Knowledge-Guided Learning for Global Carbon Flux Prediction: Integrating High-Level Remote Sensing with Bottom-Up Physical Modeling

Shuo Xu
University of Maryland
College Park, USA
shuoxu98@umd.edu

Zhihao Wang
University of Maryland
College Park, USA
zhwang1@umd.edu

Ruohan Li
University of Maryland
College Park, USA
r526li@umd.edu

Ruichen Wang
University of Maryland
College Park, USA
ruichenw@umd.edu

Lei Ma*
University of Maryland
College Park, USA
lma6@umd.edu

George C. Hurtt*
University of Maryland
College Park, USA
gchurtt@umd.edu

Xiaowei Jia
University of Pittsburgh
Pittsburgh, USA
xiaowei@pitt.edu

Yiqun Xie†
University of Maryland
College Park, USA
xie@umd.edu

Abstract

Forest ecosystems play a critical role in the global carbon cycle and Earth system, providing essential ecosystem services, including carbon sequestration, biodiversity support, and natural resources. Process-based physical models have been developed for decades and widely used to predict carbon dynamics in forest ecosystems. However, they are often based on predefined rules under simplifying assumptions, and have limited flexibility to incorporate increasingly available observations from remote sensing and in-situ networks. In particular, ecosystem models mostly rely on bottom-up subprocesses with subprocess-specific parameterization to mirror the real-world system. While remote sensing satellites provide large-scale observations of key ecosystem parameters (e.g., plant functional types) that can inform the calibration of individual subprocesses, such usage has not been considered due to the granularity mismatch: high-level, aggregated satellite observations are not applicable to individual sub-processes. Existing knowledge-guided learning approaches primarily focus on directly matched variables at the shared granularity to constrain the training process. We propose a knowledge-guided learning framework that bridges this representation gap with a decomposition-resembling design to incorporate high-level, aggregated remote sensing parameters into bottom-up, individualized sub-processes. Furthermore, to address the challenge of sparse in-situ carbon flux labels (i.e., final targets), the framework also includes uncertainty-aware refinements using probabilistic label expansion. Experimental results on global-scale benchmark datasets show that the knowledge-guided framework significantly improves prediction performance by enabling the effective use of high-level information.

CCS Concepts

• **Computing methodologies** → **Machine learning algorithms**; • **Applied computing** → **Environmental sciences**; • **Mathematics of computing** → *Time series analysis.*

Keywords

Knowledge-guided learning, carbon flux prediction, remote sensing, process-based modeling, time-series prediction

ACM Reference Format:
Shuo Xu, Zhihao Wang, Ruohan Li, Ruichen Wang, Lei Ma, George C. Hurtt, Xiaowei Jia, and Yiqun Xie. 2026. Knowledge-Guided Learning for Global Carbon Flux Prediction: Integrating High-Level Remote Sensing with Bottom-Up Physical Modeling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3818927

1 Introduction

Forest ecosystems form a fundamental component of the Earth system, exerting strong controls on the global carbon cycle through carbon sequestration, storage, and exchange with the atmosphere. In addition to being one of the largest carbon sinks (e.g., terrestrial ecosystems absorb ~1/3 of CO_2 emission), forests also support biodiversity, regulate ecosystem processes, and provide essential natural resources. As such, the ability to monitor and understand carbon fluxes in forest ecosystems at global scale is essential in Earth science to advance carbon cycle research and inform management. Advanced air-borne systems such as remote sensing satellites and in-situ platforms such as networks of monitoring stations have generated invaluable observations of the Earth systems and offer exciting opportunities in enhancing the monitoring capabilities. Such abilities are essential to improve the solutions for tackling grand challenges such as carbon neutralization, global warming, extreme events, etc. In carbon flux monitoring, satellite platforms provide global coverage of forest ecosystems, offering critical information such as forest coverage, plant function types, and canopy heights. On the other hand, in-situ sites consisting networks of carbon flux towers offer years of highly dynamic observations of key carbon variables. The increasing availability of sensing-based observations

*Lei Ma and George Hurtt are domain science experts in ecosystems, biosphere, and global carbon cycle.
†Corresponding author: Yiqun Xie (xie@umd.edu)

KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818927

has led to major advances of process-based models to better integrate this valuable information into the modeling to enhance the estimation. In forest ecosystems, the Ecosystem Demography (ED) model is a new generation of models that has been developed over the past two decades, providing the ability to incorporate these observations (e.g., billions of height measurements from NASA GEDI) at the global scale. The ED model plays important roles in major science and policy efforts, including the Global Carbon Budget [18, 19] and the NASA Carbon Monitoring System [26], and has been operationally adopted to support decision making, e.g., at the State of Maryland, US, for carbon inventory updates [2, 3].

Despite the promising potential, there are several challenges integrating sensing and process-based modeling: (1) Most of the models, including ED, follow rule-based processes. While the knowledge structure does not require extensive training, it also significantly constrains the models' ability to refine the fixed structure (e.g., structures with simplifying assumptions) using rich real observations. (2) Process-based models such as ED are often designed in a detailed, bottom-up manner, which use a collection of lower-level, sub-stream processes and gradually combine them into the full system for modeling and predictions. However, remote sensing satellites very often only collect aggregated, high-level information, that do not provide the needed observations for the sub-stream processes, significantly constraining the value of the large-scale monitoring results. For example, in the ED model, the overall ecosystem process is modeled as a combination of many sub-stream processes corresponding to different forest ages, where each experiences different competition between different plant function types (PFTs) such as deciduous trees, evergreens, or grass/shrubs. Due to the complexity of the cross-PFT competition and the interactions with different environment conditions, PFT proportions generated by ED may deviate from the true values over time, resulting in errors in its carbon variable estimates. To address this, it will be ideal if the PFT observations from remote sensing satellites can be leveraged to dynamically correct the values. However, the satellite-based observations reflect only the aggregated results from all the sub-stream processes (e.g., often tens or hundreds of them depending on the granularity) and do not contain sub-stream level information, making the high-level observations not compatible with the lower-level, bottom-up modeling. As a result, these valuable observations have not been utilized in the ED model for quality improvements. (3) The inputs and outputs from the process-based models at large scale (e.g., global-scale) are often large in data volume. As a result, domain scientists normally do not further save intermediate results from different intermediate processes and it is highly expensive to regenerate the results, making it harder to associate the observations with bottom-up processes. (4) While in-situ observations from flux towers provide direct labels for the final carbon flux variables, they are only available at limited sites. Furthermore, the temporal coverage is often highly constrained compared to the multi-decade scope of ecosystem modeling (e.g., 40 years, 100 years), due to sensor installation time, repairs, and maintenance (Fig. 1).

Existing works on knowledge-guided machine learning (ML) mainly focus on connecting process-based and data-driven methods via directly matched variables, using physical rules and simulations to constrain the training process [65]. Earlier approaches were typically trained to estimate the residual between observations and the

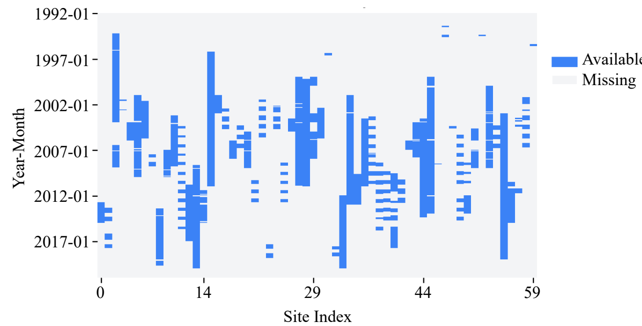

Figure 1: Example in-situ data availability from ABoVE.

outputs of physics-based simulations [16, 59, 67], which was later extended into hybrid approaches that integrate results from both physics-based and ML models [31, 47, 68]. However, in these cases the process-based and ML models still operate separately and do not exploit complementary benefits. More recently, studies have highlighted the promise of using physical knowledge to inform and guide the training of ML models. Such efforts include the design of loss functions that enforce compliance with known physical laws [15, 28, 31, 36, 48, 50, 55, 70], strategies for initializing models through knowledge transferred from simulations [25, 28, 41, 50, 56], and development of model architectures that explicitly encode physical symmetries [5, 51] and general physical relationships such as mass conservation [11, 24, 37, 44, 52, 53, 72]. These models demonstrated that ML models can acquire more generalizable abilities with limited observations by using knowledge from process-based models. However, these models focus on training with matched variables and do not consider or address the utilization of satellite-based indirect labels (i.e., high-level labels are not compatible with sub-stream processes). The models with knowledge-guided architecture also rely on intermediate outputs from smaller and less-expensive process-based models that are often unavailable in large scale problems due to the excessive storage cost as well as the expensive computation to re-run the middle outputs.

