of CLM5, selected by a sensitivity analysis, to the observations of water and carbon fluxes. Specifically, they trained their feedforward ANNs to emulate CLM5 outputs given parameters as predictors. The trained ML models were able to estimate global optimal parameter values and quantify the contribution of parameter uncertainty to the overall uncertainty in LSM simulated GPP and LH. Moreover, they were able to inspect several interpretation methods to better understand the inner workings of the ANN as emulators of CLM.

4.3. Crop Yield Prediction

Crop yield is often predicted leveraging the physics based LSMs. Early and reliable crop yield forecasts are critical for farmers and decision-makers in food security policymaking, planning, and trade. Agriculture is greatly affected by weather and climate, and hence, large-scale crop growth simulations under a changing climate in a regional and global scale remains a priority. At the same time, the climate is affected by the extension of agricultural practices due to land-atmosphere interactions. The majority of the studies implemented a ‘hybrid’ approach, combining process-based modeling and ML, to provide a better yield estimate than those could provide individually.

Everingham et al. [56] provided a framework to combine RF (their ML technique) and APSIM (a process-based crop model) to predict the annual variation of sugarcane yield in Australia. Biomass indices, an output from the crop model, were identified to be critical as a predictor for the ML model for the next month’s yield prediction. They also considered observed rainfall, temperature, radiation, and seasonal climate indices as input features for ML algorithms. This is an example of ML aiding the investigation of the relative importance of several predictors. Feng et al. [57] implemented a similar approach, with addition of climate extreme indices, to refine the prediction of wheat yield in Australia. They compared the performance of a process-based model APSIM with several combinations of the process-based model and ML models, such as APSIM+ multiple linear regression (MLR) and APSIM+ MLR+ RF. They reported an additional improvement of 19% in the yield prediction accuracy as compared to APSIM+MLR and 33% as compared to APSIM alone when the RF was included in the modeling chain.

Schlund et al. [58] developed a two-step approach to constrain the projected GPP at the end of the 21st century in the Representative Concentration Pathway 8.5 scenario with ML. They fed observational data into the ML algorithms that have been trained on Coupled Model Intercomparison Project (CMIP5) data to learn the relationships between present-day carbon estimates and the future scenario GPP (target variables). Their ML model showed superior performance to the CMIP5 ensemble mean. It also predicted an increasing GPP in the high latitudes. Folberth et al. [59] built novel crop meta-models from coarser GCMs to derive estimates of crop yield at a high spatial resolution without requiring the crop model set up at high-resolution. ML methods used in this study were RF and gradient boosting regression tree (GBRT). They were able to go to 0.25° from 0.5° and the results showed high accuracy ($R^2 > 0.96$) in predicting maize yields, ET and crop available water.

In a recent study, Shahhosseini et al. [60] showed that adding physics-based crop model variables as input features to ML models can improve the performance of ML models by 29% on average in the US Corn Belt. They compared the performances of several ML models such as RF, linear regression (LR), least absolute shrinkage and selection operator (LASSO) regression, Light Gradient Boost (LightGBM), Extreme Gradient Boost (XGBoost), and also an ensemble of them to investigate their added value individually and in combination. This study also assessed the relative importance of various variables from the process-based models as input features to different ML models. As such, they included phenology related variables, crop-related variables, and soil-related variables separately to investigate their individual added value in model predictions. The results indicated that the soil-water related variables were most important for crop yield prediction over the US corn belt.

4.4. Hybrid Simulation of Land-Surface Variables Other Than ET

Soil moisture, turbulent momentum, and heat flux are some of the critical variables of land-atmosphere interactions apart from ET. Pelissier et al. [61] combined Noah LSM and an ML technique to improve the prediction of top-layer soil moisture at nine AmeriFlux tower sites. They were able to obtain a 3-fold decrease in error metric. In particular, they highlighted the potential of learning structural error of a model using a hybrid model at a point-scale. However, they indicated that building a global scale model will need the inclusion of remote sensing data and account for uncertainty propagation through LSMs. In another study, ML was found better than the Monin-Obukhov similarity theory to predict momentum and heat fluxes [62]. The main objective of this paper was to substitute the Monin-Obukhov similarity theory for calculating fluxes in the LSM named 'Veg3d'. This technique of replacing a parameterization with an ML scheme has been extensively used in climate models [63,64]. [62] also showed that the results of ANN involved LSMs were equivalent, if not superior, to the conventional method. They indicated that implementation of an ML technique might save 5% of a central processing unit (CPU) time of a regional climate model, and this can be further improved by replacing all uncertain components of the climate models with neural network subroutines. This is a further impetus to implement ML methods for efficient prediction.

4.5. Benchmarking the LSMs

As the LSMs become increasingly complex and observational data volumes rapidly expand, there is a growing need for frequent and intensive testing and evaluating the models to fully utilize the richness of large Earth System data sets like satellite or FLUXNET measurements. Multi-scale synthesis and Terrestrial Model Intercomparison Project (MsT-MIP) [65,66] and Land Model Testbed (LMT) [67] are two such initiatives that incorporated ML techniques for this purpose. Schwalm et al. [68] quantified divergence as the spread in output from multiple models and an observational constraint. Considering inter-model spread as a measure of the approximations in physical and biogeochemical processes in the models, the ML experiments highlighted the uncertainties in the structures of the carbon

pools and advised against the hard parametric limits on ecosystem function. Their results confirmed that model intercomparison projects like MsMTIP ensemble provide a proper framework to link model skill to structure, excluding confounding factors while taking into account inter-model spread. RF helped to find model structural characteristics, which serve as important factors of skill for several MsTIP variables. LMT, with a similar objective, leveraged existing tools of International Land Model Benchmarking (ILAMB) and was able to launch thousands of ensemble simulations simultaneously while perturbing the parameters of LSMs. They developed the ML-based benchmarking workflow to increase the diagnostic capacity of LMT. New relationship metrics were designed using ML methods to benchmark the LMT produced ensemble simulations against in-situ measurements of GPP, ET, and SH. The initiative was an outstanding attempt to accelerate the model development cycle by means of careful assessment of model fidelity and analysis of model outputs.

