RESEARCH-ARTICLE

Towards Surrogate Models with Hybrid Spatial Neural Networks: A Summary of Results

SHENGYA ZHANG, University of Minnesota Twin Cities, Minneapolis, MN, United States

ARUN SHARMA, University of Minnesota Twin Cities, Minneapolis, MN, United States

MAJID FARHADLOO, University of Minnesota Twin Cities, Minneapolis, MN, United States

MINGZHOU YANG, University of Minnesota Twin Cities, Minneapolis, MN, United States

RUOLEI ZENG, University of Minnesota Twin Cities, Minneapolis, MN, United States

SUBHANKAR GHOSH, University of Minnesota Twin Cities, Minneapolis, MN, United States

View all

Open Access Support provided by:

University of Minnesota Twin Cities

Colorado State University

Published: 03 November 2025

GeoSIM '25: 8th ACM SIGSPATIAL International Workshop on Geospatial Simulation
November 3 - 6, 2025
MN, Minneapolis, USA

Conference Sponsors:
SIGSPATIAL

https://doi.org/10.1145/3764921.3770153
ISBN: 9798400721847

Towards Surrogate Models with Hybrid Spatial Neural Networks: A Summary of Results

Shengya Zhang
University of Minnesota
Minneapolis, USA
zhan9051@umn.edu

Arun Sharma
University of Minnesota
Minneapolis, USA
arunshar@umn.edu

Majid Farhadloo
University of Minnesota
Minneapolis, USA
farha043@umn.edu

Mingzhou Yang
University of Minnesota
Minneapolis, USA
yang7492@umn.edu

Ruolei Zeng
University of Minnesota
Minneapolis, USA
zeng0208@umn.edu

Subhankar Ghosh
University of Minnesota
Minneapolis, USA
ghosh117@umn.edu

Yao Zhang
Colorado State University
Colorado, USA
Yao.Zhang@colostate.edu

Mu Hong
Colorado State University
Colorado, USA
Mu.Hong@colostate.edu

Licheng Liu
University of Minnesota
Minneapolis, USA
lichengl@umn.edu

David Mulla
University of Minnesota
Minneapolis, USA
mulla003@umn.edu

Shashi Shekhar
University of Minnesota
Minneapolis, USA
shekhar@umn.edu

Abstract

The goal is to develop an efficient and accurate surrogate model for Daycent, a widely used but computationally expensive ecosystem model. This problem is important due to its societal applications in sustainable agriculture. Challenges include balancing the trade-off between prediction time and solution quality (e.g., accuracy), as well as the need to capture spatial relationships both within and across sites, while also accounting for varied crop management practices that introduce irregular and non-stationary patterns, reducing predictability. Related work on surrogate models with traditional feed-forward artificial neural networks (SM-ANN) has shown that these models have limited accuracy and often fail to capture spatial dependencies. To address these limitations, we explore novel Surrogate Models with Hybrid Spatial Neural Networks (SM-Hybrid) capable of explicitly modeling spatial autocorrelation and tele-connections. Experimental results show that the proposed SM-Hybrid is more accurate than SM-ANN and is twice as fast as the Daycent model.

CCS Concepts

• **Information systems** → Data mining; • **Computing methodologies** → Machine Learning; • **Applied Computing** → Environmental Sciences; • **Computing Methodologies** → Modeling and Simulation.

GeoSIM '25, November 3–6, 2025, Minneapolis, MN, USA

ACM ISBN 979-8-4007-2184-7/2025/11…$15.00
https://doi.org/10.1145/3764921.3770153

Keywords

Surrogate Modeling; Spatial Neural Network; Spatial Autocorrelation; Spatial Teleconnection; Sustainable Agriculture; Daycent Model

ACM Reference Format:
Shengya Zhang, Arun Sharma, Majid Farhadloo, Mingzhou Yang, Ruolei Zeng, Subhankar Ghosh, Yao Zhang, Mu Hong, Licheng Liu, David Mulla, and Shashi Shekhar. 2025. Towards Surrogate Models with Hybrid Spatial Neural Networks: A Summary of Results. In *The 8th ACM SIGSPATIAL International Workshop on Geospatial Simulation (GeoSIM '25), November 3–6, 2025, Minneapolis, MN, USA*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3764921.3770153

1 Introduction

Given a widely used process-based ecosystem model (e.g., DayCent [38]) that generates accurate yet computationally demanding outputs, the goal is to construct a more efficient yet sufficiently accurate *deep neural network*-based surrogate model. By integrating complementary models (e.g., hybrid architectures) and learning from sparse and spatially heterogeneous data, the surrogate model maintains acceptable predictive accuracy (e.g., above 90%) while significantly improving efficiency. The Daycent model takes inputs such as daily weather data (e.g., temperature, precipitation), soil texture, land management practices, and initial soil conditions, and outputs variables including soil organic carbon, nitrate levels, and greenhouse gas (GHG) fluxes (e.g., N_2O, CO_2, CH_4) [12]. Its sub-models for plant growth, soil organic matter decomposition, and other biogeochemical processes enable assessments of land-use change, agricultural management, and climate impacts on soil carbon stocks and GHG emissions.

Surrogate models are simplified statistical or machine learning approximations that emulate the input–output behavior of computationally expensive process-based models such as DayCent. They are widely used in simulation and modeling to reduce runtime while retaining sufficient predictive fidelity [18**?**]. In the context of DayCent, surrogate models can provide rapid approximations of soil organic carbon dynamics, greenhouse gas emissions, and crop yield responses, enabling large-scale scenario testing and decision support that would otherwise be computationally prohibitive.

The surrogate modeling problem is important because soil-based GHG emissions contribute significantly to climate change, and their mitigation plays a critical role in achieving environmental sustainability. For example, agriculture and land-use change account for approximately a quarter of total global GHG emissions [29]. Carbon farming practices, such as regenerative agriculture and soil conservation techniques, play a key role in sequestering soil carbon, reducing GHG emissions, and enhancing soil health. Given these efforts, it is essential to develop reliable and scalable methods for accurately quantifying carbon sequestration and GHG emissions at the field level [33]. This will enable a precise evaluation of management and climate impact, and help ensure that farmers receive fair compensation for their contributions to carbon mitigation.

The problem is challenging for several reasons. First, balancing the trade-off between computational efficiency and solution quality (e.g., acceptable accuracy) remains a significant difficulty for model development. Additional challenges arise from data sparsity in both training and validation, as measurements are often unevenly distributed across the input domain. For example, in the U.S. Midwest, observations of soil organic carbon (SOC) tend to cluster in specific locations, leaving other areas relatively under-sampled [28]. Moreover, capturing long-term temporal dependencies—from fine-grained intervals to multi-year trends—is essential for robust predictions. Finally, agricultural landscapes are further complicated by human interventions such as crop rotation, fallow cycles, and selective planting, which introduce irregular and non-stationary dynamics that reduce model predictability.

Previous surrogate modeling efforts have primarily relied on shallow artificial neural networks (ANNs) to approximate process-based models, such as DayCent [36]. However, these approaches often exhibit limited predictive accuracy. Their main shortcomings lie in neglecting spatial dependencies (e.g., autocorrelation and tele-connections) and being developed within narrowly defined geographic areas, which restricts generalization to regions with higher spatial heterogeneity or SOC and GHG emissions.

To address these limitations, we investigate Surrogate Models with Hybrid Spatial Neural Networks (SM-Hybrid), designed to capture both local spatial patterns and long-range dependencies explicitly. As illustrated in Figure 1, SM-Hybrid balances computational efficiency with predictive fidelity by combining two complementary components: a convolutional neural network (CNN) for localized feature extraction and a transformer for broader spatial interactions via attention mechanisms. Compared to conventional surrogate models based on feed-forward ANNs (SM-ANN) and the process-based DayCent model, SM-Hybrid achieves improved accuracy at a fraction of the computational cost, offering a flexible and scalable alternative for spatially explicit ecological applications.

Our contributions are as follows:

- This paper proposes Surrogate Models with Hybrid Spatial Neural Networks (SM-Hybrid) that explicitly capture spatial autocorrelation (e.g., using CNNs) and tele-connections (e.g., using Transformers).
- We conduct a comparative study with well-known neural network architectures, including standard artificial neural networks (SM-ANN) and vision transformers (SM-ViT), based on different partition strategies accounting for spatial heterogeneity.
- Experimental findings indicate that SM-Hybrid often achieves higher accuracy than SM-ANN and, in our tested settings, can be up to two orders of magnitude faster than the process-based Daycent model, depending on the dataset and task configuration.

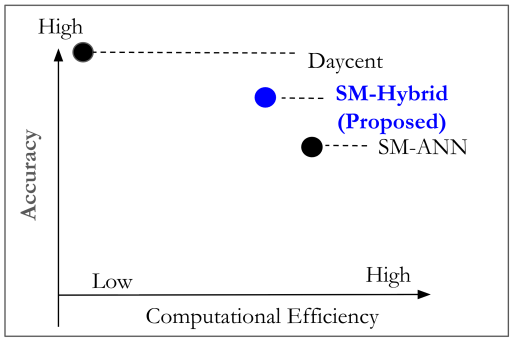


Figure 1: Limitation of Related Work

Scope: This paper presents a preliminary study demonstrating the potential of Surrogate Models with Hybrid Spatial Neural Networks (SM-Hybrid) for accelerating DayCent simulations. We restrict our contribution in this work to *Step 1* of a larger investigation framework (described in Appendix 5.4), focusing on comparing candidate surrogate models (CNN, Transformer, and Hybrid). Broader considerations such as scalability, interpretability, and end-to-end deployment are beyond the current scope. We also do not explore advanced data fusion (e.g., integrating remote sensing and ground-based observations), transfer learning, or domain adaptation for cross-ecosystem generalization. Finally, this study only evaluates shallow CNNs with two convolutional layers and does not consider deeper residual CNNs or attention-enhanced backbones tailored for geo-spatial textures.

GeoSim Workshop Relevance: This work is relevant to the GeoSim workshop because it does the following:

- It extends the process-based DayCent model with surrogate approaches to accelerate geospatial simulation of soil carbon and nitrogen dynamics, greenhouse gas (GHG) fluxes, and crop yields.
- This work leverages deep learning architectures (CNNs, Transformers, and hybrids) to capture spatial dependencies such as autocorrelation and tele-connections, consistent with GeoSim's emphasis on spatially explicit simulation.
- This work supports large-scale experiments and interactive decision support by approximating DayCent with models that are orders of magnitude faster.
- The paper highlights variability across soil textures, management practices, and climate conditions, aligning with GeoSim's focus on land use and environmental processes.