We propose a knowledge-guided framework to integrate process-based and learning-based models by enabling the use of high-level remote sensing observations at large scale. Specifically, our contributions are:

- We propose a knowledge-guided learning framework with a learned decomposition-and-resembling (DERE) process to enable high-level label integration: (1) Knowledge-aligned decomposition of end-to-end simulation data to represent intermediate, sub-stream processes of a process-based model, as a preparation for integration of indirect, higher-level labels. This is necessary as there is often no intermediate modeling results saved for large-scale process-based models, which are costly both in space and computation. (2) Knowledge-aligned resembling of decomposed intermediate, sub-stream processes to enable supervision from high-level labels available at large-scale to constrain the sub-stream process. This data-driven sub-stream process calibration can significantly enhance model generalizability using limited in-situ flux tower observations of final carbon variables.
- We propose a diffusion-based probabilistic label expansion module to increase the temporal coverage of in-situ observations

for finetuning, with explicitly learned uncertainty-awareness to incorporate the generated probabilistic labels during training.
- We carry out global-scale experiments with multiple carbon monitoring network datasets using the CarbonSense benchmark data [17]. Extensive comparisons with time-series models and their knowledge-guided extensions demonstrated the effectiveness of our DERE-based knowledge-guided learning framework.

The code is available on GitHub at https://github.com/ai-spatial/DERE.

2 Problem Definition

2.1 General formulation

Our problem is formulated with the following inputs and outputs:
Inputs: (1) Physical/environment conditions $\boldsymbol{x}_{s,t}$ that are needed to infer target output variables $\boldsymbol{y}_{s,t}$ (e.g., carbon flux) at each location s in a spatial domain $\mathcal{S}$ (e.g., global) along each time step t in a time-series $\mathcal{T} = \{1, ..., T\}$. (2) Initial states $\boldsymbol{c}_k$ (e.g., forest ages). Each location (e.g., a sub-region of a forest) consists of a set of initial states $\{\boldsymbol{c}_1, ..., \boldsymbol{c}_K\}$ and their proportions $\{\alpha_1, ..., \alpha_K\}$.
Outputs: Predicted target variables $\hat{\boldsymbol{y}}_{s,t}$ for the set of locations and time-series. The ground truth $\boldsymbol{y}_{s,t}$ from in-situ data are available at a limited number of locations $\mathcal{S}' \subset \mathcal{S}$ for a subset of temporal periods $\mathcal{T}' \subset \mathcal{T}$.

In addition, there are the following auxiliary information related to process-based modeling:

- **Physical model-based estimates of target output variables**, $(\boldsymbol{y}^P_{s,t})_k$, using a process-based model $\mathcal{M}^P$ (e.g., ED) based on the physical conditions $\boldsymbol{x}_{s,t}$ in $\mathcal{S}$ and $\mathcal{T}$ and initial state $\boldsymbol{c}_k$. Here $(\boldsymbol{y}^P_{s,t})_k$ represents the end-results from the process-based model for each initial state and does not further contain intermediate results (e.g., results from different PFTs and their competition in ecosystems) due to the excessive cost of space and computation in large-scale applications. In other words, domain scientists often do not save the intermediate results due to the storage cost and it is also too expensive to re-run the model to generate them.
- **Aggregated, high-level observations**, $z_{s,t}$ (e.g., from satellites), on outputs of the intermediate, sub-stream processes as defined in Def. 1. These observations cannot be used in the process-based model because they are mixed at the aggregated level and are not compatible with the sub-stream processes.

Definition 1 (Intermediate, sub-stream processes). *Denote $\mathcal{M}^P$ as the entire set of functions from a process-based model, with $(\boldsymbol{y}^P_{s,t})_k = \mathcal{M}^P(\boldsymbol{x}_{s,t}, \boldsymbol{c}_k)$. In process-based modeling, $\mathcal{M}^P$ often consists of a set of intermediate and sub-stream processes:*

– *An **[Intermediate Process]** generates intermediate results from part of the physical system, and the results are fed into other parts to complete the simulation:*

$$\mathcal{M}^P(\cdot) = \mathcal{M}^P_1(\mathcal{M}^P_2(...))$$

where $\mathcal{M}^P_1$ and $\mathcal{M}^P_2$ are intermediate processes. For example, the forest ecosystem process contains intermediate processes from cohorts of trees belonging to different PFTs. These trees of different PFTs then compete to yield the final outcome.

– *An intermediate process exists in different **[Sub-stream Processes]**, which are same physical processes based on different initial states $\boldsymbol{c}_k$. They run in parallel and aggregate into the complete process. For example, a forest area often contains trees with different initial ages (initial states) in different sub-plots. Formally, there exist:*

$$\mathcal{M}^P_1\big(\mathcal{M}^P_2(\boldsymbol{x}_{s,t}, \boldsymbol{c}_1)\big),\ \mathcal{M}^P_1\big(\mathcal{M}^P_2(\boldsymbol{x}_{s,t}, \boldsymbol{c}_2)\big),\ ...,\ \mathcal{M}^P_1\big(\mathcal{M}^P_2(\boldsymbol{x}_{s,t}, \boldsymbol{c}_K)\big)$$

2.2 Global carbon flux prediction

The general formulation provides basic definitions for later method description. Here we also illustrate this formulation for carbon flux prediction. Specifically, the **input physical conditions** $\boldsymbol{x}_{s,t}$ include meteorological variables (e.g., temperature, precipitation), soil properties and more, where each time step may correspond to a month or shorter for monitoring over multiple decades at the global scale. The initial state $\boldsymbol{c}_k$ corresponds to the initial age of trees at the beginning of the process. The **output carbon variables** $\boldsymbol{y}_{s,t}$ include Gross Primary Production (GPP), Net Ecosystem Exchange (NEE), Ecosystem Respiration (RECO), etc.

The auxiliary information from process-based modeling include: (1) **Estimates of carbon variables $\boldsymbol{y}^P_{s,t}$ from the ED model** [27, 40, 43] on variables such as GPP, NEE and RECO, which do not contain intermediate results due to the excessive storage and recomputation cost. (2) **High-level, aggregated observations $z_{s,t}$ from remote sensing satellites**, which provide information of forest PFTs (e.g., deciduous, evergreen, shrubs) over large geographic scale that are important for cross-PFT competition modeling. However, in ecological modeling, such natural competitions need to be modeled as multiple sub-streams of different forest processes (i.e., based on different initial ages $\boldsymbol{c}_k$), whereas the sensing-based PFTs do not contain information at the needed granularity. As a result, existing efforts have not been able to leverage these high-level observations to enhance the prediction quality. Fig. 2 shows an overview of the multi-source data involved.

Interdisciplinary Collaboration: This work is performed in collaboration with domain scientists in carbon monitoring. For example, Hurtt is an original creator of the ED model [40, 43] and the Science Team Lead of the NASA Carbon Monitoring System. This study aims to leverage AI to integrate process-based models, in-situ observations, and large-scale, high-level remote sensing data to enhance carbon flux monitoring capabilities.

3 Related Work

Time-series prediction. Deep learning models for time-series prediction are natural data-driven frameworks to model the input-output relationships in our problem setting. Transformer-based architectures have become widely adopted for tasks with long sequences thanks to their ability to capture long-range dependencies [12, 13, 58] compared to earlier models based on recurrent structures [9, 33, 63, 66]. Numerous adaptations of transformers have been introduced for further improvements, including ProbSparse self-attention in Informer [74], frequency-domain attention in FEDFormer [75], cross-feature as well as cross-time dependency modeling in Crossformer [73], exogenous feature integration in TimeXer [61], time-variable inversion in iTransformer [39], and multivariate SimpleTM [8]. While these models have shown promising performances in general forecasting tasks, they are by design data-driven methods that do not consider physical guidance from process-based

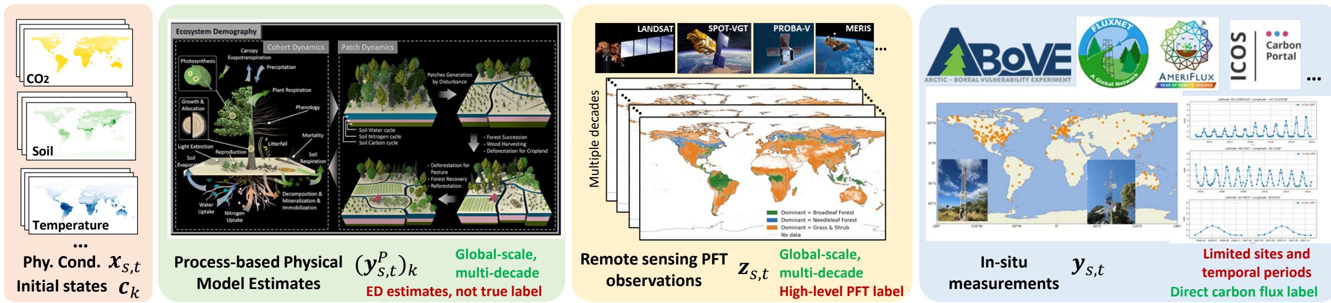

Figure 2: Examples of multi-source data, where the goal is to integrate high-level remote sensing with bottom-up processes.

models, limiting their performance when only limited in-situ observations are available for large-scale applications.

