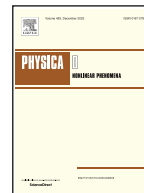




ELSEVIER

Contents lists available at ScienceDirect

Physica D

journal homepage: [www.elsevier.com/locate/physd](http://www.elsevier.com/locate/physd)

# Spatiotemporal system forecasting with irregular time steps via masked autoencoder

Kewei Zhu <sup>a,1</sup>, Yanze Xin <sup>b,1</sup>, Jinwei Hu <sup>c</sup>, Xiaoyuan Cheng <sup>d</sup>, Yiming Yang <sup>e,f</sup>,  
Sibo Cheng <sup>g,\*</sup>

<sup>a</sup> Department of Chemical Engineering, University College London, UK

<sup>b</sup> Department of Computing, Imperial College London, UK

<sup>c</sup> Department of Computer Science, University of Liverpool, UK

<sup>d</sup> Dynamic Systems Lab, University College London, UK

<sup>e</sup> Department of Civil, Environmental & Geomatic Engineering, University College London, UK

<sup>f</sup> Department of Statistical Science, University College London, UK

<sup>g</sup> CEREA, ENPC, EDF R&D, Institut Polytechnique de Paris, France

## ARTICLE INFO

Communicated by Dmitry Pelinovsky

### Keywords:

Deep learning

Self-attention

Fluid dynamics

Spatiotemporal forecasting

Climate prediction

## ABSTRACT

Predicting high-dimensional dynamical systems with irregular time steps presents significant challenges for current data-driven algorithms. These irregularities arise from missing data, sparse observations, or adaptive computational techniques, reducing prediction accuracy. To address these limitations, we propose a novel method: a Physics-Spatiotemporal Masked Autoencoder. This method integrates convolutional autoencoders for spatial feature extraction with masked autoencoders optimised for irregular time series, leveraging attention mechanisms to reconstruct the entire physical sequence in a single prediction pass. The model avoids the need for data imputation while preserving the physical integrity of the system. Here, 'physics' refers to high-dimensional fields generated from underlying dynamical systems, rather than enforcing explicit physical constraints or PDE residuals. We evaluate this approach on multiple simulated datasets and real-world ocean temperature data. The results demonstrate that our method achieves significant improvements in prediction accuracy, robustness to nonlinearities, and computational efficiency over traditional convolutional and recurrent network methods. The model shows potential for capturing complex spatiotemporal patterns without requiring domain-specific knowledge, with applications in climate modelling, fluid dynamics, ocean forecasting, environmental monitoring, and scientific computing.

## 1. Introduction

Predicting high-dimensional dynamical systems is challenging when observations occur at irregular time intervals. Such irregularities frequently arise in scientific simulations and observational datasets due to sensor failures, sparse measurement networks [1], or adaptive time

stepping in numerical solvers based on Partial Differential Equations (PDEs) [2–4]. Conventional machine learning models, such as Multi-Layer Perceptrons (MLPs) and Recurrent Neural Networks (RNNs), typically assume regularly sampled data and struggle to generalise when faced with temporal gaps or uneven sampling [5]. To compensate, many workflows rely on preprocessing techniques [6], such as resampling,

*Acronyms:* P-STMAE, Physics Spatiotemporal Masked Autoencoder; PDE, Partial Differential Equation; ARIMA, Auto-Regressive Integrated Moving Average; MLP, Multi-Layer Perceptron; CNN, Convolutional Neural Network; RNN, Recurrent Neural Network; M-RNN, Multi-directional Recurrent Neural Network; Seq2Seq, Sequence-to-Sequence; GRU, Gated Recurrent Unit; LSTM, Long Short-Term Memory Network; ConvRAE, Deep Convolutional Recurrent Autoencoder; ConvLSTM, Convolutional Long Short-Term Memory Network; FC-LSTM, Fully Connected LSTM; NLP, Natural Language Processing; BERT, Bidirectional Encoder Representations from Transformers; ROM, Reduced Order Modelling; POD, Proper Orthogonal Decomposition; DMD, Dynamic Mode Decomposition; PCA, Principal Component Analysis; CAE, Convolutional Autoencoder; MAE, Masked Autoencoder; TiMAE, Time Series Masked Autoencoder; MSE, Mean Squared Error; SSIM, Structural Similarity Index Measure; PSNR, Peak Signal-to-Noise Ratio; SWE, Shallow Water Equation; SST, NOAA Sea Surface Temperature; VAE, Variational Autoencoder; GAN, Generative Adversarial Network; MSPCNN, Multi-Scale Physics-Constrained Neural Network; PINN, Physics-Informed Neural Network; KAN, Kolmogorov-Arnold Network.