- We also discuss trade-offs between computational efficiency and predictive fidelity, with opportunities for future refinement through data fusion and domain adaptation.

Organization: Section 2 describes the application domain. Section 3 provides background on the Daycent model and formally defines the problem. Section 4 details the proposed approach and design decisions. Experiments are presented in Section 5, and Section 6 details related work. Finally, Section 7 concludes the paper with future directions.

2 Illustrative Application Domains

2.1 COMET-Planner and DayCent

COMET-Planner [17] is a tool developed by Colorado State University and USDA NRCS that provides region-specific estimates of GHG reduction potential for conservation practices. It builds on the DayCent model, which simulates soil and plant processes with field-calibrated accuracy but is computationally intensive. To remain practical, COMET-Planner uses precomputed DayCent outputs organized into lookup tables, enabling fast access but limiting flexibility when applied outside the calibration context.

Surrogate models offer a more adaptive alternative by approximating DayCent outputs with machine learning trained on simulation data. Once trained, they provide predictions in seconds, accept broader input ranges, and scale to diverse soil, management, and climate conditions. This supports more site-specific and data-driven decision-making than fixed lookup tables.

3 Problem Formulation

3.1 Background

1.Daycent Model: Process-based bio geochemical models like Daycent have become essential tools for understanding carbon and nitrogen dynamics in ecosystems. In the Daycent model, a **pool** is a conceptual compartment of organic matter (either plant litter or soil organic matter) categorized by its decomposition rate. The Daycent framework uses a hierarchical data structure to track carbon and nitrogen in ecosystems. The model divides organic matter into ***five*** main pools, ***two*** for plant litter and ***three*** for soil organic matter, each with decomposition rates ranging from months to centuries. The model's multi-pool design captures how organic matter breaks down at different speeds: some *decompose* ***quickly (active pool)***, some ***very slowly (passive pool)***, and some at ***intermediate rates (slow pool)***. The model also simulates water movement through soil layers and tracks nitrogen transformations, including losses through various pathways. Daycent operates on daily time steps to simulate plant growth, organic matter decomposition, soil-water and temperature regimes, and nutrient cycling processes across multiple soil layers. Its high computational cost is due to the complex process coupling, where multiple submodels (plant growth, soil-organic matter decomposition, water movement, and temperature dynamics) interact simultaneously and require iterative solutions at each time step. The total computational demand of the model can be conceptualized as: ***High temporal resolution × Complex processes × Multi-layer soil × Large spatial scale × Long time series = Enormous computational load***, where each factor increases the overall requirements.

This computational intensity often limits DayCent's use in large-scale or time-sensitive applications such as COMET-Planner. To overcome this, researchers have developed surrogate models that approximate DayCent's outputs with much lower resource requirements. Although DayCent provides high accuracy, its cost makes it less practical for many scenarios. However, that same accuracy makes it a strong benchmark for training surrogates. The output of DayCent can be used to teach neural networks the key relationships between environmental input and ecosystem responses.

The proposed SM-Hybrid addresses this gap. It reduces runtime substantially while maintaining predictive accuracy and enables fast parameter exploration, uncertainty analysis, and real-time decision support that would be infeasible with the original model.

2. Conceptual Overview of the Daycent Model Diagram: A conceptual representation of Daycent, shown in Figure 2, illustrates the flow of carbon and nitrogen through key ecosystem compartments and their interactions with environmental factors.

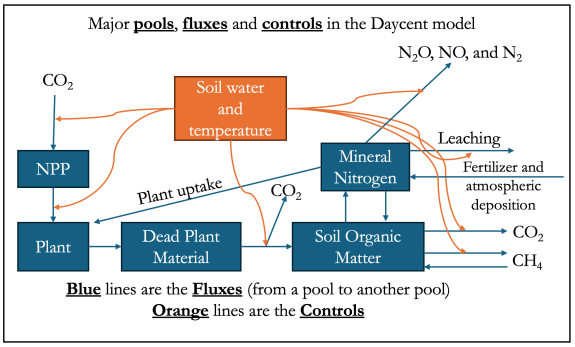


Figure 2: Conceptual Diagram of Daycent model

Inputs: Environmental drivers include weather, soil properties, vegetation, and management practices. *Outputs:* Daily estimates of CO_2, CH_4, and nitrogen gases (e.g., N_2O, NO, N_2), leaching of mineral nitrogen, soil organic carbon stocks, and crop yields, providing a dynamic view of ecosystem processes.

The model represents three primary compartments: **atmosphere** (greenhouse gas emissions from soil processes), **vegetation** (plant components that assimilate carbon and uptake nitrogen), and **soil** (water and temperature dynamics, soil organic matter pools, and dead plant material). Key fluxes include **carbon**, where NPP drives CO_2 uptake and decomposition releases carbon back to the atmosphere, and **nitrogen**, where uptake and precipitation inputs are offset by denitrification, nitrification, and leaching across soil layers.

Soil water and temperature dynamics have a direct influence on decomposition and nutrient cycling. Decomposition transfers C and N from dead-plant material to SOM pools—active (with a turnover time of 0.5–1 year), slow (10–50 years), and passive (1,000–5,000 years) — with soil type, vegetation type, and management modulating all processes.

3. Surrogate Models: Surrogate (or meta) models are lightweight statistical emulators that approximate high-fidelity numerical process-based models at a fraction of the computational cost. Whereas traditional NWP systems solve PDEs (e.g., Navier–Stokes, thermodynamics) on supercomputers for hours, surrogates learn input–output mappings from those simulations and return forecasts

in seconds [18]. Their speed enables real-time tasks, renewable energy scheduling, emergency response, and disaster management, where latency is critical. For example, Microsoft's *Aurora* delivers 10-day, high-resolution forecasts almost instantaneously, outperforming many conventional systems [51].

Beyond direct forecasting, surrogates integrate seamlessly with data assimilation. A FourCastNet graph-neural surrogate [2] was embedded in a variational framework and demonstrated skill preservation over a year, despite partial and noisy observations, illustrating how assimilation can stabilize otherwise drift-prone machine learning models. Domain-focused surrogates show similar gains: rapid flood-hazard estimates in an Andean basin [52], CNN-based regional wind-wave prediction [60], and storm-surge projections on evolving coastlines [30]. Recent studies [9, 41] argue that coupling ML with data assimilation is key to capturing both short and long-term climate dynamics while guarding against instability and data sparsity.

Table 1: Summary of Mathematical Notations and Symbols

Notation	Description
S	Training set $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N}$ generated with *Daycent*
$\mathbf{X}_i$	Input features for site i (soil, climate, management)
$\mathbf{Y}_i$	Daycent outputs for site i (SOC, GHG, yield)
f_θ	Spatial surrogate (neural network) with parameters θ
$\hat{\mathbf{Y}}_i$	Surrogate prediction for $\mathbf{X}_i$: $\hat{\mathbf{Y}}_i = f_\theta(\mathbf{X}_i)$
T	Number of time steps in the sequence
$\mathbf{X}_t$	Grid at time t, $\mathbf{X}_t \in \mathbb{R}^{H\times W\times C_0}$
H, W	Height and width of each spatial grid
C_0	Channels in the raw input grid
L	Number of convolutional blocks
k_h, k_w	Convolution kernel height and width
s_h, s_w	Convolution stride (height / width)
p_h, p_w	Zero-padding (height / width)
$Z_{i,j,c}^{(l)}$	Pre-activation at (i, j, c) in layer l with bias $b_c^{(l)}$
z_t	Global-average-pooled CNN feature at time t
e_t	Linear embedding of z_t into d_{model} dims
d_{model}	Transformer token (model) dimension
$PE(t, s)$	Positional encoding at time t, site s
α, β	Temporal / spatial scaling factors in PE
$\tilde{e}_t$	Token fed to Transformer: $e_t + PE(t, s)$
H_{head}	Number of self-attention heads
Q_h, K_h, V_h	Query, key, value matrices for head h
d_k	Query/key dimension per head
B_{space}	Learned spatial-bias matrix ($T \times T$) applied to attention logits
A_h	Normalised attention weights for head h
W^O	Output projection after concatenating heads
$W_{1,2}$, $b_{1,2}$	FFN linear weights / biases
d_{ff}	Hidden dimension in FFN
$h_t^{(n)}$	Hidden state after MHSA in layer n (superscript n indexes the Transformer layer)
$W_{\text{out}}, b_{\text{out}}$	Regression head weight / bias
d_{out}	Dimension of surrogate output vector
$\hat{y}$	Final surrogate prediction (e.g., SOC, GHG, yield)
s_0	Reference site used in spatial distance for $PE(t, s)$

3.2 Problem Statement

The goal is to develop a Surrogate Model with Spatial Neural Networks (SM-Hybrid) that efficiently approximates the outputs of a computationally expensive process-based model such as *Daycent*.

Input:

- A dataset $S = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N}$, where each sample comprises input $\mathbf{X}_i$ and output $\mathbf{Y}_i$ generated by Daycent.
- A set of candidate deep-learning architectures designed to capture spatial dependencies, including spatial autocorrelation and tele-connections.

Output:

- A surrogate model f_θ, trained on S, that approximates Daycent outputs.

Objective: Computational efficiency.

Constraints: Solution quality.

Semantic description of inputs and samples: A *site* denotes a unique field location (lat/long, elevation) associated with a soil map unit and management regime; one training *sample* corresponds to a site-scenario pair (rotation, fertilizer rate, and timing) with an input window of meteorology and management spanning the preceding T days (here, T=365 unless stated). Inputs include static features (soil texture, bulk density) and dynamic features (daily $T_{\max}/T_{\min}$, precipitation, radiation, planting/harvest operations). Outputs are Daycent-derived targets for that site–scenario, e.g., daily NEE and annual yield.

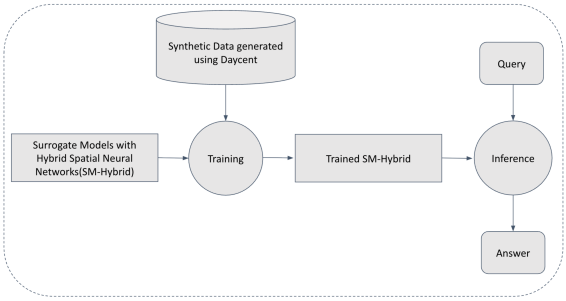


Figure 3: Conceptual Diagram of Accelerated Daycent approximations using SM-Hybrid

To address this problem, we consider four key elements (Figure 3). The dataset $S = \{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^{N}$ pairs inputs $\mathbf{X}_i$ (soil, climate, management practices) with Daycent outputs $\mathbf{Y}_i$ (SOC, GHG emissions, crop yield). Candidate architectures, including CNNs and Transformers, are used to capture spatial autocorrelation and tele-connections. The resulting surrogate f_θ maps $\mathbf{X}_i \mapsto \hat{\mathbf{Y}}_i$ with accuracy comparable to Daycent but at significantly lower cost.