Data-driven emulation of process-based model. Recent studies have also explored variants of forecasting models as learning-based emulators to approximate process-based models. For example, these models have been developed to emulate climate models [42, 49, 60, 71] and weather forecasting models [7, 32, 34] to improve the scalability for higher resolution tasks. In addition, deep learning surrogates have been widely explored for process-based simulations involving the solution of partial differential equations [30, 46, 54]. Our previous efforts also developed learning-based emulators for ecosystem and carbon simulation models at the global scale [35, 62]. However, these emulators mainly aim to enhance the computational efficiency of process-based models instead of combining physical knowledge with real observations to further improve the prediction quality.

Knowledge-guided machine learning. There has been growing efforts on incorporating physics into ML models to enhance both predictive performance and generalizability for solving scientific problems [65]. Early studies mainly consider residual modeling, where simple regression models [16, 67] or recurrent networks [59] are used to infer differences between process-based models and ground truth. However, these methods are unable to enforce physics-based constraints and can only make use of the simulation results when corresponding observations are simultaneously available. More recent strategies start focusing on deeper integration of knowledge, and common strategies include modifying layer architectures based on process-based models [4, 14, 45], pretraining with simulation data [20, 25, 50, 56, 69], and adding physics-constrained loss functions [28, 36, 48, 50, 70]. Variants have also been developed to learn from multiple process-based models [10, 29], performing model selection [9], integrating simulation data to reduce bias [23, 64], etc. However, these models focus on training with directly matched variables and do not consider high-level, indirect labels that are not readily compatible with process-based models but are often available at aggregated levels over large geographic scale. Similarly, the models that use knowledge-based architectures based on dissected physical models [23, 38, 65, 69] mainly utilize intermediate outputs from simulation models to enforce physical rules (e.g., conservation laws), and do not consider integration of external high-level labels (e.g., from remote sensing).

4 Knowledge-Guided Learning: Integrating High-Level Remote Sensing Observations

The knowledge-guided learning framework has a decomposition-and-resembling structure to enable the integration of high-level remote sensing labels at large scale to provide guidance to intermediate, sub-stream processes. Sec. 4.1 and 4.2 discuss details of the designs, and Sec. 4.3 presents a probabilistic label expansion module to increase the temporal coverage of in-situ labels and uncertainty-aware finetuning. Fig. 3 shows the overall framework.

4.1 Knowledge-aligned data decomposition

As explained in Def. 1, the entire physical process $\mathcal{M}^P$ contains intermediate, sub-stream processes. These are often not readily available for data-driven models to learn right off the simulation data, as the intermediate results are often not saved due to the excessive size at large scale and high cost of re-generation. Thus, a necessary initial step is to separate out the intermediate process of interest as a preparation to incorporate the high-level remote sensing labels (Sec. 2.1) in the later resembling step. This is the goal of knowledge-aligned decomposition. For clarity: (1) the decomposition only separates out an intermediate process and does not consider sub-stream processes that will be addressed in the resembling step; and (2) the decomposition step uses only the simulation data $\boldsymbol{y}^P_{s,t}$ from $\mathcal{M}^P$ and does not use observations.

As shown in Fig. 3(a), the decomposition step splits out an intermediate process, approximated by a learned function $\mathcal{F}^z(\boldsymbol{x}_{s,t}, \boldsymbol{c}_k)$ to generate $((\hat{\boldsymbol{x}}^z_{s,t})_k, (\hat{z}_{s,t})_k)$ before reaching the final output $(\hat{\boldsymbol{y}}_{s,t})_k$. Specifically, for ecosystem carbon models, it aims to separate intermediate processes of trees belonging to different PFTs (e.g., evergreen, deciduous) and the corresponding PFT proportions $(\hat{z}_{s,t})_k$. The carbon outcomes $(\hat{\boldsymbol{x}}^z_{s,t})_k$ of different PFT groups are then combined via competition to generate the final outputs $(\hat{\boldsymbol{y}}_{s,t})_k$ (Fig. 3 (a), bottom). This is important as the key objective is to explicitly represent $(\hat{z}_{s,t})_k$ so later on it can be connected to the high-level remote sensing labels during resembling. The functional relationship between $((\hat{\boldsymbol{x}}^z_{s,t})_k, (\hat{z}_{s,t})_k)$ and $(\hat{\boldsymbol{y}}_{s,t})_k$, denoted by "$\otimes$", can be defined either using a pre-defined, differentiable function based on the physical process (when such functions are direct and simple) or using another learned function, or a combination of the two. In our case, $(\hat{\boldsymbol{x}}^z_{s,t})_k$ are combined into final carbon variables by linear combination (i.e., a pre-defined weighted sum based on proportions $(\hat{z}_{s,t})_k$) plus non-linear competition using a learned function.

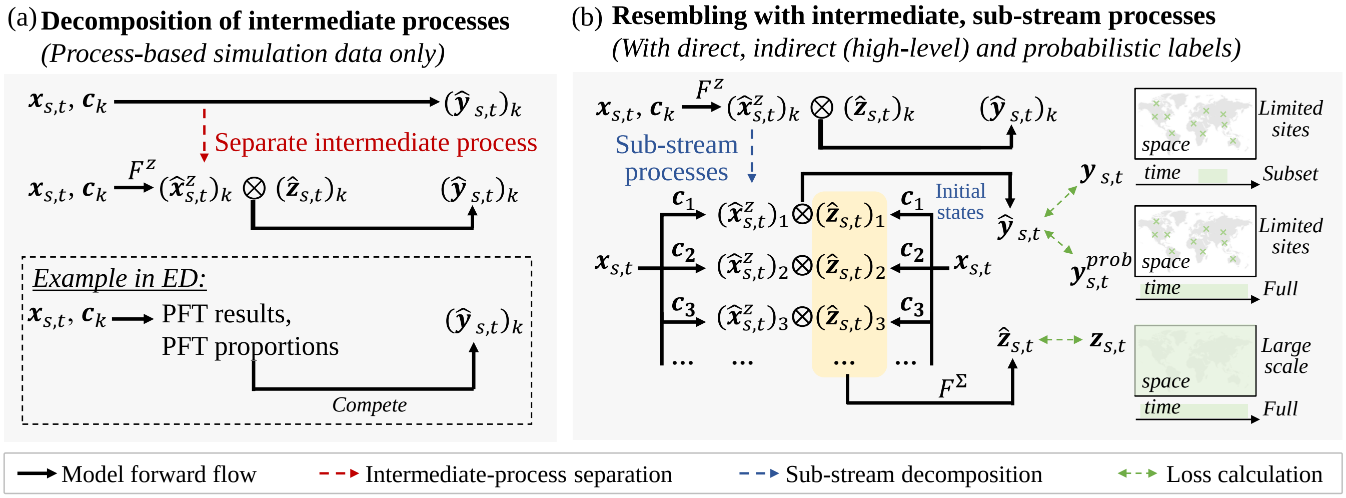

Figure 3: Overview of the decomposition and resembling framework. The left panel shows simulation-based decomposition, while the right panel shows observation-guided resembling with direct, probabilistic, and high-level remote sensing labels.

All the learned functions can be pre-trained using the ED model estimates before finetuned later using observation data.

4.2 Knowledge-aligned resembling of intermediate, sub-stream processes

The resembling process will build on the separated intermediate process and integrate it with different types of real observations. As shown in Fig. 3(b), the resembling process will first explicitly integrate all sub-stream processes, governed by different initial states $\boldsymbol{c}_k$ (i.e., forest ages). Here the function $\mathcal{F}^z(\boldsymbol{x}_{s,t}, \boldsymbol{c}_k)$ is responsible for generating all pairs of $((\hat{x}^z_{s,t})_k, (\hat{z}_{s,t})_k)$ corresponding to different initial states $\boldsymbol{c}_k$, and these sub-stream results are then combined into the final prediction $\hat{\boldsymbol{y}}_{s,t}$. The final prediction is no longer subscripted by k as all sub-streams will be combined using corresponding proportions and learned competition function, as described in Sec. 4.1.

