CO_2-Net: A Physics-Informed Spatio-Temporal Model for Global Surface CO_2 Reconstruction

Hao Zheng[1] *   Yuting Zheng[1] *   Hanbo Huang[1]   Chaofan Sun[1]   Enhui Liao[1]
Lin Liu[2]   Yi Han[2]   Hao Zhou[2]   Shiyu Liang[1] †
[1]Shanghai Jiaotong University, China   [2]National Univeristy of Science and Technology, China
{hubert.zheng, zhengyt058, hhuang417, scf024, ehliao, lsy18602808513}@sjtu.edu.cn,
{liulin16, hanyi12, zhouhao23a}@nudt.edu.cn

Abstract

Reconstructing atmospheric surface CO_2 is crucial for understanding climate dynamics and informing global mitigation strategies. Traditional inversion models achieve precise global CO_2 reconstruction but rely heavily on uncertain prior estimates of fluxes and emissions. Inspired by recent advances in data-driven weather forecasting, we explore whether data-driven models can reduce reliance on these priors. However, CO_2 reconstruction presents unique challenges, including complex spatio-temporal dynamics, periodic patterns and sparse observations. We propose CO_2-Net, a data-driven model that addresses these challenges without requiring extensive prior data. We formulate CO_2 reconstruction as solving a constrained advection-diffusion equation and derive three key components: physics-informed spatio-temporal factorization for capturing complex transport dynamics, wind-based embeddings for modeling periodic variations and a semi-supervised loss for integrating sparse CO_2 observations with dense meteorological data. CO_2-Net is designed in three sizes—small (S), base (B) and large (L)—to balance performance and efficiency. On CMIP6 reanalysis data, CO_2-Net (S) and (L) reduce RMSE by 11% and 71%, respectively, when compared to the best data-driven baseline. On real observations, CO_2-Net (L) achieves RMSE comparable to inversion models. The ablation study shows that the effectiveness of wind-based embedding and semi-supervised loss stems from their compatibility with our spatio-temporal factorization. Code is available at https://github.com/Leamonz/CORE.

1. Introduction

Atmospheric surface CO_2 is a primary driver of climate change, contributing significantly to global warming through the greenhouse effect. Comprehensive surface CO_2 data is critical for understanding the carbon cycle, improving climate predictions and supporting mitigation strategies such as carbon capture, emissions reduction and renewable energy, aligned with international climate goals like the Paris Agreement. However, historical surface CO_2 observations are limited due to sparse monitoring networks like the NOAA Federated Aerosol Network [2]. While recent satellite data and related reconstruction methods provide detailed CO_2 records for the past decade, reconstructing atmospheric surface CO_2 before the satellite era relies on limited station data. Inverse modeling approaches, including Bayesian synthesis [33], variational methods [21] and Kalman filters [32], address the sparsity by combining sparse CO_2 observations with auxiliary meteorological variables (e.g., temperature, wind, surface pressure) within physical transport models. While accurate, these methods are constrained by their reliance on uncertain prior estimates (e.g., fluxes and emissions) and are computationally expensive, limiting their scalability for real-time applications.

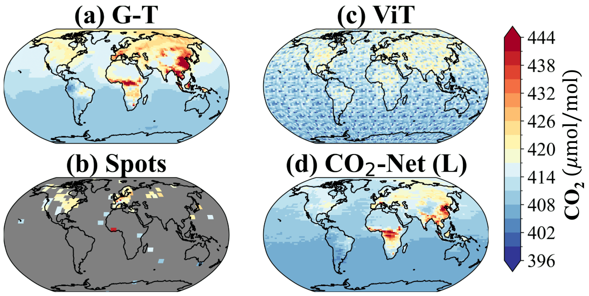

Figure 1. CO_2 concentrations (μmol/mol): (a) Ground-truth; (b) Sparse direct observations; (c) Output of ViT; (d) Output of our CO_2-Net (L). Our CO_2-Net (L) outperforms ViT in ground-based CO_2 reconstruction, and well captures CO_2 dynamics.

Recently, deep neural networks have made significant advancements in climate modeling. Models such as FourCastNet [31], ClimaX[29], ClimODE [42], Pangu [3] and Aurora [5] achieve error rates comparable to traditional medium-range weather prediction systems while offering substantial

[*] Equal contribution.
[†] Corresponding author.

improvements in computational efficiency. Despite these advances, data-driven approaches for reconstructing global surface CO_2 concentrations remain relatively underexplored. This gap raises a critical question: *Can data-driven methods reconstruct global surface CO_2 concentrations with error rates comparable to inversion models, within the same order of magnitude, without relying on extensive prior data?*

This reconstruction problem can be framed as an image inpainting task, where missing data needs to be filled. However, it introduces three unique challenges that are not typically encountered in conventional inpainting tasks. First, capturing the *spatio-temporal dynamics* of CO_2 concentration is inherently difficult. While spatio-temporal factorization techniques have been explored in video understanding [19, 43], it remains unclear how to effectively adapt them to accurately model atmospheric dynamics. Second, atmospheric patterns exhibit strong *temporal periodicity*, including seasonal (e.g., summer-winter) and diurnal (day-night) cycles, as well as *spatial periodicity* driven by phenomena like trade winds. Prior work [22, 42] incorporated spatial and temporal embeddings to model these periodic signals. However, how these embeddings improve the model's ability to capture atmospheric dynamics remains unclear. Third, CO_2 observations are *sparse* (Figure 1(b)), with limited spatial coverage (covering 0.7% of global surface) and a relatively short time span (collected after 2000). Although self-supervised training methods have shown promise in addressing sparsity in computer vision [13, 41], their application to CO_2 reconstruction remains unclear. The key challenge lies in identifying appropriate hidden representations and selecting input variables that adhere to physical principles to design an effective self-supervised reconstruction loss.

1.1. Our Contribution

In this paper, we propose CO_2-Net, a physics-informed spatio-temporal model designed specifically for CO_2 reconstruction with three key components.

CO_2-Net employs a **spatio-temporal factorization** inspired by the linear superposition of PDE solutions. We frame the reconstruction problem as a constrained advection-diffusion equation, decomposing it into a time-invariant particular solution and a time-varying homogeneous solution. Using this decomposition, we apply a spatial expert to model the particular solution, capturing spatial distribution of CO_2, and a temporal expert to model the homogeneous solution, capturing the temporal dynamics of CO_2 and its interactions with other variables. This design enables CO_2-Net to capture complex spatio-temporal dynamics.

CO_2-Net is equipped with **wind-flow based spatio-temporal embeddings** designed to capture periodic patterns in atmospheric dynamics. We theoretically prove that the homogeneous solution can be uniquely represented as a series of spatio-temporal basis functions. These functions guide the design of periodic embeddings, enabling the model to effectively represent periodic variations.

CO_2-Net incorporates a **semi-supervised loss** to address the challenge of sparse CO_2 observations while ensuring physically grounded reconstruction. The novel aspect lies in identifying the appropriate hidden representation to reconstruct the relevant input features, which enables the design of a self-supervised loss. We prove that the wind field can be exactly recovered from the homogeneous solution. Based on this, we leverage the hidden representations of the temporal expert to reconstruct the wind field, ensuring that the model outputs adhere to physical principles and effectively mitigate the sparsity issue in the CO_2 observation.

We develop CO_2-Net in three sizes, ranging from 38M to 247M, to balance performance and efficiency. Extensive experiments on reanalysis data show that CO_2-Net (S) achieves lower RMSE than all baselines, while CO_2-Net (L) further reduces it significantly, with reductions of 11% and 71%, respectively, compared to the best data-driven baseline, Vision Transformer (ViT), as shown in Figure 1(c) and (d). On real observation data, CO_2-Net (L) achieves an RMSE comparable to inversion models. The ablation study shows that the effectiveness of wind-based embedding and semi-supervised loss stems from their compatibility with our spatio-temporal factorization, highlighting the crucial interplay between model architecture, embedding design and self-supervised task formulation in CO_2 reconstruction.

2. Preliminary

Notations. Let S^2 denote the unit sphere in $\mathbb{R}^3$, parameterized with the latitude-longitude grid $(\theta, \phi) \in \Omega = [-\frac{\pi}{2}, \frac{\pi}{2}] \times [-\pi, \pi]$. For a time-dependent function $\varphi(\theta, \phi, t)$, we write $\dot{\varphi} = \frac{\partial \varphi}{\partial t}$. The symbol "$\cdot$" denotes the standard inner product on $\mathbb{R}^3$. The spherical gradient ∇ on S^2 is given in (θ, ϕ)-coordinates by $\nabla \varphi = \frac{\partial \varphi}{\partial \theta}\, \hat{e}_\theta + \frac{1}{\cos\theta} \frac{\partial \varphi}{\partial \phi}\, \hat{e}_\phi$, where $\hat{e}_\theta$ and $\hat{e}_\phi$ are the unit tangent vectors in the θ- and ϕ-directions, respectively. The divergence on S^2 is denoted by $\nabla \cdot$, and the spherical Laplacian is denoted by ∇^2.

Atmospheric Data. Reconstructing CO_2 concentrations requires integrating sparse CO_2 observations with auxiliary data, such as temperature, humidity, and pressure, which are densely sampled via satellites and sensor networks. While auxiliary variables provide near-complete spatio-temporal coverage, CO_2 data remain sparse and unevenly distributed, collected primarily from ground-based stations [46]. Prior studies [6, 20, 49, 50] demonstrate that leveraging the dense observation of auxiliary variables and the shared spatio-temporal dynamics between these variables and CO_2 enhances reconstruction precision. Building on these insights, we assume full observations of the wind field $\mathbf{w}$ and auxiliary variables $\Phi_1, \ldots, \Phi_K$ over the spatio-temporal grid $\Omega \times [0, T]$, while CO_2 data are limited to specific ground-based locations, denoted by $\mathcal{Z} \subset \Omega$.

Atmospheric Advection-Diffusion Dynamics describe the spatiotemporal transport and diffusion of CO_2 concentration, $\varphi(\theta, \phi, t)$, and K auxiliary variables, $\mathbf{\Phi} = (\Phi_1, \dots, \Phi_K)$, driven by the wind field $\mathbf{w}(\theta, \phi, t)$. These processes are governed by the advection-diffusion equation, capturing the interactions of advection, diffusion, and external sources. The dynamics are represented by the linear differential operator:

$$\mathcal{L}_C[\varphi] = \frac{\partial \varphi}{\partial t} + (\mathbf{w} \cdot \nabla)\varphi - C\nabla^2\varphi,$$

where the diffusion coefficient C varies dynamically in space and time. The governing equations for CO_2 and auxiliary variables are expressed as: $\mathcal{L}_D[\varphi] = s, \mathcal{L}_{D_k}[\Phi_k] = s_k, k = 1, \dots, K$, where D and D_k are the respective diffusion coefficients for φ and Φ_k. The terms $s(\theta, \phi, t)$ and $s_k(\theta, \phi, t)$ denote external influences, such as radiation or chemical transformations. Notably, all diffusion coefficients and source terms exhibit dynamic variability in both space and time.

Problem Formulation. The goal is to reconstruct global CO_2 concentrations from sparse observations by solving the following constrained partial differential equation (PDE):

$$\mathcal{L}_D[\varphi] = s, \quad \text{subject to } \varphi\big|_{\mathcal{Z}\times[0,T]} = f\big|_{\mathcal{Z}\times[0,T]}, \tag{1}$$

where f denotes the observed CO_2 concentrations and $f\big|_{\mathcal{Z}\times[0,T]}$ denotes its restriction to the subset $\mathcal{Z} \times [0, T]$. To address the difficulty of solving this PDE, we reformulate it as an optimization problem to minimize the mean squared error between observed and predicted CO_2 concentrations:

$$\min_F \mathcal{L}_{\text{supv.}}(F) \triangleq \|\varphi - F[\mathbf{w}, \mathbf{\Phi}, f]\|_2^2, \tag{2}$$

where the functional norm $\|\cdot\|_2$ denotes the L^2 norm over $\Omega \times [0, T]$. Here, $F[\mathbf{w}, \mathbf{\Phi}, f]$ is a reconstruction model mapping the wind field $\mathbf{w}$, auxiliary variables $\mathbf{\Phi}$, and partial CO_2 observations f to an approximation of the full CO_2 concentration φ.