5. Techniques of Machine Learning

ML methods are automated or semi-automated techniques of data inference without any prior assumptions (data-driven). ML techniques can be broadly classified into (1) supervised learning and (2) unsupervised learning. Supervised learning is based on a priori specification of output and one or more inputs. In unsupervised learning, there is only input and no outputs, and the model is aimed to discover patterns in the data. Most applications in land surface models are based on supervised learning, where neural networks or other machine learning models are trained to provide a known output from the model. The supervised ML techniques can be further classified into two categories of (a) traditional ML methods and (b) DL based methods. In this section, we discuss the specific ML techniques used by the previously mentioned papers (for different purposes), with brief details, under these two categories. Our goal was to identify the most popular ones from these two categories and include further detailed description about those. The discussion is also summarized in Table 1. Following convention, ANN (traditional) is considered as networks with single hidden layer and DL-ANNs are the networks with multiple hidden layers.

Table 1. Widely used ML algorithms used in land surface modeling.

ML Algorithm	Category	Purpose	Reference
Feedforward ANN	Traditional	Provide monthly estimates of GPP, LH and SH at a global scale	[42]
ELM and GRNN		Obtain ET ₀ from temperature data in southwest China	[44]
RT, Bagging, RF and SVM		Provide ET estimates in central Florida	[46]
SVM and GANN		Provide crop ET estimates in China	[47]
RF, RT, KRR, RS, ANN		Provide CO ₂ , LH, SH, R _n at multiple sites globally	[48]
Extra-Trees		Obtain global parameter estimates for Noah land model	[52]
MCMC		Parameter optimization of land model	[54]
RF		Improving regional crop yield prediction	[56,57,59,60]
GBRT		Constrain uncertainty in GPP estimates	[58]
DL-ANN	Deep learning	Obtain LH estimates over ocean	[43]
DL-ANN		Constructing physics-constrained ML model	[49]
DL-ANN		Constructing emulators for land modeling	[55]

5.1. Traditional ML Methods

Common traditional ML methods include SVM, CART, bootstrap aggregating (Bagging), Kernel ridge regression (KRR), RF, ANN etc. Alemohammad et al. [42] used a feedforward ANN as a supervised learning approach to retrieve monthly estimates of GPP, LE, and SH at a global scale using remotely sensed SIF estimates besides other meteorological

logical predictors like precipitation, temperature, soil moisture, snow cover, and R_n . The ANN consisted of three layers: (1) an input layer connected to input data, (2) one hidden layer, and (3) an output layer producing LE, SH, and GPP outputs. A number of neurons in the hidden layer were chosen based on the complexity of the problem. [44] compared the performance of an extreme learning machine (ELM) technique and generalized regression neural network (GRNN) model in predicting reference evapotranspiration (ET_o). Their ELM consisted of 50 hidden nodes, which performed better than the GRNN, which is similar to feedforward neural networks but based on nonlinear regression theory of function estimation. More specifically, GRNN contains four layers named summation layer, input layer, pattern layer, and output layer [68]. They were able to reproduce more accurate ET_o estimates than the conventional empirical Hargreaves model. A similar study was conducted by Gocic et al. [45] over two weather stations in Serbia. [46] compared the supervised learning approaches Regression Tree (RT), Bagging, RF, and SVM while predicting actual ET. An RT model utilizes a decision tree as a predictive model, and target variables are real values. RTs can be further advanced by gradient boosting (GBRT) [58,59]. ‘Bagging’ is a machine learning ensemble technique. RF is a set of uncorrelated simple regression trees. In contrast, SVMs are supervised learning models to deal with classification and regression problems [69]. Some studies have used genetic algorithms to optimize the ANN (GANN) [47] to predict crop-specific ET estimates.

Apart from these commonly used techniques, KRR [70] and regression splines (RS) [71] are also used by few studies to estimate ET [48]. Global model parameter estimation in [52] benefited from Extra-Trees [72]. MCMC has also been used with ML (Gaussian Process Regression) for parameter optimization [54]. RF is the most commonly used technique for crop yield estimate [56,57]. To understand the relative importance of predictors, which may provide information about the usefulness of them at predicting the target variable, ‘feature importance’ is normally computed internally by these methods. Overall, ANN and RF are the most widely used traditional ML methods in LSM applications. However, RTs with improved ‘gradient boosting’ or ‘Bagging’ can be extremely powerful to create data-driven models for land-surface applications. We discuss RF in further due to its significant vogue in past literature.

