High-Fidelity Deep Approximation of Ecosystem Simulation over Long-Term at Large Scale

Zhihao Wang
zhwang1@umd.edu
University of Maryland

Yiqun Xie*
xie@umd.edu
University of Maryland

Xiaowei Jia
xiaowei@pitt.edu
University of Pittsburgh

Lei Ma
lma6@umd.edu
University of Maryland

George Hurtt
gchurtt@umd.edu
University of Maryland

ABSTRACT

Ecosystem services, such as carbon sequestration, biodiversity, and climate regulation, play essential roles in combating climate change. Projection of ecosystem dynamics under various scenarios is critical in understanding potential impacts and informing policies and mitigation strategies. Ecosystem Demography (ED) model is a major mechanistic model for ecosystem dynamics projection, but its computational cost has been a major bottleneck in performing large-scale (e.g., global, national) projections at very high spatial resolution. We aim to approximate the ED model using deep neural networks at operational high accuracy to assist large-scale climate studies. The deep approximation is non-trivial due to challenges by long-term error accumulation (e.g., 40 years), highly diverse scenarios, and high cost in training data generation. We propose a Deep-ED approximation model to address the challenges with a multi-scale cumulative loss reduction structure, significance-based scenario partitioning, self-guided forwarding, and physics-aware active learning strategies. Experiment results in the northeastern US demonstrate the high accuracy of Deep-ED and its potential in large-scale ecosystem projection.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Information systems** → **Spatial-temporal systems**.

KEYWORDS

Deep learning, long-term, ecosystem demography, physical model

ACM Reference Format:
Zhihao Wang, Yiqun Xie, Xiaowei Jia, Lei Ma, and George Hurtt. 2023. High-Fidelity Deep Approximation of Ecosystem Simulation over Long-Term at Large Scale. In *The 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23), November 13–16, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3589132.3625577

*Corresponding author.

SIGSPATIAL '23, November 13–16, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0168-9/23/11.
https://doi.org/10.1145/3589132.3625577

1 INTRODUCTION

Climate change has led to a wide range of impacts on environment, health, and society, including global warming, sea level rise, extreme events, natural disasters, and loss of biodiversity. According to the United Nations' Sustainable Development Goals (SDGs), "climate action" is a major area of focus and it is urgent to take actions to combat climate change and reduce its impacts [12]. Ecosystem services, such as carbon sequestration, biodiversity, and climate regulation, play essential roles in combating climate change [4, 25]. In particular, short-term (e.g., 5 years) and long-term (e.g., 40 years) projections of ecosystem dynamics under different scenarios are critical for geoscientists and policymakers to envision climate change impacts and design mitigation strategies [18, 27].

Ecosystem Demography (ED) is the new generation of global ecosystem model that has been developed by the geoscience community for two decades. ED's distinctive characteristic lies in its ability to capture detailed ecological processes across vast spatial scales from local to global, and to simulate vegetation and carbon dynamics over time spans from hours to centuries [10, 18, 27]. ED can also incorporate land-use history changes and be initialized using data from remote sensing products. Consequently, ED has been an important component of major carbon monitoring missions (e.g., NASA's Carbon Monitoring System), and its high accuracy has led to deployment in real operational environments (e.g., Maryland, US [17]). However, the complex physical processes being modeled in ED lead to expensive computational cost, which has been the major bottleneck in performing projections and simulations at large scales (e.g., global, national), especially at high spatial and temporal resolutions. This significantly limits ecologists' ability to envision a diverse set of pathways under different possible climate conditions.

We aim to approximate the ED model using a deep neural network model, given its high expressive power and computational efficiency in the inference phase. However, the deep approximation is non-trivial due to the following challenges. First, ED is often used for long-term projections (e.g., 40 years) [29], where the extensive duration is modeled as a sophisticated sequence of fine-granularity steps (e.g., weekly, monthly). The strict temporal dependency between one state and its previous state necessitates that to get an ED-projected value at any timestamp (e.g., 10^{th} year), one must first run through all previous steps from the very beginning with ED. For a learning-based approximation, it means there will be no intermediate label to use for corrections in the middle of the time-series. Second, the long-term projection leads to increased error accumulations. For example, a monthly-based deep approximation will have accumulated

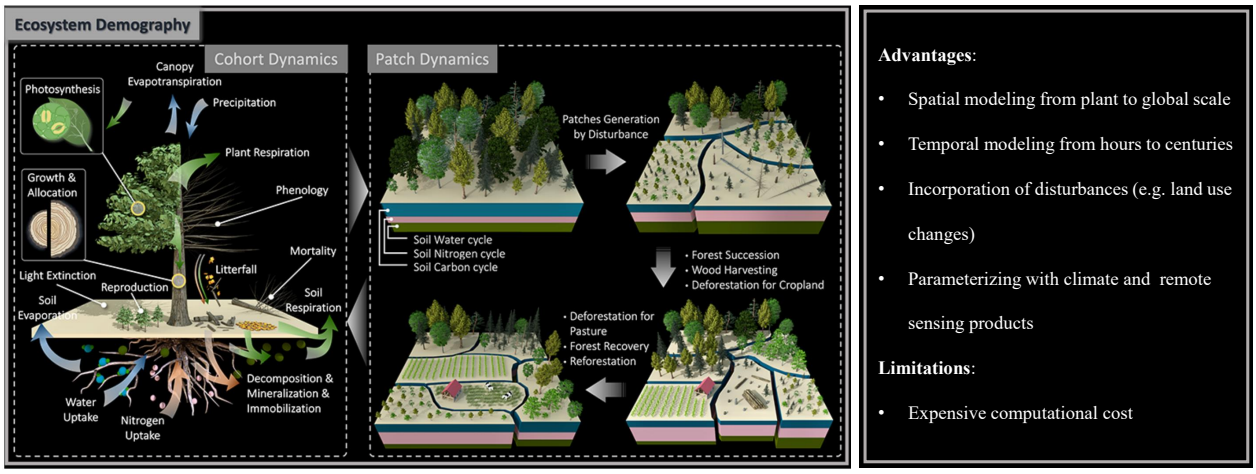


Figure 1: Schematic diagram of Ecosystem Demography (ED) model [27].

errors 480 times for a 40-year projection. The complex physical processes in ED also involve a lot of heterogeneous patterns across ecosystem variables and temporal scales, which further aggravate such error accumulations. Third, to account for climate uncertainty, the ED model is often operated to perform multiple projections with the same input data but different initial conditions (e.g., vegetation ages). When such data are used for training deep neural networks, the training process can be confused as many training sequences share the same feature values for tens of years (e.g., 100^+ features per month) and only differ by one or a few initial conditions. Finally, ED is computationally expensive, and it is time-consuming to generate a large number of samples for training or fine-tuning.

Related work. (1) Physical model approximation: Previous studies have explored machine learning based approximations for physical simulation models in several domains, such as fluid dynamics [37], wave propagation [35], amorphous carbon [6], etc. Recent approximations often utilize deep learning models (e.g., convolutional networks, encoder-decoder), with some using traditional methods such as random forests and Gaussian models [1, 15]. Neural operators (e.g., Fourier [26]) have also been developed for solving partial differential equations (e.g., fluid dynamics). The approximations have not considered plant-scale and long-term ecosystem simulation models such as ED. **(2) Error accumulation:** This phenomenon is often referred to as compounding errors in reinforcement learning (RL), and mitigation strategies often use multi-step predictions with a sequence of potential agent actions [2, 23, 38]. However, the formulations are RL-specific and not directly applicable to physical simulations or predictions. Multi-step methods have also been developed for prediction methods [8, 33, 36], such as direct strategies and Multiple-Input Multiple-Output (MIMO), which predict the next k steps without requiring each step to build on its previous one. However, these methods are often used for relatively smaller numbers of steps (e.g., 4 to 12), and the performance largely drops for longer steps [33]. They also do not consider heterogeneous patterns across variables and temporal scales. **(3) Active learning:** For sample selection, active learning has been extensively studied to query small subsets of samples that can best improve model performance [31]. A CORE-SET approach uses a core-set loss and converts it to a k-cover problem to improve sampling diversity [34]. Confidence sampling [24] and entropy-based methods [20] prioritize samples with higher uncertainty. Batch Active learning by Diverse Gradient Embeddings (BADGE) uses diverse gradient embedding to consider both diversity and uncertainty simultaneously [3]. In addition to existing objectives such as uncertainty and diversity, our approximation also needs to explicitly consider generalization over spatial and physical domains. **(4) Physics-guided machine learning (PGML):** Different from this work, PGML [7, 14, 19] mainly focuses on the cooperation between physical (e.g., conservation rules) and data-driven models.