3. Our Approach: CO_2-Net

3.1. Physics-based Spatiotemporal Factorization

To design an appropriate spatio-temporal factorization for CO_2 reconstruction, we analyze the solution of the constrained PDE in Eq. (1) over a short time interval $[\tau, \tau + \Delta t]$. The following proposition establishes that the general solution comprises two components: a time-invariant particular solution and a time-varying homogeneous solution.

Proposition 1 (**Linear Superposition of Solutions in PDE**). *Let φ_{part} be the particular solution of the equation, satisfying $\mathcal{L}_D[\varphi_{part}] = s$. Let φ_{homo} be the homogeneous solution, satisfying $\mathcal{L}_D[\varphi_{homo}] = 0$ under the constraint $\varphi_{homo} = f - \varphi_{part}$ on $\mathcal{Z} \times [0, T]$. Then, the general solution to Eq.* (1) *is $\varphi_{general} = \varphi_{homo} + \varphi_{part}$. Moreover, if the wind field, source and diffusion coefficient are constant over the interval $[\tau, \tau + \Delta t]$, the particular solution φ_{part} becomes time-invariant, depending only on the initial time τ.*

Remark: The proof in Appendix A follows directly from the linearity of the advection-diffusion operator $\mathcal{L}_C$ and the principle of linear superposition for PDE solutions.

Inspired by this, we introduce a *spatial expert* to model the particular solution and a *temporal expert* to model the homogeneous solution, with an MLP fusing the two experts to produce the final prediction. While CAST [19], a model designed for action recognition in videos, employs a similar spatio-temporal factorization, our approach is derived from underlying physical principles.

Spatial Expert. The spatial expert models the time-invariant particular solution φ_{part} by capturing spatial correlations within individual frames of the input sequence. As illustrated in Figure 2, it consists of a spatial tokenizer and a L-layer ViT. The input is a tensor of shape $B \times (K + 1 + E) \times \Delta t \times H \times W$, where B is the batch size, $K + 1 + E$ is the number of channels (including CO_2, K auxiliary variables and E additional spatio-temporal embeddings), Δt is the number of input frames, and $H \times W$ denote the spatial dimensions. The spatial expert processes even frames of the input sequence, mixing them across the sequence to discard inter-frame features. This ensures a focus solely on spatial correlations within individual frames. The spatial tokenizer divides each frame into $N = \frac{HW}{pq}$ non-overlapping patches of size $p \times q$, flattening each patch into tokens of dimension D, resulting in a tensor of shape $(B \cdot \frac{\Delta t}{2}) \times N \times D$. The L-layer ViT then processes these tokens to learn intra-frame spatial correlations while maintaining the tensor shape. The spatial expert outputs h_{spatial}, a hidden representation encoding spatial dependencies.

Temporal Expert. The temporal expert models the time-varying homogeneous solution φ_{homo} by capturing inter-frame dynamics in the input sequence. It shares the same input tensor as the spatial expert and comprises a temporal tokenizer and a L-layer ViT. Unlike the spatial expert, it tokenizes the sequence into $\frac{\Delta t}{2} \cdot N$ non-overlapping tubelets of size $2 \times p \times q$ without subsampling or rearrangement, preserving inter-frame relationships. These tubelets are projected into tokens of dimension D, resulting in a tensor of shape $B \times (\frac{\Delta t}{2} \cdot N) \times D$. The tokens are then processed by the L-layer ViT to model temporal correlations and inter-frame dynamics. The final output is a hidden representation, h_{temporal}, which encodes the temporal dependencies.

Output Fusion. The spatial and temporal expert outputs, h_{spatial} and h_{temporal}, are fused via element-wise addition: $h_{\text{fusion}} = h_{\text{spatial}} + h_{\text{temporal}}$. This fused representation is processed by a three-layer MLP with hidden dimensions of 768, and the final layer maps the output to the target dimension $H \times W$, producing the prediction.

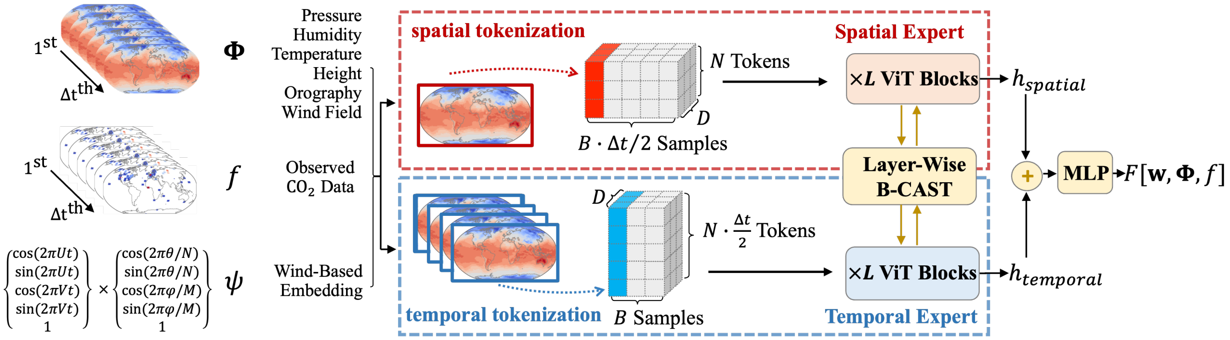

Figure 2. **Overview of CO$_2$-Net.** CO$_2$-Net integrates a spatial expert and a temporal expert, each consists of L layer ViT blocks. The two experts interact via the layer-wise B-CAST module, strengthening spatio-temporal understanding. Hidden representations from the two experts are fused through a three-layer MLP to generate the final prediction.

3.2. Wind-Flow based Spatio-Temporal Embedding

In this subsection, we design spatio-temporal embeddings grounded in the physical dynamics of the advection-diffusion equation. We prove that the homogeneous solution of the constrained equation is unique and explicitly represented as wave components in spatial (latitude and longitude) and temporal dimensions. Under a zonal flow model, we derive these components analytically, capturing periodicity and directional atmospheric CO$_2$ dynamics. These insights guide the design of embeddings that incorporate periodicity and transport dynamics for modeling atmospheric processes.

We formalize these ideas in the following theorem, which establishes the existence, uniqueness, and wave-based representation of the homogeneous solution to the constrained advection-diffusion equation.

Theorem 2 **(Existence and Uniqueness of the Real Analytical Solution).** *Let $\mathcal{Z} \subset S^2$ be a non-empty open subset. Let source s and coefficient ν be two real analytic functions defined on $S^2 \times [0, T]$. Let $f(\cdot, t)$ be real analytic on $\mathcal{Z}$ for each t. The PDE $\mathcal{L}_\nu[\varphi] = s$, subject to $\varphi = f$ on $\mathcal{Z} \times \{0\}$, admits a unique analytical solution,*

$$\varphi(\theta, \phi, t) = \sum_{n \geq 1} \sum_{|m| \leq n} A_{n,m}(t) e^{im\phi} P_n^m(\cos\theta), \quad (3)$$

where $A_{n,m}(t)$ are time-dependent coefficients and P_n^m are associated Legendre functions. Furthermore, assuming a constant zonal wind flow $\mathbf{w} = (0, U)$, a time-invariant source term s, and a constant diffusion coefficient ν, the expansion coefficients $A_{n,m}(t)$ take the form $A_{n,m}(t) = B_{n,m} e^{-\nu n(n+1)t + imUt} + C_{n,m}$, where $B_{n,m}$ and $C_{n,m}$ are time-independent constants determined by the initial condition and the source term.

Remark: The proof, shown in Appendix B, establishes the existence and uniqueness of the solution using the Cauchy–Kovalevskaya theorem. However, a direct application is not possible, since the initial condition is given only on an open subset $\mathcal{Z}$, instead of S^2. Therefore, we first use the identity theorem for real-analytic functions to show that the initial condition is uniquely determined and real-analytic, allowing us to apply the Cauchy–Kovalevskaya theorem rigorously.

By this theorem, the unique solution is a power series of the basis functions 1, $e^{i\phi}$, $e^{i\theta}$, e^{iUt}, and e^{iVt}, capturing periodic patterns in time and space. Using Euler's formula, we define the temporal embedding as

$$\psi(t) = \{\cos(2\pi Ut), \sin(2\pi Ut), \cos(2\pi Vt), \sin(2\pi Vt), 1\},$$

where U and V denote the zonal (east-west) and meridional (north-south) wind speeds, respectively. The spatial terms $e^{i\phi}$ and $e^{i\theta}$ can be expressed using trigonometric functions,

$$\psi(\theta, \phi) = \left\{\cos\frac{2\pi\theta}{N}, \sin\frac{2\pi\theta}{N}, 1\right\} \times \left\{\cos\frac{2\pi\phi}{M}, \sin\frac{2\pi\phi}{M}, 1\right\},$$

where N and M denote the grid resolutions for latitude and longitude. Combining the temporal terms and spatial terms leads to the **joint spatio-temporal embedding:** $\psi(\theta, \phi, t) = \psi(\theta, \phi) \times \psi(t)$. Following ClimODE [42], we adopt the same spatial embeddings but incorporate wind-flow dynamics into the temporal embeddings, replacing its daily and yearly approach for improved CO$_2$ reconstruction. In Appendix D, we present an additional theorem demonstrating that, with joint spatio-temporal embedding, the solution in Eq. (3) can be implemented as a neural network with an architecture closely resembling CO$_2$-Net.

3.3. Layer-wise Spatio-Temporal Connection

Prior research [19, 37] has shown that exchanging hidden representations between spatial and temporal experts in early layers significantly improves model performance compared to simple summation at the output layer. Lee et al. [19] introduced the Bottleneck Cross-Attention in Space and Time (B-CAST) module, which facilitates interaction between spatial and temporal models through early-layer representations, enabling more balanced spatio-temporal learning. Similarly, Skean et al. [37] demonstrated that embeddings from intermediate layers are more effective for downstream tasks than those from the final layer. Building on these findings, we adopt a layer-wise connection module based on B-CAST to efficiently exchange representations between the spatial and temporal experts, enhancing spatio-temporal understanding.

3.4. Semi-supervised Physics-informed Loss

Traditional supervised learning minimizes the loss $\mathcal{L}_{\text{supv.}}$ to fit reconstructed outputs to ground truth labels. While effective in vision tasks with complete inputs, such as classification [28, 51] and recognition [45, 47], this approach is inadequate for CO_2 reconstruction due to observation sparsity, which can impede training. Inspired by the effectiveness of self-supervised learning in mitigating data sparsity [7, 13, 41], we leverage hidden representations from intermediate layers for self-supervised reconstruction.

A key challenge in self-supervised reconstruction is selecting appropriate hidden representations and defining suitable reconstruction objectives [8, 15]. This challenge intensifies with multiple input variables and the involvement of spatiotemporal experts, making it unclear which variables to reconstruct and which representations to use. To address this, we propose a theorem demonstrating that the wind field can be precisely reconstructed from the temporal expert when it corresponds to the homogeneous solution, providing insights into the design of self-supervised objectives.

Theorem 3. *Let $\varphi : S^2 \times [0, T] \to \mathbb{R}$ be a smooth function with non-vanishing gradient and satisfy $\frac{\partial \varphi}{\partial t} + (\mathbf{w} \cdot \nabla)\varphi = D\nabla^2\varphi$, where $\mathbf{w}$ is a non-zero time-dependent smooth vector field on the sphere S^2. Then $\mathbf{w}$ is uniquely determined by φ.*

Remark: The proof in Appendix C uses the fact that if $\mathbf{v}$ satisfies $(\mathbf{v} \cdot \nabla)\varphi = 0$ in a domain where $\nabla\varphi$ is non-vanishing, then $\mathbf{v} = 0$. Hence, if the homogeneous solution φ has a non-vanishing gradient in a domain, this theorem guarantees that the wind field $\mathbf{w}$ is uniquely determined.

Based on this result, we define the wind field reconstruction loss as a self-supervised objective. We implement it using a linear perceptron applied to the output representation of the temporal expert, h_{temporal}. The loss is given by

$$\mathcal{L}_{\text{wind}} = \min_{W} \|\mathbf{w} - W \cdot h_{\text{temporal}}[\mathbf{w}, \mathbf{\Phi}, f]\|_2^2,$$

where $\mathbf{w}$ is the input wind field and W is the weight matrix of the reconstruction module. The norm $\|\cdot\|_2$ denotes the L^2 norm over $\Omega \times [0, T]$. The final **semi-supervised loss** is defined as the weighted sum of $\mathcal{L}_{\text{wind}}$ and the supervised loss $\mathcal{L}_{\text{supv.}}$ defined in Eq. (2), i.e., $\mathcal{L}_{\text{semi-supv.}} = \mathcal{L}_{\text{supv.}} + \lambda \cdot \mathcal{L}_{\text{wind}}$, where λ balances the two terms. This loss ensures alignment with supervised labels while enforcing physical constraints from wind-driven dynamics.