Random Forests (RF)

RF [73–75] are supervised non-parametric ML algorithms, which are ensemble-learning methods using decision trees as base learners. This is a fast, flexible and robust approach to high-dimensional data mining that relies on aggregating the results of an ensemble of simple estimators. Decision trees are intuitive ways to classify objects or label objects by binary splitting. Variable space is divided into a set of boundaries, and a model is fitted to each set (which can be a constant in the simplest case). However, overfitting tendency of such trees can be reduced by an ensemble of randomized trees and that motivates the usage of ‘bagging’ or RF technique (equipped with some additional randomization techniques reducing the correlation between the trees). In LSM applications, RF is mostly used within the context of regression, i.e.; for continuous prediction rather than categorical. Users generally provide the number of trees in the forest and a criterion to measure the quality of a split, such as mean square error. Hyperparameters are also provided to improve the model performance and control overfitting. In general, statistical learning has two main purposes: (1) prediction and (2) inference. In the case of RF, interpreting the model can be done either by looking at a single tree in the forest or by examining the relative importance of independent variables. Overall, RF method is found to be (1) consistent (2) reducing the bias and variance simultaneously, (3) achieving convergence at high rate and (4) adaptive to sparsity, which makes it applicable even with noisy predictor variables [74].

RF is often superior to other traditional ML methods like SVM or CART because it (1) considers an ensemble of decision trees and does not need multiple models like others (e.g.; SVM) to provide a probabilistic prediction, (2) provides more generalized solution,

(3) needs no feature scaling (4) can handle high dimensional spaces as well as large number of training examples.

5.2. Deep Learning (DL) methods

Some commonly used DL methods include DL-ANN, CNN, LSTM, variational auto-encoders (VAN) and generative adversarial networks (GAN). However, till date, DL-ANNs are the only deep learning tools used with land modeling. For example, [43] used a neural network with two hidden layers containing fifteen neurons in total and LH, and SH were output from the model. Physics guided machine learning models have also used DL-ANN as the ML technique [49]. DL-ANNs are also useful in constructing emulators of LSMs in to characterize the uncertainties and parameter optimization [55]. Dew point temperature is an important parameter to assess surface humidity conditions, critical for agricultural purposes. Constantin et al. [76] realistically predicted dew point temperatures at sub-daily scale using other meteorological variables as input to a DL-ANN consisting of three hidden layers. We discuss ANN and DL in further detail below.

Artificial Neural Networks (ANN)

ANN is the supervised learning approach to approximate relationships between input and output using interconnecting units called neurons. The development of ANNs was motivated by the functioning of the human brain. Like the brain, the connections between neurons decide the function of the network. Multiple inputs given to the neuron are assigned with different weights, calculated by minimizing a loss function, that collectively determines the importance of input signals [77]. This part is termed as ‘training’ and some popular training methods are backpropagation, evolutionary algorithm and gradient descent. Output signal is produced by the summation block, which adds all of the weighted inputs algebraically. The input and output layers are called ‘visible layers’ as they are directly connected to inputs and outputs, respectively. Hidden layers are the layers in the middle (see Figure 2). ‘Activations’ are the outputs sent to the $(i + 1)^{\text{th}}$ layer from the i^{th} layer, which is calculated based on the transformation functions (such as sigmoidal, tanh, rectified linear units or ReLU) of combined input from $(i - 1)^{\text{th}}$ layer. ‘Width’ is defined by the number of neurons on a layer and ‘depth’ is defined by the number of layers in the network.

Deep Learning (DL) is a suite of tools based on carefully designed large ANNs with multiple hidden layers. Compared to non-deep networks and traditional earlier-generation ML methods, DL is characterized by an abundance of layers to process the complex information in big data, addition of unsupervised learning units and effective regularization techniques. These allow automatic extraction of features from input data. By contrast, traditional ML models, such as CART and SVM, need human construction of relevant input features for best results, which is often the most tedious step. Many of these conventional methods are also not suitable for generalization purposes and working with large data sets that DL can handle. Though DL-ANN are the only DL techniques used with land modeling so far, LSMs can be further benefitted from some of these techniques previously in water science, especially hydrology [78,79] and envisage more potential uses for those in the future. For example, CNN can be used to estimate crop yields by merging LSM and satellite data. However, DL can face some of the issues like other ML methods, such as convergence of the learning process dependent on training data, hidden layers acting as a black box, etc.

In addition to DL techniques, gaussian process emulation can be used for the simple representation of complex land-surface features and test different potential values of parameters. High-resolution land cover data can be obtained using Kernel Regression. Wavelet transform can also be used to quantify the information transfer between different model components of an LSM.

6. Possible future directions

Combining the applications of ML in past literature and keeping in mind the still existing problems in land surface modeling, we provide some future directions to improve the current status of land modeling leveraging machine learning techniques.