The most important part of the resembling is the integration of direct labels $\boldsymbol{y}_{s,t}$ (e.g., observations from in-situ flux towers on carbon variables) and high-level remote sensing labels $z_{s,t}$ (e.g., aggregated PFT proportions across ages $\boldsymbol{c}_k$ from satellite observations). In particular, the key is to enable the use of high-level labels $z_{s,t}$ that are often available at much larger scales, both geographically and temporally, compared to direct labels $\boldsymbol{y}_{s,t}$ as shown by the illustrative maps in Fig. 3(b) as well as Fig. 2; for example, satellite-based $z_{s,t}$ tends to have global coverage. With the decomposition of intermediate processes and the resembling of sub-stream processes, the high-level labels $z_{s,t}$ can be compared with the integrated $\mathcal{F}^\Sigma(\{(\hat{z}_{s,t})_k\}_{k=1\ldots K})$ from all the sub-stream processes. In the ecosystem carbon case, $\mathcal{F}^\Sigma$ can be prefixed by a simple sum, $\sum_i^K((\hat{z}_{s,t})_k \cdot \alpha_k)$, where α_k is the weight of an initial state $\boldsymbol{c}_k$ that is used to aggregate the PFT proportions across the states.

Furthermore, the final predictions $\hat{\boldsymbol{y}}_{s,t}$ are constrained by the direct labels $\boldsymbol{y}_{s,t}$ as a regular part of a training process, which are often available at a limited number of locations (e.g., carbon flux towers that are expensive to build). We will also enhance it with a probabilistic label tuning method in the next section. During training, each batch contains samples intended both for the high-level label comparison and direct label comparison.

4.3 Probabilistic label expansion and uncertainty-aware finetuning

As direct labels $\boldsymbol{y}_{s,t}$ are often limited in the temporal domain (i.e., major data gaps in carbon flux towers), we further develop a probabilistic label expansion strategy and uncertainty-aware tuning to enrich the usable labels. For example, Fig. 4 shows several examples of observations at carbon flux tower sites, where the measurements are only available for a subset of years (more general availability distributions in Fig. 1). To address this, we propose a simpler prediction task to expand the labels, where the goal is to predict the missing labels at each site leveraging existing labels as additional inputs. Comparing to the original task that aims to predict $\boldsymbol{y}_{s,t}$ using $\boldsymbol{x}_{s,t}$ and $\boldsymbol{c}_k$, this task has significantly reduced difficulty as it only aims to make predictions at sites where a set of labels is already known in a time window, and those labels are given as part of the inputs. This makes it feasible to leverage these predictions to facilitate the model tuning in our original task. Furthermore, we make the predictions probabilistic so the finetuning step can explicitly utilize the uncertainty to determine whether or not a prediction should be used. Specifically, we adopt conditional diffusion [57] to generate the missing measurements, where the existing observations from the same time-series can be added as conditions, and the variance can be obtained. During training, we set a subset of observations as part of the given conditions while using the remaining observations for loss evaluations. Denote $\boldsymbol{y}^m_{s,t}$ as the masked observations for prediction, $\boldsymbol{y}^{m'}_{s,t}$ as those given as conditions, and j as the position in the denoising sequence, we have:

$$p_\theta((\boldsymbol{y}^m_{s,t})_{j-1} \mid (\boldsymbol{y}^m_{s,t})_j, \boldsymbol{y}^{m'}_{s,t}) = \mathcal{N}\Big((\boldsymbol{y}^m_{s,t})_{j-1}; \mu_\theta\big((\boldsymbol{y}^m_{s,t})_j, \boldsymbol{y}^{m'}_{s,t}\big), \sigma_j^2 I\Big)$$

Once trained, we generate 10 samples per site to estimate the variance and confidence interval. The training of the diffusion-based

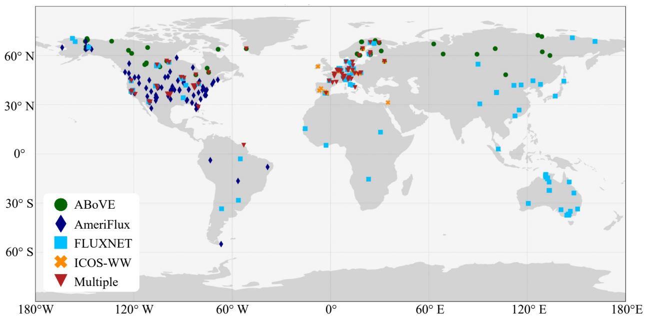

Figure 4: Distribution of in-situ sites.

label expansion only involves the training sites in the evaluation (Sec. 5.1), ensuring strict separation from the test sites.

Given the predictions with variance, we include an uncertainty-aware tuning module as part of the resembling process, where a learned sub-network is used to adaptively determine the weight of each predicted label based on its value and variance. Overall, our DERE model is trained with direct labels (limited spatio-temporal coverage), probabilistic labels, and indirect, high-level labels (full spatio-temporal coverage), as shown in the right side of Fig. 3.

5 Experiments

5.1 Data

We conducted extensive experiments at the global scale using the collection of datasets from the CarbonSense benchmark data [17] for carbon flux prediction. Specifically, CarbonSense includes various datasets representing different in-situ carbon flux observation networks under different conditions, including AmeriFlux, FLUXNET, ICOS-2023, ICOS-WW, and a Multiple dataset spanning the networks. Through data inspection, we found that ICOS-2023 only has a few sites with one year of data. This leaves very few data points for testing that can cause highly unstable results. Thus, we replaced it (ICOS-WW still contains data from the network) with the recent ABoVE dataset covering the broad Arctic region [6]. Fig. 4 shows the geographic distribution of the sites from different datasets. We used 3 key variables GPP, RECO, and NEE for the evaluation, which are available across all the datasets. For model training and testing, we spatially split each dataset with 80% sites for training and 20% sites for testing (i.e., site-level separation between training and testing). The diffusion-based label-expansion component also involves only training sites.

For data on the process-based model side, we used the simulations from the ED model [27, 40, 43], which has global coverage and the simulation results fully covered the temporal range of the in-situ observations. The data is available in our recent benchmark dataset [62] focusing on physical model emulation (did not consider observations). For the high-level, remote sensing labels, we used the satellite-derived PFT information from the ESA CCI PFT dataset [22], which also has global coverage for the temporal range. All the datasets are publicly available.

5.2 Candidate methods

We consider four different categories of candidate methods: (1) The process-based model ED; (2) Learning-based time-series models as listed below; (3) Knowledge-guided extensions of the time-series models (more details later); (4) Our proposed method DERE. For the training, we used the recommended settings from the papers, and trained till convergence via patience checks on validation data (10% of training data; independent from test data).

- **Transformer** [58]: The vanilla Transformer serves as the fundamental sequence modeling framework with self-attention to help capture long-range temporal dependencies.
- **Informer** [74]: A Transformer variant that employs sparse attention and generative-style decoding to enhance long-sequence prediction efficiency.
- **FEDFormer** [75]: A Transformer variant using frequency-domain attention to strengthen series decomposition and boost forecasting accuracy.
- **iTransformer** [39]: A Transformer variant that adopts an inverted design, emphasizing feature dimensions over time steps and helping to exploit correlations among input variables.
- **Timexer** [61]: A Transformer variant that refines attention by integrating inter-target and input-target relations to better capture temporal dependencies.
- **SimpleTM** [8]: A lightweight time series forecasting model that streamlines architecture and improves computational efficiency while maintaining strong predictive performance.

For each forecasting method, we consider two variants: a default data-driven version (baseline) and an extension with knowledge-guided machine learning (KGML). The baseline models were trained directly on in-situ observations as target values, whereas the KGML variants include two generally used strategies: (1) They are pre-trained on simulation data from ED and then finetuned with in-situ observations; and (2) The training included physics-constrained loss functions based on the carbon mass balance (e.g., $NEE = RECO - GPP$), in addition to the regular MSE loss based on in-situ label values. Both are used in our DERE model as well. Finally, DERE is implemented using Transformer as the backbone.

5.3 Results

Overall Evaluation. Tables 1 to 3 present the overall testing performance using in-situ GPP, RECO, and NEE data, evaluated with MAE and RMSE. We provide the top-2 counts (i.e., number of times a method ranked in top-2 across all columns) for convenience. The proposed DERE method shows the best performance compared to the others. The pure data-driven baselines did not outperform the process-based ED model, potentially due to the generalization challenge with limited observations. As a reference, ED is a fairly strong model that has been used in major systems including NASA Carbon Monitoring System. In contrast, KGML-based methods and the proposed approach outperform the baseline, underscoring the importance of incorporating physics knowledge. The enhanced performance of the proposed DERE model can be attributed to its integration of direct, indirect (high-level) and probabilistic labels in addition to the simulation data, which will be further illustrated via ablation studies later. The results show the promising potential of integrating high-level remote sensing labels that are available at

Table 1: GPP result comparison (top results in bold; runner-ups with underlines).