4. Experiment

In this section, we conduct experiments to address the following research questions:

Table 1. Overview of datasets.

Dataset	Auxiliary Variables							CO_2 Data $\mu\text{mol}/\text{mol}$
	PS (Pa)	HUSS (%)	TAS (K)	UAS (m/s)	VAS (m/s)	GPH (m)	ORO (m)	
CarbonTracker(CT)	✓	✓	✓	✓	✓	✓	✓	Spot Records
CMIP6	✓	✓	✓	✓	✓	×	×	Monthly Avg.

Table 2. Overview of dataset divisions by year.

Dataset	Train	Validation	Test
CarbonTracker(CT)	2002-2018	2001, 2019	2000, 2020
CMIP6	1865-1999	1860-1864, 2000-2004	1850-1859, 2005-2014

RQ1: How does CO_2-Net compare in (1) global CO_2 reconstruction accuracy against baselines and (2) spot reconstruction accuracy against inversion models?

RQ2: Does CO_2-Net perform better in reconstructing both long-term global dynamics and short-term local events?

RQ3: How do spatio-temporal factorization, wind-based embedding, layer-wise connections and the semi-supervised loss function individually enhance performance?

4.1. Experimental Setups

In this subsection, we introduce the experimental setups, including the data, baselines, evaluation metrics and implementation details. We provide details in Appendix E.

Reanalysis Data. We use two datasets, CarbonTracker (CT) [16] and CMIP6 [11], both with a spatial resolution of $2° \times 3°$. We use the CT[†] released in 2022, and for CMIP6, we use data from historical experiments conducted by the CanESM5[†] model. CT provides high temporal resolution data at 3-hour intervals, while CMIP6 offers coarser, monthly averaged data. CT includes seven surface variables (Table 1), such as surface air pressure (PS), specific humidity (HUSS), air temperature (TAS), and wind components (UAS, VAS), while CMIP6 excludes geopotential height (GPH) and orography (ORO). Dataset splits are shown in Table 2, with validation and testing sets covering the first and last years to evaluate long-term CO_2 trends.

Real Observation Data. We utilize real atmospheric CO_2 observation data from GLOBALVIEWplus (GV+) [36], provided by NOAA. GV+ comprises global CO_2 measurements collected from a network of surface spots worldwide. It includes 3-hour interval CO_2 measurements from 96 surface spots. We randomly select 72 spots and employ observations from these spots for training, while data from the remaining 24 spots are used to assess the capability of the model to reconstruct actual CO_2 concentration and compare its performance with inversion models.

Baselines. We evaluate three types of baselines: inversion, numerical and data-driven models. Inversion models include 4D-Variation [16] and COLA [23]. Numerical models include Kriging interpolation [30] with spherical and

[†] https://gml.noaa.gov/ccgg/carbontracker/

[†] https://climate-scenarios.canada.ca/?page=cmip6-scenarios

Table 3. RMSE(↓, in $\mu\text{mol}/\text{mol}$) and ACC (↑) comparison of different models across datasets and regions. $*$ indicates spatio-temporal reconstruction models, while others are static reconstruction models. AS-EU represents Asia-Europe. Details are presented in Appendix F.3.

Dataset	Methods	Sizes (M)	Global		Ocean		AS-EU	
			RMSE	ACC	RMSE	ACC	RMSE	ACC
CMIP6	Senseiver	0.11	36.47$_{\pm 6.81}$	0.59$_{\pm 0.16}$	31.94	0.79	40.52	0.43
	SwinLSTM*	3.3	27.42$_{\pm 0.09}$	0.52$_{\pm 0.01}$	28.41	0.52	27.70	0.51
	CycleGAN	28	51.48$_{\pm 0.36}$	0.12$_{\pm 0.04}$	53.34	0.12	52.01	0.11
	ViT	76	18.18$_{\pm 1.01}$	0.49$_{\pm 0.02}$	18.86	0.49	18.38	0.49
	CO_2-Net* (S)	38	16.18$_{\pm 1.19}$	0.73$_{\pm 0.07}$	22.48	0.72	16.27	0.79
	CO_2-Net* (B)	95	9.91$_{\pm 0.93}$	0.98$_{\pm 0.03}$	10.27	**0.99**	9.91	**0.98**
	CO_2-Net* (L)	247	**5.31**$_{\pm 0.49}$	**0.99**$_{\pm 0.00}$	**5.49**	**0.99**	**5.36**	0.97
CT	Spherical	–	7.41$_{\pm 0.40}$	0.12$_{\pm 0.01}$	5.55	0.14	9.80	0.21
	Exponential	–	7.40$_{\pm 0.37}$	0.11$_{\pm 0.01}$	5.56	0.13	9.83	0.20
	Senseiver	0.11	6.39$_{\pm 0.45}$	0.43$_{\pm 0.10}$	4.74	0.28	9.70	0.50
	SwinLSTM*	3.3	5.31$_{\pm 0.01}$	0.61$_{\pm 0.00}$	2.90	0.45	6.33	0.65
	CycleGAN	28	4.70$_{\pm 0.01}$	0.71$_{\pm 0.00}$	1.57	0.88	8.39	0.73
	ViT	76	5.42$_{\pm 0.21}$	0.50$_{\pm 0.04}$	2.27	0.72	9.43	0.63
	CO_2-Net* (S)	38	3.59$_{\pm 0.06}$	0.72$_{\pm 0.01}$	1.32	0.84	6.08	0.72
	CO_2-Net* (B)	95	3.41$_{\pm 0.04}$	0.77$_{\pm 0.01}$	**1.03**	0.90	5.58	0.77
	CO_2-Net* (L)	247	**3.36**$_{\pm 0.04}$	**0.85**$_{\pm 0.00}$	1.09	**0.94**	**5.39**	**0.85**

Table 4. RMSE(↓) in $\mu\text{mol}/\text{mol}$ comparison on real observations.

Methods	4D-Var.	COLA	Sphe.	Expo.	Sens.	ViT	Cycle.	Swin.	CO_2-Net (L)
RMSE	**7.80**	7.94	8.46	8.40	8.75	9.10	9.45	8.34	7.81

exponential variogram models. Data-driven baselines include Vision Transformer (ViT) [1], SwinLSTM [40], CycleGAN [34], and the Implicit Neural Representation (INR) model Senseiver [35].

Metrics. We use latitude-weighted Root Mean Squared Error (RMSE) with unit $\mu\text{mol}/\text{mol}$, and Anomaly Correlation Coefficient (ACC) to evaluate model performance as in ClimODE [42]. Latitude-weighted RMSE measures reconstruction accuracy while accounting for Earth's curvature, whereas ACC evaluates degree of consistency between model reconstruction and groud-truth with respect to their anomaly patterns. Lower RMSE and higher ACC indicate better performance.

Implementation Details. CO_2-Net is implemented in PyTorch, based on the CAST codebase [19], with all parameters trained from scratch. The model inputs include spot CO_2 data, auxiliary variables and wind-based embeddings. Both spatial and temporal experts incorporate a L-layer ViT as backbone. We implement CO_2-Net in three sizes: small (S), base (B) and large (L), with L set to $1, 4, 12$, respectively. A three-layer MLP with a hidden dimension of 768 serves as the output head. We use AdamW optimizer [24] (momentum betas 0.9 and 0.999) and adopt cosine annealing for learning rate scheduling. We identify an optimal combination of learning rate and weight decay through grid search within the sets $\{5e-5, 1e-4, 3e-4, 5e-4, 8e-4\}$ and $\{1e-4, 5e-4, 8e-4, 1e-3\}$, respectively. We find that the best learning rate and weight decay are $5e-4$ and $8e-4$. The model is trained for 250 epochs by default. Experiments are conducted on two NVIDIA RTX 6000 Ada GPUs with a batch size of 32 per GPU.

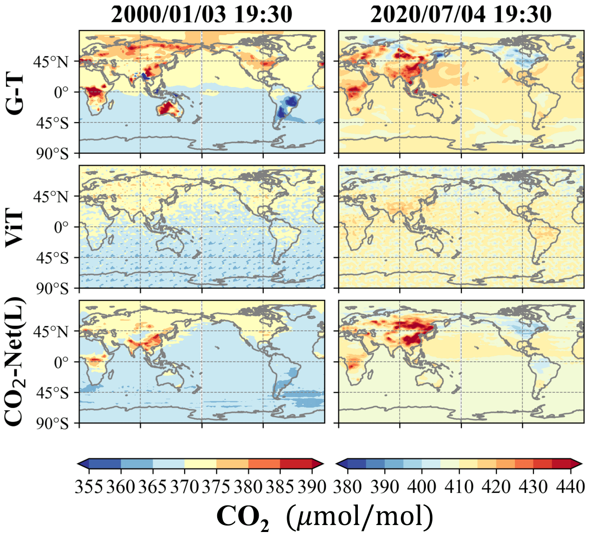

Figure 3. Ground-truth and model reconstructed results of CO_2 concentration ($\mu\text{mol}/\text{mol}$) on CarbonTracker in 2000 and 2020.

4.2. Global CO_2 Reconstruction Results (RQ1)

We compare the global and real observation reconstruction performance of CO_2-Net with baseline methods, as summarized in Table 3 and Table 4.

Obs 1: On both reanalysis datasets, CO_2-Net outperforms all baselines in global CO_2 reconstruction, achieving the lowest RMSE and highest ACC (Table 3). We observe that even the CO_2-Net (S) surpasses all baselines, while CO_2-Net (L) further reduces RMSE to 5.31 $\mu\text{mol}/\text{mol}$ on CMIP6, representing a 71% decrease compared to the best baseline. Computational costs, including time per epoch and VRAM usage per iteration, are provided in Appendix F.1. Compared to the baselines, CO_2-Net (S) achieves a lower RMSE while maintaining comparable computational efficiency. Larger CO_2-Net sizes improve performance, highlighting the effectiveness and scalability of our architecture.

Furthermore, the RMSE reduction is more significant on CMIP6 compared to CT. We hypothesize that this difference is due to CMIP6's longer time span of 165 years and coarser temporal resolution with monthly averages, resulting in greater data fluctuations. Additionally, our model achieves the lowest RMSE for Asia-Europe region and the Ocean, yet a notable disparity exists between these areas. We attribute this disparity to the relatively low spatial variation of CO_2 over oceans, whereas industrial emissions over land increase regional deviations.

Obs 2: On real observation data, CO_2-Net achieves RMSE comparable to inversion models and outperforms all data-driven baselines (Table 4). On the GLOBALVIEW-plus dataset, CO_2-Net (L) achieves a reconstruction RMSE of 7.81 $\mu\text{mol}/\text{mol}$, on par with the inversion model 4D-Var (7.80 $\mu\text{mol}/\text{mol}$). In comparison, data-driven baselines like ViT and CycleGAN exhibit RMSEs 16.7% and 21% higher

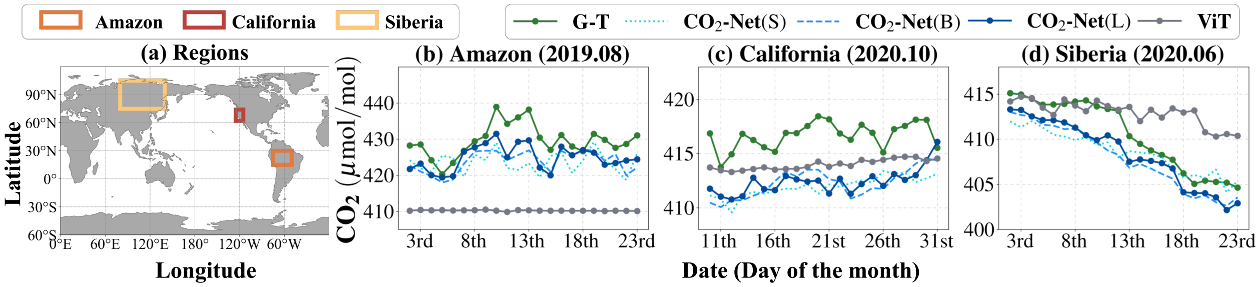

Figure 4. Ground-truth and reconstructed results of CO_2 concentration in local regions where wildfires or forest carbon absorption appears.