1. ML techniques can be a suitable way to reduce the process complexity of the LSMs. Currently, the complex interactive processes of the terrestrial system and way they are represented in conventional LSMs makes them intractable. It is often difficult to assess the added value of a complex process given its cost of burdening the model. ML can help in this regard to take a modular approach, where modelers can represent a complex process in multiple ways and test the model performance easily. This will help in reducing the structural uncertainty of the models as well. The model intercomparison projects (e.g.; MsTMIP) are often constrained by the under-sampling of the potential range of model configurations. Such technique will also help us in collaborative development of the model, rather than adding processes on a specific need basis.
2. Processes in the model which have little physics knowledge or complex calculation, but more data availability, could be replaced by a ML-based surrogate. As mentioned in Section 4.4, there have been some attempts to predict important land surface properties by a hybrid modeling approach, but we are still lacking in exploring some of the more fundamental variables which can be easily provided by ML applications as the amount of observations increase in recent days. For example, hydrology, phenology and snow cover fraction. Hydrologic observations are increasing all over the globe with advanced velocimetry techniques and phenology is easily obtained now from satellite remote sensing data. Improving these components leveraging ML will help upgrade the overall LSM performance.
3. Parametric uncertainty can be investigated, and parameters can be better optimized following Dagon et al. [55]. However, we need to consider longer data records (now possible with the availability of computing resources) and more output variables. Good quality globally gridded observations are now available. For example, Atmospheric CO₂ from National Oceanic and Atmospheric Administration (NOAA), Commonwealth Scientific and Industrial Research Organisation (CSIRO), gridded FLUXNET from Max Planck Institute for Biogeochemistry (MPI-BCG), river flow (GRDC), albedo (MODIS), and so on (ILAMB project, etc.).
4. Data assimilation is proved to be an important tool to improve model simulations at different earth system modeling applications [80,81]; and advancement in ML can further enhance that. Conventional data assimilation methods include Kalman filter and variational approaches. These methods have underlying assumptions of normality, Markovian processes, zero error covariance and similar ML algorithms, being completely data-driven, may improve the assimilation in terms of speed, accuracy, and efficiency.
5. There are several avenues for better crop yield prediction. Soil hydrology related variables such as soil moisture or drought indices (obtained from either remote sensing, in-situ measurements, or process-based models) can further improve the ML predictions on crop yield. Precision agriculture is another advanced recent agricultural technology which includes sensors, robotics, and AI to assemble 'big data' which can further be processed with ML models to develop a sustainable agricultural practice with enhanced yield. We can use such data and the extracted information to assess the yield variability among regions to make informed decisions [82]. As such, the yield stability maps, included in the landscape, can provide information about environment friendly areas and constant low yield areas. We should focus on intensifying the high yield areas sustainably. On the other hand, low yield zones can be improved by perennial bioenergy crops and recycling of plant available nutrients.
6. Interpretability of ML models [83] has potential to reveal some of the hidden links and physics between different parts of the terrestrial hydrology. Relative importance

of the predictors to predict a specific outcome could be more rigorously analyzed for similar interpretations.

7. Conclusions

The implementation of ML techniques in earth system modeling has been widely accepted by many researchers in recent years. Direct application of such techniques, particularly on LSM, is quite recent. LSM plays an important role in simulating the feedback and interactions between the land surface and the atmosphere. As we escalate toward rapid changes in the climate and land use, the future of the terrestrial biosphere and hydrology remains a global concern. LSMs are the fundamental tools to simulate and predict the state of the terrestrial surface and play a critical role in optimizing the amount of carbon dioxide to limit global warming. This provides motivation for improving the performance of LSMs.

The complex nature of linked land-surface processes makes it difficult to create a tractable LSM. The interactive components include atmosphere, urban environment, canopy processes, agriculture, hydrology, soil dynamics, snow physics, soil processes and so on. Recently with the need for hydrologists, ecologists, biogeochemists, physiologists, the structure of LSMs is extremely complex. We argue that ML methods can help us achieve better performance of LSMs, make it efficient and also create a flexible structure of the models. With the help of ML and adequate data in recent ages, some critical applications of LSM have already been upgraded.

The main improvements were in the areas of ET estimation, parameter calibration, crop yield, model benchmarking, and uncertainty quantification. ET estimation is important in earth system modeling as ET is an important part of both water and energy cycles. ML methods were able to use satellite data and improve global ET estimates. The major complexity of the LSMs come from a significant number of parameterized processes within LSMs. ML surrogates were able to replace some of these processes to improve the model performance and reduce the uncertainty. Crop yield prediction majorly benefitted from the 'hybrid' model approach where process-based models and ML models were combined to achieve better yield predictions. Model benchmarking is a relatively newer concept but an extremely important part of LSM improvement, and ML has been helpful in this context too.

The widely used ML techniques along with land surface modeling are ANN, RF, and modified versions of Gradient trees. Supervised learning is often used with the help of these techniques to improve the LSM performance. Other ML techniques are also often used along with these, such as genetic algorithms, SVM, GBRT, 'Bagging,' MCMC, etc. We foresee that there is scope to apply the recent DL methods like CNN and LSTM regression to improve several parts of the land surface modeling, following other applications in water science and earth system modeling. For example, high-resolution data preparation and improved data assimilation techniques can leverage these advanced methods.

Additional improvements are possible in the areas of complexity reduction of LSMs, multiple parameter calibration, structure revision, uncertainty reduction, and sustainable crop yield estimate. Interpretability of ML models should be carefully investigated to further explore the underlying physics and processes behind the yet less-known land surface-related processes.

Author Contributions: Conceptualization, S.P.; P.S.; methodology, S.P.; P.S.; writing—original draft preparation, S.P.; writing—review and editing, P.S. and S.P.; visualization, S.P.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to thank the Department of Atmospheric Science of University of Illinois at Urbana-Champaign for providing us with resources for the study. We gratefully thank the three anonymous reviewers for their valuable comments and suggestions, which helped to improve the paper. We also thank the editor for editorial guidance.