\* Corresponding author.

E-mail address: [sibo.cheng@enpc.fr](mailto:sibo.cheng@enpc.fr) (S. Cheng).

<sup>1</sup> These authors contributed equally to this work.

<https://doi.org/10.1016/j.physd.2026.135189>

Received 3 October 2025; Received in revised form 24 December 2025; Accepted 17 March 2026

Available online 19 March 2026

0167-2789/© 2026 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

**Main Notations**

$x_t$	Observation state in the full space at time step $t$
$\hat{x}_t$	Predicted state in the full space at time step $t$
$z_t$	Observation state in the latent space at time step $t$
$\hat{z}_t$	Predicted state in the latent space at time step $t$
$T_{in}$	Set of input time steps
$T_{out}$	Set of forecasting time steps
$T_{miss}$	Set of missing time steps of the input
$T_{obs}$	Set of observed time steps of the input
$t_{in}$	Number of input time steps
$t_{out}$	Number of forecasting time steps
$\mathbf{X}_T$	Physical states for the set of time steps $T$
$\hat{\mathbf{X}}$	Reconstructed sequence in the full space
$\mathbf{Z}_T$	Latent states for the set of time steps $T$
$\hat{\mathbf{Z}}$	Reconstructed sequence in the latent space
$\Phi_x$	Placeholders in the full space
$\phi_x$	Physical placeholder state
$\Phi_z$	Placeholders in the latent space
$\phi_z$	Latent placeholder state
$d_x$	Dimension of the physical state
$d_z$	Dimension of the latent state
$d_k$	Dimension of the attention vector
$\mathbf{Q}, \mathbf{K}, \mathbf{V}$	Attention matrices of transformer blocks
$\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$	Trainable weight matrices of transformer blocks
$\mathbf{A}$	Attention weights matrix
$\mathbf{O}$	Output matrix from transformer block
$\mathbf{O}'$	Final output matrix from transformer block
$f_E$	Encoder function mapping physical to latent space
$f_D$	Decoder function mapping latent to physical space
$\theta_E$	Parameters of the encoder
$\theta_D$	Parameters of the decoder
$\delta_t$	Positional embedding at time step $t$
$\mathcal{L}$	Overall loss function
$\lambda$	Weighting coefficient for the latent loss

interpolation [7], or data assimilation [8] to produce uniformly spaced sequences. However, these procedures can introduce bias, increase computational cost, and obscure the true temporal dynamics of the system [9,10]. There is a clear need for models that can directly learn from irregular time series without preprocessing while accurately capturing the underlying spatiotemporal structure of the physical system.

Traditional approaches to handling temporally irregular observations include time series models like Auto-Regressive Integrated Moving Average (ARIMA) [11] and data assimilation algorithms such as Kalman Filters [12]. While ARIMA models are effective for univariate, stationary time series, they face limitations with high-dimensional, nonlinear systems [13,14]. Kalman Filters estimate the state of a linear system from incomplete measurements but rely on assumptions of linearity and normality that often do not hold in complex, high-dimensional dynamics [15]. Extended and Unscented Kalman Filters [16–19] attempt to address nonlinearities but still struggle with high dimensionality and irregular time steps common in real-world data.

More recently, deep learning models like Convolutional Neural Networks (CNNs) and RNNs have shown advantages in surrogate modelling of time series problems by leveraging their ability to capture spatial and temporal patterns, respectively. Deep Convolutional Recurrent Autoencoder (ConvRAE) [20] combines CNNs with Long Short-Term Memory Networks (LSTMs) to capture both spatial and temporal patterns. However, it inherits RNNs' drawbacks, such as vanishing and exploding gradients [21], which become detrimental for long sequences in high-dimensional dynamical systems [22]. Convolutional Long Short-Term Memory Network (ConvLSTM) [23] combines convolutional operations

with LSTM cells to directly model spatiotemporal relations. Nevertheless, these RNN-based approaches rely on regularly sampled time series, limiting their applicability to irregular time steps. Alternative methods [24] require interpolation to estimate time points that do not align with the established time grid. These limitations prompt the need for models capable of handling irregular time steps and incomplete data while preserving the integrity of physical processes (see Fig. 1).