This framework enables rapid, scalable inference for interactive planning, large-scale simulations, and decision support. For example, given a no-till corn–soybean rotation on silty-clay-loam soil in central Iowa under projected 2030 conditions, the surrogate can return SOC change (+0.15 t ha^{-1}), and expected corn yield (9.3 t ha^{-1}) in milliseconds, versus several seconds for Daycent. For context, annual SOC stock changes in row-crop systems are typically on the order of $\pm(0.1–0.5)$ t ha^{-1}, so +0.15 t ha^{-1} represents a modest but practically meaningful gain. Efficiency gains must not sacrifice accuracy; the surrogate must preserve Daycent's predictive quality to ensure reliable decisions.

4 Proposed Surrogate Model with Hybrid Spatial Neural Networks (SM-Hybrid)

We present a hybrid architecture that combines a CNN and a transformer for spatiotemporal modeling, leveraging the CNN for local feature extraction and the transformer for long-range dependencies.

4.1 Spatial Feature Extraction

The CNN module starts by processing data that represents a spatial map over time, much like a series of snapshots of a field. Each snapshot, denoted as $\mathbf{X}_t$, captures information such as soil carbon (C), moisture, or nitrogen (N) levels across different locations and soil depths (e.g., 0–1 cm, 1–4 cm). The CNN analyzes this grid by focusing on small patches at a time using a convolution process to identify patterns, such as areas with high soil moisture or SOC. This process involves sliding a small window (called a filter) over the grid, combining the data within the window to highlight important features. The filter learns to detect patterns through training, adjusting its weights to recognize key characteristics better.

After identifying these patterns, the CNN applies a ReLU function to emphasize the most significant features, while ignoring less important ones. Then, max-pooling simplifies the data by keeping only the most prominent features in each small region, reducing the grid's size while preserving its essence.

After several layers of this process, the CNN condenses the grid into a compact set of features, denoted as z_t, which summarizes the spatial patterns at each time step. This summary captures the local relationships in the data, such as how soil properties in one part of the field relate to those in nearby areas.

The CNN processes input tensors $\{\mathbf{X}_t\}_{t=1}^{T}$, where $\mathbf{X}_t \in \mathbb{R}^{H\times W\times C_0}$ is a spatial grid at time t.

For the l-th convolutional layer, with input $H^{(l-1)} \in \mathbb{R}^{H_{l-1}\times W_{l-1}\times C_{l-1}}$ (where $H^{(0)} = \mathbf{X}_t$), the output is

$$Z^{(l)}_{i,j,c} = \sum_{p=0}^{k_h-1}\sum_{q=0}^{k_w-1}\sum_{c'=0}^{C_{l-1}-1} W^{(l)}_{p,q,c',c} H^{(l-1)}_{i+p,j+q,c'} + b^{(l)}_c, \tag{1}$$

where $W^{(l)} \in \mathbb{R}^{k_h\times k_w\times C_{l-1}\times C_l}$ are filter weights and $b^{(l)}_c$ is the bias. Output dimensions are

$$H_l = \left\lfloor \frac{H_{l-1}+2p_h-k_h}{s_h} \right\rfloor + 1, \quad W_l = \left\lfloor \frac{W_{l-1}+2p_w-k_w}{s_w} \right\rfloor + 1. \tag{2}$$

ReLU activation is applied:

$$H^{(l)}_{i,j,c} = \max\bigl(0, Z^{(l)}_{i,j,c}\bigr). \tag{3}$$

Max-pooling reduces dimensions:

$$P^{(l)}_{i,j,c} = \max_{p=0}^{m_h-1}\max_{q=0}^{m_w-1} H^{(l)}_{s^p_h i+p,\, s^p_w j+q,\, c}. \tag{4}$$

After L layers, the feature map $H^{(L)} \in \mathbb{R}^{H_L\times W_L\times C_L}$ is globally averaged:

$$z_t = \frac{1}{H_L W_L}\sum_{i=1}^{H_L}\sum_{j=1}^{W_L} H^{(L)}_{i,j,:} \in \mathbb{R}^{C_L}. \tag{5}$$

Next, an embedding step adjusts the size of the features summarized by the CNN to match the requirements of the transformer, ensuring they are transformed into a format it can understand. The sequence $\{z_t\}_{t=1}^{T}$ is linearly embedded:

$$e_t = W_e z_t + b_e, \tag{6}$$

with $W_e \in \mathbb{R}^{C_L\times d_{\text{model}}}$ and $b_e \in \mathbb{R}^{d_{\text{model}}}$.

We adopt the same padding/stride across layers and apply global average pooling to stabilize gradients under variable spatial supports; residual connections are deferred to future ablations. Teleconnections refer to long-range spatial dependencies (e.g., the influence of soil moisture conditions in one region on nitrogen cycling in another), while spatial autocorrelation refers to local dependencies among nearby sites. These concepts are central to ecological simulation but are often neglected in prior surrogate modeling approaches.

4.2 Spatially Aware Positional Encoding

We capture topological relationships from the Daycent models (Figure 2) by employing a spatially-aware positional encoder. The encoder helps the transformer understand the topological interactions and provides it with special labels that indicate when and where each piece of data originates. For example, in Daycent, time embedding might indicate the specific day and depth of the soil layer where the measurement was taken. This is crucial because soil properties can vary significantly at different depths (e.g., 0-1 cm vs. 15-30 cm) or at different times (e.g., during a wet season vs. a dry season). By adding these labels, we ensure that the transformer can account for these factors, thereby improving its ability to model Daycent processes, such as N cycling or soil organic matter decomposition.

Positional encodings $PE(t,s) \in \mathbb{R}^{d_{\text{model}}}$ combine temporal and spatial terms:

$$PE(t,s)_{2i} = \sin\Bigl(\frac{t}{\alpha^{2i/d_{\text{model}}}}\Bigr) + \sin\Bigl(\frac{\text{dist}(s,s_0)}{\beta^{2i/d_{\text{model}}}}\Bigr), \tag{7}$$

$$PE(t,s)_{2i+1} = \cos\Bigl(\frac{t}{\alpha^{2i/d_{\text{model}}}}\Bigr) + \cos\Bigl(\frac{\text{dist}(s,s_0)}{\beta^{2i/d_{\text{model}}}}\Bigr), \tag{8}$$

where $\text{dist}(s,s_0)$ is the spatial distance from reference s_0, and α, β are scaling factors. The Transformer input is

$$\tilde{e}_t = e_t + PE(t,s_t). \tag{9}$$

4.3 Temporal and Spatial Modeling

The proposed spatially aware positional encoding also captures long-range relationships (e.g., how a rainy season might affect soil moisture months later, or how soil carbon in one area influences nearby regions) via self-attention. For example, it might notice that high rainfall in the spring leads to increased nitrogen gas emissions in the summer. To make this process more effective, we incorporate a spatial bias that enables the transformer to prioritize relationships based on physical distance, such as the interactions between soil layers at different depths. This is followed by multiple like feed-forward networks to refine its understanding of the data. After several layers of processing, the proposed SM-Hybrid produces a rich representation of the data that captures both short-term and long-term patterns in soil C, moisture, and GHG fluxes.

SM-Hybrid processes $\{\tilde{e}_t\}_{t=1}^{T}$ with N layers of multi-head self-attention (MHSA) and a feed-forward network (FFN). For head h,

$$Q_h = \tilde{E}W^Q_h, \quad K_h = \tilde{E}W^K_h, \quad V_h = \tilde{E}W^V_h, \tag{10}$$

where $\tilde{E} = [\tilde{e}_1,\ldots,\tilde{e}_T]^T$ and $W^{Q,K,V}_h \in \mathbb{R}^{d_{\text{model}}\times d_k}$. Attention scores include spatial bias:

$$A_h = \text{softmax}\Bigl(\frac{Q_h K_h^T}{\sqrt{d_k}} + B_{\text{space}}\Bigr), \tag{11}$$

with $B_{\text{space}} \in \mathbb{R}^{T\times T}$. The MHSA output is

$$\text{MHSA}(\tilde{E}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H)W^O, \tag{12}$$

where $W^O \in \mathbb{R}^{d_{\text{model}}\times d_{\text{model}}}$. The FFN is

$$\text{FFN}(x) = W_2\ \text{ReLU}(W_1 x + b_1) + b_2, \tag{13}$$

with $W_1 \in \mathbb{R}^{d_{\text{model}}\times d_{\text{ff}}}$ and $W_2 \in \mathbb{R}^{d_{\text{ff}}\times d_{\text{model}}}$. Layer normalisation:

$$h_t^{(n)} = \text{LayerNorm}\big(\tilde{e}_t^{(n-1)} + \text{MHSA}(\tilde{e}_t^{(n-1)})\big), \tag{14}$$

$$\tilde{e}_t^{(n)} = \text{LayerNorm}\big(h_t^{(n)} + \text{FFN}(h_t^{(n)})\big). \tag{15}$$

Finally, the transformer's output is summarized to produce the final predictions. We average the information over time to get a single, concise result, which is then transformed into the specific outputs we're interested in, such as the amount of CO_2 emitted, the soil moisture level, or the concentration of N gases. This step ensures that the model's predictions are directly usable for understanding soil dynamics in the Daycent model, supporting applications like climate change mitigation or agricultural planning. The transformer output $h^{(N)} \in \mathbb{R}^{T\times d_{\text{model}}}$ is mean-pooled and linearly transformed:

$$\hat{y} = W_{\text{out}}\left(\frac{1}{T}\sum_{t=1}^{T} h_t^{(N)}\right) + b_{\text{out}}, \tag{16}$$

where $W_{\text{out}} \in \mathbb{R}^{d_{\text{model}}\times d_{\text{out}}}$ and $b_{\text{out}} \in \mathbb{R}^{d_{\text{out}}}$. We utilize a lightweight transformer (two encoder layers, four heads) to minimize computations while capturing long-range dependencies, along with an ablation study that quantifies the contributions of the CNN, transformer, and spatial bias.