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

References

- Huntingford, C.; Jeffers, E.S.; Bonsall, M.B.; Christensen, H.M.; Lees, T.; Yang, H. Machine learning and artificial intelligence to aid climate change research and preparedness. *Environ. Res. Lett.* **2019**, *14*, 124007. [CrossRef]
- Zhang, P.; Zhang, L.; Leung, H.; Wang, J. A deep-learning based precipitation forecasting approach using multiple environmental factors. In *2017 IEEE International Congress on Big Data (BigData Congress)*; IEEE: New York, NY, USA, 2017; pp. 193–200. [CrossRef]
- Shen, C. A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res.* **2018**, *54*, 8558–8593. [CrossRef]
- Schmidt, L.; Heße, F.; Attinger, S.; Kumar, R. Challenges in applying machine learning models for hydrological inference: A case study for flooding events across Germany. *Water Resour. Res.* **2020**, *56*, e2019WR025924. [CrossRef]
- Mou, L.; Ghamisi, P.; Zhu, X.X. Deep recurrent neural networks for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [CrossRef]
- Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [CrossRef]
- Koster, R.D.; Dirmeyer, P.A.; Guo, Z.; Bonan, Z.; Chan, E.; Cox, P.; Gordon, C.T.; Kanae, S.; Kowalczyk, E.; Lawrence, D.; et al. Regions of strong coupling between soil moisture and precipitation. *Science* **2004**, *305*, 1138–1140. [CrossRef] [PubMed]
- Guo, Z.; Dirmeyer, P.A.; Koster, R.D.; Sud, Y.C.; Bonan, G.; Oleson, K.W.; Chan, E.; Verseghy, D.; Cox, P.; Gordon, C.T.; et al. GLACE: The Global Land Atmosphere Coupling Experiment. 2. Analysis. *J. Hydrometeor.* **2006**, *7*, 611–625. [CrossRef]
- Dirmeyer, A.P.; Schlosser, C.A.; Brubaker, K.L. Precipitation, recycling and land memory: An integrated analysis. *J. Hydrometeor.* **2009**, *10*, 278–288. [CrossRef]
- Abramowitz, G.; Leuning, R.; Clark, M.; Pitman, A. Evaluating the Performance of Land Surface Models. *J. Clim.* **2008**, *21*, 5468–5481. [CrossRef]
- Fisher, R.A.; Koven, C.D. Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. *J. Adv. Model. Earth Syst.* **2020**, *12*, e2018MS001453. [CrossRef]
- Richardson, L.F. *Weather prediction by numerical process* Cambridge University Press. *Q. J. Royal Meteorol. Soc.* **1922**, *48*, 282–284. [CrossRef]
- Charney, J.G.; Quirk, W.J.; Chow, S.H.; Kornfield, J. A comparative study of the effects of albedo change on drought in semiarid regions. *J. Atmos. Sci.* **1977**, *34*, 1366–1385. [CrossRef]
- Matthews, E. Global vegetation and land use: New high-resolution data bases for climate studies. *J. Clim. Appl. Meteorol.* **1983**, *22*, 474–487. [CrossRef]
- Nicholson, S.E. Land surface-atmosphere interaction: Physical processes and surface changes and their impact. *Prog. Phys. Geogr.* **1988**, *12*, 36–65. [CrossRef]
- Garratt, J.R. Sensitivity of Climate Simulations to Land-Surface and Atmospheric Boundary-Layer Treatments-A Review. *J. Clim.* **1993**, *6*, 419–448. [CrossRef]
- Koster, R.D.; Suarez, M.J.; Higgins, R.W.; Van den Dool, H.M. Observational evidence that soil moisture variations affect precipitation. *Geophys. Res. Lett.* **2003**, *30*, 1241. [CrossRef]
- Dirmeyer, P.; Randal, A.; Koster, D.; Guo, Z. Do Global Models Properly Represent the Feedback between Land and Atmosphere? *J. Hydrometeorol.* **2006**, *7*, 1177–1198. Available online: <http://www.jstor.org/stable/24910939> (accessed on 25 November 2020). [CrossRef]
- Seneviratne, S.I.; Thierry, C.; Edouard, L.D.; Hirschi, M.; Jaeger, E.B.; Lehner, I.; Orlowsky, B.; Teuling, J.A. Investigating soil moisture–climate interactions in a changing climate: A review. *Earth Sci. Rev.* **2010**, *99*, 125–161. [CrossRef]
- Sellers, P.J.; Mintz, Y.; Sud, Y.C.; Dalcher, A. A simple biosphere model (SIB) for use within general circulation models. *J. Atmos. Sci.* **1986**, *43*, 505–531. [CrossRef]
- De Kauwe, M.G.; Kala, J.; Lin, Y.-S.; Pitman, A.J.; Medlyn, B.E.; Duursma, R.A.; Abramowitz, G.; Wang, Y.-P.; Miralles, D.G. A test of an optimal stomatal conductance scheme within the CABLE land surface model. *Geosci. Mod. Dev.* **2015**, *8*, 431–452. [CrossRef]
- Ding, R.; Kang, S.; Du, T.; Hao, X.; Zhang, Y. Scaling Up Stomatal Conductance from Leaf to Canopy Using a Dual-Leaf Model for Estimating Crop Evapotranspiration. *PLoS ONE* **2014**, *9*, e95584. [CrossRef]
- Liang, X.; Wood, E.F.; Lettenmaier, D.P. Surface soil moisture parameterization of the VIC-2L model: Evaluation and modification. *Global Planet. Chang.* **1996**, *13*, 195–206. [CrossRef]
- Takata, K.; Emori, S.; Watanabe, T. Development of the minimal advanced treatments of surface interaction and runoff. *Glob. Planet. Chang.* **2003**, *38*, 209–222. [CrossRef]
- Chen, F.; Xie, Z. Effects of crop growth and development on regional climate: A case study over East Asian monsoon area. *Clim. Dyn.* **2012**, *38*, 2291–2305. [CrossRef]