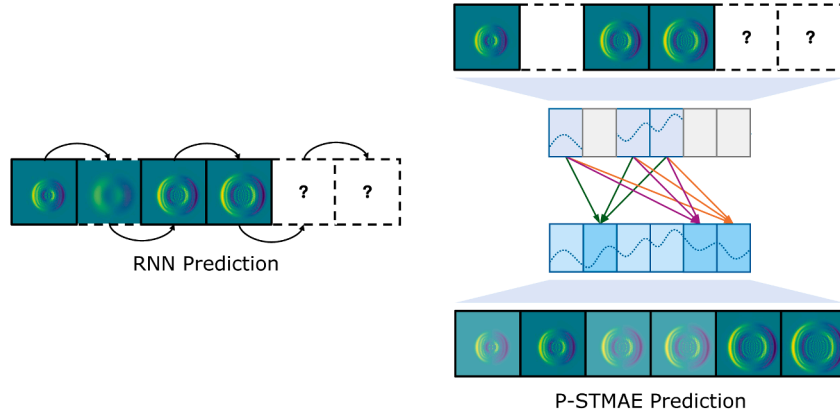
Recent advancements in transformers [25] have introduced a powerful architecture for time series modelling for handling irregular and incomplete data [26,27]. Originally developed for Natural Language Processing (NLP) [28], transformers are well-suited for sequences of variable lengths and missing elements due to their self-attention mechanism. Unlike RNNs that process inputs sequentially, this mechanism enables simultaneous attention to various parts of the input sequence and is highly parallelisable, thereby capturing long-range dependencies even when parts of the data are missing or unevenly spaced. Methods like Bidirectional Encoder Representations from Transformers (BERT) [29] in language understanding and Masked Autoencoder (MAE) [30] in image recognition demonstrate the efficacy of masking strategies for learning robust representations. Building on these successes, Time Series Masked Autoencoder (TiMAE) [31] has demonstrated the utility of self-supervised learning and masked modelling for reconstructing missing data points in time series prediction. However, this approach has largely been applied in low-dimensional systems, such as financial [32] or healthcare [33] time series, and has not been fully extended to more complex, high-dimensional dynamical systems governed by physical processes. Modelling these systems presents challenges due to computational complexity and memory usage associated with high dimensionality, while maintaining spatial and temporal coherence under irregular observations remains difficult for traditional deep learning models and common transformer variants [34].

Existing approaches for irregular time series modelling in dynamical systems exhibit fundamental limitations. Neural ordinary differential equation (Neural ODE)-based methods require continuous-time solvers and are often sensitive to stiffness and numerical instability in high-dimensional PDE systems. Interpolation-based transformers and recurrent models rely on resampling or imputation, which can distort the true temporal dynamics. In contrast, masked reconstruction enables direct modelling of irregularly sampled sequences without explicit interpolation or continuous-time integration.

To address these challenges, we propose a novel model called Physics Spatiotemporal Masked Autoencoder (P-STMAE). Unlike previous transformer models primarily focused on low-dimensional data, P-STMAE is specifically designed for modelling high-dimensional dynamical systems, incorporating both spatial and temporal dependencies in a unified framework. The core innovation lies in combining a convolutional autoencoder for spatial feature extraction with a masked autoencoder optimised for irregular time series prediction. The convolutional autoencoder compresses high-dimensional physical data into a low-dimensional latent space, thereby reducing computational complexity while retaining essential spatiotemporal features. In the latent space, a masked autoencoder uses the transformer's self-attention mechanism to predict future states. The framework introduces placeholder and masking strategies to handle temporal dependencies among partially observed sequences, with positional encodings preserving the temporal order under irregular sequences. The training adopts a purely data-driven approach [35], optimising a combination of physical and latent space losses without domain-specific knowledge.

We conduct numerical experiments on three datasets: two simulated scenarios from PDEBench [36] (Shallow Water equations and Diffusion Reaction equations) and one real-world ocean fluid dataset, NOAA Sea Surface Temperature (SST) [37], obtained from satellite and ship-based observations. This combination ensures that P-STMAE adheres to scientific standards and generalises to diverse challenges in physical systems.

In summary, this paper makes the following key contributions:



**Fig. 1.** Comparison of Sequence-to-Sequence (Seq2Seq) prediction methods in dynamical systems of irregular time steps. **Left:** Traditional RNN-based models feature step-wise rolling out with necessary data imputation for handling missing steps, which may introduce biases and cumulative errors. **Right:** Our model performs element-wise predictions in the latent space by adaptive attention mechanism to reconstruct the complete sequence in a single pass.