	Methods	ABoVE		AmeriFlux		FLUXNET		ICOS-WW		Multiple		Top-2
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	count
Physical	ED	0.427	0.703	0.883	1.419	0.822	1.110	0.713	0.968	0.678	1.010	0/10
Baseline	Transformer	0.502	0.684	1.030	1.405	0.850	1.147	0.955	1.235	0.736	0.997	0/10
	Informer	0.523	0.729	1.093	1.465	0.882	1.180	0.859	1.096	0.817	1.065	0/10
	FEDformer	0.451	0.626	0.917	1.347	0.768	1.025	1.154	1.524	0.790	1.080	0/10
	iTransformer	0.561	0.662	1.007	1.480	0.887	1.162	1.009	1.340	0.851	1.149	0/10
	TimeXer	0.685	1.001	1.250	1.939	1.156	1.602	1.451	1.860	1.100	1.547	0/10
	SimpleTM	0.631	0.782	1.057	1.549	0.972	1.251	1.132	1.434	1.007	1.346	0/10
KGML	Transformer	0.352	0.541	0.651	1.033	0.699	0.951	0.711	1.014	0.636	0.919	2/10
	Informer	0.373	0.555	**0.604**	**0.947**	0.756	1.011	0.839	1.141	0.618	**0.888**	**4/10**
	FEDformer	0.369	0.531	0.769	1.115	0.726	0.926	0.830	1.121	0.701	1.003	0/10
	iTransformer	0.363	0.508	0.678	1.027	0.654	0.885	**0.641**	0.919	0.621	0.909	3/10
	TimeXer	0.375	0.566	0.657	1.036	0.649	0.879	0.791	1.050	0.678	0.990	1/10
	SimpleTM	0.406	0.587	0.737	1.102	0.669	**0.855**	0.651	0.916	0.680	0.976	3/10
Proposed	DERE	**0.302**	**0.485**	0.663	1.068	**0.598**	0.863	0.682	**0.888**	**0.585**	0.890	**7/10**

Table 2: RECO result comparison (top results in bold; runner-ups with underlines).

	Methods	ABoVE		AmeriFlux		FLUXNET		ICOS-WW		Multiple		Top-2
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	count
Physical	ED	0.431	0.657	0.603	0.855	0.603	0.833	**0.464**	**0.620**	0.457	0.649	**3/10**
Baseline	Transformer	0.602	0.857	0.804	1.040	0.559	0.743	0.698	0.886	0.488	0.709	0/10
	Informer	0.498	0.719	0.720	0.941	0.546	0.749	0.697	0.927	0.530	0.740	0/10
	FEDformer	0.431	0.611	0.616	0.867	0.482	0.638	0.832	1.054	0.574	0.804	0/10
	iTransformer	0.412	0.526	0.543	0.774	0.510	0.662	0.679	0.969	0.535	0.752	0/10
	TimeXer	0.607	0.854	0.946	1.279	0.713	0.960	1.386	1.770	0.784	1.031	0/10
	SimpleTM	0.503	0.660	0.620	0.865	0.561	0.709	0.862	1.126	0.626	0.859	0/10
KGML	Transformer	0.328	0.467	0.526	0.705	0.438	0.583	0.586	0.756	0.466	0.655	**3/10**
	Informer	0.354	0.495	0.448	**0.614**	0.475	0.626	0.559	0.716	0.463	0.638	**3/10**
	FEDformer	**0.319**	**0.443**	0.533	0.736	0.506	0.645	0.582	0.781	0.541	0.727	2/10
	iTransformer	0.348	0.492	**0.439**	0.650	0.445	0.585	0.595	0.796	0.537	0.733	2/10
	TimeXer	0.388	0.543	0.591	0.769	0.512	0.640	0.739	0.914	0.555	0.744	0/10
	SimpleTM	0.385	0.555	0.492	0.694	0.487	0.643	0.519	0.707	0.545	0.733	0/10
Proposed	DERE	0.320	0.472	0.465	0.671	**0.393**	**0.557**	0.502	0.639	**0.442**	**0.627**	**7/10**

Table 3: NEE result comparison (top results in bold; runner-ups with underlines).

	Methods	ABoVE		AmeriFlux		FLUXNET		ICOS-WW		Multiple		Top-2
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	count
Physical	ED	0.276	0.386	0.599	0.920	0.613	0.889	0.530	0.691	0.537	0.770	0/10
Baseline	Transformer	0.413	0.500	0.615	0.875	0.575	0.790	0.489	0.665	0.532	0.726	0/10
	Informer	0.217	0.287	0.626	0.860	0.561	0.787	**0.466**	**0.589**	0.554	0.741	2/10
	FEDformer	0.201	0.266	0.546	0.805	0.548	0.769	0.535	0.680	0.529	0.735	0/10
	iTransformer	0.365	0.427	0.711	1.026	0.697	0.940	0.585	0.762	0.625	0.855	0/10
	TimeXer	1.288	2.209	0.714	1.129	0.770	1.126	1.340	1.644	0.843	1.474	0/10
	SimpleTM	0.334	0.415	0.692	0.997	0.664	0.912	0.640	0.827	0.655	0.894	0/10
KGML	Transformer	0.177	0.244	0.507	0.733	0.504	0.722	0.511	0.721	0.489	**0.691**	**4/10**
	Informer	0.186	0.252	**0.456**	**0.700**	0.509	0.721	0.519	0.730	0.501	0.699	3/10
	FEDformer	0.197	0.268	0.506	0.741	0.522	0.735	0.475	0.627	0.492	0.698	2/10
	iTransformer	0.179	**0.239**	0.540	0.770	0.600	0.834	0.485	0.641	0.534	0.737	1/10
	TimeXer	0.192	0.264	0.529	0.800	0.532	0.760	0.563	0.735	0.505	0.714	0/10
	SimpleTM	0.187	0.255	0.582	0.842	0.588	0.790	0.527	0.709	0.492	0.698	0/10
Proposed	DERE	**0.175**	0.244	0.496	0.713	**0.478**	**0.683**	0.505	0.696	**0.476**	0.692	**8/10**

Table 4: Seasonal comparison of different methods (top results in bold; runner-ups with underlines). Both the baseline and KGML models are the Transformer versions.

	Methods	Spring		Summer		Fall		Winter		Top-2 Count
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	
GPP	Baseline	0.831	1.060	1.229	1.611	0.590	0.765	0.612	0.856	0/8
	KGML	<u>0.693</u>	<u>0.946</u>	<u>0.950</u>	**1.312**	<u>0.508</u>	<u>0.690</u>	<u>0.361</u>	<u>0.609</u>	8/8
	DERE	**0.644**	**0.883**	**0.915**	<u>1.325</u>	**0.446**	**0.622**	**0.338**	**0.570**	8/8
RECO	Baseline	0.562	0.759	0.830	1.105	0.514	0.682	0.473	0.677	0/8
	KGML	<u>0.441</u>	<u>0.613</u>	<u>0.682</u>	**0.876**	<u>0.433</u>	<u>0.569</u>	<u>0.321</u>	<u>0.452</u>	8/8
	DERE	**0.372**	**0.514**	**0.678**	<u>0.883</u>	**0.404**	**0.552**	**0.264**	**0.390**	8/8
NEE	Baseline	0.554	0.717	0.784	1.066	0.458	0.586	0.381	0.524	0/8
	KGML	**0.482**	**0.661**	<u>0.685</u>	<u>0.952</u>	<u>0.417</u>	<u>0.555</u>	<u>0.293</u>	**0.448**	8/8
	DERE	<u>0.505</u>	<u>0.698</u>	**0.634**	**0.908**	**0.393**	**0.513**	**0.291**	<u>0.457</u>	8/8

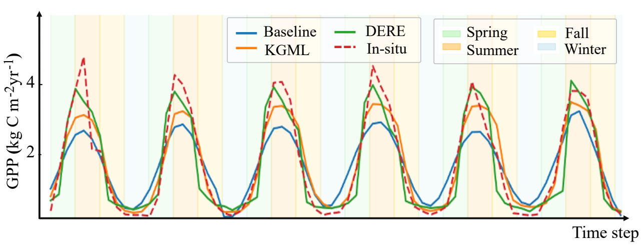

Figure 5: Comparison of GPP over time steps with seasonal backgrounds at an example site. Both the baseline and KGML models here are the Transformer versions.

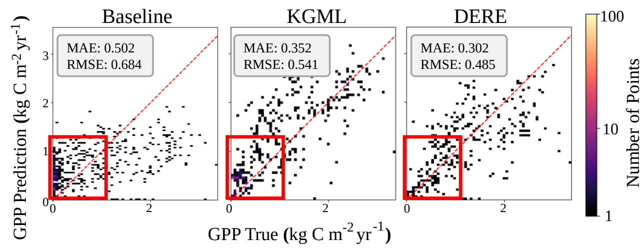

Figure 6: GPP true vs. prediction for ABoVE. Both the baseline and KGML model here are from the Transformer.

large scale. Since DERE used Transformer as the backbone, which is not necessarily optimal for every dataset, especially AmeriFlux where Informer showed stronger performances (the KGML version). Thus, we further tested DERE with Informer to confirm that it can inherit improvements from the backbone models, with the results included in the Appendix. In addition, the Appendix also included results about non-time-series models.