5 Experimental Evaluation

The goal of the experiments was to evaluate the benefits of a deep neural network for developing a fast surrogate model with accuracy comparable to Daycent, with research questions (RQ) as follows:

RQ1: How do our proposed methods compare to Daycent and baseline ANN on both prediction and training time?

RQ2: How do different configurations of our proposed model perform when predicting NEE and crop yield?

RQ3: How do the candidate surrogate methods perform under different spatial and temporal partitioning strategies?

Experimental Design: Figure 4 shows the overall experimental design. The Data Division component involves various data-partitioning strategies used to determine the training, testing, and validation datasets. For each partition strategy, we compare CNN, Transformer, and Hybrid Surrogate Model with Hybrid of convolution and Transformer (SM-Hybrid). Finally, the fourth component is based on MSE-Loss and execution time.

Implementation Details: The experiments were conducted using Google Colab Pro, leveraging an NVIDIA A100 GPU for both training and evaluation. This configuration provided the computational power necessary to train models with high complexity, such as convolutional neural networks (CNNs) and transformers. All models were optimized using Adam [**?**] with default learning-rate scheduling. Zero-padding and stride values were held constant across layers to maintain consistency. Global average pooling was used to stabilize gradient updates under variable spatial supports.

We evaluated four models: CNN, Transformer, and our proposed Surrogate Model with CNN and Transformers (SM-Hybrid), each optimized for balanced complexity and generalization. All models were optimized using Adam. The details are as follows:

- **Surrogate Model with Convolutional Neural Networks (SM-CNN):** Consists of two convolutional layers (filters: 32, 64; kernel size: 3×3; activation: ReLU), followed by MaxPooling, dropout (0.2–0.4), and a dense layer with 64 neurons.
- **Surrogate Model with a Transformer (SM-T):** Uses two convolutional layers (filters: 32, 64; kernel size: 3×3; activation: ReLU) for feature extraction, followed by MaxPooling, dropout (0.2–0.4), and two Transformer encoder layers (feedforward dimension: 64, dropout: 0.2). The output is projected through a dense layer with 64 neurons.
- **Surrogate Model with a Hybrid Architecture (SM-Hybrid):** Combines CNN-based local feature extraction (two 1D convolutional layers with GlobalAveragePooling) and Transformer-based sequence modeling (two encoder layers, feedforward dimension: 64, dropout: 0.2). This design balances locality and long-range dependencies, with a total of 32,002 trainable parameters.
- **Surrogate Model with Vision Transformer (SM-ViT):** Adapts the vision transformer architecture by partitioning input sequences into patches, each linearly projected into embeddings. These embeddings pass through two transformer encoder layers (multi-head self-attention, feedforward dimension: 64, dropout: 0.2). The output tokens are aggregated using GlobalAveragePooling and projected through a dense layer with 64 neurons. This architecture emphasizes global spatial dependencies but often underperforms in capturing fine-scale spatial heterogeneity compared to CNN or hybrid variants.

Choice of Candidates: We considered several architectures for our surrogate models. Each candidate was motivated by different strengths. CNN-based models are effective at extracting local spatial features such as neighborhood patterns, while transformer models are well-suited for capturing long-range temporal dependencies. Our hybrid design combines these two capabilities, enabling the model to balance fine-scale locality with broader contextual information. We also considered the *Swin Transformer*, which excels in vision tasks with stable spatial hierarchies, but performs less effectively in our domain. Agricultural landscapes are strongly influenced by human interventions, such as crop rotation, planting legumes for soil enrichment, or leaving fields fallow. These decisions introduce irregular, non-stationary dynamics that reduce the predictability of models like Swin, which assume more consistent spatial structures. While further domain-specific tuning could potentially improve its performance, our experiments indicate that CNNs, transformers, and their hybrid form provide a better match to the spatiotemporal complexities of the data.

Dataset Description: The Daycent dataset consists of simulation results from 2,562 sites within the U.S. Midwest, spanning the period 2000–2020. In this work, we focused on only two key output variables: Net Ecosystem Exchange (NEE), which represents daily CO_2 fluxes, and an approximation of total Soil Organic Carbon (SOC) change. Crop Yield will be added (more details in Appendix A). Figure 5 illustrates two different spatial partitioning strategies for data distribution across *Illinois (IL)*, *Iowa (IA)*, and *Indiana (IN)*. This comparison highlights the impact of spatial awareness in training AI models. The structured spatial partitioning method Figure 5

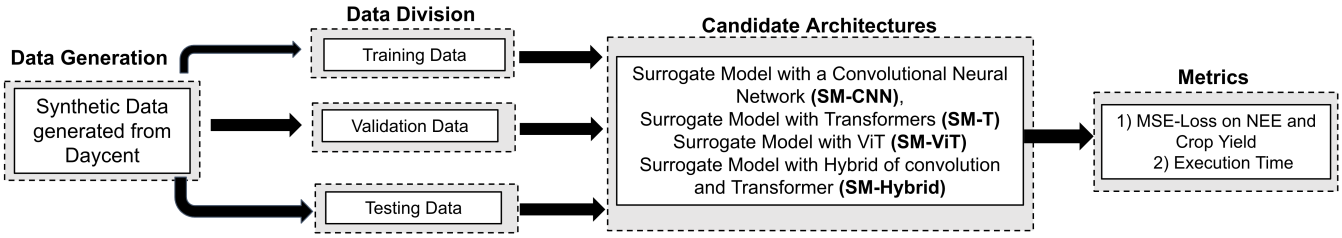
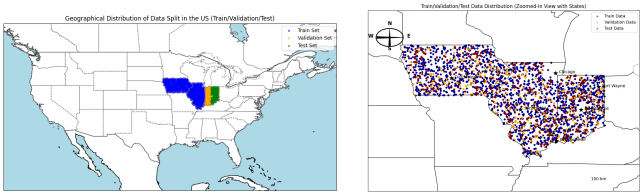


Figure 4: Experiment Design

(a) preserves spatial locality, enabling models to better learn regional characteristics, whereas the random partitioning approach Figure 5 (b) does not explicitly account for geographical divisions.



(a) Longitudinal Partitioning **(b) Random Partitioning**

Figure 5: Spatial Partitioning of Synthetic Data

We evaluated out-of-region generalization using three complementary holdout protocols: **(a)**a **spatial (longitudinal) block holdout** that withholds entire, contiguous geographic regions, **(b)** an IID **random split** that primarily measures interpolation, and **(c)** a **temporal block holdout** that withholds contiguous years. Unless noted otherwise, we adopted a **70/15/15** train/validation/test split along the relevant axis: For the spatial block, one region is held out for testing. In contrast, the remaining regions were used for training and validation. For the temporal block, the most recent period was reserved for testing, and earlier periods were withheld for training/validation. This design quantifies how well the model **transfers to unseen geographies or unknown years**, in contrast to the IID split and temporal block, which mixes spatial contexts. As shown in Section 5.1, the **accuracy and efficiency differ across protocols**, indicating that the choice of split materially affects conclusions. These findings underscore the need to account for **spatial and temporal dependence** when training and evaluating surrogate models across heterogeneous domains.

(a) Longitudinal Partitioning: Figure 5 (a) presents a structured geographical data split based on longitude, where Illinois, Iowa, and Indiana are divided into three spatial segments: 70% for training (blue), 15% validation (orange), and 15% testing (green). This spatial partitioning ensures that the model learns regional spatial dependencies and generalizes well across different geographies.

(b) Random Partitioning: Figure 5 depicts a purely random allocation of data points across the same region, with no enforcement of spatial boundaries. Here, training, validation, and test samples are randomly distributed throughout the area, leading to a mix of regional influences in all sets.

(c) Temporal Partitioning: Temporal partitioning enables models to learn temporal dependencies and assess performance on future time windows, enhancing generalizability for long-term trend predictions. The upper panel in Figure 6 shows Net Ecosystem Exchange (i.e., NEE) in gC/m²/day, with intense seasonal cycles, segmented into training, validation, and testing periods using distinct background colors. The lower panel depicts annual crop yield (ton/ha) over the same period, ensuring consistent temporal splits across variables.

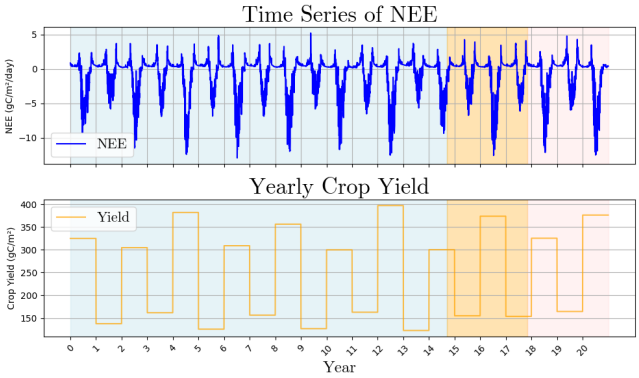


Figure 6: Temporal Partitioning of Synthetic Data

Synthetic Data Generation: The dataset was generated using Daycent version 279. The simulation sites were restricted to annual croplands in all counties of Iowa, Illinois, and Indiana, using USDA Cropland Data Layer datasets (https://croplandcros.scinet.usda.gov/). Site-specific soil characteristics required for Daycent were derived from SSURGO [50]. Daily weather data were from gridMET [1]. Common rain-fed corn–soybean rotations were simulated for 21 years in the three states with state-specific planting dates. Forty-two scenarios were simulated, namely two control scenarios (corn–soybean and soybean–corn rotations without fertilization) and combinations of two rotations (corn–soybean or soybean–corn), ten fertilization rates (3.36 to 33.6 g N/m2), and two application timings (on corn's planting date or 30 days afterwards).

The crop yield provided in the Daycent data is the harvested grain carbon amount each year (one value per year). It is not the average value of all crop yields over a field, but only specific to one crop type, either "corn" or "soybean". The crop type is a variable in the input file, named PLANTT (see Appendix A for a description), which has two categories: 0 represents soybeans and 1 represents corn. Together, the input variables FERTZRN, PDOY, and PLANTT (described in Appendix A) contain information on fertilizer rate, plant date, crop type, and rotation information, which allows for the reference of a specific yield to a particular management scenario, resulting in separate yields for different fertilizer rates.