- A spatiotemporal masked autoencoder for latent dynamics modelling.
- Placeholder-based attention for handling irregular and missing time steps.
- A unified framework for sequence reconstruction and forecasting.
- The proposed model outperforms ConvLSTM and ConvRAE with improved efficiency and interpretability.

Our approach provides computational advantages over traditional physics-based PDE solvers. Traditional time-stepping methods often require many small steps due to stability constraints and may involve iterative linear solves, whereas P-STMAE learns an approximation to the flow map and produces forecasts with a limited number of GPU-based forward passes [38,39], thereby reducing inference time and energy consumption while preserving spatiotemporal fidelity.

To avoid potential ambiguity, we clarify that the notion of "physical consistency" in this work is relative rather than constraint-based. P-STMAE does not explicitly enforce PDE residuals or conservation laws. Instead, inspired by masked autoencoder approaches for time-series dynamics such as TS-MAE [40], it models the temporal evolution of physical fields in a latent space and improves the stability and coherence of the learned dynamics without introducing hard physics constraints. Related latent-space forecasting approaches have also been explored and validated in independent studies [41,42].

To the best of our knowledge, our proposed model is among the first to unify Convolutional Autoencoder (CAE)-based spatial compression with masked temporal modelling using transformers in the latent space, specifically targeting high-dimensional, irregularly sampled dynamical systems.

The remainder of this paper is organised as follows: [Section 2](#) provides related work on reduced order modelling and deep learning-based approaches for high-dimensional dynamical systems. [Section 3](#) introduces the proposed P-STMAE model, detailing its architecture and methodology. [Section 4](#) presents the numerical experiments conducted on both synthetic and real-world datasets to evaluate the model's performance. Finally, [Section 5](#) discusses limitations and concludes the paper.

## 2. Related work

### 2.1. Traditional approaches to irregular time series

Traditional approaches to handling temporally irregular observations include time series models like ARIMA [11] and data assimilation algorithms such as Kalman Filters [12]. While ARIMA models are effective for univariate, stationary time series, they face limitations with high-dimensional, nonlinear systems [13,14]. Kalman Filters es-

timate the state of a linear system from incomplete measurements but rely on assumptions of linearity and normality that often do not hold in complex, high-dimensional dynamics [15]. Extended and Unscented Kalman Filters [16,17] attempt to address nonlinearities but still struggle with high dimensionality and irregular time steps common in real-world data. These limitations motivate the development of deep learning approaches capable of handling high-dimensional, nonlinear dynamics with irregular temporal sampling.

### 2.2. Reduced order modelling and autoencoders

Reduced Order Modelling (ROM) aims to reduce the computational cost of simulating high-fidelity dynamical systems by constructing efficient surrogate models that preserve essential system dynamics [43–47]. These models enable faster predictions by approximating the original system in a lower-dimensional space while retaining relevant features for downstream tasks [48,49].

Traditional projection-based methods, such as Proper Orthogonal Decomposition (POD) [50] or Dynamic Mode Decomposition (DMD), project data onto optimal linear subspaces that explain most of the variance. However, their effectiveness is limited when the system exhibits strong nonlinearity and time-varying behaviour. Moreover, they are often intrusive, requiring access to the governing equations or system operators during the reduction process [51].

Deep learning offers a non-intrusive and flexible alternative for ROM. In particular, autoencoders can learn nonlinear manifolds from data alone, making them effective in compressing and reconstructing high-dimensional spatial features [52,53]. CAEs are widely used in spatiotemporal modelling, where spatial patterns can be compressed into latent representations and later decoded to reconstruct the original fields.

Formally, a CAE consists of an encoder and decoder:

$$\mathbf{z}_t = f_E(\mathbf{x}_t; \theta_E), \quad (1)$$

$$\hat{\mathbf{x}}_t = f_D(\mathbf{z}_t; \theta_D), \quad (2)$$

where  $t \in \mathbb{N}$  represents each valid time step, and  $f_E : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_z}$  represents the encoding function that maps the input physical state  $\mathbf{x}_t$  to the compressed latent space  $\mathbf{z}_t$  using the parameters  $\theta_E$ , commonly  $d_z \ll d_x$ . Similarly,  $f_D : \mathbb{R}^{d_z} \rightarrow \mathbb{R}^{d_x}$  denotes the decoding function that reconstructs the original physical state  $\hat{\mathbf{x}}_t \in \mathbb{R}^{d_x}$  using the parameters  $\theta_D$ .