Contributions. To address the limitations, we propose Deep-ED, a deep network based approximation, to generate high-accuracy approximations of ED over long-term:

- We propose a coupled multi-scale multi-branch network structure to address the large discrepancy between annual and monthly patterns for long-term projections.
- We present a self-guided strategy to handle heterogeneous temporal patterns among ecosystem variables. We further propose a significance-based network partitioning to handle conflicting responses from different ecosystem variables.
- We develop a light-weight long-term cumulative loss with signed de-sequencing. We also show the importance of batch diversity for addressing high sample similarity.
- We incorporate active learning with additional spatial and physical considerations to improve sample generation.

We carried out experiments in the northeastern US. The results show that Deep-ED is able to improve long-term approximation quality compared to the baseline methods.

2 PROBLEM DEFINITION

In the following, we provide a general description of the ED model and define the deep approximation problem.

Ecosystem Demography (ED) model: ED is an ecosystem model that has been widely adopted in carbon monitoring research and operations. It simultaneously models a wide array of factors and

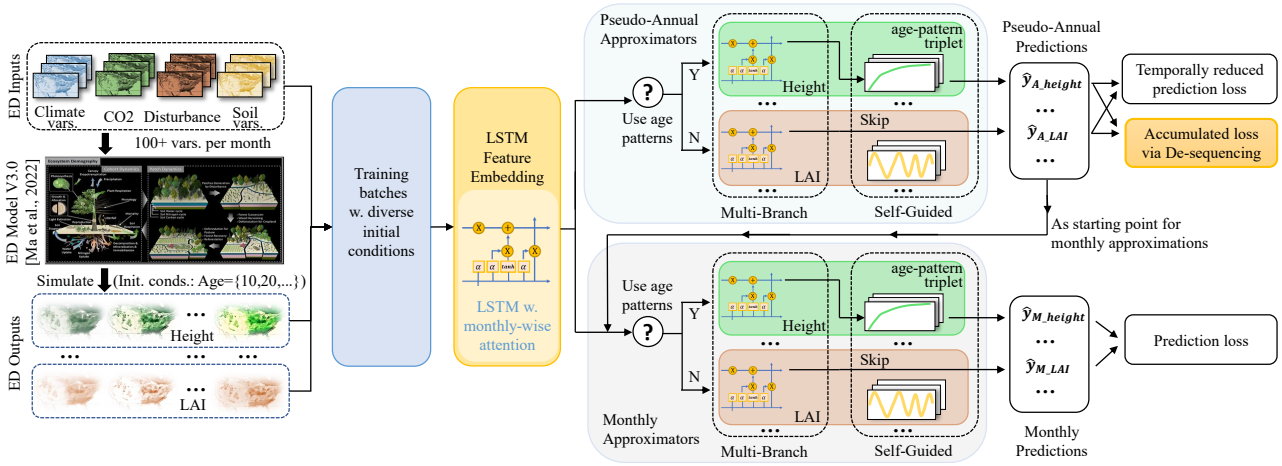


Figure 2: The overall framework of Deep-ED.

simulates plant dynamics by considering growth, mortality, reproduction, hydrology, carbon cycle and soil biogeochemistry [27, 30]. Fig. 1 shows a schematic diagram of ED from [27]. According to Fig. 1, ED is able to model ecosystem processes at the level of individual plants, using vegetation structure attributes such as height and diameter, along with physiological processes such as leaf photosynthesis and phenology. It further simulates cross-plant interactions based on the mechanical rules (e.g., competition among plants for growth resources such as light, water, and nutrients).

ED's distinctive characteristic lies in ability to capture detailed ecological processes across vast spatial scales from local to global, and to simulate vegetation and carbon dynamics over time spanning from hours to century-scales [27, 30]. Importantly, ED can also incorporate land-use history changes and be initialized using data from satellite remote sensing products. With the rapid development in satellite platforms, remote sensing advancements including higher resolution of spectral imagery and the vast availability of LiDAR data have made it possible to provide accurate initial conditions for ED [28]. Given its flexibility and capabilities, ED has been broadly used to simulate ecosystem dynamics and estimate future carbon sequestration potential under the impact of climate change to facilitate policy-making [11, 16, 17]. The model has also been deployed in real operational environments (e.g., Maryland, US [17]).

Problem formulation: Given input features $\mathbf{X}$ (e.g., soil properties and dynamic climate variables) and simulated outputs $\mathbf{Y}$ (e.g., vegetation height, aboveground biomass, leaf area index, etc.) of the ED model at the given temporal scale t (i.e., the step-size for ED outputs), we aim to learn a deep learning model $\mathcal{F}(\mathbf{\Theta})$, which can approximate the complex process – represented by a sequence of physical functions describing the ecosystem – by a set of network layers and parameters $\mathbf{\Theta}$. **Temporal scale:** ED includes physiological processes that operate at various temporal scales such as photosynthesis at hourly, phenology at monthly, and reproduction at yearly. Currently, ED export all variables at monthly scale. In this paper, the step-size t is also one-month by default. **ED outputs:** We use 7 outputs selected by ecologists: vegetation height, aboveground biomass (AGB), soil temperature, leaf area index (LAI), gross primary production (GPP), net primary production (NPP), and heterotrophic respiration (Rh). The selected variables are commonly inter-compared between models and evaluated against reference datasets.

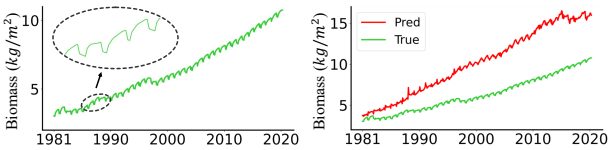


Figure 3: Examples of monthly variation and error accumulation.

3 METHOD

We propose a Deep-ED framework to approximate the process-based ED model. Fig. 2 shows the overall framework. The general architecture uses the long-short-term-memory (LSTM) layers to simulate the sequential steps in the ED model, where each recurrent step represents a month.

3.1 Multi-Scale Multi-Branch Network Structure for Long-Term Projection Approximation

3.1.1 Multi-Scale Structure. Given a time-series of inputs $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_T\}$ and outputs $\mathbf{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_T\}$ with T time steps, where each time step represents one month, a typical way to train the LSTM is just to build the samples by separating the time-series into overlapping or non-overlapping time windows and let the model approximate the labels at each time step. While this maximizes the use of training labels, the large discrepancy between monthly and annual patterns can substantially decrease the quality of approximations over the long term (e.g., 20 or 40 years). Specifically, Fig. 3 (left) shows an example of the monthly trajectory of aboveground biomass over 40 years at a single location, as simulated by the ED model. While biomass is consistently growing over the years, we can clearly see the monthly variations or seasonality embedded in the trajectory. Fig. 3 (right) shows a LSTM-based approximation trained with full $\mathbf{X}$ and $\mathbf{Y}$. While the approximation is within a reasonable

range at the start, the deviation gradually grows as error accumulates over a long period.