Table 2: Performance comparison of different neural-network methods

Methods	Accuracy	Prediction Time	Training Time
Daycent (Original)	1.000	2 s/site	N/A
ANN	0.790	0.00032 s/site	118.81 s
Proposed Surrogate Model with Spatial Neural Networks (SM-Hybrid)			
SM-CNN	0.900	0.00475 s/site	77.92 s
SM-T	0.950	0.01026 s/site	846.15 s
SM-ViT	0.966	0.01096 s/site	73.30 s
SM-Hybrid	0.954	**0.00343** s/site	**62.97** s

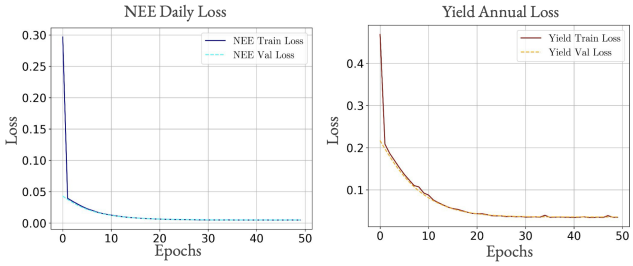
5.1 Computational Efficiency (RQ1)

Evaluation metric: While DayCent requires approximately 2 seconds per site, our SM-Hybrid reduces inference time to 0.00343 s/site, representing an ≈600× speedup. For perspective, running DayCent on 10,000 sites would require nearly 5.5 hours, whereas SM-Hybrid would finish in less than a minute. This large gain highlights the practical impact of surrogate modeling for regional and cross-country scale assessments. Unless otherwise noted, we report *Accuracy* as the coefficient of determination R^2 on the hold-out test set (per target), averaged across runs and reported alongside time metrics. For loss curves, we plot MSE. We will include 95% bootstrap CIs for R^2 in the Appendix. Table 2 presents training time (seconds) and epochs for different architectures (CNN, Transformer, and Hybrid) across three data partitioning strategies (Random, Longitude, and Yearly). The transformer requires the longest training time, reaching 1553.02s (Random), 846.15s (Longitude), and 410.83s (Yearly), approximately 5–10 times that of CNN, which exhibits the highest efficiency at 45.44s under the Yearly split. The Hybrid model balances computational complexity and efficiency. Structured data partitioning (Longitude/Yearly) significantly reduces training time, with CNN showing an 84.4% decrease from Random to Yearly splits, indicating that spatial and temporal structure enhances learning efficiency. Additionally, convergence speed varies across architectures, with the Hybrid model reaching early stopping the fastest (10 epochs, Yearly split), while CNN completes all 50 epochs in Random splits, highlighting differences in model adaptability to data distributions.

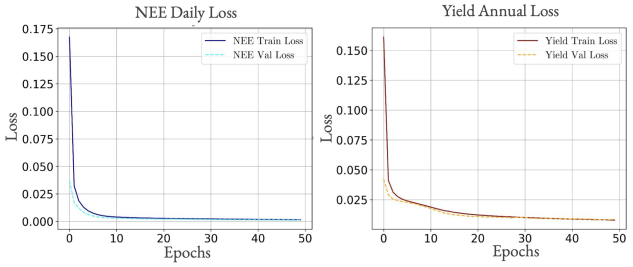
Limitations: NN model comparisons face limitations due to variations in the hidden layer dimensions (64–512) and network depth (6–12 layers), which introduce biases in the training time. Additionally, the dynamic early stopping mechanism (**patience=5, restore best weights=True**) affects convergence while preventing overfitting and improving efficiency; it also introduces comparison bias, as the models stop at different epochs. These factors complicate direct comparisons of training time and convergence speed across architectures. In addition, future revisions will report R^2 with 95% CIs via site-wise bootstrap and paired significance tests (e.g., Wilcoxon signed-rank) for model comparisons.

5.2 Comparative Analysis (RQ2 and RQ3)

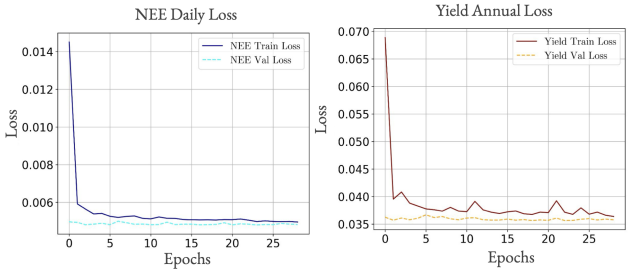
We evaluated our proposed surrogate models on three partitionings of the NEE and the Crop Yield data by comparing SM-CNN, SM-T, and SM-Hybrid based on three splits: (a) Random Split, (b) Spatial Data Split, and (c) Temporal Splits on both NEE and Crop Yield. To improve comparability, we merge the three loss-curve figures into a single multi-panel figure with *aligned y-axes* for viewing training/validation loss across splits.



(a) CNN performance for NEE and Crop Yield MSE Loss



(b) Transformer performance for NEE and Crop Yield MSE Loss

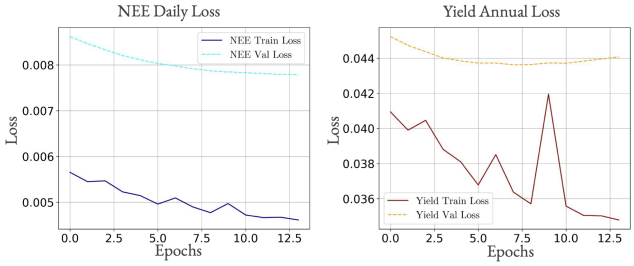


(c) Hybrid (CNN+Transformers) for NEE and Crop Yield MSE Loss

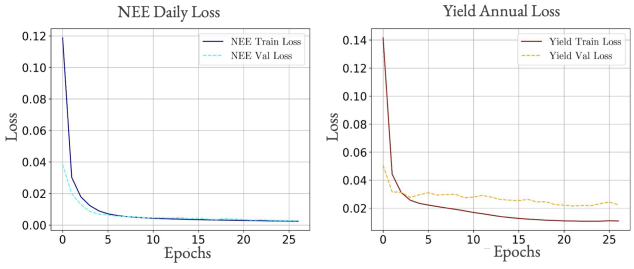
Figure 7: Performance Comparison on Random Data Split

Random Data Split: As shown in Figure 7, the CNN model shows a sharp decline in training loss for both NEE and crop yield, stabilizing at a low value with validation loss following closely, indicating strong generalization. This is expected since CNNs are less sensitive to spatial autocorrelation and perform well under uniformly distributed patterns. The Transformer also converges quickly but cannot fully exploit long-range dependencies due to the disrupted spatial structure of random splits. The hybrid CNN-Transformer achieves comparable convergence with stable validation loss, though its advantages are less pronounced. Overall, CNN performs best in this setting, while the loss of spatial and temporal continuity limits the Transformer-based models.

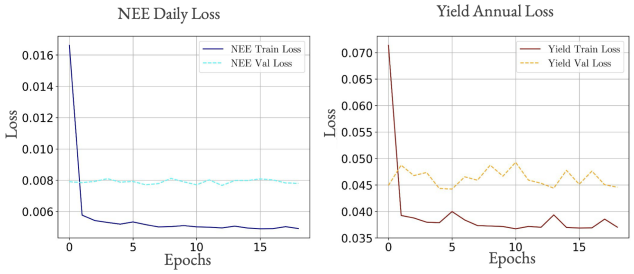
Spatial Data Split: Figure 8 shows that CNN training loss for NEE and yield declines steadily, but validation loss remains elevated, indicating limited generalization to unseen regions. This reflects CNN's reliance on local features, which limits its ability to capture spatial autocorrelation and long-range teleconnections. By contrast, the Transformer converges with lower validation loss, leveraging self-attention to model distant dependencies. The hybrid CNN-Transformer achieves the most balanced performance, with stable



(a) CNN performance for NEE and Crop Yield MSE Loss

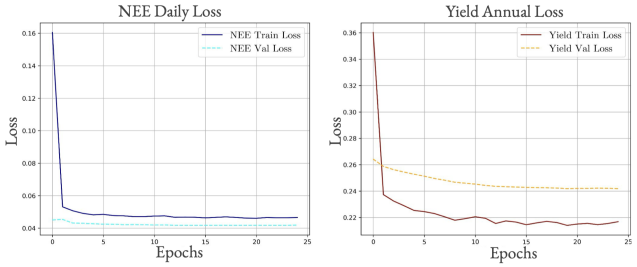


(b) Transformer performance for NEE and Crop Yield MSE Loss

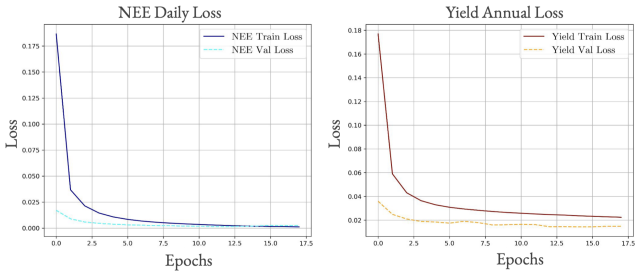


(c) Hybrid (CNN+Transformers) for NEE and Crop Yield MSE Loss

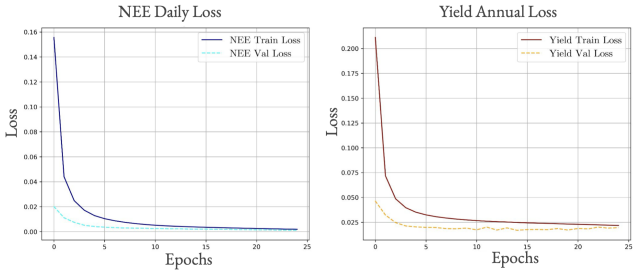
Figure 8: Performance Comparison on Spatial Data Split



(a) CNN performance for NEE and Crop Yield MSE Loss



(b) Transformer performance for NEE and Crop Yield MSE Loss



(c) Hybrid (CNN+Transformers) for NEE and Crop Yield MSE Loss

Figure 9: Performance Comparison of DNN architectures based on Temporal Data Split

training and validation loss, effectively combining local feature extraction with global dependency modeling.