The training goal is to minimise the reconstruction loss, which measures the mean squared error between the input and output physical states averaged over total time steps  $T$ :

$$\theta_E^*, \theta_D^* = \arg \min_{\theta_E, \theta_D} \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[ \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 \right]. \quad (3)$$

A notable advantage of using deep autoencoders in ROM is their ability to learn nonlinear manifolds, for high-dimensional and nonlinear dynamics, which are common in fluid mechanics, climate systems, and biological simulations.

### 2.3. Sequence modelling in latent space

RNNs [54] are a type of neural networks tailored for handling sequence data, making them effective for time series and sequential tasks. Unlike traditional feed-forward neural networks, RNNs capture information from previous states through internal hidden memories, enabling them to maintain and process temporal dependencies [55]. Specifically, LSTMs use three information gates, including input, forget, and output gates, to regulate what information should be added, retained, or output from the cell state, thus maintaining long-term dependencies more effectively [56]. They are widely used in reduced-order spatiotemporal system modelling in computational physics [57,58].

#### 2.3.1. Latent sequence forecasting via ConvRAE and ConvLSTM

One notable approach to employs RNNs for high-dimensional dynamical system modelling is the ConvRAE [20]. The model first employs a deep convolutional autoencoder to compress a sequence of physical fields  $\{(x_t, t)\}_{t=1}^T$  into the latent space  $\{(z_t, t)\}_{t=1}^T$ . The compressed representation is then sent into an LSTM network to model its temporal evolution, thus preserving future predictive capabilities by autoregressive rollouts.

Mathematically, it predicts future states  $\{(\hat{z}_t, t)\}_{t=T_{in}+1}^{T_{in}+T_{out}}$  based on input states  $\{(z_t, t)\}_{t=1}^{T_{in}}$ . For each forecasting step, it predicts the next state  $\hat{z}_{t+1}$  in an autoregressive manner based on the previous output  $\hat{z}_t$  and a hidden memory  $\mathbf{h}_t$ :

$$\hat{z}_{t+1}, \mathbf{h}_{t+1} = \text{LSTM-Cell}(\hat{z}_t, \mathbf{h}_t). \quad (4)$$

Common loss such as Mean Square Error (MSE) can be used to minimise the distance between predicted physical states and the ground truth, i.e.

$$\hat{\mathbf{x}}_t = f_D(\hat{z}_t; \theta_D), \quad (5)$$

$$\mathcal{L}_{\text{ConvRAE}} = \frac{1}{T_{\text{out}}} \sum_{t=T_{in}+1}^{T_{in}+T_{out}} \mathbb{E} \left[ \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 \right]. \quad (6)$$

ConvRAE effectively integrates the strengths of CNNs in capturing localised spatial features and LSTMs in preserving temporal dependencies. However, it inherits common drawbacks associated with RNNs, including susceptibility to vanishing and exploding gradients [21], which become problematic when dealing with long sequences typical in high-dimensional dynamical systems [22].

Another influential model is the ConvLSTM [23], which is specifically designed for spatiotemporal sequence forecasting in a Seq2Seq framework. ConvLSTM extends the conventional Fully Connected LSTM (FC-LSTM) by incorporating convolutional operations directly into the cell of the LSTM to model temporal transitions at the full physical space. Unlike ConvRAE, which operates as a two-stage method, ConvLSTM offers an end-to-end approach that simultaneously captures both spatial and temporal dependencies within a unified architecture. However, both ConvRAE and ConvLSTM require regularly sampled time series, which limits their applicability to irregular time steps without preprocessing.

#### 2.3.2. Challenges with irregular time series

While RNNs and their variants excel in modelling regularly sampled time series, they face significant challenges when dealing with irregular time steps [59] and high dimensionality [60]. Standard RNN-based Seq2Seq models inherently assume consistent sampling intervals, making it difficult to handle missing or unevenly spaced observations without additional preprocessing steps.

To address this, many RNN-based models rely on explicit imputation or interpolation as a preprocessing step [6], including statistical interpolation [7,61], nearest neighbour search [62], and time-aware data filling [63]. However, these preprocessing procedures can introduce bias, increase computational cost, and obscure the true temporal dynamics of the system [9,10], thereby limiting their effectiveness in capturing complex spatiotemporal dynamics under high sparsity.