It has been recognized in time-series analysis that it is important to remove cyclic local variations from the synoptic longer-term trend. Unlike many existing studies that model the local patterns as noises [40], the local variations in ED are meaningful responses to physical environments (e.g., temperature). Climate-related studies have shown the detrending of these seasonal variations is beneficial for the analysis [9, 41], and example methods include wavelet transforms and Kolmogorov-Zurbenko filter. However, these methods are intended for trend analysis and do not consider complex interactions with external features in the prediction setting. An alternative method is to detrend the data using historical averages [32]. However, in ecological long-term projection, past averages do not represent future averages (e.g., biomass), and the cyclic trends are non-stationary given the climate variables and current states.

We use a two-phase multi-scale structure to reduce the effects of monthly variations when producing approximations in the long term, while keeping the model's monthly prediction ability. Specifically, our structure consists of two components:

- **Pseudo-annual approximator with loss-reduced LSTMs:** This aims to provide stable predictions of annual changes of target variables. We call it a pseudo-annual prediction model mainly because it still uses a LSTM that runs at the monthly scale. However, instead of evaluating losses using labels at all monthly time steps, we only use one label every 12 steps (i.e., one per year). The reasons are that: (1) While providing full labels gives a model more information to learn, we find that it also makes the learning more difficult. For example, it can be much more challenging to accurately predict the monthly variations, which are more sensitive to short-term meteorological changes, than to predict the aggregated effects over the entire year. The attempt to close the gap to the harder-to-predict monthly values leads to increased errors at the annual scale (results in experiments). (2) We keep the monthly LSTM instead of switching to an annual step size because ED is a sequential model, which means there are temporal dependencies between monthly features. If all features within each year are fed in as one mixed set, it can be difficult to learn such dependencies (results in experiments). The following is an example of the temporally-reduced loss with MSE on one sample and one target variable:

$$\mathcal{L}_{mse}(\mathbf{X}, \mathbf{Y}) = \frac{||(\mathcal{F}_{lstm}(\mathbf{X}) - \mathbf{Y}) \odot \mathbf{m}||_2^2}{\mathbf{m}^T\mathbf{m}} \quad (1)$$

  where $\mathbf{X} \in \mathbb{R}^{T\times d}$ and $\mathbf{Y} \in \mathbb{R}^T$ are features and labels for one sample, $\odot$ is the Hadamard product, and $\mathbf{m} \in \mathbb{R}^T$ is a mask where $\mathbf{m}_i = 1$ if $i\%12 = 0$ and 0 otherwise.
- **Monthly approximator:** This component only aims to make predictions of monthly values within one year given an initial value at the beginning, which is the prediction from the pseudo-annual model during the test. In other words, the monthly approximator only focuses on one year and does not need to consider long-term prediction. It should be noted that later methods on cumulative error reduction (e.g., Sec. 3.3) are irrelevant for this part.

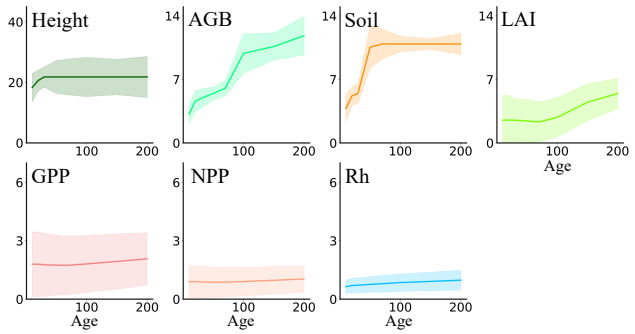


Figure 4: Heterogeneous patterns of target variables.

3.1.2 Multi-Branch Structure. Since ED simulates a variety of ecosystem variables, which may have different functional relationships with the input features, by default, we consider a multi-head output structure for the LSTM component. In the multi-branch structure, we further increase the separation between the paths from $\mathbf{X}$ to different $\mathbf{Y}$ (i.e., separate LSTM layers after initial embedding as shown in Fig. 2) mainly to increase the extendability of a trained model. In practice, different users of ED may only consider certain ecosystem variables. In addition, as ED is a sequential model, all target variables need an initial value at the beginning of the time-series as an input. Thus, different training datasets may have different inputs and outputs. The multi-branch structure allows separate training components to affect only a subset of parameters. When new target variables are added, the multi-branch structure can also help avoid the impact of new training on existing targets.

3.2 Heterogeneous Patterns: Self-Guided Forwarding & Significance-Based Partitioning

The multi-branch structure separates the approximation of functional relationships between different target variables. In this section, we further reduce the impact caused by the heterogeneous responses from different target variables to time (or age) by a combination of self-guided forwarding and significance-based partitioning.

The mean curves in Fig. 4 show the patterns of value changes in the seven target variables (Sec. 2) over 200 years. They are aggregated averages over all training samples generated by ED and the background shows the distribution. One interesting pattern we can observe is that the rate of height growth pattern (top-left) has a large shift around the 30^{th} year. This is mainly caused by the phenology of trees: after certain ages, trees tend to expand the diameter of their trunks rather than continue to grow in height. Based on the observation, it may reduce the difficulty of training if such knowledge can be incorporated into the learning process. Additionally, we also observe that the temporal patterns of different variables may vary a lot. For example, after trees slow down on height growth, they may continue to expand in diameters and canopy sizes, which will lead to steady increases in aboveground biomass. Moreover, the changes over age present very different behaviors for different target variables. We can observe that some variables (e.g., height, aboveground biomass) show stronger monotonic patterns whereas some others (e.g., GPP, NPP, Rh) show more variations around a more horizontal mean line.

Based on these insights, we use a two-step approach with self-guided forwarding and significance-based training partitioning. In

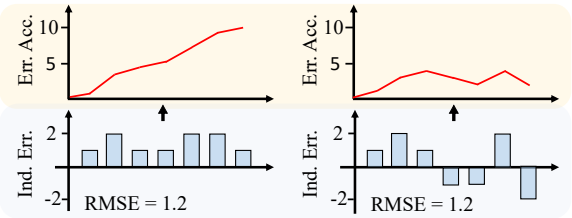


Figure 5: An illustrative example of the de-seq. loss.

the first step, the self-guided forwarding utilizes the time-series trends from Fig. 4 and applies the target-specific trends as input guidance to all target variables. Basically, based on the initial age a_0 of the simulation and the number of years needed in the projection, a triplet $[a_i, \mu_i^j, \Delta_i^j]$ will be formed as an input for the i^{th} year and j^{th} target, where $a_i = a_0 + i$, μ_i^j is the mean of the j^{th} target at age a_i, and Δ_i^j is the change rate on the mean curve from a_i to a_{i+1}. The change rate provides additional guidance on the growth speed of a target at the age. Since we can see that the mean curves tend to flatten out after the maximum age shown in Fig. 4, we keep the mean stationary after that age. Age-based guidance is added as inputs both at the first layer and before the output layer.

Age-based guidance tends to be helpful for target variables that present a stronger year-over-year trend (e.g., strong monotonicity) and are relatively less sensitive to environmental conditions. However, they may provide spurious information for targets with weaker annual trends and higher sensitivity to physical changes. Thus, we separate the network branches into two partitions: with and without age-based trends. In the initial step, all branches will not use the age-based trends. Then, after the initial training converges, we apply the age-based trends to all targets and use a significance-based test to automatically determine the partition assignment of a branch (i.e., with or without the age-based trends in prediction). Specifically, we use the dependent T-test to determine if the prediction quality on a target variable has significantly improved after applying the age-based trends:

$$\tau = \mu_{dif} / \left(\sigma_{dif} \cdot (\sqrt{DF_{val} + 1})^{-1}\right) \tag{2}$$

where dif is the error difference after applying the age-based trends (i.e., "treatment"), N_{val} is the number of validation samples used for the test, $\mu_{dif} = \left(\sum_{i=1}^{N_{val}} dif_i\right)/N_{val}$, $\sigma_{dif} = \left(\sum_{i=1}^{N_{val}} (dif_i - \mu_{dif})^2\right)/N_{val}$, and $DF = N_{val} - 1$ is the degree of freedom. For simplicity, the test here is for a single target variable. To determine significance, we compare τ with the critical values from the T-test's look-up table for a dependent test case (i.e., using the same set of samples before and after the "treatment". We use 0.01 as the significance level by default.