Table 5. **Ablation study (RMSE (↓) in $\mu\mathrm{mol/mol}$ |ACC (↑)).** (a) Impact of the physics-informed spatio-temporal factorization and layer-wise connections; (b) Impact of the wind-flow based spatio-temporal embedding; (c) Impact of the semi-supervised loss function.

(a) **S-T Factorization and Connection.**
(Settings: w/ wind-based emb., w/o semi-sup. loss)

Dataset	CT	CMIP6
spatial only	3.51 |0.84	14.28 |0.96
temporal only	3.51 |**0.85**	16.63 |0.99
S-T (w/o layer con.)	3.45 |**0.85**	13.01 |**1.00**
S-T (w/ layer con.)	**3.43** |0.84	**10.83** |0.98

(b) **Wind-Flow based Embedding.**
(Settings: CMIP6 dataset, w/o semi-sup. loss)

Embedding	Wind	ClimODE	None
Cycle-GAN	28.43 |**0.92**	29.27 |0.85	**28.36** |0.74
SwinLSTM	27.80 |0.50	27.97 |0.40	**27.42** |**0.52**
ViT	21.51 |0.43	22.16 |**0.50**	**18.18** |0.49
CO_2-Net (L)	**10.83** |**0.98**	27.06 |0.78	23.40 |0.98

(c) **Semi-supervised Loss**
(Settings: CMIP6 dataset, w/ wind-based emb.)

Semi-sup. loss	w/	w/o
Cycle-GAN	28.45 |**0.94**	**28.43** |0.92
SwinLSTM	28.22 |0.48	**27.80** |**0.50**
ViT	**19.78** |**0.54**	21.51 |0.43
CO_2-Net (L)	**5.31** |**0.99**	10.83 |0.98

than CO_2-Net (L), respectively. Additional results of CO_2-Net (S) and (B) are provided in Appendix F.3.

4.3. Reconstructing Long-Term Global Dynamics and Short-Term Local Events (RQ2)

We evaluate CO_2-Net in reconstructing long-term global dynamics and short-term local events, comparing its performance with ViT.

Obs3: CO_2-Net (L) reconstructs long-term global warming trends more precisely, capturing peak and bottom values accurately (Figure 3). Global warming, driven by rising CO_2 levels, is evident in the increasing CO_2 concentrations from 2000 to 2020. Both models capture this trend, but CO_2-Net (L) achieves more precise reconstruction. On the CT dataset, it closely matches peak ground-truth values in East Asia and Central Africa (January 2000, July 2020), where ViT underestimates them. It also better reconstructs low values in South America (January 2000) and North America (July 2020) than ViT. Additional results of CO_2-Net (S) and (B) are provided in Appendix F.3.

Obs4: CO_2-Net (L) outperforms baseline models in tracking rapid CO_2 concentration increases from local wildfires and decreases from carbon absorption events (Figure 4). For example, during the Amazon wildfire in August 2019, CO_2-Net (L) accurately tracks the sharp rise in CO_2 concentration from 420 μmol/mol to 440 μmol/mol over five days, while ViT fails to capture the trend and underestimates the values. Similarly, during the Siberian carbon absorption event in June 2020, CO_2-Net closely follows the rapid decline in CO_2 concentration, while ViT shows a gentler slope with larger deviations. All three variants of CO_2-Net accurately reflect short-term changes, with CO_2-Net (S) and CO_2-Net (B) showing greater deviations from the ground truth compared to CO_2-Net (L).

4.4. Ablation: Factorization, Embedding and Loss Function (RQ3)

Now we analyze the contribution of each key component to performance gains, based on CO_2-Net (L).

Obs5 (Factorization): The spatial and temporal experts outperform all baselines individually, their fusion improves results, and layer-wise connections provide further enhancement (Table 5a). On CMIP6, the spatial and temporal experts alone achieve RMSEs of 14.28 and 16.63 $\mu\mathrm{mol/mol}$, respectively, demonstrating their effectiveness. Simple fusion of their hidden representations reduces the RMSE to 13.01 $\mu\mathrm{mol/mol}$, while adding layer-wise connections further reduces it to 10.83 $\mu\mathrm{mol/mol}$. For fair comparison, the semi-supervised loss is excluded, as it requires knowledge of the homogeneous solution and is not applicable to single-expert settings. A similar trend is observed on CT, although the improvement is less pronounced.

Obs6 (Embedding): Wind-based embeddings enhance the performance of CO_2-Net (L) but show limited or negative effectiveness in baseline models (Table 5b). We compare the proposed wind-based embedding with the ClimODE embedding and observe that wind-based embedding significantly strengthens CO_2-Net (L), achieving an RMSE of 10.83 $\mu\mathrm{mol/mol}$ compared to 27.06 $\mu\mathrm{mol/mol}$ with ClimODE embedding and 23.40 $\mu\mathrm{mol/mol}$ without embedding. However, our embedding negatively impacts baseline models, underscoring the need for careful embedding design. In contrast, our model benefits from the embedding, as the unique PDE solution can be expressed as a combination of these embeddings and a network similar to CO_2-Net (i.e., Theorem 2 and 4), a property not shared by baselines, highlighting the complexity of embedding design.

Table 6. **Ablation Study (RMSE (↓) in $\mu\text{mol/mol}$ |ACC (↑)).** (a) Comparison of varying number of historic data points; (b) Evaluation of different temporal resolutions; (c) Assessment of the impact auxiliary variables. Experiments are conducted using CO_2-Net (L).

(a) **Historic Data Points.**

# point	CT	# point	CMIP6
0h	3.55 |0.83	0m	20.0 |0.92
12h	3.47 |0.83	1m	17.3 |0.87
24h	3.36 |0.85	3m	14.8 |0.99
48h	**3.28** |**0.86**	4m	**5.31** |**0.99**
96h	3.43 |0.84	5m	8.03 |1.00

(b) **Resolution.**

Res.	CT	Res.	CT
3h	**3.36** |**0.85**	18h	5.34 |0.26
6h	4.88 |0.42	24h	5.57 |0.15
9h	5.04 |0.30	36h	5.72 |0.09
12h	4.95 |0.30	48h	5.67 |0.08

(c) **Contribution of Auxiliary Variables.**

Removed Var.	UAS	VAS	Both	HUSS	PS	TAS	All	GPH	ORO	Both
CT (RMSE)	3.48	3.45	3.50	3.43	3.45	3.43	3.45	3.38	3.40	3.43
CT (ACC)	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.85	0.85	0.84
CMIP6 (RMSE)	31.3	40.0	44.7	12.3	13.2	8.47	31.7	—	—	—
CMIP6 (ACC)	0.83	0.38	0.43	0.88	0.98	0.72	0.35	—	—	—

Obs7 (Loss): Semi-supervised loss improves the reconstruction precision of CO_2-Net (L) but is less effective for baseline models (Table 5c). On the CMIP6 dataset, CO_2-Net (L) achieves a smaller RMSE of 5.31 $\mu\text{mol/mol}$ with the wind reconstruction loss, compared to 10.83 $\mu\text{mol/mol}$ without it. These results demonstrate the compatibility of the spatio-temporal factorization with the wind reconstruction loss, consistent with Theorem 3. While the loss is ineffective without explicit factorization, showing the entanglement between the model architecture and the design of self-supervised task in CO_2 reconstruction.

4.5. Ablation: Hyper-parameters and Auxiliaries

We analyze the effects of historic data points, temporal resolution, and auxiliary variables on performance, based on CO_2-Net (L). Ablation study on the loss coefficient λ is provided in Appendix F.3. **Historic Data Points:** Moderate-length historical sequences improve performance, while excessively long sequences degrade it (Table 6a). This reflects a trade-off between capturing relevant temporal features and introducing less relevant data. **Temporal Resolution:** Higher temporal resolution generally improves performance (Table 6b). For example, on CarbonTracker, a 3-hour resolution achieves the lowest RMSE and highest ACC, consistent with prior findings [4]. **Auxiliary Variables:** Wind flow variables (UAS, VAS) and dynamical variables (HUSS, PS, TAS) significantly affect reconstruction performance, while static variables (GPH, ORO) have minimal impact (Table 6c). Removing wind flow variables leads to the largest performance drop, highlighting their importance in modeling CO_2 dynamics, constructing spatio-temporal embedding, and enabling semi-supervised loss function.

5. Related Work

Global Reconstruction of Carbon Dioxide. Traditional methods for global surface CO_2 reconstruction like Bayesian synthesis [12], variational approaches [21] and Kalman filters [32] rely on computationally intensive transport models and extensive priors. Although accurate, their computational demand and dependence on priors limit their ability to achieve high-resolution modeling. Recently, data-driven methods like ClimaX [29], ClimODE [42] and Aurora [5] have revolutionized weather prediction by accelerating computation and improving accuracy. However, their application to CO_2 reconstruction remains largely unexplored.

Data-driven Methods for Data Reconstruction. There are three main techniques for data reconstruction: Super Resolution (SR), using methods like SinSR [44] and MeshfreeFlowNet [10], upscales low-resolution images, but it is unsuitable here due to the extreme sparsity of observations. Neural Inpainting, using GANs [17, 34], ViTs [1] and diffusion models [38], produces semantically meaningful inpainting. Implicit Neural Representations (INRs) [9, 27, 35] parameterize sensor or density domains and rely on transformations for supervision. However, these methods only focus on static reconstruction, overlooking temporal dynamics essential for spatio-temporal applications.

Spatio-Temporal Models. Spatio-temporal models enhance dynamic system performance by capturing spatial and temporal relationships, typically following three approaches. Two-stream architectures, like Spatio-Temporal Side Tuning [43] and CAST [19], separate spatial and temporal processing. Token-based encoding, as in Valley [26], encodes inputs into distinct spatial and temporal tokens for independent processing. Integrated spatio-temporal representations, such as Spatio-Temporal Representation Learning [39, 48], combine spatial and temporal features into a unified model using deep learning. However, these methods lack a formal mathematical proof for factorization, particularly its physical interpretation and connection to scientific principles.

6. Broader Impact and Future Work

Conclusion and Broader Impact. We propose CO_2-Net for global surface CO_2 reconstruction, achieving state-of-the-art performance on reanalysis data and comparable results to inversion models on real observations. Built on the advection-diffusion equation, which also governs other tracers like temperature and atmospheric gases (e.g., nitrogen dioxide and methane), our CO_2-Net can be extended beyond CO_2 reconstruction, offering a versatile framework for climate analysis and environmental monitoring. Our preliminary results in Appendix F suggest its potential applicability to reconstructing other atmospheric variables, encouraging further exploration in broader environmental contexts.

Future Work. Our model does not yet fully incorporate physical laws, such as conservation principles, which are crucial for ensuring physical consistency. Future research could integrate these constraints to improve the fidelity of reconstructions. We hope this work inspires further advancements in data-driven atmospheric modeling.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 62306179), the National Key Research and Development Program of China (2023YFC2808802), Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) (nos. SML2023SP219), the Ocean Negative Carbon Emissions (ONCE) Program, the National Key Research and Development Program of China (2024YFB3310000) and the Youth Independent Innovation Science Fund of the National University of Defense Technology (ZK24-53).