More recent work has used generative models such as Variational Autoencoders (VAEs) [64] and Generative Adversarial Networks (GANs) [65,66] to impute missing values. However, these approaches may suffer from data inefficiency, instability in long-term predictions, and non-uniqueness of outputs, limiting their use in physical system modelling [64,67,68]. While probabilistic generative models are highly effective for noisy or underdetermined real-world systems and can explicitly represent predictive uncertainty [69], this work focuses on deterministic reconstruction to enable clarity and controlled analysis of irregularly sampled latent dynamics in deterministic PDE simulation settings, where the governing equations define a single-valued flow map [70].

### 2.4. Masked transformers and physics-informed latent modelling

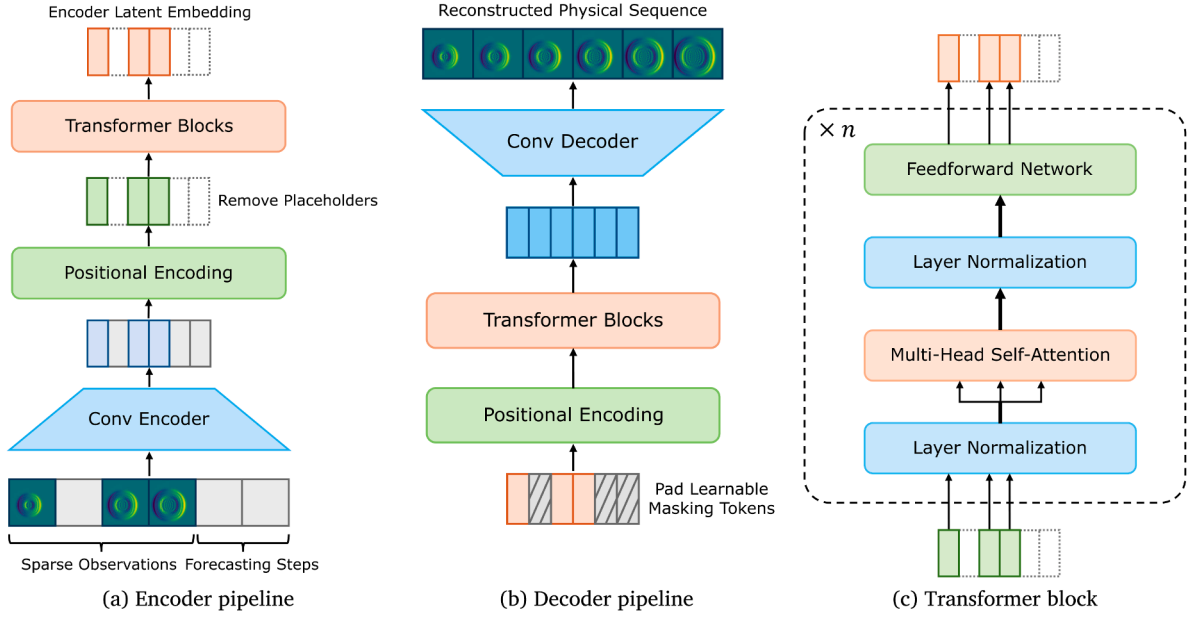
Existing approaches for irregular time series modelling in dynamical systems exhibit fundamental limitations. Neural ordinary differential equation (Neural ODE) based methods require continuous-time solvers and are often sensitive to stiffness and numerical instability in high-dimensional PDE systems. Interpolation-based transformers and recurrent models rely on resampling or imputation, which can distort the true temporal dynamics. Masked reconstruction offers an alternative by enabling direct modelling of irregularly sampled sequences without explicit interpolation or continuous-time integration.

Recent transformer-based models have shown promise in modelling long-range dependencies and handling irregular sequences directly through masked attention and learned time encodings. Works like Time-MAE and masked sequence models [71,72] have successfully addressed missing data without interpolation. Building on these successes, TiMAE [31] has demonstrated the utility of self-supervised learning and masked modelling for reconstructing missing data points in time series prediction. However, these approaches have primarily been applied to low-dimensional systems, such as financial [32] or healthcare [33] time series, and have not been fully extended to the more complex, high-dimensional dynamical systems governed by physical processes. Modelling these systems presents challenges due to computational complexity and memory usage associated with high dimensionality, while maintaining spatial and temporal coherence under irregular observations remains difficult for traditional deep learning models and common transformer variants [34].