3.3 Enforcing Cumulative Loss via De-Sequencing

Since ED is commonly used for long-term projection, having accurate predictions at annual or short-term scales is insufficient, so the proposed Deep-ED approximator must overcome the error accumulation effects. While it is possible to use a loss function that considers error accumulation, such loss is very expensive to compute and easily overloads the memory due to the length of the chain. For example, a loss function can often be:

$$\begin{aligned}\mathcal{L}_{acc} &= ||\mathbf{Y}_T - \hat{\mathbf{Y}}_T||_2^2 = ||\mathbf{Y}_T - \mathcal{F}(\mathbf{X}_T, \hat{\mathbf{Y}}_{T-1})||_2^2 \\ &= ||\mathbf{Y}_T - \mathcal{F}(\mathbf{X}_T, \mathcal{F}(\mathbf{X}_{T-1}, \mathcal{F}(...\mathcal{F}(\mathbf{X}_0, \mathbf{Y}_0)...)))||_2^2\end{aligned}$$

where T is the number of time-steps in the long-term projection (e.g., 480 months for 40 years), and $\mathbf{Y}_0$ contains initial target values for the simulation (e.g., initial vegetation height). As we can see, the chaining of steps largely affects the practicality of this loss for error accumulation.

To mitigate the problem, we propose a de-sequenced loss, in which the full sequence S is decomposed into non-overlapping independent sub-sequences $\{s_1, s_2, ..., s_K\}$ that do not incur computational problems (e.g., 12 months). For each S, all its sub-sequences $s_i \in S$ will be included in the same batch for an epoch. Then, a de-sequenced loss $\mathcal{L}_{dsq}(S)$ will estimate the *signed* error at the final step of each $s_i \in S$, and enforce a zero-sum loss on the sum of the *signed* errors:

$$\mathcal{L}_{dsq}(S) = \sum_{s_i \in S} \mathcal{L}_{signed}(s_i) = \sum_{s_i \in S} (\mathbf{Y}_{T_i} - \hat{\mathbf{Y}}_{T_i}) \tag{3}$$

where T_i denotes the final step of the i^{th} sub-sequence s_i. Note that in the interest of training efficiency and feasibility, the initial target value (e.g., height) for each s_i is an input from training labels rather than a predicted output from the previous step as in $\mathcal{L}_{acc}$ above (i.e., the forward process in each $s_i \in S$ is independent).

Different from traditional regression losses such as RMSE, here it is important to keep the signs of the errors, as $\mathcal{L}_{dsq}$ aims to centralize the sum of the individual errors to zero. Fig. 5 shows a simplified example of the de-sequenced loss for error accumulation reduction. As we can see, $\mathcal{L}_{dsq}$ tries to create a more balanced distribution of sub-sequence predictions around the truth for each full sequence. With the independence among $\{s_i \in S\}$, $\mathcal{L}_{dsq}$ can be efficiently evaluated, and we find it effective for reducing error accumulation in experiments. In training, $\mathcal{L}_{dsq}$ will be combined with a regular MAE or MSE loss for the independent sub-sequences:

$$\mathcal{L}_{com} = \alpha \cdot \mathcal{L}_{dsq} + \sum_{s_i \in S} \mathcal{L}_{mae}(s_i) \tag{4}$$

where α is a weight for scale adjustment and set as $\alpha = 1$ in our tests.

3.4 Training with Initial Condition Diversity

To make long-term projections more useful, ED often simulates under a variety of initial conditions to generate a distribution of possible ecosystem pathways instead of a single value for each target variable. As mentioned in challenges, this leads to many samples having the same value of most features and only differing by one or a few initial conditions, which causes confusion during training. Through examination, we find that it is very important to form training batches with a high diversity of initial conditions, otherwise, it will significantly decrease the approximation quality. Specifically, we consider the full sequence (e.g., 40 years) at one location as a unit time-series sample (for $\mathcal{L}_{dsq}$ calculation). Then, in each batch, we try to maximize the number of initial conditions (e.g., different initial ages) for each unit sample under the batch-size limit. Comparisons will be shown in the experiments.

3.5 Active Learning with Geo-Physical Diversity

We additionally include an active learning model to assist with the generation of new training samples in practice. Our approach is based on the BADGE paradigm as it simultaneously considers uncertainty and diversity through gradient embedding (gradients from the second to last layer). BADGE projects the points into a gradient space as a representation for uncertainty and uses the k-means++ seeding algorithm (KSA) to sample a batch of diverse points [3]. However, BADGE was designed for classification. The algorithm cannot be directly used here as its gradients are based on losses from hypothetical class labels, which are set to $\arg\max_c p_c$ (p_c is the predicted class probability) for the unlabeled samples, which are unavailable for regression.

We extend BADGE for Deep-ED in two ways: (1) **Near-term gradient embedding:** Since ED is a sequential model for long-term projection, we use near-term projections (i.e., the first 3 years in our implementation) to represent each time-series, where the later year labels are unknown. The cheaper near-term labels enable gradient calculation for points in the sample pool. If a point at a location is selected for label-query, ED will run through the rest of the projection length (e.g., 40 years) for the point to get full labels for the time-series. (2) **Geo- and physical-diversity:** Denote n as the number of points to select for label-query at each round. The original BADGE will sample all n points in the gradient space with KSA. However, this does not consider the variability of environmental properties over space [43] and the diversity of model's initial conditions (i.e., age of vegetation). As both are important for generalizability, we split n into 3 subsets, where the first $\lceil n/3 \rceil$ are selected by KSA in the gradient space. After that, we re-project all the points – including selected ones (existing centers) – to the geographical space, and resume KSA to sample the next $\lceil n/3 \rceil$ points. Finally, we re-project the points to the space of initial physical conditions (i.e., 1D vegetation age), where KSA will select the rest of the points (Alg. 1).

Algorithm 1 Active learning w. geo-phy. diversity (1 round)

Require: • Near-term gradients $\mathbf{D}_{grad}$, geo-coordinates $\mathbf{D}_{loc}$ and initial conditions $\mathbf{D}_{cond}$ of samples in the pool; • Number of samples to select n.

```
Query set Q = init_empty_set()
Q.add(KSA(D_grad, n_center=⌈n/3⌉, existing_seeds=Q))
Q.add(KSA(D_loc, n_center=⌈n/3⌉, existing_seeds=Q))
Q.add(KSA(D_cond, n_center=n − |Q|, existing_seeds=Q))
return query_labels_ED(Q)
```

4 EXPERIMENTS

4.1 Data

We simulated 40 years of ED data for 8 different initial conditions, representing vegetation ages in set $\{10, 20, 30, 50, 70, 100, 150, 200\}$. The geographic area of the simulation covers the northeastern US (35°N~40°N, 75°W~80°W), and the temporal range of the auxiliary information (e.g., climate variables) used is from 1980 to 2020. The input data includes meteorology from NASA Dayment and MERRA2, soil properties from the POLARIS dataset, and CO_2 concentration from NOAA CarbonTracker. The quality of the simulation in the geographic region has been extensively evaluated in [27]. Our data includes 320,000 simulated samples, and we randomly split train and test datasets as 50% and 50% for the following experiments. Specifically, train and test datasets are both spatially and temporally non-overlapping.

4.2 Candidate methods and measures

We consider the following candidate methods in the evaluation. All the following models use one month as the step (or one sample for non-time-series methods) unless explicitly stated otherwise. To avoid confusion, when the models make predictions for long-term, they iterate over their own predictions from the previous timestamp (e.g., using t's output as input for $t+1$) and they do not use any intermediate information generated by ED for correction. ED outputs for the test data are only used for evaluation.