26. Kucharik, C.J.; Brye, K.R. Integrated Biosphere Simulator (IBIS) Yield and Nitrate Loss Predictions for Wisconsin Maize Receiving Varied Amounts of Nitrogen Fertilizer. *J. Environ. Qual.* **2003**, *32*. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Chen, F.; Yates, D.N.; Nagai, H.; LeMone, M.A.; Ikeda, K.; Grossman, R.L. Land surface heterogeneity in the cooperative atmosphere surface exchange study (CASES-97). Part I: Comparing modeled surface flux maps with surface-flux tower and aircraft measurements. *J. Hydrometeorol.* **2003**, *4*, 196–218. [\[CrossRef\]](#)
28. Dickinson, R.E.; Shaikh, M.; Bryant, R.; Graumlich, L. Interactive canopies for a climate model. *J. Clim.* **1998**, *11*, 2823–2836. [\[CrossRef\]](#)
29. Ivanov, V.Y.; Bras, R.L.; Vivoni, E.R. Vegetation-hydrology dynamics in complex terrain of semiarid areas: 2. Energy-water controls of vegetation spatiotemporal dynamics and topographic niches of favorability. *Water Resour. Res.* **2008**, *44*, 1–20. [\[CrossRef\]](#)
30. Lipson, M.; Hart, M.; Thatcher, M. Efficiently modelling urban heat storage: An interface conduction scheme in an urban land surface model (aTEB v2.0). *Geosci. Model Dev.* **2017**, *10*, 991–1007. [\[CrossRef\]](#)
31. Lawrence, D.M.; Fisher, R.A.; Koven, C.D.; Oleson, K.W.; Swenson, S.C.; Bonan, G.; Collier, N.; Ghimire, B.; van Kampenhout, L.; Kennedy, D.; et al. The Community Land Model Version 5: Description of new features, benchmarking, and impact of forcing uncertainty. *J. Adv. Model. Earth Syst.* **2017**, *11*, 245–4287. [\[CrossRef\]](#)
32. Decharme, B.; Delire, C.; Minvielle, M.; Colin, J.; Vergnes, J.; Alias, A.; Saint-Martin, D.; Séférian, R.; Sénézi, S.; Voldoire, A. Recent changes in the ISBA-CTRIP land surface system for use in the CNRM-CM6 climate model and in global off-line hydrological applications. *J. Adv. Model. Earth Syst.* **2019**, *11*, 1207–1252. [\[CrossRef\]](#)
33. Fisher, R.A.; Koven, C.D.; Anderegg, W.R.; Christoffersen, B.O.; Dietze, M.C.; Farrior, C.E.; Holm, J.A.; Hurtt, G.C.; Knox, R.G.; Lawrence, P.J.; et al. Vegetation demography in Earth System Models: A review of progress and priorities. *Glob. Chang. Biol.* **2018**, *24*, 35–54. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Miguez-Macho, G.; Fan, Y.; Weaver, C.P.; Walko, R.; Robock, A. Incorporating water table dynamics in climate modeling: 2. Formulation, validation, and soil moisture simulation. *J. Geophys. Res.* **2007**, *112*, D13108. [\[CrossRef\]](#)
35. Yao, Q.; Li, Z.; Song, Y.; Wright, S.J.; Guo, X.; Tringe, S.G.; Tfaily, M.M.; Paša-Tolić, L.; Hazen, T.C.; Turner, B.L.; et al. Community proteogenomics reveals the systemic impact of phosphorus availability on microbial functions in tropical soil. *Nat. Ecol. Evol.* **2018**, *2*, 499–509. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Boone, A.; Samuelsson, P.; Gollvik, S.; Napoly, A.; Jarlan, L.; Brun, E.; Decharme, B. The interactions between soil-biosphere-atmosphere land surface model with a multi-energy balance (ISBA-MEB) option in SURFEXv8-Part 1: Model description. *Geosci. Model Dev.* **2017**, *10*. [\[CrossRef\]](#)
37. Nabel, J.E.M.S.; Naudts, K.; Pongratz, J. Accounting for forest age in the tile-based dynamic global vegetation model JSBACH4 (4.20p7; git feature/forests)—A land surface model for the ICON-ESM. *Geosci. Model Dev.* **2020**, *13*, 185–200. [\[CrossRef\]](#)
38. Wiltshire, A.J.; Duran Rojas, M.C.; Edwards, J.M.; Gedney, N.; Harper, A.B.; Hartley, A.J.; Hendry, M.A.; Robertson, E.; Smout-Day, K. JULES-GL7: The Global Land configuration of the Joint UK Land Environment Simulator version 7.0 and 7.2. *Geosci. Model Dev.* **2020**, *13*, 483–505. [\[CrossRef\]](#)
39. Smith, B.; Samuelsson, P.; Wramneby, A.; Rummukainen, M. A model of the coupled dynamics of climate, vegetation and terrestrial ecosystem biogeochemistry for regional applications. *Tellus A* **2010**, *63*, 87–106. [\[CrossRef\]](#)
40. Niu, G.Y.; Yang, Z.L.; Mitchell, K.E.; Chen, F.; Ek, M.B.; Barlage, M.; Kumar, A.; Manning, K.; Niyogi, D.; Rosero, E.; et al. The community Noah land surface model with multiparameterization options (NoahMP): 1. Model description and evaluation with local scale measurements. *J. Geophys. Res.* **2011**, *116*, D12109. [\[CrossRef\]](#)
41. Jones, C.G.; Lawton, J.H.; Shachak, M. Organisms as ecosystem engineers. *JSTOR* **1994**, *69*, 373–386. [\[CrossRef\]](#)
42. Alemohammad, S.H.; Fang, B.; Konings, A.G.; Aires, F.; Green, J.K.; Kolassa, J.; Miralles, D.; Prigent, C.; Gentine, P. Water, Energy, and Carbon with Artificial Neural Networks (WECANN): A statistically based estimate of global surface turbulent fluxes and gross primary productivity using solar induced uorescence. *Biogeosciences* **2017**, *14*, 4101–4124. Available online: <https://bg.copernicus.org/articles/14/4101/2017/> (accessed on 6 February 2021). [\[CrossRef\]](#) [\[PubMed\]](#)
43. Bourras, D.; Eymard, L.; Liu, W.T. A neural network to estimate the latent heat flux over oceans from satellite observations. *Int. Remote Sens. J.* **2002**, *23*, 2405–2424. [\[CrossRef\]](#)
44. Feng, Y.; Peng, Y.; Cui, N.; Gong, D.; Zhang, K. Modeling reference evapotranspiration using extreme learning machine and generalized regression neural network only with temperature data. *Comput. Electron. Agric.* **2017**, *136*, 71–78. [\[CrossRef\]](#)
45. Gocic, M.; Petković, D.; Shamshirband, S.; Kamsin, A. Comparative analysis of reference evapotranspiration equations modelling by extreme learning machine. *Comput. Electron. Agric.* **2016**, *127*, 56–63. [\[CrossRef\]](#)
46. Granata, F. Evapotranspiration evaluation models based on machine learning algorithms—A comparative study. *Agric. Water Manag.* **2019**, *217*, 303–315. [\[CrossRef\]](#)
47. Tang, D.; Feng, Y.; Gong, D.; Hao, W.; Cui, N. Evaluation of artificial intelligence models for actual crop evapotranspiration modeling in mulched and non-mulched maize croplands. *Comput. Electron. Agric.* **2018**, *152*, 375–384. [\[CrossRef\]](#)
48. Tramontana, G.; Jung, M.; Schwalm, C.R.; Ichii, K.; Camps-Valls, G.; Ráduly, B.; Reichstein, M.; Arain, M.A.; Cescatti, A.; Kiely, G.; et al. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences* **2016**, *13*, 4291–4313. [\[CrossRef\]](#)
49. Zhao, W.L.; Gentine, P.; Reichstein, M.; Zhang, Y.; Zhou, S.; Wen, Y.; Lin, C.; Li, X.; Qiu, G.Y. Physics constrained machine learning of evapotranspiration. *Geophys. Res. Lett.* **2019**, *46*, 14496–14507. [\[CrossRef\]](#)
50. Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195–204. [\[CrossRef\]](#) [\[PubMed\]](#)