Simultaneously, there is a growing interest in combining physical priors with latent representation learning [73]. Physics-informed neural networks (PINNs) and hybrid data-physics models have shown effectiveness in preserving physical consistency and improving generalisation under small data regimes.

## 3. Methodology

P-STMAE addresses irregular time series prediction using a masked modelling strategy by combining a convolutional autoencoder for spatial representation and a masked transformer for temporal modelling. The framework directly handles irregular time steps through masked reconstruction, eliminating the need for preprocessing or regular sampling that can introduce bias and computational overhead. By leveraging transformer self-attention instead of sequential RNNs processing, P-STMAE captures long-range dependencies without suffering from vanishing gradients or requiring regular input sequences. In contrast to Neural ODE methods that require continuous-time solvers, P-STMAE learns a discrete flow map in a latent space, enabling efficient inference without



**Fig. 2.** Architecture of the proposed P-STMAE framework. (a) Convolutional encoder compresses physical states into latent representations. Positional encodings are added, and a masked transformer captures temporal dependencies in latent space. (b) Learnable masking tokens are padded at missing and future time steps. Transformer blocks process the sequence, and the convolutional decoder reconstructs the complete physical fields. (c) Each transformer block consists of layer normalization, multi-head self-attention, and a feedforward network. Self-attention operates only on observed latent states.

numerical integration. To address the computational challenges of applying transformers to high-dimensional systems, P-STMAE operates in a compressed latent space and reduces memory usage while maintaining spatiotemporal coherence. This section outlines the overall framework, the encoder and decoder designs, the attention-based temporal model, and the training loss formulation.

### 3.1. Overall framework

Mathematically, we consider a physical sequence consisting of input states  $\mathbf{X}_{T_{\text{in}}}$  defined on input steps  $T_{\text{in}} = \{1, 2, \dots, t_{\text{in}}\}$ , followed by forecasting states  $\mathbf{X}_{T_{\text{out}}}$  on output steps  $T_{\text{out}} = \{t_{\text{in}} + 1, \dots, t_{\text{in}} + t_{\text{out}}\}$ . We denote:

$$\mathbf{X}_{T_{\text{in}}} = \{(\mathbf{x}_t, t) \mid t \in T_{\text{in}}\}, \quad \mathbf{X}_{T_{\text{out}}} = \{(\mathbf{x}_t, t) \mid t \in T_{\text{out}}\}, \quad (7)$$

where  $\mathbf{x}_t \in \mathbb{R}^{d_x}$  is the physical state at time  $t$ .

To simulate irregularity, we randomly split  $T_{\text{in}}$  into two disjoint sets:  $T_{\text{obs}}$  (observed steps) and  $T_{\text{miss}}$  (missing steps), such that  $T_{\text{obs}} \cup T_{\text{miss}} = T_{\text{in}}$ . This design enables learning from partially observed sequences without data imputation or interpolation, thereby addressing the limitations inherent in preprocessing-based approaches. Missing and future steps are replaced with placeholder variables  $\Phi_x$ :

$$\Phi_x = \{(\phi_x, t) \mid t \in T_{\text{miss}} \cup T_{\text{out}}\}, \quad \phi_x \in \mathbb{R}^{d_x}. \quad (8)$$

The model then reconstructs the complete sequence  $\hat{\mathbf{X}}$  from  $\mathbf{X}_{T_{\text{obs}}}$  and placeholders  $\Phi_x$ :

$$\hat{\mathbf{X}} = \text{P-STMAE}(\mathbf{X}_{T_{\text{obs}}}, \Phi_x) = \{(\hat{\mathbf{x}}_t, t) \mid t \in T_{\text{in}} \cup T_{\text{out}}\}. \quad (9)$$

Fig. 2 illustrates the overall pipeline of the proposed P-STMAE framework.

### 3.2. Spatial encoder: Convolutional autoencoder

To reduce spatial redundancy and improve computational efficiency, we adopt a CAE to project high-dimensional inputs  $\mathbf{x}_t \in \mathbb{R}^{d_x}$  into compact latent representations  $\mathbf{z}_t \in \mathbb{R}^{d_z}$ , where  $d_z \ll d_x$ . This compression mitigates the computational complexity and memory challenges of applying transformers directly to high-dimensional physical fields, facilitating efficient temporal modelling in the latent space.