- **RF:** Random Forest regression with 100 trees [5].
- **S-LSTM:** A single-head LSTM model for all ED outputs [13].
- **M-LSTM:** A multi-head LSTM model where different heads are used for different ED outputs to capture cross-variable heterogeneity.
- **MY-LSTM:** M-LSTM that is trained with each step being a year instead of a month, where all features over 12 months are concatenated as inputs at the annual level. This annual-step model is used to show the need for learning with a monthly sequence. It can also be considered as a multi-step model where the model directly predicts the 12^{th} step.
- **MA-LSTM:** M-LSTM with attention layers to prioritize the most informative periods in the time series [21, 39].
- **LSTNet:** A time-series model extracting short-term local dependency using CNN and long-term temporal patterns using RNN [22].
- **Seq2seq:** A sequence-to-sequence model with additional encoder and decoder modeling, where both are implemented with LSTMs [42].
- **Deep-ED:** Our proposed approach.

4.3 Results

Comparisons to candidate methods. Table 1 shows the prediction performance on target variables for 6 different initial conditions. The proposed Deep-ED outperforms the other candidate methods in the vast majority of target variables and initial conditions. The results show the effectiveness of Deep-ED in alleviating heterogeneous patterns over time and target variables. For example, vegetation height tends to have a more monotonic trend over the time steps whereas variables such as LAI and GPP tend to have more seasonal variations. We can see that the MY-LSTM did not perform well, which shows the importance of following the temporal sequence of the physical simulation. Similarly, the traditional regressors Ridge and RF had higher errors as they were not effective in capturing the temporal relationships.

Improvements on long-term error accumulations. Fig. 6 and 7 show the error accumulation (log-scale) over 40 years for two example variables (height, AGB) on all different initial ages. Aligning with our expectations, the results show that Deep-ED was able to maintain a lower error with the designs on error accumulation reduction. Although some candidate methods have relatively small RMSE over all 40 years due to the effects of averaging, the prediction errors

Table 1: Prediction RMSE on different initial conditions (* denotes RMSEs within 0.01 compared to the best).

	Age 10							Age 20						
Model	Height	AGB	Soil	LAI	GPP	NPP	Rh	Height	AGB	Soil	LAI	GPP	NPP	Rh
Ridge	472.46	964.38	201.88	435.19	77.2	38.6	8.67	489.4	988.93	205.75	446.8	79.3	39.65	8.81
RF	12.13	6.96	4.63	1.42	0.75	0.38	0.22	9.49	6.31	3.69	1.39	0.74	0.37	0.23
S-LSTM	3.88	0.94	1.52	0.26	0.15	0.07	0.10	2.73	0.79	0.90	0.22	0.13	**0.06***	0.08
M-LSTM	1.18	0.62	0.56	**0.20**	**0.10**	**0.05**	**0.06***	1.14	0.79	0.60	0.23	**0.10**	**0.05**	**0.06***
MY-LSTM	1.6	0.34	0.72	0.27	0.16	0.08	**0.06***	2.57	0.64	1.05	0.27	0.16	0.08	0.07
MA-LSTM	1.84	0.37	0.65	0.27	0.12	**0.06***	0.07	1.56	0.35	0.54	0.24	**0.11***	**0.06***	**0.06***
LSTNet	2.72	0.57	0.68	0.26	0.14	**0.07***	0.07	2.00	0.53	0.58	0.24	0.13	**0.06***	0.07
Seq2seq	5.02	1.27	1.85	0.31	0.12	**0.06***	0.09	4.22	1.05	1.54	0.33	0.13	**0.06***	0.08
Deep-ED	**0.39**	**0.16**	**0.15**	**0.21***	**0.11***	**0.06***	**0.05**	**0.60**	**0.18**	**0.15**	**0.19**	**0.11***	**0.06***	**0.05**

	Age 30							Age 50						
Model	Height	AGB	Soil	LAI	GPP	NPP	Rh	Height	AGB	Soil	LAI	GPP	NPP	Rh
Ridge	466.1	956.16	201.06	431	76.43	38.21	8.62	458.07	948.18	200.87	426.70	75.65	37.83	8.60
RF	7.19	5.69	2.99	1.4	0.73	0.36	0.24	4.39	4.79	2.11	1.41	0.67	0.34	0.26
S-LSTM	2.83	1.04	0.67	0.25	0.13	**0.06***	0.07	3.70	1.68	0.84	0.40	0.15	**0.07***	0.07
M-LSTM	1.05	1.06	0.83	0.33	0.12	**0.06***	0.07	0.77	1.23	1.17	0.48	0.14	**0.07***	0.10
MY-LSTM	2.74	0.73	1.03	0.29	0.16	0.08	0.07	2.78	0.92	1.09	0.36	0.18	0.09	0.07
MA-LSTM	1.28	0.37	0.41	0.21	**0.10**	**0.05**	**0.05**	1.23	0.88	0.51	0.49	**0.13***	**0.07***	**0.06***
LSTNet	1.87	0.67	0.75	0.23	0.13	**0.06***	0.07	1.62	0.89	1.17	0.29	**0.13***	**0.07***	0.08
Seq2seq	3.39	0.97	1.33	0.39	0.14	0.07	0.07	2.11	1.05	0.91	0.48	0.14	**0.07***	0.07
Deep-ED	**0.80**	**0.15**	**0.14**	**0.19**	**0.11***	**0.06***	**0.05**	**0.71**	**0.18**	**0.12**	**0.20**	**0.12**	**0.06**	**0.05**

	Age 70							Age 100						
Model	Height	AGB	Soil	LAI	GPP	NPP	Rh	Height	AGB	Soil	LAI	GPP	NPP	Rh
Ridge	476.51	970.35	202.66	438.05	77.71	38.85	8.71	481.81	978.2	203.73	441.6	78.35	39.18	8.73
RF	3.38	4.07	1.72	1.22	0.65	0.33	0.27	2.94	3.23	1.42	0.93	0.69	0.35	0.3
S-LSTM	3.73	2.69	1.22	0.73	0.22	0.10	**0.07***	3.68	4.31	1.79	1.35	0.33	0.16	0.09
M-LSTM	0.68	0.85	1.23	0.36	**0.12***	**0.06***	0.10	0.63	0.76	1.31	0.46	0.13	0.06	0.10
MY-LSTM	2.4	1.08	0.96	0.52	0.19	0.10	0.08	2.57	0.94	0.82	0.55	0.19	0.09	0.08
MA-LSTM	1.17	0.99	0.73	0.52	0.14	0.07	**0.07***	1.07	0.99	0.75	0.44	0.13	0.07	0.07
LSTNet	1.27	1.63	1.59	0.56	0.17	0.09	0.10	1.14	2.31	1.72	0.94	0.23	0.11	0.12
Seq2seq	1.73	1.37	0.75	0.60	0.16	0.08	**0.07***	1.65	2.03	0.82	0.96	0.21	0.11	0.09
Deep-ED	**0.54**	**0.25**	**0.16**	**0.22**	**0.11**	**0.05**	**0.06**	**0.57**	**0.20**	**0.13**	**0.22**	**0.11**	**0.05**	**0.05**

	Age 150							Age 200						
Model	Height	AGB	Soil	LAI	GPP	NPP	Rh	Height	AGB	Soil	LAI	GPP	NPP	Rh
Ridge	462.40	952.60	200.93	429.05	76.08	38.04	8.60	498.91	1003.78	208.63	454.02	80.64	40.32	8.92
RF	2.80	2.52	1.35	0.91	0.75	0.38	0.33	2.88	2.35	1.28	0.89	0.73	0.37	0.35
S-LSTM	2.83	4.91	1.77	1.56	0.37	0.18	0.11	1.97	3.8	1.4	1.23	0.29	0.14	0.11
M-LSTM	0.64	1.24	1.48	0.71	0.16	0.08	0.11	0.73	1.4	1.32	0.68	0.14	0.07	0.10
MY-LSTM	2.12	1.67	0.74	0.61	0.17	0.09	0.09	1.91	1.52	0.74	0.42	0.14	0.07	0.09
MA-LSTM	1.20	0.89	0.85	0.37	**0.11***	**0.05**	0.08	1.11	0.81	0.62	0.26	0.08	**0.04**	0.07
LSTNet	1.64	2.82	1.65	0.97	0.24	0.12	0.12	1.68	2.82	1.84	0.99	0.24	0.12	0.13
Seq2seq	1.55	2.35	0.68	0.89	0.20	0.10	0.08	1.48	1.88	0.54	0.58	0.13	0.07	**0.07***
Deep-ED	**0.50**	**0.28**	**0.16**	**0.20**	**0.10**	**0.05**	**0.06**	**0.52**	**0.28**	**0.22**	**0.21**	**0.09**	**0.04**	**0.06**

accumulated over time, which limits the use of these models for long-term projections.