51. Pan, S.; Pan, N.; Tian, H.; Friedlingstein, P.; Sitch, S.; Shi, H.; Arora, V.K.; Haverd, V.; Jain, A.K.; Kato, E. Evaluation of global terrestrial evapotranspiration by state-of-the-art approaches in remote sensing, machine learning, and land surface models. *Hydrol. Earth Syst. Sci. Discuss* **2019**, 1–51. [\[CrossRef\]](#)
52. Chaney, N.W.; Herman, J.D.; Ek, M.B.; Wood, E.F. Deriving global parameter estimates for the Noah land surface model using FLUXNET and machine learning. *J. Geophys. Res. Atmos.* **2016**, *121*, 13218–13235. [\[CrossRef\]](#)
53. Beven, K.; Freer, J. Equinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *J. Hydrol.* **2001**, *249*, 11–29. [\[CrossRef\]](#)
54. Sawada, Y. Machine Learning Accelerates Parameter Optimization and Uncertainty Assessment of a Land Surface Model. *J. Geophys. Res. Atmos.* **2020**. [\[CrossRef\]](#)
55. Dagon, K.; Sanderson, B.M.; Fisher, R.A.; Lawrence, D.M. A machine learning approach to emulation and biophysical parameter estimation with the Community Land Model, version 5. *Adv. Stat. Clim. Meteorol. Oceanogr.* **2020**, *6*, 223–244. [\[CrossRef\]](#)
56. Everingham, Y.; Sexton, J.; Skocaj, D.; Inman-Bamber, G. Accurate prediction of sugarcane yield using a random forest algorithm. *Agron. Sustain. Dev.* **2016**, *36*, 27. [\[CrossRef\]](#)
57. Feng, P.; Wang, B.; Liu, D.; Waters, C.; Xiao, D.; Shi, L.; Yu, Q. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. *Agric. Forest Meteorol.* **2020**, 285–286, 107922, ISSN 0168-1923. [\[CrossRef\]](#)
58. Schlund, M.; Eyring, V.; Camps-Valls, G.; Friedlingstein, P.; Gentine, P.; Reichstein, M. Constraining uncertainty in projected gross primary production with machine learning. *J. Geophys. Res. Biogeosci.* **2020**, *125*, e2019JG005619. [\[CrossRef\]](#)
59. Christian, F.; Baklanov, A.; Balkovič, J.; Skalsky, R.; Khabarov, N.; Obersteiner, M. Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. *Agric. Forest Meteorol.* **2019**, *264*, 1–15. [\[CrossRef\]](#)
60. Shahhosseini, M.; Guiping Hu, S. Archontoulis and Isaiah Huber. Coupling Machine Learning and Crop Modeling Improves Crop Yield Prediction in the US Corn Belt. *arXiv Preprint* **2020**, arXiv:abs/2008.04060.
61. Pelissier, C.; Frame, J.; Nearing, G. Combining parametric land surface models with machine learning. *arXiv Preprint* **2020**, arXiv:abs/2002.06141.
62. Leufen, L.H.; Schädler, G. Calculating the Turbulent Fluxes in The Atmospheric Surface Layer with Neural Networks. *Geosci. Model Dev.* **2019**, *12*, 2033–2047. [\[CrossRef\]](#)
63. Gentine, P.; Pritchard, M.; Rasp, S.; Reinaudi, G.; Yacalis, G. Could machine learning break the convection parameterization deadlock? *Geophys. Res. Lett.* **2018**, *45*, 5742–5751. [\[CrossRef\]](#)
64. Rasp, S.; Pritchard, M.S.; Gentine, P. Deep learning to represent sub-grid processes in climate models. *Proc. Natl. Acad. Sci. USA* **2018**, *39*, 9684–9689. [\[CrossRef\]](#) [\[PubMed\]](#)
65. Huntzinger, D.N.; Schwalm, C.R.; Wei, Y.; Cook, R.B.; Michalak, A.M.; Schaefer, K.; Jacobson, A.R.; Arain, M.A.; Ciais, P.; Fisher, J.B.; et al. *NACP MsTMIP: Global 0.5-deg Terrestrial Biosphere Model Outputs (Version 1) in Standard Format*; ORNL DAAC: Oak Ridge, TN, USA, 2016. [\[CrossRef\]](#)
66. Schwalm, C.R.; Schaefer, K.; Fisher, J.B.; Huntzinger, D.; Elshorbany, Y.; Fang, Y.; Hayes, D.; Jafarov, E.; Michalak, A.M.; Piper, M. Divergence in land surface modeling: Linking spread to structure. *Environ. Res. Commun.* **2019**, *1*, 111004. [\[CrossRef\]](#)
67. Sreepathi, S.; Xu, M.; Collier, N.; Kumar, J.; Mao, J.; Hoffman, F.M. Land Model Testbed: Accelerating Development, Benchmarking and Analysis of Land Surface Models. *OSFHOME* **2020**. [\[CrossRef\]](#)
68. Specht, D.F. A general regression neural network. *IEEE Trans. Neural Netw.* **1991**, *2*, 568–576. [\[CrossRef\]](#)
69. Ren, Y.Z.; Xia, K.W.; Wang, Y.; Shi, J. Application on Network Traffic Prediction Based on Least Squares Support Vector Machine. *Appl. Mech. Mater.* **2010**, *20–23*, 364–369. [\[CrossRef\]](#)
70. Shawe-Taylor, J.; Cristianini, N. *Kernel Methods for Pattern Analysis*; Cambridge University Press: Cambridge, UK, 2004. [\[CrossRef\]](#)
71. Alonso Fernández, J.R.; García Nieto, P.J.; Díaz Muñoz, C.; Álvarez Antón, J.C. Modeling eutrophication and risk prevention in a reservoir in the Northwest of Spain by using multivariate adaptive regression splines analysis. *Ecol. Eng.* **2014**, *68*, 80–89. [\[CrossRef\]](#)
72. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [\[CrossRef\]](#)
73. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
74. Tyrallis, H.; Papacharalampous, G.; Langousis, A. A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. *Water* **2019**, *11*, 910. [\[CrossRef\]](#)
75. Nguyen, L.H.; Joshi, D.R.; Clay, D.E.; Henebry, G.M. Characterizing land cover/land use from multiple years of Landsat and MODIS time series: A novel approach using land surface phenology modeling and random forest classifier. *Remote Sens. Environ.* **2018**, 111017. [\[CrossRef\]](#)
76. Constantin, I.; Lungu, M.L.; Panaitescu, L.; Ilie, M.; Simulating Lungu, D.; Nita, S. Simulating for predicting the hourly dew point temperature using artificial neural networks. *J. Environ. Prot. Ecol.* **2014**, *15*, 1101–1109. Available online: https://www.researchgate.net/publication/285526297_Simulating_for_predicting_the_hourly_dew_point_temperature_using_artificial_neural_networks (accessed on 11 March 2021).
77. Park, Y.-S.; Lek, S. Artificial Neural Networks: Multilayer Perceptron for Ecological Modeling. *Dev. Environ. Model.* **2016**, *28*, 123–140. [\[CrossRef\]](#)
78. Fang, K.; Shen, C.; Kifer, D.; Yang, X. Prolongation of SMAP to Spatiotemporally Seamless Coverage of Continental, U.S. Using a Deep Learning Neural Network. *Geophys. Res. Lett.* **2017**, *44*, 11030–11039. [\[CrossRef\]](#)