Temporal Data Split: The results in Figure 9 show that the CNN model exhibits a sharp initial decline in both NEE and yield loss during training. In contrast, the validation loss remains significantly higher, suggesting poor generalization across different years. Since CNNs primarily rely on local feature extraction, they struggle in yearly splits where temporal dependencies are crucial for capturing trends over time. Transformers exhibits a rapid drop in both training and validation loss, ultimately stabilizing at a significantly lower value, which highlights its strength in modeling long-range temporal dependencies and learning trends effectively across multiple years. The hybrid CNN-Transformer model falls between the two, benefiting from both local feature extraction (CNN) and long-term sequence modeling (Transformer). This allows it to perform better than CNN but slightly worse than a pure Transformer in this setting, as CNN components may still struggle with temporal generalization. Overall, transformers perform best in yearly splits due to their ability to capture long-range dependencies. The hybrid CNN-Transformer performs moderately well by balancing feature extraction and sequence modeling, while the CNN struggles the most due to its limited ability to learn across temporal scales.

In summary, the transformer and hybrid models outperform CNN in spatial and temporal splits, as they better generalize across regions and time periods. The Transformer model is robust in capturing tele-connections, making it ideal for capturing long-range dependencies. In contrast, the hybrid CNN-Transformer model benefits from both localized feature extraction and long-range spatial-temporal relationships, resulting in a more balanced approach.

5.3 Discussion

This work introduces a hybrid surrogate model that combines CNNs and Transformers to capture spatial dependencies. CNNs effectively model local autocorrelation but struggle with long-range interactions. Transformers complement this with attention mechanisms that represent tele-connections, where distant regions influence one another. Advances such as the Swin Transformer further extend this capacity through hierarchical, multi-scale structures, underscoring progress toward architectures tailored for complex geospatial processes. While our model is intentionally compact, future extensions will incorporate domain priors (e.g., soil or climate embeddings), residual CNN backbones, and specialized attention for spatiotemporal fields.

Persistent challenges remain in modeling spatial data. Spatial variability—where environmental characteristics differ across regions—limits the effectiveness of uniform models. Prior studies [14, 15, 22] show that one-size-fits-all approaches often fail to capture local context, with consequences for applications such as precision agriculture, climate forecasting, and disaster response. Edge effects also reduce accuracy near spatial boundaries due to limited context, introducing systematic bias. Addressing these issues is essential for models that generalize across diverse environments. SM-Hybrid explicitly models local and long-range dependencies, but further work is needed to mitigate variability, boundary effects, and region-specific complexities.

5.4 Investigative Framework

We propose a three-step framework to predict soil carbon, moisture, and greenhouse gas (GHG) fluxes while balancing temporal forecasting accuracy and spatial generalization. This paper focuses on *Step 1* (synthetic DayCent training) and leaves Steps 2–3 (auxiliary data fusion and refinement with domain adaptation) for the future.

Step 1: Training on Synthetic Data: Convolutional Neural Networks are used to capture localized spatial patterns, enabling the model to learn fundamental relationships among soil carbon stocks, moisture, and greenhouse gas fluxes.

Step 2: Incorporating Auxiliary Data. To improve generalization in data-sparse regions, auxiliary high-resolution datasets (e.g., SSURGO soil characteristics, DEMs, weather, evapotranspiration) are added. These features are integrated via feature augmentation (extra channels) or multi-branch architectures (separate subnetworks merged before the sequence model). The architecture transitions from a CNN to a hybrid CNN–LSTM or CNN–Transformer, with self-attention capturing both local and global dependencies.

Step 3: Refinement and Spatiotemporal Generalization. Models are fine-tuned with ground-truth observations (field campaigns, remote sensing, long-term monitoring) to capture seasonal and climate-driven variability. Domain adaptation enhances transferability across regions:

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{j=1}^{n} \|y_j - f_\theta(x_j, z_j)\|^2 + \Omega(\theta), \tag{17}$$

with task-specific weights:

$$\mathcal{L} = w_c \mathrm{MSE}(c) + w_m \mathrm{MSE}(m) + w_f \mathrm{MSE}(f). \tag{18}$$

In this paper, we focus on Step 1 and demonstrate the promise of incorporating spatial relationships derived from the Daycent model (shown in Figure 2) into a spatial-temporal aware surrogate model that speeds up predictions with reasonable accuracy.

6 Related Work

Surrogate models have been widely applied beyond ecology, accelerating domains such as numerical weather prediction, hydrology, and storm surge forecasting [30, 41]. These examples highlight their general utility and contextualize our contribution to advancing DayCent modeling. Within ecology, surrogate approaches range from Gaussian processes [40] to neural networks [24, 42], though many overlook spatial heterogeneity and autocorrelation [5]. Recent spatiotemporal hybrids (e.g., RF–CNNs [4]) address these gaps, underscoring the need for systematic baselines and ablations.

Machine learning has also been used to approximate biogeochemical processes, with Random Forests and SVMs applied to soil moisture, carbon fluxes, and crop yields [6, 55], and deep learning applied to SOC dynamics and GHG emissions [26]. Yet these methods often struggle with scalability and spatiotemporal structure [14, 15, 19, 20]. Hybrid approaches that integrate process knowledge with ML—for example, linking soil carbon pools to measurable fractions [10] or surrogate-assisted optimization [7]—have been validated against long-term field data and remote sensing [11, 25]. Building on this, our work develops a neural surrogate for DayCent that explicitly models spatial heterogeneity through CNNs and transformers. Recent advances in generative and sequence models further expand geo-simulation. Grid-based methods [16, 31, 34] and pixel-level synthesis [8, 32] prioritize efficiency but yield coarse outputs. Variational autoencoders (VAEs) and GANs model complex trajectory distributions [23, 27, 39, 43–49, 53, 54, 58], yet often rely on low-resolution encodings such as grids [37, 59] or image-like formats [21, 56], reducing spatiotemporal fidelity [35].

7 Conclusion and Future Work

Purely data-driven methods often fall short in scientific domains, as they depend on large labeled datasets and may fail to respect physical laws, leading to implausible predictions in areas such as climate science and biology [3, 13, 57]. To address this, we embed spatial and physical reasoning through a mass-balance index to ensure statistical coherence and improve computational efficiency. This paper presents Surrogate Models with Hybrid Spatial Neural Networks (SM-Hybrid), which explicitly capture spatial autocorrelation and tele-connections in agricultural data. Experiments show that SM-Hybrid achieves higher accuracy than SM-ANN and is two orders of magnitude faster than DayCent, outperforming CNNs and transformers in both efficiency and spatial fidelity.

Future Work: We will extend the three-step framework by (i) expanding datasets with additional variables (e.g., soil GHG emissions) and soil-type–based splits, (ii) exploring advanced deep learning approaches (new architectures, knowledge-guided ML, transfer learning) to reduce overfitting and cost, and (iii) broadening outputs beyond NEE and yield (e.g., N_2O, leaching, soil water content) with confidence intervals. Beyond interpolation, we plan to study out-of-distribution generalization using physics-informed constraints and domain adaptation. Additional directions include ablation studies on padding, stride, and residual connections, clarifying highlighted inputs in the appendix by linking them to DayCent variables, and richer spatial encodings using agronomic priors and great-circle distances. Finally, we aim to connect SM-Hybrid outputs to decision-support tools such as COMET-Planner, translating efficiency gains into applied agricultural management.

Acknowledgments

This material is based on work supported by the USDA under Grant No. 2023-67021-39829, the National Science Foundation under Grants No. 2118285, 2040459, 1737633, 1901099, and 1916518, the USDOE Office of Energy Efficiency and Renewable Energy under FOA No. DE-FOA0002044, USDA under Grant No. 2021-51181-35861, and USDOD under Grant No. HM04762010009. We would also like to thank Kim Koffolt and the Spatial Computing Research Group for their helpful comments and refinements.

References

[1] John T Abatzoglou. 2013. Development of gridded surface meteorological data for ecological applications and modelling. *International journal of climatology* 33, 1 (2013), 121–131.

[2] Melissa Adrian, Daniel Sanz-Alonso, and Rebecca Willett. 2024. Data Assimilation with Machine Learning Surrogate Models: A Case Study with FourCastNet. *arXiv preprint arXiv:2405.13180* (2024). https://arxiv.org/abs/2405.13180

[3] Mark Alber, Adrian Buganza Tepole, William R Cannon, Suvranu De, Salvador Dura-Bernal, Krishna Garikipati, George Karniadakis, William W Lytton, Paris Perdikaris, Linda Petzold, et al. 2019. Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *NPJ digital medicine* 2, 1 (2019), 115.

[4] John Barry and Robert Gichuhi. 2024. Hybrid Machine Learning Approaches for Spatial Heterogeneity in Ecological Modeling. *Environmental Modeling & Software* 147 (2024), 105633. https://doi.org/10.1016/j.envsoft.2023.105633

[5] Keith Beven. 2018. *Rainfall-Runoff Modelling: The Primer* (3rd ed.). John Wiley & Sons.

[6] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/A:1010933404324

[7] E. E. Campbell, S. A. Williams, J. M. Antle, B. Basso, and K. H. Paustian. 2021. Integrating machine learning with process-based models to improve soil carbon quantification: A DayCent case study. In *Proceedings of the ASA-CSSA-SSSA International Annual Meeting*. Abstract only.

[8] Chu Cao and Mo Li. 2021. Generating mobility trajectories with retained data utility. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. ACM, Singapore, 2610–2620.

[9] A. Carrassi. 2023. Surrogate modeling for the climate sciences dynamics with machine learning and data assimilation. *Frontiers in Applied Mathematics and Statistics* 9 (2023). https://www.frontiersin.org/journals/applied-mathematics-and-statistics/articles/10.3389/fams.2023.1290936/full

[10] S. R. S. Dangal, C. R. Schwalm, M. A. Cavigelli, H. T. Gollany, V. L. Jin, and J. Sanderman. 2022. Improving soil carbon estimates by linking conceptual pools against measurable carbon fractions in the DAYCENT model (Version 4.5). *Journal of Advances in Modeling Earth Systems* 14, 5 (2022), e2021MS002622.

[11] S. J. Del Grosso, S. M. Ogle, W. J. Parton, and F. J. Breidt. 2010. Estimating uncertainty in N_2O emissions from U.S. cropland soils. *Global Biogeochemical Cycles* 24 (2010), GB1009.

[12] S. J. Del Grosso, D. S. Ojima, W. J. Parton, A. R. Mosier, G. A. Peterson, and D. S. Schimel. 2002. Simulated effects of dryland cropping intensification on soil organic matter and greenhouse gas exchanges using the DAYCENT ecosystem model. *Environmental Pollution* 116 (2002), S75–S83.

[13] J H Faghmous and Vipin Kumar. 2014. A big data guide to understanding climate change: Case for theory-guided data science. *Big data* 2, 3 (2014), 155–163.