Effects of active learning with geo-physical diversity. Fig. 8 shows the effectiveness of our active learning methods in querying new training samples. Here we consider three candidates: (1) Random sampling; (2) BADGE-NG: BADGE with near-term regression gradients; and (3) BADGE-SP: BADGE-NG with geo-physical diversity considerations. We can see that the integration of near-term gradients allowed the method to leverage the gradient space for sampling. Compared to random selections, our method converges more quickly and robustly as the newly generated samples are considered as harder-to-train given the current model performance. The integration of near-term gradient embeddings with spatial- and physical-diversity achieves the best performance by taking into consideration the domain knowledge in the model learning process.

4.4 Self-Analysis

We carry out a self-analysis to evaluate the effects of different components in Deep-ED. To emphasize the performance of long-term predictions, all models are trained and evaluated on an annual scale in this self-analysis. Specifically, we consider the following versions of the Deep-ED model: (1) **MM**: Multi-branch structure with monthly loss; (2) **MMA**: Multi-scale multi-branch structure with pseudo annual approximation; (3) **MMA-S**: MMA with self-guided forwarding; (4) **MMA-SS**: MMA-S with significance-based partitioning; (5) **MMA-SSD**: MMA-SS with de-sequencing loss function; (6) **Full**: MMA-SSD with conditional diversity for training.

As shown in Fig. 9, each component of Deep-ED gradually improves the model performance in predicting target variables. We can see that the effects of different components vary a bit across target variables, but the full model almost always has the best performance by the integration.

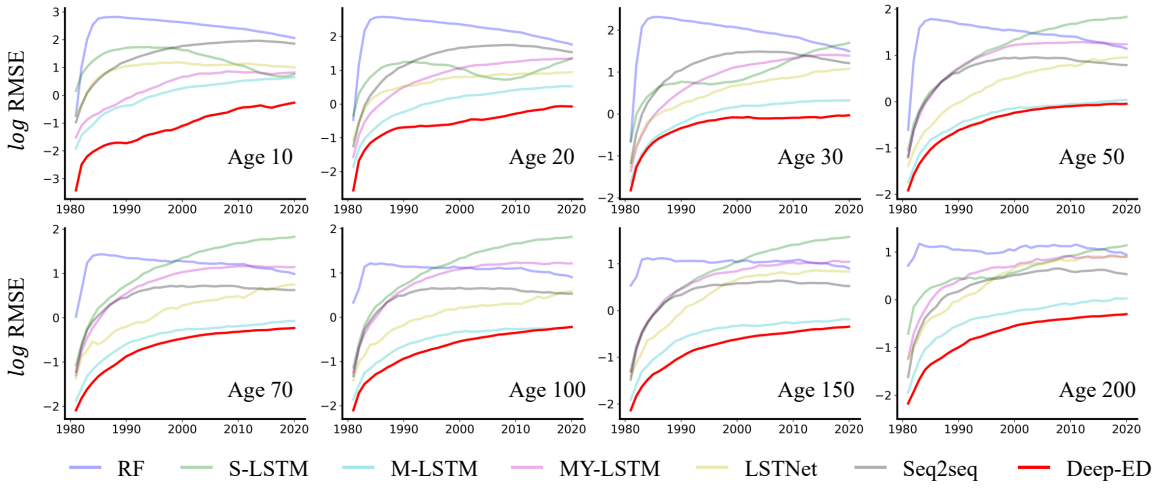
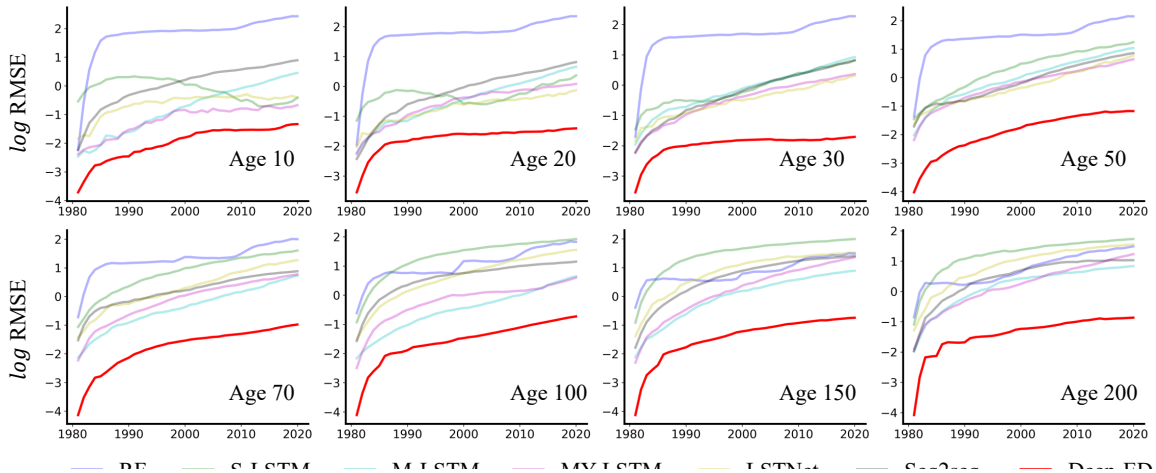


Figure 6: Accumulated errors over 40 years on height for all ages



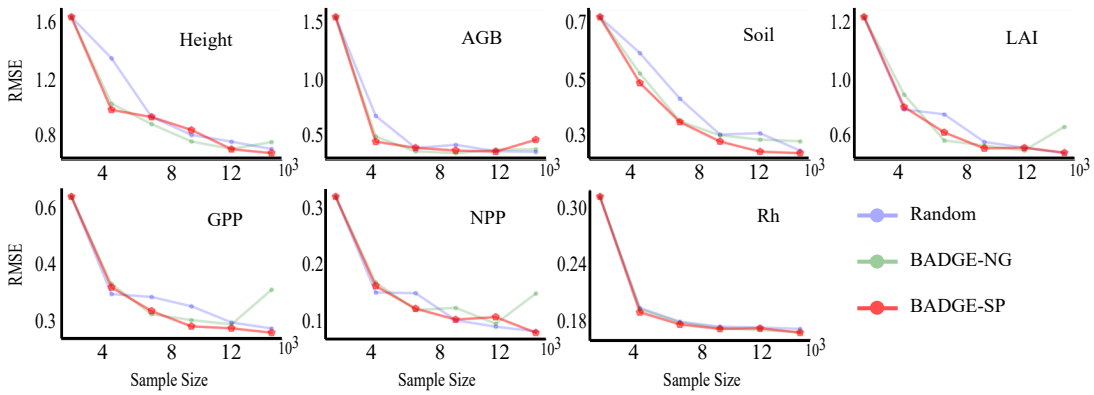


Figure 8: Active learning performances.