-
79. Rahmani, F.; Lawson, K.; Ouyang, W.; Appling, A.; Oliver, S.; Shen, C. Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. *Environ. Res. Lett.* **2021**. [[CrossRef](#)]
 80. Li, Y.; Ryu, D.; Western, A.W.; Wang, Q. Assimilation of stream discharge for flood forecasting. Updating a semidistributed model with an integrated data assimilation scheme. *Water Resour. Res.* **2015**. [[CrossRef](#)]
 81. Pal, S.; Dominguez, F.; Dillon, M.E.; Alvarez, J.; Garcia, C.M.; Nesbitt, S.W.; Gochis, D. Hydrometeorological Observations and Modeling of an Extreme Rainfall Event using WRF and WRF-Hydro during the RELAMPAGO Field Campaign in Argentina. *J. Hydrometeor.* **2021**, *22*, 331–351. [[CrossRef](#)]
 82. Basso, B.; Antle, J. Digital agriculture to design sustainable agricultural systems. *Nat. Sustain.* **2020**, *3*, 254–256. [[CrossRef](#)]
 83. Kohlbrenner, M.; Bauer, A.; Nakajima, S.; Binder, A.; Samek, W.; Lapuschkin, S. Towards best practice in explaining neural network decisions with LRP. In Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; Available online: <https://arxiv.org/abs/1910.09840> (accessed on 6 February 2021).