References

[1] Dosovitskiy Alexey. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*, 2020. 6, 8, 4

[2] Elisabeth Andrews, Patrick J. Sheridan, John A. Ogren, Derek Hageman, Anne Jefferson, Jim Wendell, Andrés Alástuey, Lucas Alados-Arboledas, Michael Bergin, Marina Ealo, A. Gannet Hallar, András Hoffer, Ivo Kalapov, Melita Keywood, Jeongeun Kim, Sang-Woo Kim, Felicia Kolonjari, Casper Labuschagne, Neng-Huei Lin, AnneMarie Macdonald, Olga L. Mayol-Bracero, Ian B. McCubbin, Marco Pandolfi, Fabienne Reisen, Sangeeta Sharma, James P. Sherman, Mar Sorribas, and Junying Sun. Overview of the noaa/esrl federated aerosol network. *Bulletin of the American Meteorological Society*, 100(1):123 – 135, 2019. 1

[3] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556*, 2022. 1

[4] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023. 8

[5] Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan Weyn, Haiyu Dong, Anna Vaughan, et al. Aurora: A foundation model of the atmosphere. *arXiv preprint arXiv:2405.13063*, 2024. 1, 8

[6] Kun Cai, Liuyin Guan, Shenshen Li, Shuo Zhang, Yang Liu, and Yang Liu. Full-coverage estimation of co2 concentrations in china via multisource satellite data and deep forest model. *Scientific Data*, 11(1):1231, 2024. 2

[7] Chenjie Cao, Qiaole Dong, and Yanwei Fu. Learning prior feature and attention enhanced image inpainting. In *European conference on computer vision*, pages 306–322. Springer, 2022. 5

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020. 5

[9] Pan Du, Meet Hemant Parikh, Xiantao Fan, Xin-Yang Liu, and Jian-Xun Wang. Conditional neural field latent diffusion model for generating spatiotemporal turbulence. *Nature Communications*, 15(1):10416, 2024. 8

[10] Soheil Esmaeilzadeh, Kamyar Azizzadenesheli, Karthik Kashinath, Mustafa Mustafa, Hamdi A Tchelepi, Philip Marcus, Mr Prabhat, Anima Anandkumar, et al. Meshfreeflownet: A physics-constrained deep continuous space-time super-resolution framework. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15. IEEE, 2020. 8

[11] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016. 5, 3

[12] Kevin Robert Gurney, Rachel M Law, A Scott Denning, Peter J Rayner, Bernard C Pak, David Baker, Philippe Bousquet, Lori Bruhwiler, Yu-Han Chen, Philippe Ciais, et al. Transcom 3 inversion intercomparison: Model mean results for the estimation of seasonal carbon sources and sinks. *Global Biogeochemical Cycles*, 18(1), 2004. 8

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 5

[14] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 646–661. Springer, 2016. 4

[15] Lianghua Huang, Yu Liu, Bin Wang, Pan Pan, Yinghui Xu, and Rong Jin. Self-supervised video representation learning by context and motion decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13886–13895, 2021. 5

[16] Jacobson, A. R. and Schuldt, K. N. et al. CarbonTracker CT2022. Technical Report https://doi.org/10.25925/Z1GJ-3254, NOAA Global Monitoring Laboratory, https://doi.org/10.25925/Z1GJ-3254, 2023. 5, 3

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019. 8

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 4

[19] Dongho Lee, Jongseo Lee, and Jinwoo Choi. Cast: cross-attention in space and time for video action recognition. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 4, 6, 8

[20] Jie Li, Kun Jia, Xiangqin Wei, Mu Xia, Zhulin Chen, Yunjun Yao, Xiaotong Zhang, Haiying Jiang, Bo Yuan, Guofeng Tao, et al. High-spatiotemporal resolution mapping of spatiotemporally continuous atmospheric co2 concentrations over the global continent. *International Journal of Applied Earth Observation and Geoinformation*, 108:102743, 2022. 2

[21] Junjie Liu, Kevin W Bowman, Meemong Lee, Daven K Henze, Nicolas Bousserez, Holger Brix, G James Collatz, Dimitris Menemenlis, Lesley Ott, Steven Pawson, et al. Carbon monitoring system flux estimation and attribution: impact

of acos-gosat xco2 sampling on the inference of terrestrial biospheric sources and sinks. *Tellus B: Chemical and Physical Meteorology*, 66(1):22486, 2014. 1, 8

[22] Peiyuan Liu, Tian Zhou, Liang Sun, and Rong Jin. Mitigating time discretization challenges with weatherode: A sandwich physics-driven neural ode for weather forecasting. *arXiv preprint arXiv:2410.06560*, 2024. 2

[23] Zhiqiang Liu, Ning Zeng, Yun Liu, Eugenia Kalnay, Ghassem Asrar, Bo Wu, Qixiang Cai, Di Liu, and Pengfei Han. Improving the joint estimation of co 2 and surface carbon fluxes using a constrained ensemble kalman filter in cola (v1. 0). *Geoscientific Model Development*, 15(14):5511–5528, 2022. 5

[24] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6, 4

[25] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 4

[26] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023. 8

[27] Xihaier Luo, Wei Xu, Yihui Ren, Shinjae Yoo, and Balu Nadiga. Continuous field reconstruction from sparse observations with implicit neural networks. *arXiv preprint arXiv:2401.11611*, 2024. 8

[28] Meike Nauta, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. Pip-net: Patch-based intuitive prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2744–2753, 2023. 5

[29] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023. 1, 8

[30] Margaret A Oliver and Richard Webster. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System*, 4(3): 313–332, 1990. 5, 6

[31] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022. 1

[32] Wouter Peters, Andrew R Jacobson, Colm Sweeney, Arlyn E Andrews, Thomas J Conway, Kenneth Masarie, John B Miller, Lori MP Bruhwiler, Gabrielle Pétron, Adam I Hirsch, et al. An atmospheric perspective on north american carbon dioxide exchange: Carbontracker. *Proceedings of the National Academy of Sciences*, 104(48):18925–18930, 2007. 1, 8

[33] C Rödenbeck, S Houweling, Michael Gloor, and Michael Heimann. Co 2 flux history 1982–2001 inferred from atmospheric data using a global inversion of atmospheric transport. *Atmospheric Chemistry and Physics*, 3(6):1919–1964, 2003. 1

[34] Veit Sandfort, Ke Yan, Perry J Pickhardt, and Ronald M Summers. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. *Scientific reports*, 9(1):16884, 2019. 6, 8, 4

[35] Javier E Santos, Zachary R Fox, Arvind Mohan, Daniel O'Malley, Hari Viswanathan, and Nicholas Lubbers. Development of the senseiver for efficient field reconstruction from sparse observations. *Nature Machine Intelligence*, 5 (11):1317–1325, 2023. 6, 8, 4

[36] K. N. Schuldt, J. Mund, T. Aalto, J. B. Abshire, K. Aikin, G. Allen, M. Andrade, A. Andrews, F. Apadula, S. Arnold, B. Baier, P. Bakwin, L. Bäni, J. Bartyzel, G. Bentz, P. Bergamaschi, A. Beyersdorf, T. Biermann, S. C. Biraud, and Miroslaw Zimnoch. Multi-laboratory compilation of atmospheric carbon dioxide data for the period 1957-2023; obspack_co2_1_globalviewplus_v10.0_2024-09-26. Data set, 2024. NOAA Global Monitoring Laboratory. 5, 3

[37] Oscar Skean, Md Rifat Arefin, Yann LeCun, and Ravid Shwartz-Ziv. Does representation matter? exploring intermediate layers in large language models. *arXiv preprint arXiv:2412.09563*, 2024. 4

[38] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023. 8

[39] Jiahao Su, Wonmin Byeon, Jean Kossaifi, Furong Huang, Jan Kautz, and Anima Anandkumar. Convolutional tensor-train lstm for spatio-temporal learning. *Advances in Neural Information Processing Systems*, 33:13714–13726, 2020. 8

[40] Song Tang, Chuang Li, Pu Zhang, and RongNian Tang. Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13470–13479, 2023. 6, 4

[41] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 2, 5

[42] Yogesh Verma, Markus Heinonen, and Vikas Garg. Climode: Climate and weather forecasting with physics-informed neural odes. *arXiv preprint arXiv:2404.10024*, 2024. 1, 2, 4, 6, 8

[43] Xiao Wang, Qian Zhu, Jiandong Jin, Jun Zhu, Futian Wang, Bo Jiang, Yaowei Wang, and Yonghong Tian. Spatio-temporal side tuning pre-trained foundation models for video-based pedestrian attribute recognition. *arXiv preprint arXiv:2404.17929*, 2024. 2, 8

[44] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25796–25805, 2024. 8

[45] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer.

Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13587–13597, 2022. 5

[46] Debra Wunch, Geoffrey C Toon, Jean-François L Blavier, Rebecca A Washenfelder, Justus Notholt, Brian J Connor, David WT Griffith, Vanessa Sherlock, and Paul O Wennberg. The total carbon column observing network. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 369(1943):2087–2112, 2011. 2

[47] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3333–3343, 2022. 5

[48] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6548–6557, 2020. 8

[49] Lingfeng Zhang, Tongwen Li, and Jingan Wu. Deriving gapless co2 concentrations using a geographically weighted neural network: China, 2014–2020. *International Journal of Applied Earth Observation and Geoinformation*, 114:103063, 2022. 2

[50] Lingfeng Zhang, Tongwen Li, Jingan Wu, and Hongji Yang. Global estimates of gap-free and fine-scale co2 concentrations during 2014–2020 from satellite and reanalysis data. *Environment International*, 178:108057, 2023. 2

[51] Zhaodi Zhang, Zhiyi Xue, Yang Chen, Si Liu, Yueling Zhang, Jing Liu, and Min Zhang. Boosting verified training for robust image classifications via abstraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16251–16260, 2023. 5

CO$_2$-Net: A Physics-Informed Spatio-Temporal Model for Global Surface CO$_2$ Reconstruction

Supplementary Material

A. Proof of Proposition 1

We first recall some important notations. Let $\mathcal{L}_\nu[\cdot]$ be the linear advection–diffusion operator

$$\mathcal{L}_\nu[\varphi] \;=\; \frac{\partial\varphi}{\partial t} \;+\; (\mathbf{w}\cdot\nabla)\,\varphi \;-\; \nu\,\nabla^2\,\varphi,$$

where $\mathbf{w}(\theta,\phi,t)$ is the wind field, $\nu(\theta,\phi,t)$ is the diffusion coefficient, and ∇^2 is the spherical Laplacian on S^2. Suppose we have a constrained PDE:

$$\mathcal{L}_\nu[\varphi] = s(\theta,\phi,t), \quad \varphi\big|_{\mathcal{Z}\times[0,T]} = f\big|_{\mathcal{Z}\times[0,T]},$$

where s is a given source term and $\varphi = f$ on the subset $\mathcal{Z}\times[0,T]$ (representing observational data). The goal is to prove:

(1) Superposition. The general solution φ_{general} to the constrained PDE is the sum of a particular solution φ_{part} and a homogeneous solution φ_{homo}.

(2) Time-Invariance of φ_{part}. If $\mathbf{w}$, ν and s are constant over the time interval $[\tau,\tau+\Delta t]$, then φ_{part} becomes time-invariant in that interval, depending only on the initial time τ.

Proof. **(1) Proof of Superposition**
Step 1. Define the particular solution. Consider a particular solution φ_{part} satisfying

$$\mathcal{L}_\nu[\varphi_{\text{part}}] = s.$$

Such a solution does not necessarily match the observation f on $\mathcal{Z}\times[0,T]$; it only accounts for the source s.
Step 2. Construct the homogeneous solution. Let φ be a solution of the same PDE with the constraint $\varphi = f$ on $\mathcal{Z}\times[0,T]$. Define

$$\varphi_{\text{homo}} = \varphi - \varphi_{\text{part}}.$$

Because $\mathcal{L}_\nu$ is a linear operator,

$$\mathcal{L}_\nu[\varphi_{\text{homo}}] = \mathcal{L}_\nu[\varphi] - \mathcal{L}_\nu[\varphi_{\text{part}}] = s - s = 0.$$

Hence, φ_{homo} solves the homogeneous PDE.
Step 3. Enforce the boundary constraint. On $\mathcal{Z}\times[0,T]$, where $\varphi = f$, we have

$$\varphi_{\text{homo}} = \varphi - \varphi_{\text{part}} = f - \varphi_{\text{part}},$$

which forces φ_{homo} to match that boundary constraint in conjunction with φ_{part}. Since φ was an arbitrary solution, all solutions can be decomposed as

$$\varphi = \varphi_{\text{part}} + \varphi_{\text{homo}}.$$

This proves the superposition principle.
(2) Proof of Time-Invariance of φ_{part}
Step 1. Assume time-invariant source field, wind field and diffusion coefficients. Assume the wind field, diffusion coefficient and source term are constant for $t \in [\tau,\tau+\Delta t]$,

$$\mathbf{w}(\theta,\phi,t) = \mathbf{w}(\theta,\phi), \nu(\theta,\phi,t) = \nu(\theta,\phi), s(\theta,\phi,t) = s(\theta,\phi).$$

Thus, the operator

$$\mathcal{L}_\nu[\varphi] = \frac{\partial\varphi}{\partial t} + (\mathbf{w}\cdot\nabla)\varphi - \nu\,\nabla^2\varphi$$

has no explicit time-dependence over $[\tau,\tau+\Delta t]$.
Step 2. Derive the steady-state equation. Assume a particular solution φ_{part} satisfying

$$(\mathbf{w}\cdot\nabla)\,\varphi_{\text{part}} - \nu\nabla^2\varphi_{\text{part}} = s.$$

Substituting into $\mathcal{L}_\nu[\varphi_{\text{part}}] = s$ yields

$$\frac{\partial\varphi_{\text{part}}}{\partial t} + (\mathbf{w}\cdot\nabla)\,\varphi_{\text{part}} - \nu\nabla^2\varphi_{\text{part}} = s.$$

Subtracting the first equation from the second equation gives

$$\frac{\partial\varphi_{\text{part}}}{\partial t} = 0.$$

Hence, φ_{part} is time-invariant in the interval $[\tau,\tau+\Delta t]$.
Step 3. Analyze the dependence on the Initial Time. Once $\varphi_{\text{part}}(\theta,\phi,\tau)$ is specified at $t=\tau$, it remains the same for $t \in [\tau,\tau+\Delta t]$. Thus it depends only on τ (and spatial boundary conditions). Therefore, φ_{part} can be taken as a time-invariant solution whenever the PDE coefficients remain constant in the time interval. □

B. Proof of Theorem 2

Proof idea for existence and uniqueness of the analytical solution. The Cauchy-Kovalevskaya Theorem for second-order parabolic equations guarantees the existence and uniqueness of an analytic solution, provided the initial value g is analytic on the entire domain S^2. Thus, we first establish the existence of a unique function g that satisfies $f = g$ on $\mathcal{Z}$ and is analytic on S^2. With this, the theorem directly ensures the existence and uniqueness of the solution.