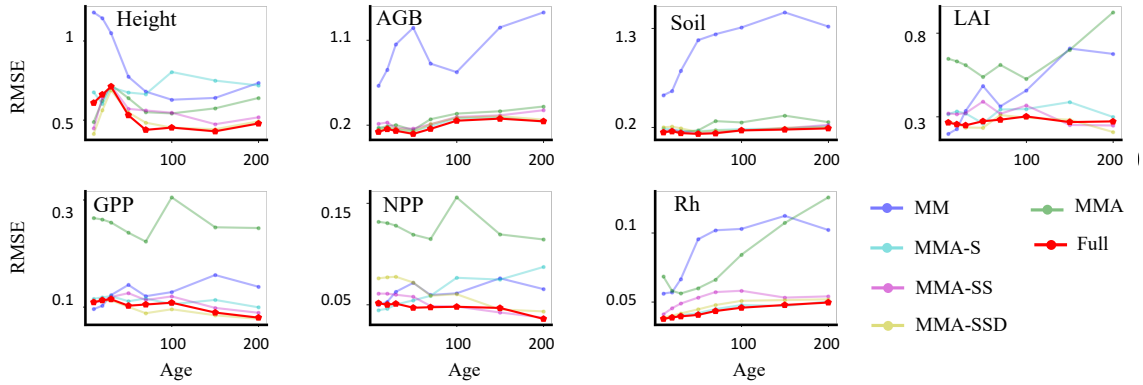


Figure 9: Improvement of each component in Deep-ED.

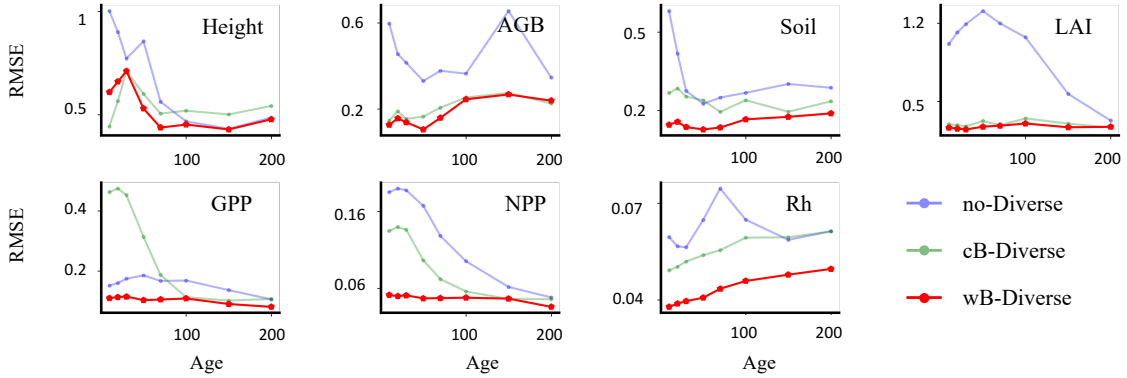


Figure 10: Importance of diversity in an iteration.

Effects of training strategies (Sec. 3.4). Fig. 10 compares three strategies for batch formation (i.e., same dataset but different batches), where the X-axis shows different initial ages and the Y-axis shows the RMSE on test data averaged for each initial age. The blue curve uses different samples with the same initial age in a batch (**No-Diverse**), and the batches with the same initial ages are made adjacent in an epoch. The green curve forms the batches in the same way, but the order of batches is random, showing the cross-batch diversity (**CB-Diverse**). Finally, the red curve is our strategy with diversified initial ages within a batch (**WB-Diverse**). We can clearly see the large impact of batch formation and the importance of having diverse initial conditions in a batch.

Execution time. Compared to ED, Deep-ED was able to reduce the simulation time by orders of magnitude. Table 2 shows the execution time of Deep-ED's simulation with a single RTX A4500 GPU. The annual column means the simulation only uses the pseudo-annual branch of Deep-ED to generate annual projections, whereas the overall includes the monthly branch as well for finer-granularity projections. The additional time from the monthly branch is shown in the monthly column. In contrast, ED's run time is highly expensive, which is about 192 CPU hours (AMD EPYC Processor) for a sample size of 320,000 (1,000 sites with 8 initial ages for 40 years). This shows Deep-ED's promising potential for projecting long-term ecosystem changes at both large scale and high resolution.

Table 2: Execution time (Unit: hour)

Sample Size	Annual	Monthly	Overall
320,000	0.01	0.01	0.02
8,000,000	0.22	0.20	0.42
16,000,000	0.46	0.38	0.84

5 CONCLUSIONS

We proposed a deep learning based model – Deep-ED – to approximate the ED model with a multi-scale cumulative loss reduction structure, significance-based scenario partitioning, self-guided forwarding, and geo-physics-aware active learning strategies. Our results showed that Deep-ED achieved high-quality approximations in long-term projection tasks (i.e., 40 years), demonstrating its potential for ecosystem simulation at large scale and high resolution. In future work, we will incorporate Deep-ED in real application scenarios and thoroughly evaluate how well it can help reproduce existing downstream analysis results that were computed with ED.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2105133, 2126474 and 2147195; NASA under Grant No. 80NSSC22K1164 and 80NSSC21K0314; USGS under Grant No. G21AC10207; Google's AI for Social Good Impact Scholars program; the DRI award and the Zaratan supercomputing

cluster at the University of Maryland; and Pitt Momentum Funds award and CRC at the University of Pittsburgh.

REFERENCES

[1] AA Masrur Ahmed, Ravinesh C Deo, Qi Feng, Afshin Ghahramani, Nawin Raj, Zhenliang Yin, and Linshan Yang. 2021. Deep learning hybrid model with Boruta-Random forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity. *Journal of Hydrology* 599 (2021), 126350.

[2] Kavosh Asadi, Dipendra Misra, Seungchan Kim, and Michel L Littman. 2019. Combating the compounding-error problem with a multi-step model. *arXiv preprint arXiv:1905.13320* (2019).

[3] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671* (2019).

[4] Elena M Bennett, Garry D Peterson, and Line J Gordon. 2009. Understanding relationships among multiple ecosystem services. *Ecology letters* 12, 12 (2009), 1394–1404.

[5] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[6] Miguel A Caro, Gábor Csányi, Tomi Laurila, and Volker L Deringer. 2020. Machine learning driven simulated deposition of carbon films: From low-density to diamondlike amorphous carbon. *Physical Review B* 102, 17 (2020), 174201.

[7] Shengyu Chen, Yiqun Xie, Xiang Li, Xu Liang, and Xiaowei Jia. 2023. Physics-Guided Meta-Learning Method in Baseflow Prediction over Large Regions. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, 217–225.

[8] Haibin Cheng, Pang-Ning Tan, Jing Gao, and Jerry Scripps. 2006. Multistep-ahead time series prediction. In *Advances in Knowledge Discovery and Data Mining: 10th Pacific-Asia Conference, PAKDD 2006, Singapore, April 9-12, 2006. Proceedings 10*. Springer, 765–774.

[9] Robert E Eskridge, Jia Yeong Ku, S Trivikrama Rao, P Steven Porter, and Igor G Zurbenko. 1997. Separating different scales of motion in time series of meteorological variables. *Bulletin of the American Meteorological Society* 78, 7 (1997), 1473–1484.

[10] Rosie A Fisher, Charles D Koven, William RL Anderegg, Bradley O Christoffersen, Michael C Dietze, Caroline E Farrior, Jennifer A Holm, George C Hurtt, Ryan G Knox, Peter J Lawrence, et al. 2018. Vegetation demographics in Earth System Models: A review of progress and priorities. *Global change biology* 24, 1 (2018), 35–54.