Seasonal Analysis. Since the carbon variables exhibit strong seasonality patterns, Table 4 further reports the seasonal performance of different methods. Seasons are defined as spring (Mar-May), summer (Jun-Aug), fall (Sep-Nov), and winter (Dec-Feb). Given the space constraints, the results are aggregated over the entire dataset and the baseline and KGML models are based on the Transformer backbone, which is the same as DERE. We can see the proposed DERE achieves the best performance in most cases compared to the baseline and KGML models, showing stable performance under different seasonal conditions. In addition, error magnitudes are strongly correlated with seasonal carbon flux variations: summer often shows higher errors (e.g., 0.915 for GPP MAE for DERE) due to higher carbon flux values and variations, and winter often shows the lowest errors (e.g., 0.338 for GPP MAE for DERE) due to reduced vegetation activity. These results further show that DERE can capture seasonal carbon flux patterns beyond the overall averaged evaluation.

Qualitative visualization. Figure 5 shows the temporal dynamics of GPP from the baseline, KGML, the proposed DERE method, and in-situ observations at an example site to visualize the qualitative performance. The Transformer backbone is used for the baseline and KGML here. As we can see, the proposed method aligns more closely with in-situ observations than the other models. Both the KGML and proposed methods demonstrate more stable and consistent temporal patterns than the baseline model. The seasonal backgrounds help visualize the seasonal pattern alignment, including the generally higher GPP in summer and lower GPP in winter observed in the in-situ data. Fig. 6 presents more detailed scatter plots of GPP predictions versus ground-truth observations across ABoVE network. The proposed DERE method is able to correct large number of deviations highlighted by the red boxes, leading to improved performance.

Probabilistic Label Expansion. Figure 7 visualizes examples of label expansion on the in-situ data with conditional diffusion, where the effect on final predictions are included in the ablation study. The green line indicates the mean of the sampled predictions. In masked regions, the predicted series closely match the hidden in-situ data, demonstrating accurate and reliable GPP imputation. The green shaded region represents the 5%-95% confidence interval across the imputed time series. The variance from the predictions can be leveraged by the uncertainty-aware tuning module to improve the usability of the probabilistic labels. The patterns remain fairly stable with varying proportions of missing data, potentially benefiting from auxiliary environment inputs.

Ablation Study. Table 5 provides the ablation study results. The proposed DERE model achieves the best top-2 count across all variables, with stepwise improvements from the baseline to KGML, then KGML with PFT, and finally DERE. Extending KGML with high-level remote sensing labels helps significantly reduce

Table 5: Ablation study (top results in bold; runner-ups with underlines).

	Methods	ABoVE		AmeriFlux		FLUXNET		ICOS-WW		Multiple		Top-2
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	Count
GPP	Baseline	0.502	0.684	1.030	1.405	0.850	1.147	0.955	1.235	0.736	0.997	0/10
	KGML	0.352	0.541	**0.651**	**1.033**	0.699	0.951	0.711	1.014	0.636	0.919	2/10
	KGML + indirect labels	0.321	0.536	0.677	1.084	0.619	**0.850**	0.685	**0.888**	**0.573**	**0.888**	8/10
	DERE	**0.302**	**0.485**	0.663	1.068	**0.598**	0.863	**0.682**	0.888	0.585	0.890	**10/10**
RECO	Baseline	0.602	0.857	0.804	1.040	0.559	0.743	0.698	0.886	0.488	0.709	0/10
	KGML	0.328	**0.467**	0.526	0.705	0.438	0.583	0.586	0.756	0.466	0.655	3/10
	KGML + indirect labels	0.334	0.526	0.526	0.721	0.409	0.567	0.508	0.639	**0.402**	**0.596**	7/10
	DERE	**0.320**	0.472	**0.465**	**0.671**	**0.393**	**0.557**	**0.502**	**0.639**	0.442	0.627	**10/10**
NEE	Baseline	0.413	0.500	0.615	0.875	0.575	0.790	**0.489**	**0.665**	0.532	0.726	2/10
	KGML	0.177	0.244	0.507	0.733	0.504	0.722	0.511	0.721	0.489	**0.691**	2/10
	KGML + indirect labels	0.179	**0.242**	**0.475**	0.723	**0.471**	**0.677**	0.505	0.696	0.482	0.698	7/10
	DERE	**0.175**	0.244	0.496	**0.713**	0.478	0.683	0.505	0.696	**0.476**	0.692	**9/10**

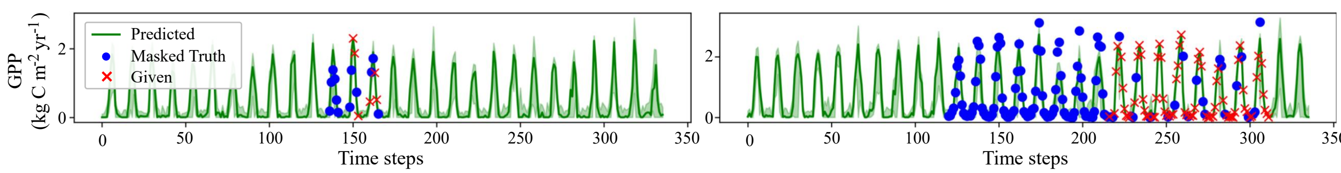

Figure 7: Probabilistic label expansion examples. The green shaded regions represent sample-based 5%-95% confidence interval.

the errors. The full DERE with probabilistic labels and uncertainty-aware tuning yielded the best overall performance.

6 Limitations and Ethical Considerations

First, on data privacy/ethical/bias considerations: all the datasets in the study from publicly available sources, including in-situ data (e.g., ICLR'24 CarbonSense [17]), process-based model data (both inputs conditions and outputs from NeurIPS'25 CarbonGlobe [62], which is our prior work focusing on physical model emulation without considering observations), and remote sensing PFT labels from ESA CCI [1, 21], with detailed documentation. The code is available on GitHub at https://github.com/ai-spatial/DERE. The in-situ dataset has a denser distribution in US and Europe, due to the flux tower availability, and has reduced coverage in regions such as Africa, where the results could be less representative. Second, on the other limitations, we have not considered scenario-based evaluations, e.g., for different climate zones and forest types. The prediction task primarily focuses on estimation (i.e., filling in gaps not covered by in-situ sensors) and has not considered future forecasts.

7 Conclusion

We presented a knowledge-guided learning framework with a decomposition-and-resembling approach that bridges process-based models and learning models to leverage high-level, indirect labels from remote sensing platforms that are available at large scale. We further developed a probabilistic label expansion module to extend the temporal coverage of sparse limited observations with explicit uncertainty awareness. Extensive experiments on global carbon monitoring datasets, including the CarbonSense benchmark, demonstrate that our proposed DERE-based knowledge-guided learning framework can effectively improve prediction quality compared with existing methods.

Future work will explore extensions to cover prediction tasks such as future forecasting, scenario-based evaluations under different climate zones and forest types, and methods tackling anomalous or rare conditions.

GenAI Disclosure

GenAI tools were used only to seek suggestions for minor language refinement (e.g., word/phrase choices) for a minimal portion of the paper (i.e., fewer than 10 sentences).

Acknowledgments

Shuo Xu, Zhihao Wang, Ruohan Li, Ruichen Wang, and Yiqun Xie are supported in part by the NSF under Grant No. 2126474, 2147195, 2425844, and 2530610; NASA under grant 80NSSC25K0013 and 80NSSC25K7221; Google's AI for Social Good Impact Scholars program; and the Zaratan cluster at the University of Maryland. George Hurtt is supported by NASA under Grant No. 80NSSC25K0013, 80NSSC25K7221 and 80NSSC22K1733. Lei Ma is supported by the NASA under Grant No. 80NSSC25K0013 and 80NSSC24K1632. Xiaowei Jia was partially supported by the NSF under Grant No. 2239175, 2147195, 2316305, 2425845, 2530609, 2203581; NASA under Grant No. 80NSSC24K1061 and 80NSSC25K0013; the USGS awards G21AC10564 and G22AC00266; and Pitt Momentum Funds and CRC at the University of Pittsburgh. We would like to acknowledge high-performance computing support from the Derecho system (doi:10.5065/qx9a-pg09) provided by the NSF National Center for Atmospheric Research (NCAR), sponsored by the National Science Foundation.