Proof idea for finding the explicit form. Since the solution φ is real-analytic, it admits an expansion as an infinite series of spherical harmonics. Furthermore, we demonstrate that when the wind flow is zonal, the expansion coefficients take a simplified form.

Proof. **(1) Existence and uniqueness of the analytical solution.** Since $\mathcal{Z}$ is an open set on which the real–analytic functions f and g agree, the identity theorem for real–analytic functions implies that f and g must coincide on the entire sphere (which is connected). Consequently, our PDE problem reduces to a standard initial–value problem with an analytic initial condition. Furthermore, because the advection–diffusion equation is a second–order parabolic PDE and its coefficients ν, s, and the wind $\mathbf{w}$ are all analytic, the Cauchy–Kovalevskaya theorem ensures that this PDE admits a unique real–analytic solution.

(2) Explicit form in terms of infinite series of spherical harmonics. Let $S^2 = \{\mathbf{x} \in \mathbb{R}^3 : \|\mathbf{x}\| = 1\}$ denote the unit sphere. In spherical coordinates (θ, ϕ), where $\theta \in [0, \pi]$ (polar angle) and $\phi \in [0, 2\pi)$ (azimuthal angle), the space $L^2(S^2)$ of square-integrable functions on S^2 is spanned by the set of **spherical harmonics** $\{Y_n^m(\theta, \phi)\}$, which form a complete orthonormal system. Consequently, for each fixed t, any function $\varphi : S^2 \times \mathbb{R} \to \mathbb{R}$ admits a unique expansion in terms of spherical harmonics:

$$\varphi(\theta, \phi, t) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} A_{n,m}(t) Y_n^m(\theta, \phi),$$

where the coefficients $A_{n,m}(t)$ are given by the inner product:

$$A_{n,m}(t) = \int_{S^2} \varphi(\theta, \phi, t) Y_n^m(\theta, \phi)\, d\Omega,$$

with $d\Omega = \sin\theta\, d\theta\, d\phi$ being the standard measure on the sphere. The series converges in the $L^2(S^2)$-sense for each t, and if φ is smooth in (θ, ϕ), then the convergence is uniform together with all derivatives. The spherical harmonics $Y_n^m(\theta, \phi)$ are defined as

$$Y_n^m(\theta, \phi) = N_{n,m} e^{im\phi} P_n^m(\cos\theta),$$

where $P_n^m(x)$ are the **associated Legendre functions** of degree n and order m. $N_{n,m}$ is a normalization factor ensuring orthonormality with respect to the inner product:

$$\int_0^{\pi} \int_0^{2\pi} Y_n^m(\theta, \phi) Y_{n'}^{m'*}(\theta, \phi) \sin\theta\, d\theta\, d\phi = \delta_{nn'} \delta_{mm'}.$$

Each Y_n^m is an eigenfunction of the **Laplacian operator** ∇^2 on the sphere:

$$\nabla^2 Y_n^m(\theta, \phi) = -n(n+1) Y_n^m(\theta, \phi).$$

By the completeness of the set $\{Y_n^m\}$ in $L^2(S^2)$, it follows that any function $\varphi(\theta, \phi, t)$ (for each fixed t) can be expressed as an infinite linear combination of spherical harmonics. If φ is sufficiently smooth, then the series converges uniformly, and differentiation can be performed term by term. Thus, the function $\varphi(\theta, \phi, t)$ can always be represented as an infinite series of spherical harmonics.

Furthermore, let the source function $s(\theta, \phi)$ have an expansion in terms of spherical harmonics

$$s(\theta, \phi) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} S_{n,m} Y_n^m(\theta, \phi),$$

where the coefficients

$$S_{n,m} = \int_{S^2} s(\theta, \phi) Y_n^m(\theta, \phi)\, d\Omega.$$

Using the eigenvalue property of spherical harmonics and the zonal wind, the equation for φ transforms into an infinite system of coupled ordinary differential equations (ODEs) for $A_{n,m}(t)$:

$$\frac{dA_{n,m}}{dt} + imU A_{n,m} = -\nu n(n+1) A_{n,m} + S_{n,m}.$$

The general solution to this first-order linear ODE is obtained using the integrating factor $e^{(imU + \nu n(n+1))t}$, given by

$$A_{n,m}(t) = e^{-(imU + \nu n(n+1))t} (A_{n,m}(0) + \int_0^t S_{n,m} e^{(imU + \nu n(n+1))\tau} d\tau).$$

We can rewrite the the formula in the following form

$$A_{n,m}(t) = B_{n,m} e^{-\nu(n+1)nt + imUt} + C_{n,m},$$

where

$$B_{n,m} = A_{n,m}(0) - \frac{S_{n,m}}{\nu n(n+1) - imU},$$

$$C_{n,m} = \frac{S_{n,m}}{\nu n(n+1) - imU}.$$

□

C. Proof of Theorem 3

Recall that we consider the advection-diffusion equation without source term (the solution is homogeneous) on the unit sphere S^2,

$$\frac{\partial \varphi}{\partial t} + (\mathbf{w} \cdot \nabla)\varphi = D \nabla^2 \varphi.$$

We want to argue that if we know the evolution of φ over time (for appropriate choices of initial conditions), then the field $\mathbf{w}$ can be uniquely determined.

Proof. **Step 1. Assume two wind fields.** Suppose there exist two smooth, time-dependent vector fields $\mathbf{w}_1$ and $\mathbf{w}_2$ on S^2 satisfies the equation without source term, giving:

$$\frac{\partial \varphi}{\partial t} + (\mathbf{w}_1 \cdot \nabla)\varphi = D\, \nabla^2 \varphi,$$

and

$$\frac{\partial \varphi}{\partial t} + (\mathbf{w}_2 \cdot \nabla)\varphi = D\,\nabla^2 \varphi.$$

Step 2. Subtract the two equations. Subtracting the second equation from the first gives:

$$(\mathbf{w}_1 - \mathbf{w}_2) \cdot \nabla \varphi = 0.$$

This equality holds for every point on S^2 and for all $t \in [0, T]$, proving the pointwise orthogonality between the vector field $\mathbf{w}_1 - \mathbf{w}_2$ and the gradient $\nabla \varphi$.

Step 3. Non-vanishing gradient. Given $\nabla \varphi \neq 0$ everywhere in the domain, φ acts locally as a coordinate function with well-defined level sets, which are smooth curves on S^2. Consequently, the condition:

$$(\mathbf{w}_1 - \mathbf{w}_2) \cdot \nabla \varphi = 0$$

implies that $\mathbf{w}_1 - \mathbf{w}_2$ is orthogonal to $\nabla \varphi$ at every point, meaning $\mathbf{w}_1 - \mathbf{w}_2$ must be tangent to the level sets of φ. On a connected domain, the only smooth vector field that is tangent to the level sets of a smooth function with a non-vanishing gradient is the zero vector field. Therefore, we must have

$$\mathbf{w}_1 - \mathbf{w}_2 = 0,$$

which implies

$$\mathbf{w}_1 = \mathbf{w}_2,$$

proving the uniqueness of $\mathbf{w}$.

□

D. Neural Representation of the Solution

By Theorem 2 , we have shown that the solution to the constrained PDE can be represented in the following form:

$$\varphi(\theta, \phi, t) = \sum_{n \geq 1} \sum_{|m| \leq n} A_{n,m}(t) e^{im\phi} P_n^m(\cos\theta). \quad (4)$$

Now we are going to present the following theorem to show that this solution can be represented by a neural network shown in Figure 5.

Theorem 4. *For any point $z \in S^2$, let (θ, φ) denote the polar coordinate of the point z. Define*

$$Z_k(\varphi, \theta) = \{e^{im\phi} P_k^m(\cos\theta)\}_{m=-k}^{k}$$

and

$$T_k(t) = \{A_{k,m}(t)\}_{m=-k}^{k}.$$

Therefore, the solution (4) can be implemented by the neural network of the architecture shown in Figure 5.

Proof. Clearly, the general solution can be easily represented by the the linear combination of Z_k and T_k. Each function takes the input from the input of the entire model which consists of the coordinate, temporal embedding and the spatial embedding. □

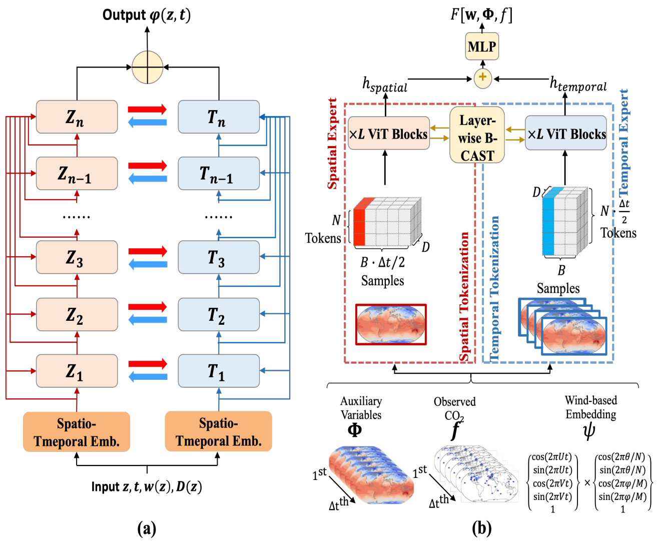
Output $\varphi(z,t)$
Z_n
Z_{n-1}
......
Z_3
Z_2
Z_1
T_n
T_{n-1}
......
T_3
T_2
T_1
Spatio-Temporal Emb.
Spatio-Temporal Emb.
Input $z, t, w(z), D(z)$
(a)
$F[\mathbf{w}, \mathbf{\Phi}, f]$
MLP
$h_{spatial}$
$h_{temporal}$
Spatial Expert
Temporal Expert
×L ViT Blocks
Layer-wise B-CAST
×L ViT Blocks
N Tokens
D
$B \cdot \Delta t/2$ Samples
D
$N \cdot \frac{\Delta t}{2}$ Tokens
B Samples
Spatial Tokenization
Temporal Tokenization
Auxiliary Variables $\mathbf{\Phi}$
Observed CO2 f
Wind-based Embedding ψ
1st
Δtth
1st
Δtth
$\begin{Bmatrix}\cos(2\pi Ut)\\ \sin(2\pi Ut)\\ \cos(2\pi Vt)\\ \sin(2\pi Vt)\\ 1\end{Bmatrix} \times \begin{Bmatrix}\cos(2\pi\theta/N)\\ \sin(2\pi\theta/N)\\ \cos(2\pi\varphi/M)\\ \sin(2\pi\varphi/M)\\ 1\end{Bmatrix}$
(b)

Figure 5. (a) The neural architecture for the solution in theorem 4. (b) Overview of CO$_2$-Net.

E. Experiment Details

In this section, we present a detailed description of the datasets in E.1, specify the hyper-parameters for all models in E.2, and elaborate on the evaluation metrics in E.3.