[11] JP Fisk, GC Hurtt, JQ Chambers, H Zeng, KA Dolan, and RI Negrón-Juárez. 2013. The impacts of tropical cyclones on the net carbon balance of eastern US forests (1851–2000). *Environmental Research Letters* 8, 4 (2013), 045017.

[12] Francesco Fuso Nerini, Benjamin Sovacool, Nick Hughes, Laura Cozzi, Ellie Cosgrave, Mark Howells, Massimo Tavoni, Julia Tomei, Hisham Zerriffi, and Ben Milligan. 2019. Connecting climate action with other Sustainable Development Goals. *Nature Sustainability* 2, 8 (2019), 674–680.

[13] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*. Ieee, 6645–6649.

[14] Paul C Hanson, Aviah B Stillman, Xiaowei Jia, Anuj Karpatne, Hilary A Dugan, Cayelan C Carey, Jemma Stachelek, Nicole K Ward, Yu Zhang, Jordan S Read, et al. 2020. Predicting lake surface water phosphorus dynamics using process-guided machine learning. *Ecological Modelling* 430 (2020), 109136.

[15] Weiwei Huo, Weier Li, Zehui Zhang, Chao Sun, Feikun Zhou, and Guoqing Gong. 2021. Performance prediction of proton-exchange membrane fuel cell based on convolutional neural network and random forest feature selection. *Energy Conversion and Management* 243 (2021), 114367.

[16] GC Hurtt, Stephen Wilson Pacala, Paul R Moorcroft, J Caspersen, E Shevliakova, RA Houghton, and B Moore Iii. 2002. Projecting the future of the US carbon sink. *Proceedings of the National Academy of Sciences* 99, 3 (2002), 1389–1394.

[17] G Hurtt, M Zhao, R Sahajpal, A Armstrong, R Birdsey, E Campbell, Katelyn Dolan, R Dubayah, JP Fisk, S Flanagan, et al. 2019. Beyond MRV: high-resolution forest carbon modeling for climate mitigation planning over Maryland, USA. *Environmental Research Letters* 14, 4 (2019), 045013.

[18] George C Hurtt, Paul R Moorcroft, Stephen W Pacala And, and Simon A Levin. 1998. Terrestrial models and global change: challenges for the future. *Global Change Biology* 4, 5 (1998), 581–590.

[19] Xiaowei Jia, Yiqun Xie, Sheng Li, Shengyu Chen, Jacob Zwart, Jeffrey Sadler, Alison Appling, Samantha Oliver, and Jordan Read. 2021. Physics-guided machine learning from simulation data: An application in modeling lake and river systems. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 270–279.

[20] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. 2009. Multi-class active learning for image classification. In *2009 ieee conference on computer vision and pattern recognition*. IEEE, 2372–2379.

[21] Avinash Kumar, Vishnu Teja Narapareddy, Veerubhotla Aditya Srikanth, Aruna Malapati, and Lalita Bhanu Murthy Neti. 2020. Sarcasm detection using multi-head attention based bidirectional LSTM. *Ieee Access* 8 (2020), 6388–6397.

[22] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. 2018. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 95–104.

[23] Nathan Lambert, Albert Wilcox, Howard Zhang, Kristofer SJ Pister, and Roberto Calandra. 2021. Learning accurate long-term dynamics for model-based reinforcement learning. In *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2880–2887.

[24] Mingkun Li and Ishwar K Sethi. 2006. Confidence-based active learning. *IEEE transactions on pattern analysis and machine intelligence* 28, 8 (2006), 1251–1261.

[25] Yang Li, Yanlan Liu, Gil Bohrer, Yongyang Cai, Aaron Wilson, Tongxi Hu, Zhihao Wang, and Kaiguang Zhao. 2022. Impacts of forest loss on local climate across the conterminous United States: Evidence from satellite time-series observations. *Science of the Total Environment* 802 (2022), 149651.

[26] Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, Anima Anandkumar, et al. 2021. Fourier Neural Operator for Parametric Partial Differential Equations. In *International Conference on Learning Representations*.

[27] Lei Ma, George Hurtt, Lesley Ott, Ritvik Sahajpal, Justin Fisk, Rachel Lamb, Hao Tang, Steve Flanagan, Louise Chini, Abhishek Chatterjee, et al. 2022. Global evaluation of the Ecosystem Demography model (ED v3. 0). *Geoscientific Model Development* 15, 5 (2022), 1971–1994.

[28] L MA, GC HURTT, H TANG, R LAMB, AJ LISTER, LP CHINI, RO DUBAYAH, J ARMSTON, E CAMPBELL, L DUNCANSON, et al. 2023. Global Forest Aboveground Carbon Stocks and Fluxes from GEDI and ICESat-2, 2018-2021. *ORNL DAAC* (2023).

[29] Nate G McDowell, Craig D Allen, Kristina Anderson-Teixeira, Brian H Aukema, Ben Bond-Lamberty, Louise Chini, James S Clark, Michael Dietze, Charlotte Grossiord, Adam Hanbury-Brown, et al. 2020. Pervasive shifts in forest dynamics in a changing world. *Science* 368, 6494 (2020), eaaz9463.

[30] Paul R Moorcroft, George C Hurtt, and Stephen W Pacala. 2001. A method for scaling vegetation dynamics: the ecosystem demography model (ED). *Ecological monographs* 71, 4 (2001), 557–586.

[31] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. 2021. A survey of deep active learning. *ACM computing surveys (CSUR)* 54, 9 (2021), 1–40.

[32] Filipe Rodrigues, Ioulia Markou, and Francisco C Pereira. 2019. Combining time-series and textual data for taxi demand prediction in event areas: A deep learning approach. *Information Fusion* 49 (2019), 120–129.

[33] Debashis Sahoo, Naveksha Sood, Usha Rani, George Abraham, Varun Dutt, and AD Dileep. 2020. Comparative analysis of multi-step time-series forecasting for network load dataset. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 1–7.

[34] Ozan Sener and Silvio Savarese. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489* (2017).

[35] Wilhelm E Sorteberg, Stef Garasto, Alison S Pouplin, Chris D Cantwell, and Anil A Bharath. 2018. Approximating the solution to wave propagation using deep neural networks. *arXiv preprint arXiv:1812.01609* (2018).

[36] Souhaib Ben Taieb, Gianluca Bontempi, Amir F Atiya, and Antti Sorjamaa. 2012. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert systems with applications* 39, 8 (2012), 7067–7083.

[37] Jonathan Tompson, Kristofer Schlachter, Pablo Sprechmann, and Ken Perlin. 2017. Accelerating eulerian fluid simulation with convolutional networks. In *International Conference on Machine Learning*. PMLR, 3424–3433.

[38] Harm Vanseijen and Rich Sutton. 2015. A deeper look at planning as learning from replay. In *International conference on machine learning*. PMLR, 2314–2322.

[39] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418* (2019).

[40] Simon DP Williams, Yehuda Bock, Peng Fang, Paul Jamason, Rosanne M Nikolaidis, Linette Prawirodirdjo, Meghan Miller, and Daniel J Johnson. 2004. Error analysis of continuous GPS position time series. *Journal of Geophysical Research: Solid Earth* 109, B3 (2004).

[41] Zhaohua Wu, Norden E Huang, Steven R Long, and Chung-Kang Peng. 2007. On the trend, detrending, and variability of nonlinear and nonstationary time series. *Proceedings of the National Academy of Sciences* 104, 38 (2007), 14889–14894.

[42] Zhongrun Xiang, Jun Yan, and Ibrahim Demir. 2020. A rainfall-runoff model with LSTM-based sequence-to-sequence learning. *Water resources research* 56, 1 (2020), e2019WR025326.

[43] Yiqun Xie, Erhu He, Xiaowei Jia, Han Bao, Xun Zhou, Rahul Ghosh, and Praveen Ravirathinam. 2021. A statistically-guided deep network transformation and moderation framework for data with spatial heterogeneity. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 767–776.