References

- [1] Centre for Environmental Data Analysis (CEDA) [n. d.]. *ESA CCI Land Cover: Global Plant Functional Types (PFT) v2.0.8 Data Access.* Centre for Environmental Data Analysis (CEDA). https://data.ceda.ac.uk/neodc/esacci/land_cover/data/pft/v2.0.8/
- [2] 2022. *Reducing Greenhouse Gas Emissions in Maryland: A Progress Report.* Technical Report. Maryland Department of the Environment. https://mde.maryland.gov/programs/air/ClimateChange/Documents/GGRA%20PROGRESSS%20REPORT%202022.pdf Accessed: 2024-12-19.
- [3] 2023. *Maryland Tree and Forest Carbon Flux: Data and Methodology Documentation.* Technical Report. Maryland Department of the Environment and Maryland Department of Natural Resources. https://mde.maryland.gov/programs/air/ClimateChange/Documents/VIMAL/MD_ForestCarbon_Flux_Methodology_01.06.23.pdf Accessed: 2024-12-19.
- [4] Brandon Anderson, Truong Son Hy, and Risi Kondor. 2019. Cormorant: Covariant molecular neural networks. In *Advances in Neural Information Processing Systems.* 14510–14519.
- [5] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. 2022. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications* 13, 1 (2022), 2453.
- [6] KE Bill, C Dieleman, JL Baltzer, GE Degre-timmons, MC Mack, SG Cumming, XJ Walker, and MR Turetsky. 2023. Arctic-Boreal Vulnerability Experiment (ABoVE). *ORNL Distributed Active Archive Center (DAAC) dataset 10.3334/ORNLDAAC/2235 (2023* (2023), 2235.
- [7] Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. 2023. Spherical fourier neural operators: Learning stable dynamics on the sphere. In *International conference on machine learning.* PMLR, 2806–2823.
- [8] Hui Chen, Viet Luong, Lopamudra Mukherjee, and Vikas Singh. 2025. SimpleTM: A Simple Baseline for Multivariate Time Series Forecasting. In *The Thirteenth International Conference on Learning Representations.*
- [9] Shengyu Chen, Nasrin Kalanat, Simon Topp, Jeffrey Sadler, Yiqun Xie, Zhe Jiang, and Xiaowei Jia. 2023. Meta-transfer-learning for time series data with extreme events: An application to water temperature prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management.* 266–275.
- [10] 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.
- [11] A Daw, RQ Thomas, CC Carey, JS Read, AP Appling, and A Karpatne. 2019. Physics-Guided Architecture (PGA) of Neural Networks for Quantifying Uncertainty in Lake Temperature Modeling. *arXiv:1911.02682* (2019).
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR* (2021).
- [14] Dapeng Feng, Jiangtao Liu, Kathryn Lawson, and Chaopeng Shen. 2022. Differentiable, learnable, regionalized process-based models with multiphysical outputs can approach state-of-the-art hydrologic prediction accuracy. *Water Resources Research* 58, 10 (2022), e2022WR032404.
- [15] Ferdinando Fioretto, Terrence WK Mak, and Pascal Van Hentenryck. 2020. Predicting AC optimal power flows: Combining deep learning and lagrangian dual methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 630–637.
- [16] Urban Forssell and Peter Lindskog. 1997. Combining Semi-Physical and Neural Network Modeling: An Example ofIts Usefulness. *IFAC Proceedings Volumes* 30, 11 (1997), 767–770.
- [17] Matthew Fortier, Mats Leon Richter, Oliver Sonnentag, and Christopher Pal. 2025. CarbonSense: A Multimodal Dataset and Baseline for Carbon Flux Modelling. In *The Thirteenth International Conference on Learning Representations.*
- [18] Pierre Friedlingstein, Michael O'sullivan, Matthew W Jones, Robbie M Andrew, Dorothee CE Bakker, Judith Hauck, Peter Landschützer, Corinne Le Quéré, Ingrid T Luijkx, Glen P Peters, et al. 2023. Global carbon budget 2023. *Earth System Science Data* 15, 12 (2023), 5301–5369.
- [19] Pierre Friedlingstein, Michael O'sullivan, Matthew W Jones, Robbie M Andrew, Judith Hauck, Peter Landschützer, Corinne Le Quéré, Hongmei Li, Ingrid T Luijkx, Are Olsen, et al. 2024. Global carbon budget 2024. *Earth System Science Data Discussions* 2024 (2024), 1–133.
- [20] Yoo-Geun Ham, Jeong-Hwan Kim, and Jing-Jia Luo. 2019. Deep learning for multi-year ENSO forecasts. *Nature* 573, 7775 (2019), 568–572.
- [21] KL Harper, C Lamarche, A Hartley, P Peylin, C Ottlé, V Bastrikov, RS Martín, SI Bohnenstengel, G Kirches, M Boettcher, et al. 2023. ESA land cover climate change initiative (Land_Cover_ cci): Global plant functional types (PFT).[Dataset], v2. 0.8. *NERC EDS Centre for Environmental Data Analysis* (2023).
- [22] K. L. Harper, C. Lamarche, A. Hartley, et al. 2023. A 29-year time series of annual 300 m resolution plant-functional-type maps for climate models. *Earth System Science Data* 15 (2023), 1465–1499. doi:10.5194/essd-15-1465-2023
- [23] Erhu He, Yiqun Xie, Licheng Liu, Weiye Chen, Zhenong Jin, and Xiaowei Jia. 2023. Physics guided neural networks for time-aware fairness: an application in crop yield prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 14223–14231.
- [24] Kethmi Hirushini Hettige, Jiahao Ji, Shili Xiang, Cheng Long, Gao Cong, and Jingyuan Wang. 2024. Airphynet: Harnessing physics-guided neural networks for air quality prediction. *arXiv preprint arXiv:2402.03784* (2024).
- [25] David Menéndez Hurtado, Karolis Uziela, and Arne Elofsson. 2018. Deep transfer learning in the assessment of the quality of protein models. *arXiv preprint arXiv:1804.06281* (2018).
- [26] 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.
- [27] 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.
- [28] Xiaowei Jia, Jared Willard, Anuj Karpatne, Jordan S Read, Jacob A Zwart, Michael Steinbach, and Vipin Kumar. 2020. Physics-guided machine learning for scientific discovery: An application in simulating lake temperature profiles. *arXiv preprint arXiv:2001.11086* (2020).
- [29] 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.
- [30] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. 2021. Physics-informed machine learning. *Nature Reviews Physics* 3, 6 (2021), 422–440.
- [31] A Karpatne, W Watkins, J Read, and V Kumar. 2017. Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling. *arXiv:1710.11431* (2017).
- [32] Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. 2023. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the platform for advanced scientific computing conference.* 1–11.
- [33] 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.
- [34] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. 2023. Learning skillful medium-range global weather forecasting. *Science* 382, 6677 (2023), 1416–1421.
- [35] Ruohan Li, Yiqun Xie, Zhihao Wang, Xiaowei Jia, Gengchen Mai, Lei Ma, George C. Hurtt, Quan Shen, and Zhili Li. 2026. EcoDiffusion: Uncertainty-Aware Emulation of Ecosystem Processes with Conditional Diffusion for Long Sequences with Single-Step Initialization. In *Proceedings of the AAAI Conference on Artificial Intelligence.*
- [36] Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. 2024. Physics-informed neural operator for learning partial differential equations. *ACM/IMS Journal of Data Science* 1, 3 (2024), 1–27.
- [37] J Ling, A Kurzawski, and J Templeton. 2016. Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. *J. Fluid Mech* (2016).
- [38] Licheng Liu, Wang Zhou, Kaiyu Guan, Bin Peng, Shaoming Xu, Jinyun Tang, Qing Zhu, Jessica Till, Xiaowei Jia, Chongya Jiang, et al. 2024. Knowledge-guided machine learning can improve carbon cycle quantification in agroecosystems. *Nature communications* 15, 1 (2024), 357.
- [39] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. 2023. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625* (2023).
- [40] 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.
- [41] Michael McCabe, Bruno Régaldo-Saint Blancard, Liam Holden Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Geraud Krawezik, Francois Lanusse, et al. 2023. Multiple physics pretraining for physical surrogate models. *arXiv preprint arXiv:2310.02994* (2023).
- [42] Griffin Mooers, Michael Pritchard, Tom Beucler, Jordan Ott, Galen Yacalis, Pierre Baldi, and Pierre Gentine. 2021. Assessing the potential of deep learning for