[14] Majid Farhadloo, Arun Sharma, Jayant Gupta, Alexey Leontovich, Svetomir N Markovic, and Shashi Shekhar. 2024. Towards spatially-lucid ai classification in non-euclidean space: An application for mxif oncology data. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*. SIAM, 616–624.

[15] Majid Farhadloo, Arun Sharma, Alexey Leontovich, Svetomir N Markovic, and Shashi Shekhar. 2025. Spatially-Delineated Domain-Adapted AI Classification: An Application for Oncology Data. In *Proceedings of the 2025 SIAM International Conference on Data Mining (SDM)*. SIAM, 487–496.

[16] Jie Feng, Zeyu Yang, Fengli Xu, Haisu Yu, Mudan Wang, and Yong Li. 2020. Learning to simulate human mobility. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. ACM, New York, NY, USA, 3426–3433.

[17] National Center for Appropriate Technology. 2020. Understanding the COMET-Planner Tool. https://attra.ncat.org/understanding-the-comet-planner-tool/. Accessed: June 1, 2025.

[18] A. I. J. Forrester, A. Sóbester, and A. J. Keane. 2008. *Engineering Design via Surrogate Modelling: A Practical Guide*. Wiley.

[19] Subhankar Ghosh, Jayant Gupta, Arun Sharma, Shuai An, and Shashi Shekhar. 2024. Reducing false discoveries in statistically-significant regional-colocation mining: A summary of results. *arXiv preprint arXiv:2407.02536* (2024).

[20] Subhankar Ghosh, Arun Sharma, Jayant Gupta, Aneesh Subramanian, and Shashi Shekhar. 2024. Towards kriging-informed conditional diffusion for regional sea-level data downscaling: A summary of results. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*. 372–383.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., Montreal, Canada, 2672–2680.

[22] Jayant Gupta, Carl Molnar, Yiqun Xie, Joe Knight, and Shashi Shekhar. 2021. Spatial variability aware deep neural networks (svann): A general approach. *ACM Transactions on Intelligent Systems and Technology (TIST)* 12, 6 (2021), 1–21.

[23] Nils Henke, Shimon Wonsak, Prasenjit Mitra, Michael Nolting, and Nicolas Tempelmeier. 2023. Condtraj-gan: Conditional sequential gan for generating synthetic vehicle trajectories. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Springer, Cham, 79–91.

[24] William W. Hsieh. 2009. *Machine Learning Methods in the Environmental Sciences: Neural Networks and Kernels*. Cambridge University Press.

[25] M. K. Jarecki, T. B. Parkin, A. S. K. Chan, J. L. Hatfield, and R. Jones. 2008. Comparison of DAYCENT-simulated and measured nitrous oxide emissions from a corn field. *Journal of Environmental Quality* 37, 5 (2008), 1685–1690.

[26] Amit Kapoor, Priti Sharma, and Sudipta Roy. 2022. Comparison of CNN and Fourier Neural Operators for Surrogate Modeling of Heterogeneous Media. *arXiv preprint arXiv:2204.12345* (2022).

[27] K Vimal Kumar, Divakar Yadav, and Arun Sharma. 2015. Graph based technique for hindi text summarization. In *Information Systems Design and Intelligent Applications: Proceedings of Second International Conference INDIA 2015, Volume 1*. Springer, 301–310.

[28] Sandeep Kumar. 2015. Estimating spatial distribution of soil organic carbon for the Midwestern United States using historical database. *Chemosphere* 127 (2015), 49–57.

[29] David Laborde, Abdullah Mamun, Will Martin, Valeria Piñeiro, and Rob Vos. 2021. Agricultural subsidies and global greenhouse gas emissions. *Nature communications* 12, 1 (2021), 2601.

[30] S. Li et al. 2024. Using surrogate modeling to predict storm surge on evolving landscapes under climate change. *npj Natural Hazards* 1, 1 (2024). https://www.nature.com/articles/s44304-024-00002-9

[31] Siyu Li, Toan Tran, Haowen Lin, John Krumm, Cyrus Shahabi, Lingyi Zhao, Khurram Shafique, and Li Xiong. 2024. Geo-Llama: Leveraging LLMs for Human Mobility Trajectory Generation with Spatiotemporal Constraints. *arXiv preprint arXiv:2408.13918* xx, yy (2024), xx–yy.

[32] Yan Lin, Huaiyu Wan, Jilin Hu, Shengnan Guo, Bin Yang, Youfang Lin, and Christian S Jensen. 2023. Origin-destination travel time oracle for map-based services. *Proceedings of the ACM on Management of Data* 1, 3 (2023), 1–27.

[33] 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.

[34] Qingyue Long, Huandong Wang, Tong Li, Lisi Huang, Kun Wang, Qiong Wu, Guangyu Li, Yanping Liang, Li Yu, and Yong Li. 2023. Practical Synthetic Human Trajectories Generation Based on Variational Point Processes. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, New York, NY, USA, 4561–4571.

[35] Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. 2021. A survey on deep learning for human mobility. *ACM Computing Surveys (CSUR)* 55, 1 (2021), 1–44.

[36] Trung H Nguyen, Duy Nong, and Keith Paustian. 2019. Surrogate-based multi-objective optimization of management options for agricultural landscapes using artificial neural networks. *Ecological Modelling* 400 (2019), 1–13.

[37] Kun Ouyang, Reza Shokri, David S Rosenblum, and Wenzhuo Yang. 2018. A non-parametric generative model for human trajectories.. In *IJCAI*, Vol. 18. International Joint Conferences on Artificial Intelligence, California, USA, 3812–3817.

[38] William J Parton, Melannie Hartman, Dennis Ojima, and David Schimel. 1998. DAYCENT and its land surface submodel: description and testing. *Global and planetary Change* 19, 1-4 (1998), 35–48.

[39] Jinmeng Rao, Song Gao, Yuhao Kang, and Qunying Huang. 2020. LSTM-TrajGAN: A deep learning approach to trajectory privacy protection. *arXiv preprint arXiv:2006.10521* 1, 1 (2020), –.

[40] Carl Edward Rasmussen and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press.

[41] S. Razavi, B. A. Tolson, and D. H. Burn. 2012. Review of surrogate modeling in water resources. *Water Resources Research* 48, 7 (2012), W07401. https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2011WR011527

[42] Markus Reichstein, Gustavo Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. 2019. Deep learning and process understanding for data-driven Earth system science. *Nature* 566 (2019), 195–204. https://doi.org/10.1038/s41586-019-0912-1

[43] Arun Sharma et al. 2022. Towards a tighter bound on possible-rendezvous areas: preliminary results. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*. 1–11.

[44] Arun Sharma, Majid Farhadloo, Yan Li, Jayant Gupta, Aditya Kulkarni, and Shashi Shekhar. 2022. Understanding Covid-19 effects on mobility: A community-engaged approach. *AGILE: GIScience Series* 3 (2022), 14.

[45] Arun Sharma, Jayant Gupta, and Shashi Shekhar. 2022. Abnormal trajectory-gap detection: A summary (short paper). In *15th International Conference on Spatial Information Theory (COSIT 2022)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 26–1.

[46] Arun Sharma, Zhe Jiang, and Shashi Shekhar. 2022. Spatiotemporal data mining: A survey. *arXiv preprint arXiv:2206.12753* (2022).

[47] Arun Sharma and Shashi Shekhar. 2022. Analyzing trajectory gaps to find possible rendezvous region. *ACM Transactions on Intelligent Systems and Technology (TIST)* 13, 3 (2022), 1–23.

[48] Arun Sharma, Xun Tang, Jayant Gupta, Majid Farhadloo, and Shashi Shekhar. 2020. Analyzing trajectory gaps for possible rendezvous: A summary of results. In *11th International Conference on Geographic Information Science (GIScience 2021)-Part I (2020)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 13–1.

[49] Arun Sharma, Syed Mohammed Arshad Zaidi, Varun Chandola, Melissa R Allen, and Budhendra L Bhaduri. 2018. WebGIobe-A cloud-based geospatial analysis framework for interacting with climate data. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*. 42–46.

[50] S Staff. 2020. Gridded Soil Survey Geographic (gSSURGO) Database for the Conterminous United States. *United States Department of Agriculture, Natural Resources Conservation Service* (2020).

[51] The New York Times. 2025. AI Weather Models Are Transforming Forecasting. *The New York Times* (21 May 2025). https://www.nytimes.com/2025/05/21/climate/ai-weather-models-aurora-microsoft.html

[52] M. Toro et al. 2020. Forecasting flood hazards in real time: a surrogate model for hydrometeorological events in an Andean watershed. *Natural Hazards and Earth System Sciences* 20, 11 (2020), 3261–3278. https://nhess.copernicus.org/articles/20/3261/2020/

[53] Pawan Kumar Upadhyay, Satish Chandra, and Arun Sharma. 2016. A novel approach of adaptive thresholding for image segmentation on GPU. In *2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)*. IEEE, 652–655.

[54] Pawan Kumar Upadhyay, Arun Sharma, et al. 2018. A Novel Approach of Intuitive K-means Clustering for Renal Calculi Detection in Ultrasound Images. *International Journal on Electrical Engineering & Informatics* 10, 1 (2018).

[55] Vladimir N. Vapnik. 1998. *Statistical Learning Theory*. Wiley.

[56] Xingrui Wang, Xinyu Liu, Ziteng Lu, and Hanfang Yang. 2021. Large scale GPS trajectory generation using map based on two stage GAN. *Journal of Data Science* 19, 1 (2021), 126–141.

[57] 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.

[58] Tianqi Xia, Xuan Song, Zipei Fan, Hiroshi Kanasugi, QuanJun Chen, Renhe Jiang, and Ryosuke Shibasaki. 2018. Deeprailway: a deep learning system for forecasting railway traffic. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, IEEE, Piscataway, NJ, 51–56.

[59] Yuan Yuan, Jingtao Ding, Huandong Wang, Depeng Jin, and Yong Li. 2022. Activity trajectory generation via modeling spatiotemporal dynamics. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 4752–4762.

[60] Y. Zhang et al. 2022. A regional wind wave prediction surrogate model based on CNN deep learning network. *Ocean Engineering* 256 (2022), 111523. https://www.sciencedirect.com/science/article/abs/pii/S0029801822010133

A Daycent Model

Daycent [38] provides a comprehensive foundation for analyzing the complex interactions between climate, soil properties, and agricultural practices, which are critical for advancing sustainable agriculture and climate change mitigation. Daily climate records facilitate the modeling of crop growth and soil processes, while crop growth simulations capture crop responses to varying climatic and edaphic conditions. Inter-annual yield data offer insights into productivity trends, and SOC changes inform assessments of C sequestration and soil health. The dataset's extensive spatial coverage across diverse soil types, climate zones, and land management regimes enhances model robustness and generalizability. Furthermore, the temporal range (2000–2020) enables the examination of long-term trends and climate-induced changes in Ag-Systems.