E.1. Data

We conduct experiments using three datasets: CarbonTracker [16], the Coupled Model Intercomparison Project Phase 6 (CMIP6) [11] and GLOBALVIEWplus (GV+) [36]. These datasets provide comprehensive information essential for accurate CO$_2$ reconstruction and analysis of atmospheric dynamics.

CarbonTracker is a sophisticated CO$_2$ measurement and modeling system developed by National Oceanic and Atmospheric Administration (NOAA). Its primary objective is to monitor global CO$_2$ uptake and emissions over time. CarbonTracker integrates atmospheric CO$_2$ observations with simulated atmospheric transport models to estimate surface fluxes of CO$_2$. We use CarbonTracker released in 2022, which provides global estimates of CO$_2$ concentrations with a spatial resolution of $2° \times 3°$ and temporal coverage from January 2000 to December 2020.

CMIP6 is an international collaborative project that encompasses outputs from approximately 100 climate models developed by research institutions worldwide. CMIP6 provides comprehensive access to a wide variety of climate variables, including surface pressure, temperature, humidity, wind speed and CO$_2$ concentration, offering a multidimensional perspective on the physical and chemical processes that govern the climate. We use data from historical experiments conducted by the CanESM5 model, with a temporal coverage from January 1850 to December 2014. We regrid it to a spatial resolution of $2° \times 3°$, ensuring consistency with

CarbonTracker.

GLOBALVIEWplus is a comprehensive data platform developed by NOAA. GV+ integrates atmospheric and oceanic observations from monitoring systems around the world to support climate research and weather forecasting. We use CO_2 observations from GV+ version 10.0 (September 26th, 2024), which offers CO_2 measurements at 3-hour intervals from 96 surface spots, thereby ensuring temporal resolution consistent with CarbonTracker.

E.2. Model Hyper-parameters

CO_2-Net. We detail the default hyper-parameters used for CO_2-Net in Table 7. We set the patch size to 9×12, preserving the aspect ratio consistent with the spatial dimensions of the datasets. We use AdamW [24] optimizer for training, and set momentum betas to 0.9 and 0.999. We set the learning rate and weight decay to $5e-4$ and $8e-4$ respectively, which are carefully selected from the sets $\{5e-5, 1e-4, 3e-4, 5e-4, 8e-4\}$ and $\{1e-4, 5e-4, 8e-4, 1e-3\}$ via grid search. We adopt a cosine sheculer [25] for learning rate scheduling.

Table 7. Hyper-parameters used for CO_2-Net.

Hyper-parameter	**Meaning**	**Value**
p	Patch size	9, 12
D	Embedding dimension	768
# Blocks	Number of ViT blocks	Small: 1 Base: 4 Large: 12
# Heads	Number of attention heads	12
MLP dimension	The hidden dimension of the MLP layers	3072
Prediction depth	Number of layers in the output head	3
Drop path	For stochastic depth [14]	0.2
Dropout	Dropout rate	0.2

Vision Transformer (ViT). We configure the hyper-parameters for ViT [1] as described in Table 8. We employ AdamW optimizer with the momentum betas set to 0.9 and 0.999. We set the learning rate and weight decay to $1e-4$ and $5e-4$ for both datasets.

CycleGAN. We use the hyper-parameters in Table 9 for CycleGAN [34] in all our experiments. We use Adam [18] optimizer when training CycleGAN, and set the momentum betas to 0.5 and 0.999. We train the model with a learning rate of $1e-4$.

Senseiver. We employ the hyper-parameters in Table 10 for Senseiver [35]. We use Adam optimizer with default momentum betas, which are 0.9 and 0.999. We train the model with a learning rate of $1e-4$ and adopt a cosine scheduler with warm up for 5 epochs.

Table 8. Hyper-parameters used for Vision Transformer (ViT).

Hyper-parameter	**Meaning**	**Value**
p	Patch size	9, 12
D	Embedding dimension	768
# Blocks	Number of ViT blocks	12
# Heads	Number of attention heads	16
MLP dimension	The hidden dimension of the MLP layers	2048
Dropout	Dropout rate	0.1

Table 9. Hyper-parameters used for CycleGAN.

(a) Hyper-parameters for Generator

Hyper-parameter	**Meaning**	**Value**
Kernel size	Kernel size of residual block	3
Stride	Stride of residual block	2
Padding size	Padding size of residual block	1
Padding type	Padding mode of residual block	Reflection
Residual blocks	Number of residual blocks	9
Dropout	Dropout rate	0.5
# Filters	Number of filters in the last convolution layer	64

(b) Hyper-parameters for Discriminator

Hyper-parameter	**Meaning**	**Value**
Kernel size	Kernel size of each convolution layer	4
Stride	Stride of each convoluti layer	2
Padding size	Padding size of each convolution layer	1
Padding type	Padding mode of each convolution layer	Zeros
# Layers	Number of layers	3
# Filters	Number of filters in the last convolution layer	64

Table 10. Hyper-parameters used for Senseiver.

Hyper-parameter	**Meaning**	**Value**
N_s	Number of sensor observations	8
N_f	Number of frequency bands for positional encoding	32
N_c	Hidden dimension	64
Depth	Number of encoder blocks	3
L	Number of self attention layers in each block	3

SwinLSTM. We configure SwinLSTM [40] with the hyper-parameters in Table 11. We adopt AdamW optimizer with the momentum betas set to 0.9 and 0.999. We set the learning rate to $1e-4$ and employ a cosine scheduler with warm up for 100 epochs.

Table 11. Hyper-parameters used for SwinLSTM.

Hyper-parameter	**Meaning**	**Value**
p	Patch size	9, 12
D	Embedding dimension	126
# Blocks	Number of ViT blocks	12
# Heads	Number of attention heads in different layers	4, 8
Window size	Window size of Swin Transformer layer	2
MLP dimension	The hidden dimension of the MLP layers	2048
Dropout	Dropout rate	0.0

E.3. Metrics

We assess the model performance using latitude-weighted RMSE and Anomaly Correlation Coefficient (ACC).

Latitude-weighted RMSE quantifies the average error between the reconstructed results and the ground-truth values. A lower RMSE indicates higher accuracy. It is calculated as follows:

$$\text{RMSE} = \frac{1}{N}\sum_{t=1}^{N}\sqrt{\frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W}\alpha(h)(y_{thw}-u_{thw})^2},$$

where N is the total number of time points, H is the number of latitude grid points, and W is the number of longitude grid points, forming a grid over the Earth's surface. The index t refers to a specific time point, while h and w represent specific latitude and longitude indices, respectively. The observed value at a given time t, latitude h, and longitude w is denoted by y_{thw}, and the corresponding reconstructed value is u_{thw}. The term $\alpha(h)$ is the latitude weight, which accounts for the curvature of the Earth, and is defined as $\alpha(h) = \cos(h)/\left(\frac{1}{H}\sum_{h'}\cos(h')\right)$. The expression $(y_{thw} - u_{thw})^2$ represents the squared difference between the observed and reconstructed values at each grid point. The summations over t, h, and w aggregate the errors over all dimensions. The square root ensures that the RMSE is expressed in the same units as the original variable.

Anomaly Correlation Coefficient (ACC) assesses the similarity between the reconstructed and observed patterns of anomalies, emphasizing relative variations over exact values. Higher ACC values indicate more accurate reconstruction of anomaly patterns. It is defined by the following formula:

$$\text{ACC} = \frac{\sum_{t,h,w}\alpha(h)\tilde{y}_{thw}\tilde{u}_{thw}}{\sqrt{\sum_{t,h,w}\alpha(h)\tilde{y}_{thw}^2\sum_{t,h,w}\alpha(h)\tilde{u}_{thw}^2}},$$

where the observed anomaly, $\tilde{y}_{thw}$, is defined as the difference between the observed value and the empirical mean C of the observed values, i.e., $\tilde{y}_{thw} = y_{thw} - C$. Similarly, the reconstructed anomaly is $\tilde{u}_{thw} = u_{thw} - C$. The latitude weight $\alpha(h)$ is the same as in the RMSE formula and adjusts for the varying grid sizes due to the Earth's curvature. The numerator, $\sum_{t,h,w}\alpha(h)\tilde{y}_{thw}\tilde{u}_{thw}$, is the weighted sum of the products of observed and reconstructed anomalies. The terms $\sum_{t,h,w}\alpha(h)\tilde{y}_{thw}^2$ and $\sum_{t,h,w}\alpha(h)\tilde{u}_{thw}^2$ are the weighted sums of squared observed and reconstructed anomalies, respectively. The square root in the denominator normalizes the anomaly products, ensuring that ACC is dimensionless.

Table 12. Size, efficiency and global reconstruction performance comparison of different models on the CT dataset. * indicates dynamic reconstruction model, while others are static reconstruction models.

Model	Sizes	$\text{Time}_{\text{Train}}$	$\text{VRAM}_{\text{Train}}$	$\text{Time}_{\text{Infer}}$	$\text{VRAM}_{\text{Infer}}$	Global	
	(M)	(s/epoch)	(GB)	(s)	(GB)	RMSE	ACC
Senseiver	0.11	377	1.00	3	5.04	$6.39_{\pm 0.45}$	$0.43_{\pm 0.10}$
SwinLSTM*	3.3	527	6.48	38	0.61	$5.31_{\pm 0.01}$	$0.61_{\pm 0.00}$
CycleGAN	28	366	9.36	14	5.42	$4.70_{\pm 0.01}$	$0.71_{\pm 0.00}$
ViT	76	79	5.40	6	1.21	$5.42_{\pm 0.21}$	$0.50_{\pm 0.04}$
CO_2-Net* (S)	38	154	9.15	25	4.36	$3.59_{\pm 0.06}$	$0.72_{\pm 0.01}$
CO_2-Net* (B)	95	407	17.82	30	4.59	$3.41_{\pm 0.04}$	$0.77_{\pm 0.01}$
CO_2-Net* (L)	247	1212	38.72	73	5.18	$\mathbf{3.36}_{\pm 0.04}$	$\mathbf{0.85}_{\pm 0.00}$

F. Additional Experimental Results

In this section, we provide more comprehensive experimental results, including discussion on computational efficiency in F.1, evaluation of the extendibility of CO_2-Net to reconstruct other variables in F.2, reconstruction performance across all regions in F.3, as well as a comparison of different variogram models used in kriging interpolation F.4.

F.1. Computational Efficiency

We evaluate the computational efficiency of our CO_2-Net against other data-driven baselines on two NVIDIA RTX 6000 Ada GPUs under the same batch size of 32 per GPU. We experiment with Senseiver[†], SwinLSTM[†] and CycleGAN[†] based on the official implementations from their github repositories. We experiment with ViT based on our own implementation. Table 12 displays the model size (number of parameters), training time (time per epoch), training memory (maximum GPU VRAM usage during training), inference time (time for processing the test set) and inference memory (maximum GPU VRAM usage during inference). Our smallest model CO_2-Net (S) achieves lower RMSE and higher ACC compared to other baselines, while showing satisfactory efficiency. Furthermore, CO_2-Net (B) and (L) achieve better performance than CO_2-Net (S) at the cost of computational efficiency, providing choices to balance performance and efficiency.

[†] https://github.com/OrchardLANL/Senseiver

[†] https://github.com/SongTang-x/SwinLSTM

[†] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

F.2. Implication For Broader Impact

We assess the model ability to reconstruct other variables governed by the advection-diffusion equation (e.g., PS, HUSS and TAS) using the CarbonTracker dataset, and compare its performance with other baselines, the results are presented in Table 13. The task is formulated as reconstructing the target variables using the spot observations, wind flow and auxiliary variables, in alignment with CO_2 reconstruction. For spot observations, we assume the target variables are only observed at the same monitoring spots where CO_2 is measured (i.e., 96 surface spots from the GLOBALVIEWplus). As in CO_2 reconstruction, we use data from 72 spots for training, while the rest 24 spots are kept for testing. CO_2-Net outperforms all baselines across all variables, which could be attributed to its physics-informed architectural design. These findings highlight the versatility and effectiveness of CO_2-Net in reconstructing atmospheric variables beyond CO_2.

Table 13. RMSE (↓) and ACC (↑) comparison of different models for reconstructing three atmospheric variables. $*$ indicates spatio-temporal reconstruction models, others are static reconstruction models.