emulating cloud superparameterization in climate models with real-geography boundary conditions. *Journal of Advances in Modeling Earth Systems* 13, 5 (2021), e2020MS002385.
[43] 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.
[44] Nikhil Muralidhar, Jie Bu, Ze Cao, Long He, Naren Ramakrishnan, Danesh Tafti, and Anuj Karpatne. 2020. PhyNet: Physics Guided Neural Networks for Particle Drag Force Prediction in Assembly. In *Proceedings of the 2020 SIAM International Conference on Data Mining*. SIAM, 559–567.
[45] N Muralidhar, MR Islam, M Marwah, A Karpatne, and N Ramakrishnan. 2018. Incorporating prior domain knowledge into deep neural networks. In *IEEE Big Data*. IEEE.
[46] Octavi Obiols-Sales, Abhinav Vishnu, Nicholas Malaya, and Aparna Chandramowliswharan. 2020. CFDNet: A deep learning-based accelerator for fluid simulations. In *Proceedings of the 34th ACM international conference on supercomputing*. 1–12.
[47] Roberto Paolucci, Filippo Gatti, Maria Infantino, Chiara Smerzini, Ali Güney Özcebe, and Marco Stupazzini. 2018. Broadband Ground Motions from 3D Physics-Based Numerical Simulations Using Artificial Neural NetworksBroadband Ground Motions from 3D PBSs Using ANNs. *Bulletin of the Seismological Society of America* 108, 3A (2018), 1272–1286.
[48] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics* 378 (2019), 686–707.
[49] Stephan Rasp, Michael S Pritchard, and Pierre Gentine. 2018. Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences* 115, 39 (2018), 9684–9689.
[50] Jordan S Read, Xiaowei Jia, Jared Willard, Alison P Appling, Jacob A Zwart, Samantha K Oliver, Anuj Karpatne, Gretchen JA Hansen, Paul C Hanson, William Watkins, et al. 2019. Process-guided deep learning predictions of lake water temperature. *Water Resources Research* 55, 11 (2019), 9173–9190.
[51] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. 2021. E (n) equivariant graph neural networks. In *International conference on machine learning*. PMLR, 9323–9332.
[52] Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. 2017. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in neural information processing systems*. 991–1001.
[53] Chaopeng Shen, Alison P Appling, Pierre Gentine, Toshiyuki Bandai, Hoshin Gupta, Alexandre Tartakovsky, Marco Baity-Jesi, Fabrizio Fenicia, Daniel Kifer, Li Li, et al. 2023. Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth & Environment* 4, 8 (2023), 552–567.
[54] Justin Sirignano, Jonathan F MacArt, and Jonathan B Freund. 2020. DPM: A deep learning PDE augmentation method with application to large-eddy simulation. *J. Comput. Phys.* 423 (2020), 109811.
[55] Russell Stewart and Stefano Ermon. 2017. Label-free supervision of neural networks with physics and domain knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.
[56] Mohammad M Sultan, Hannah K Wayment-Steele, and Vijay S Pande. 2018. Transferable neural networks for enhanced sampling of protein dynamics. *Journal of chemical theory and computation* 14, 4 (2018), 1887–1894.
[57] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems* 34 (2021), 24804–24816.
[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[59] ZY Wan, P Vlachas, P Koumoutsakos, and T Sapsis. 2018. Data-assisted reduced-order modeling of extreme events in complex dynamical systems. *PloS one* (2018).
[60] Xin Wang, Yilun Han, Wei Xue, Guangwen Yang, and Guang J Zhang. 2022. Stable climate simulations using a realistic general circulation model with neural network parameterizations for atmospheric moist physics and radiation processes. *Geoscientific Model Development* 15, 9 (2022), 3923–3940.
[61] Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. 2024. Timexer: Empowering transformers for time series forecasting with exogenous variables. *Advances in Neural Information Processing Systems* 37 (2024), 469–498.
[62] Zhihao Wang, Lei Ma, George Hurtt, Xiaowei Jia, Yanhua Li, Ruohan Li, Zhili Li, Shuo Xu, and Yiqun Xie. 2025. CarbonGlobe: A Global-Scale, Multi-Decade Dataset and Benchmark for Carbon Forecasting in Forest Ecosystems. *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track* (2025).
[63] 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 *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*. 1–10.
[64] Zhihao Wang, Yiqun Xie, Zhili Li, Xiaowei Jia, Zhe Jiang, Aolin Jia, and Shuo Xu. 2024. SimFair: Physics-guided fairness-aware learning with simulation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22420–22428.
[65] Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. 2022. Integrating scientific knowledge with machine learning for engineering and environmental systems. *Comput. Surveys* 55, 4 (2022), 1–37.
[66] Lei Xu, Xihao Zhang, Hongchu Yu, Zeqiang Chen, Wenying Du, and Nengcheng Chen. 2024. Incorporating spatial autocorrelation into deformable ConvLSTM for hourly precipitation forecasting. *Computers & Geosciences* 184 (2024), 105536.
[67] Tianfang Xu and Albert J Valocchi. 2015. Data-driven methods to improve baseflow prediction of a regional groundwater model. *Computers & Geosciences* 85 (2015), 124–136.
[68] Kun Yao, John E Herr, David W Toth, Ryker Mckintyre, and John Parkhill. 2018. The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chemical science* 9, 8 (2018), 2261–2269.
[69] Runlong Yu, Chonghao Qiu, Robert Ladwig, Paul Hanson, Yiqun Xie, and Xiaowei Jia. 2025. Physics-guided foundation model for scientific discovery: An application to aquatic science. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 28548–28556.
[70] Runlong Yu, Chonghao Qiu, Robert Ladwig, Paul C Hanson, Yiqun Xie, Yanhua Li, and Xiaowei Jia. 2024. Adaptive process-guided learning: An application in predicting lake do concentrations. In *2024 IEEE International Conference on Data Mining (ICDM)*. IEEE, 580–589.
[71] Sungduk Yu, Walter Hannah, Liran Peng, Jerry Lin, Mohamed Aziz Bhouri, Ritwik Gupta, Björn Lütjens, Justus C Will, Gunnar Behrens, Julius Busecke, et al. 2024. ClimSim: A large multi-scale dataset for hybrid physics-ML climate emulation. *Advances in Neural Information Processing Systems* 36 (2024).
[72] Linfeng Zhang, Jiequn Han, Han Wang, Wissam Saidi, Roberto Car, and E Weinan. 2018. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. In *Advances in Neural Information Processing Systems*. 4436–4446.
[73] Yunhao Zhang and Junchi Yan. 2023. Crossformer: Transformer Utilizing Cross-Dimension Dependency for Multivariate Time Series Forecasting. In *International Conference on Learning Representations*.
[74] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 11106–11115.
[75] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*. PMLR, 27268–27286.

Appendix: Additional Results

Comparison with non-sequential settings. To verify the necessity of temporal modeling, we further include a non-sequential per-timestep baseline with vanilla fully-connected networks, as each point becomes a vector to vector mapping task. In other words, this becomes a pure conditional regression problem that maps environmental inputs to carbon flux independently at each time step without historical context. This per-timestep model is included only as a diagnostic non-sequential setting to verify the need for temporal modeling, rather than as a main competing baseline. As shown in Table 6, the per-timestep model has a lower performance compared to the Transformer-based baseline in most cases (e.g., the MAE increased from 0.502 to 1.114 for GPP on ABoVE). The time-series set-up is more suitable for GPP and RECO because their values depend not only on the current environmental inputs, but also on previous ecosystem states, such as accumulated vegetation growth and seasonal development. In contrast, the per-timestep MLP removes this history and therefore mainly tests how much carbon flux can be predicted from instantaneous conditions alone.

Backbone Generalization. DERE is a modular framework and is flexible with other backbones. Given Informer's better baseline performance in some cases over Transformer, we tried it with DERE over all the datasets. As shown in Table 7, integrating DERE consistently improves performance over the Informer baseline in

Table 6: MLP baseline results.

Variables	ABoVE		AmeriFlux		FLUXNET		ICOS-WW		Multiple	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
GPP	1.114	1.169	1.071	1.201	0.975	1.193	0.971	1.064	1.055	1.173
RECO	0.970	1.031	0.825	0.935	0.774	0.980	0.725	0.817	0.749	0.858
NEE	0.269	0.332	0.335	0.414	0.313	0.398	0.342	0.407	0.369	0.453

most cases (e.g., 0.17 improvement in MAE for GPP for the ABoVE dataset). These results confirm that the performance gains provided by DERE are not limited to the Transformer structure. Given the potential variability in optimal backbone for different types of physical conditions, as suggested by the results on different in-situ observation networks, future work can explore adaptive selection of backbone choices for different scenarios. DERE is designed as a general framework and can be used with different network architectures.

Table 7: DERE result with Informer Backbone.

Variables	ABoVE		AmeriFlux		FLUXNET		ICOS-WW		Multiple	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
GPP	0.351	0.521	0.624	0.992	0.702	0.992	0.618	0.844	0.550	0.870
RECO	0.360	0.499	0.438	0.632	0.492	0.657	0.498	0.690	0.408	0.602
NEE	0.184	0.246	0.445	0.674	0.576	0.818	0.354	0.451	0.471	0.688

Statistical Significance. We conducted statistical significance analyses between DERE and the runner-up method (KGML Informer). The t-test over the entire dataset shows that DERE achieves statistically significant improvements across all variables at the 0.01 significance level.