The dataset used in this study comprises input and output variables for the Daycent model, a process-based ecosystem model designed to simulate carbon (C) and nitrogen (N) dynamics in terrestrial ecosystems. The input variables include weather parameters such as maximum and minimum temperature (TMAX, TMIN), precipitation (PREC), solar radiation (RADN), humidity, and wind speed. Soil properties are represented by parameters such as soil texture (TCSAND, TCSILT), bulk density (TBKDS), total soil carbon (TSOC), and soil moisture characteristics (TFC, TWP, TKSat). Management practices such as fertilization (FERTZR_N), planting and harvesting days (PDOY, HDOY), and crop type (PLANTT) are also included. The dataset incorporates spatial information through latitude, longitude, and elevation values. The output variables primarily capture fluxes (e.g., NEE, gross primary production (GPP), and Ecosystem respiration), crop yield, N cycling (e.g., nitrous oxide emissions (N_2O), ammonium (NH_4), and nitrate (NO_3) at different soil depths), soil water content (SWC), evapotranspiration (ET), and soil temperature (Tsoil). Notably, variables highlighted in red indicate either missing outputs from Daycent simulations or input variables not directly utilized by the model but available for potential enhancements. Blue-labeled inputs signify variables compiled and provided by the Daycent model.

B Transformers

B.1 Preliminaries

A Transformer layer maps an input sequence $X \in \mathbb{R}^{n \times d_{\text{model}}}$ to an output of the same shape via (i) multi-head self-attention (MHSA) and (ii) a position-wise feed-forward network (FFN), each wrapped with residual connections and layer normalization:

$$\tilde{X} = \mathrm{LN}(X + \mathrm{MHSA}(X)), \tag{19}$$

$$Y = \mathrm{LN}\Big(\tilde{X} + \mathrm{FFN}(\tilde{X})\Big). \tag{20}$$

For queries Q, keys K, and values V of shape $n \times d_{\text{model}}$, a single scaled dot-product attention head is

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + B\right) V, \tag{21}$$

where B is an optional bias (e.g., relative position or spatial bias; cf. Sec. B.2) and d_k is the key dimension.

B.2 Positional Encodings for Spatial & Temporal Context

Standard encodings. Let $PE : \mathbb{N} \to \mathbb{R}^{d_{\text{model}}}$ be sinusoidal encodings with frequencies spanning geometric scales:

$$PE_{(2i)}(t) = \sin(t/\omega_i), \qquad PE_{(2i+1)}(t) = \cos(t/\omega_i), \tag{22}$$

with $\omega_i = 10^{4i/d_{\text{model}}}$.

Geospatial extensions. We enrich PE with spatial structure by concatenating multi-resolution features derived from latitude/longitude and ancillary geodata:

$$\begin{aligned} PE_{\text{space}}(t, \text{lat}, \text{lon}) \;=\; \Big[& PE(t) \;\oplus\; g_{\text{wavelet}}(\text{lat}, \text{lon}) \\ & \oplus\; f_{\text{flow}}(\text{lat}, \text{lon}) \;\oplus\; k_{\text{kriging}}(\text{lat}, \text{lon}) \Big] \in \mathbb{R}^{d_{\text{model}}}. \end{aligned} \tag{23}$$

Wavelet features g_{wavelet} capture multi-scale spatial variation by evaluating a set of 2D wavelets (e.g., Haar/Daubechies) on coordinates projected to a suitable planar domain; stacking coefficients across scales yields translation- and scale-aware components.

Flow/Laplacian features f_{flow} encode spatial dependencies via a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ whose nodes are sites and edges represent adjacency (e.g., hydrologic or ecoregional neighbors). Using the (normalized) Laplacian L, we define diffusion coordinates by truncated eigenpairs $LU = U\Lambda$:

$$f_{\text{flow}}(\mathbf{x}) = U_{1:r}(\mathbf{x})\Lambda_{1:r}^{-\alpha}, \quad \alpha \in [0, 1], \tag{24}$$

which provide smooth, low-frequency embeddings aligned with spatial flows.

Kriging/GP features k_{kriging} use a stationary covariance kernel $k_\theta(\mathbf{x}, \mathbf{x}')$ (e.g., Matérn on the sphere with great-circle distance d_g):

$$k_\theta(\mathbf{x}, \mathbf{x}') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\, d_g(\mathbf{x}, \mathbf{x}')}{\rho} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\, d_g(\mathbf{x}, \mathbf{x}')}{\rho} \right). \quad (25)$$

Given anchor points $\{\mathbf{z}_j\}_{j=1}^m$, we form features

$$k_{\text{kriging}}(\mathbf{x}) = \big[k_\theta(\mathbf{x}, \mathbf{z}_1), \ldots, k_\theta(\mathbf{x}, \mathbf{z}_m) \big], \quad (26)$$

which approximate spatial correlations and can be learned end-to-end via θ and the anchor set.

Relative/rotary encodings. To preserve translation equivariance and enable long-context extrapolation, the bias B in Eq. (21) can encode relative distances: $B_{ij} = b(r_{ij})$, with r_{ij} the (signed) temporal offset and/or great-circle distance. Rotary positional encodings (RoPE) apply a complex rotation to (Q, K) that embeds relative offsets implicitly, improving stability for long sequences.

B.3 Multi-Head Attention for Spatial Heterogeneity

We factor attention into h heads, each with its own projection matrices:

$$\text{head}_\ell(Q, K, V) = \text{Attn}\Big(QW_\ell^Q, KW_\ell^K, VW_\ell^V \Big), \quad \ell = 1, \ldots, h, \quad (27)$$

and

$$\text{MHSA}(Q, K, V) = \text{Concat}\big(\text{head}_1, \ldots, \text{head}_h\big) W^O. \quad (28)$$

To capture distinct geographic drivers, we can specialize heads by conditioning the projections on feature groups (elevation/land use/climate), or by gating:

$$\text{head}_\ell^{(\text{type})} = \sigma\big(G_\ell^{(\text{type})}\big) \cdot \text{head}_\ell, \qquad G_\ell^{(\text{type})} = \phi\Big(Z_{\text{type}} W_\ell^G \Big), \quad (29)$$

where Z_{type} summarizes a variable class (e.g., climate normals), ϕ is a small MLP, and σ a sigmoid gate. This yields interpretable, type-aware heads aligned with spatial heterogeneity.

B.4 Complexity & Sparsity

Self-attention has $O(n^2 d)$ time and $O(n^2)$ memory in sequence length n. For gridded spatiotemporal data, locality or block-sparse patterns (e.g., neighborhood or halo attention) reduce cost to $O(nkd)$ with $k \ll n$, while often improving inductive bias by emphasizing local interactions.

B.5 Parameter Settings & Training Guidance

Let d_{model} be the channel width, h heads, $d_k = d_v = d_{\text{model}}/h$, L layers.

- **Default sizes.** For medium-scale geospatial tasks: $d_{\text{model}} \in [256, 512]$, $h \in \{4, 8\}$, $L \in [4, 8]$, FFN width $d_{\text{ff}} \in [4, 8] \times d_{\text{model}}$.
- **Dropout & regularization.** Dropout $p \in [0.1, 0.3]$ on attention and FFN; weight decay 10^{-4}–10^{-2}; stochastic depth (0–0.1) for deeper stacks. Consider *feature-group dropout* to improve robustness across heterogeneous sites.
- **Learning rate schedule.** AdamW with learning rate 2×10^{-4}–5×10^{-4}, warmup (2–5k steps), then cosine decay. Use gradient clipping (1.0).
- **Normalization.** Pre-LN is more stable for deep stacks; consider RMSNorm for reduced variance in heterogeneous feature scales.
- **Initialization.** Use scaled initialization for W^Q, W^K to keep $\|QK^\top\|$ in a numerically stable range; apply head-wise normalization for PE_{space} to avoid dominance by any single block in (23).
- **Masking.** For temporal forecasting, apply causal masks; for spatial grids, optional radius masks reflect finite propagation speeds or domain physics.

B.6 Incorporating Physical Reasoning

Physics-informed inductive biases can be introduced via: (i) *attention biases* B_{ij} that penalize long-range interactions exceeding feasible transport limits; (ii) *constraint losses* (e.g., mass/energy balance) added to the objective; and (iii) *structured encodings* (Sec. B) that embed geodesic distances, hydrologic connectivity, or ecoregional partitions. For a mass-balance index $M(\cdot)$, a soft constraint

$$\mathcal{L}_{\text{phys}} = \lambda \left\| M(\hat{y}) - M(y) \right\|_1 \quad (30)$$

encourages consistency without hard constraints.

B.7 Interpretability

Head-wise attributions follow from $\text{softmax}(QK^\top/\sqrt{d_k})$ maps; aggregating by feature group (elevation/land/climate) yields spatially meaningful explanations. Attention rollout and gradient-based saliency on PE_{space} help diagnose which geospatial scales and neighborhoods drive predictions.

B.8 Practical Recipe (Summary)

Use PE_{space} (Eq. 23) with (a) low-frequency Laplacian coordinates, (b) a modest set of GP anchors for kriging features, and (c) a few wavelet scales. Start with d_{model}=384, h=6, L=6, d_{ff}=4d_{model}, dropout 0.1, AdamW 3×10^{-4}, warmup 3k, cosine decay, and block-sparse spatial attention with neighborhood size $k \approx 64$. Calibrate uncertainty with MC dropout or deep ensembles, and report confidence intervals alongside standard metrics.

Table 3: Hyperparameter Settings for Hybrid-CNN

Hyperparameter	Setting Value	Reference Range
Kernel Size	3×3	$3 \sim 7$
Number of Filters	64	$32 \sim 256$
Convolutional Blocks	4	$2 \sim 6$
Residual Connections	Yes	–
Pooling Type	Max Pooling	Max / Avg
Dropout Rate	0.2	$0.1 \sim 0.5$
Batch Size	256	≥ 64
Learning Rate	0.001	$10^{-4} \sim 10^{-2}$
Activation Function	ReLU	ReLU / GELU
Input Length	180	$100 \sim 200$
Optimizer	AdamW	Adam / SGD / RMSProp