Methods	PS		HUSS		TAS	
	RMSE (Pa)	ACC	RMSE ($\times 10^{-3}\%$)	ACC	RMSE (K)	ACC
Spherical	9787.14	0.00	6.93	0.01	26.97	0.02
Exponential	9199.71	0.00	6.81	0.01	25.04	0.03
Senseiver	3019.52	0.00	2.80	0.06	7.65	0.04
CycleGAN	3674.32	0.07	2.34	0.50	4.97	0.56
SwinLSTM*	5445.14	0.58	3.40	0.88	8.90	0.85
ViT	729.59	**0.97**	1.96	**0.92**	3.51	0.95
CO_2-Net*	**191.82**	0.96	**0.97**	0.89	**1.34**	**0.97**

F.3. Additional Results

Global Reconstruction Results of CO_2-Net (S) and (B) (Figure 6). Both CO_2-Net (S) and (B) capture the trend of global warming, whereas CO_2-Net (B) reconstructs CO_2 concentration with higher precision than CO_2-Net (S).

Regional Reconstruction Results (Table 14). CO_2-Net outperforms all baselines on both reanalysis data across different regions, showing its capability in regional reconstruction. We include results from Table 3 for direct comparison.

Reconstruction Results on Real Observation Data (Table 15). Even our smallest model CO_2-Net (S) achieves RMSE comparable to inversion models while outperforming other baselines. CO_2-Net (B) and (L) further reduce RMSE as model size increases.

Ablation Study of Key Components on CarbonTracker (Table 16a and Table 16b). We find that integrating wind-based embedding and the semi-supervised loss improves the performance of CO_2-Net (L). However, the integration either worsens performance or brings limited improvement on other baselines, which is similar on CMIP6.

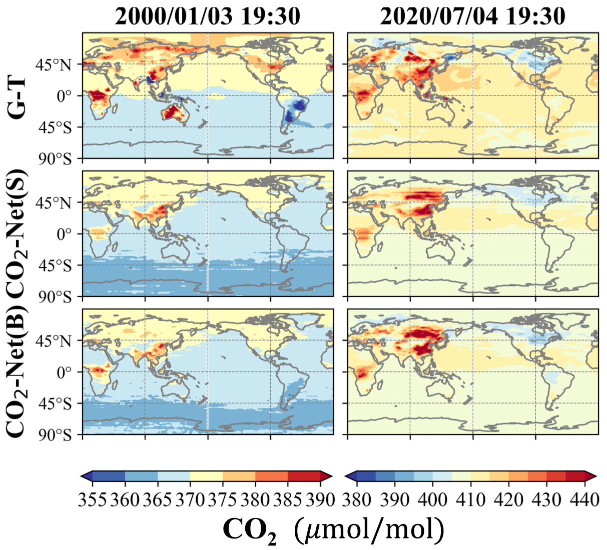

Figure 6. Ground-truth and model reconstructed results of CO_2 concentration (μmol/mol) on CarbonTracker in 2000 and 2020.

Ablation Study of Loss Coeffcient (Table 16c). We find the optimal value of λ through grid search within the set $0.1, 0.2, 0.5, 1, 2$. We find the best value of λ to be 0.2 and 1 on CarbonTracker and CMIP6, respectively.

F.4. Numerical Methods

We adopt kriging interpolation [30] as the numerical baseline for comparison. Kriging is a widely used geostatistical method that reconstructs unobserved values from sparse observational data. A key component of kriging interpolation is the variogram model, which quantifies the spatial correlation between points. We evaluate five distinct variogram models: `Linear`, `Power`, `Gaussian`, `Spherical` and `Exponential`. We apply kriging interpolation to reconstruct atmospheric CO_2 concentrations from spot observational data using five different variogram models. The results are summarized in Table 17, indicating that the `Spherical` model yields the best performance among the variogram models. However, despite this, its performance remains inferior when compared to data-driven methods.

Table 14. RMSE (↓) in $\mu\mathrm{mol}/\mathrm{mol}$ and ACC (↑) comparison of different models across datasets and regions. $*$ indicates spatio-temporal reconstruction models, others are static reconstruction models.

Dataset	Methods	Sizes (M)	Global		Ocean		Asia-Europe		North-America		South-America	
			RMSE	ACC	RMSE	ACC	RMSE	ACC	RMSE	ACC	RMSE	ACC
CMIP6	Senseiver	0.11	36.47±6.81	0.59±0.16	31.94	0.79	40.52	0.43	38.25	0.42	65.21	0.07
	SwinLSTM*	3.3	27.42±0.09	0.52±0.01	28.41	0.52	27.70	0.51	25.16	0.48	33.33	0.49
	CycleGAN	28	51.48±0.36	0.12±0.04	53.34	0.12	52.01	0.11	47.24	0.11	62.63	0.12
	ViT	76	18.18±1.01	0.49±0.02	18.86	0.49	18.38	0.49	16.96	0.48	21.78	0.49
	CO_2-Net* (S)	38	16.18±1.19	0.73±0.07	22.48	0.72	16.27	0.79	19.60	0.77	19.59	0.74
	CO_2-Net* (B)	95	9.91±0.93	0.98±0.03	10.27	**0.99**	9.91	**0.98**	9.06	**0.98**	13.55	0.93
	CO_2-Net* (L)	247	**5.31**±0.49	**0.99**±0.00	**5.49**	**0.99**	**5.36**	0.97	**4.87**	0.93	**6.46**	**0.97**
Carbon Tracker	Spherical	–	7.41±0.40	0.12±0.01	5.55	0.14	9.80	0.21	8.23	0.12	14.27	0.20
	Exponential	–	7.40±0.37	0.11±0.01	5.56	0.13	9.83	0.20	8.18	0.11	14.20	0.20
	Senseiver	0.11	6.39±0.45	0.43±0.10	4.74	0.28	9.70	0.50	6.92	0.58	13.72	0.30
	SwinLSTM*	3.3	5.31±0.01	0.61±0.00	2.90	0.45	6.33	0.65	8.83	0.65	12.20	0.50
	CycleGAN	28	4.70±0.01	0.71±0.00	1.57	0.88	8.39	0.73	5.68	0.78	11.21	0.70
	ViT	76	5.42±0.21	0.50±0.04	2.27	0.72	9.43	0.63	6.89	0.54	13.32	0.19
	CO_2-Net* (S)	38	3.59±0.06	0.72±0.01	1.32	0.84	6.08	0.72	4.25	0.72	9.36	0.66
	CO_2-Net* (B)	95	3.41±0.04	0.77±0.01	**1.03**	0.90	5.58	0.77	**3.91**	0.79	8.68	0.75
	CO_2-Net* (L)	247	**3.36**±0.04	**0.85**±0.00	1.09	**0.94**	**5.39**	**0.85**	3.94	**0.83**	**8.42**	**0.80**

Dataset	Methods	Sizes (M)	Asia		Europe		Africa		Australia		Antarctica	
			RMSE	ACC	RMSE	ACC	RMSE	ACC	RMSE	ACC	RMSE	ACC
CMIP6	Senseiver	0.11	64.04	0.02	27.85	0.83	38.57	0.80	59.90	0.10	23.70	0.36
	SwinLSTM*	3.3	28.17	0.51	25.94	0.51	33.55	0.01	32.68	0.49	13.79	0.49
	CycleGAN	28	52.89	0.12	48.70	0.11	63.07	0.12	61.39	0.12	25.80	0.12
	ViT	76	18.62	0.49	17.38	0.48	21.43	0.48	21.03	0.50	9.28	0.49
	CO_2-Net* (S)	38	17.17	0.73	18.62	0.74	29.75	0.77	8.42	0.71	28.32	0.76
	CO_2-Net* (B)	95	10.08	0.98	9.16	**0.98**	12.03	**0.99**	11.73	**0.99**	14.32	0.70
	CO_2-Net* (L)	247	**5.47**	**1.00**	**5.02**	0.96	**6.51**	0.97	**6.31**	**0.99**	**2.63**	**1.00**
Carbon Tracker	Spherical	–	9.92	0.22	7.99	0.29	10.68	0.18	7.71	0.14	3.42	0.14
	Exponential	–	9.95	0.21	8.05	0.28	10.62	0.18	7.70	0.15	3.41	0.13
	Senseiver	0.11	10.97	0.29	8.82	0.35	10.81	0.30	5.16	0.04	2.71	0.37
	SwinLSTM*	3.3	8.82	0.65	7.54	0.65	8.34	0.64	4.07	0.06	1.61	0.07
	CycleGAN	28	8.44	0.72	6.80	0.81	7.63	0.76	3.14	0.62	0.33	0.98
	ViT	76	9.23	0.52	7.69	0.58	9.38	0.34	3.75	0.46	0.68	0.92
	CO_2-Net* (S)	38	6.04	0.73	5.13	0.70	5.58	0.71	**2.31**	0.37	**0.13**	0.98
	CO_2-Net* (B)	95	5.75	0.77	5.31	0.75	5.42	0.79	2.37	0.53	0.15	0.99
	CO_2-Net* (L)	247	**5.66**	**0.86**	**5.03**	**0.78**	**5.26**	**0.86**	**2.31**	**0.67**	0.16	**1.00**

Table 15. RMSE(↓) in $\mu\mathrm{mol}/\mathrm{mol}$ comparison on real observations.

Methods	4D-Var.	COLA	Sphe.	Expo.	Sens.	ViT	Cycle.	Swin.	CO_2-Net (S)	CO_2-Net (B)	CO_2-Net (L)
RMSE	**7.80**	7.94	8.46	8.40	8.75	9.10	9.45	8.34	8.00	7.86	7.81

Table 16. **Ablation study (RMSE (↓, in $\mu\mathrm{mol}/\mathrm{mol}$) |ACC (↑)).** on CarbonTracker dataset. (a) Influence of the wind-flow based spatio-temporal embedding; (b) Impact of the semi-supervised physics-informed loss function; (c) Evaluation of varying λ values for the semi-supervised loss coefficient.

(a) **Wind-Flow based Embedding.**
(Settings: CarbonTracker, w/o semi-sup. loss)

Methods	Wind	ClimODE	None
Cycle-GAN	5.00 |0.65	4.74 |0.70	**4.70** |**0.71**
SwinLSTM	5.32 |0.61	5.32 |0.61	**5.31** |**0.61**
ViT	**5.25** |0.54	5.29 |**0.55**	5.42 |0.50
CO_2-Net (L)	**3.43** |**0.84**	3.53 |0.84	3.55 |0.83

(b) **Semi-supervised Loss.**
(Settings: CarbonTraker, w/ wind-based emb.)

Methods	w/	w/o
Cycle-GAN	5.15 |0.61	**5.00** |**0.65**
SwinLSTM	5.33 |0.61	**5.32** |**0.61**
ViT	**5.17** |**0.56**	5.25 |0.54
CO_2-Net (L)	**3.36** |**0.85**	3.43 |0.84

(c) **Loss Coeffcient λ.**
(Settings: w/ wind-based emb., w/ semi-sup. loss)

λ	CT	CMIP6
0.1	3.38 |0.85	36.89 |0.60
0.2	**3.36** |**0.85**	19.51 |0.97
0.5	3.47 |0.85	13.17 |**0.99**
1	3.40 |0.85	**5.31** |**0.99**
2	3.45 |0.84	18.25 |0.98

Table 17. Comparison of RMSE (↓, in $\mu mol/mol$) and ACC (↑) for different variogram models used in kriging interpolation across various regions.

Regions	Linear		Power		Gaussian		Spherical		Exponential	
	RMSE	ACC	RMSE	ACC	RMSE	ACC	RMSE	ACC	RMSE	ACC
Global	7.59	**0.13**	7.50	0.12	6382.26	0.12	7.41	0.12	**7.40**	0.11
Ocean	5.92	**0.17**	5.76	0.16	6620.94	0.14	**5.55**	0.14	5.56	0.13
North America	**7.83**	**0.18**	8.07	0.14	875.58	0.13	8.23	0.12	8.18	0.11
South America	**14.05**	0.19	14.07	**0.20**	3134.77	0.19	14.27	**0.20**	14.20	**0.20**
Asia	9.84	**0.22**	**9.82**	**0.22**	7398.96	0.20	9.92	**0.22**	9.95	0.21
Europe	**7.89**	**0.30**	7.91	0.29	7568.58	0.26	7.99	0.29	8.05	0.28
Australia	8.80	**0.15**	8.55	**0.15**	12979.02	0.12	7.71	0.14	**7.70**	**0.15**
Africa	10.59	0.18	**10.55**	**0.19**	5493.61	0.14	10.68	0.18	10.62	0.18
Antarctica	3.50	**0.14**	3.48	**0.14**	877.74	0.09	3.42	**0.14**	**3.41**	0.13