During encoding, physical placeholders  $\Phi_x$  are converted into latent placeholders  $\Phi_z$ :

$$\Phi_z = \{(\phi_z, t) \mid t \in T_{\text{miss}} \cup T_{\text{out}}\}, \quad \phi_z \in \mathbb{R}^{d_z}. \quad (10)$$

These placeholders are excluded from backpropagation and later masked out in the transformer blocks.

The placeholder states  $\phi_x$  and  $\phi_z$  are fixed tensors (set to zero after normalisation), excluded from gradient backpropagation, and serve solely as positional anchors for masking and attention mechanisms.

### 3.3. Temporal modelling: Masked autoencoder

The masked autoencoder learns latent temporal dynamics from partially observed sequences. By operating directly on irregularly sampled latent states without preprocessing, this approach circumvents the bias and distortion that interpolation methods typically introduce. It processes the latent inputs  $\mathbf{Z}_{T_{\text{obs}}} = \{(\mathbf{z}_t, t) \mid t \in T_{\text{obs}}\}$  using transformer blocks.

#### 3.3.1. Masked transformer blocks

Transformer blocks (see Fig. 2(c)) compute self-attention over observed latent states. In contrast to RNNs that process sequences sequentially and are prone to vanishing gradients, self-attention enables parallel processing and direct access to all observed time steps, capturing long-range dependencies without gradient degradation. Given the hidden and attention dimension, the attention weights are computed as:

$$\mathbf{Q} = \mathbf{Z}_{T_{\text{obs}}} \mathbf{W}_Q, \quad \mathbf{K} = \mathbf{Z}_{T_{\text{obs}}} \mathbf{W}_K, \quad \mathbf{V} = \mathbf{Z}_{T_{\text{obs}}} \mathbf{W}_V, \quad (11)$$

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d_k}}\right), \quad \mathbf{O} = \mathbf{A} \cdot \mathbf{V}, \quad \mathbf{O}' = \text{Linear}(\mathbf{O}) + \mathbf{Q}. \quad (12)$$

#### 3.3.2. Decoder and masking tokens

The decoder uses a lighter transformer stack to reconstruct the full latent sequence  $\hat{\mathbf{Z}} = \{(\hat{\mathbf{z}}_t, t) \mid t \in T_{\text{in}} \cup T_{\text{out}}\}$  by attending to encoded representations and learnable masks on missing steps. This enables parallel, non-autoregressive prediction, which eliminates the error accumulation inherent in autoregressive RNNs and facilitates efficient single-pass inference.

### 3.3.3. Positional embeddings

We inject sine-cosine positional embeddings  $\delta_t$  into all latent inputs:

$$\delta(t, 2i) = \sin\left(\frac{t}{10000^{\frac{2i}{d_z}}}\right), \quad \delta(t, 2i+1) = \cos\left(\frac{t}{10000^{\frac{2i}{d_z}}}\right), \quad (13)$$

$$\mathbf{z}_t \leftarrow \mathbf{z}_t + \delta_t, \quad t \in T_{\text{in}} \cup T_{\text{out}} \quad (14)$$

where  $\delta_t$  is non-trainable and ensure temporal consistency across irregular steps.

We note that the sinusoidal positional encoding is used solely to encode relative temporal ordering among irregular time steps, rather than representing absolute physical time scales.

Unlike RNN-based models, our masked transformer supports single-step inference over the entire sequence, reducing latency and eliminating autoregressive error accumulation.

This masked reconstruction paradigm is advantageous for irregular time series. Rather than relying on explicit interpolation or resampling that may distort temporal dynamics, the model learns to infer missing or unevenly spaced observations directly from surrounding context. By training on partially observed sequences, P-STMAE naturally develops robustness to irregular sampling patterns. Operating in the latent space further enhances this capability, as the model can exploit global spatiotemporal dependencies while avoiding artifacts introduced by preprocessing, thereby addressing the computational challenges of high-dimensional systems. This explains why masked autoencoders demonstrate stronger adaptability in irregular settings compared to RNN-based approaches that require regularised inputs.

### 3.4. Loss function and training objective

We jointly minimise the prediction errors in both physical and latent spaces:

$$\mathcal{L} = \frac{1}{T_{\text{in}} + T_{\text{out}}} \sum_{t=1}^{T_{\text{in}}+T_{\text{out}}} \left( \mathbb{E} \left[ \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 \right] + \lambda \cdot \mathbb{E} \left[ \|\hat{\mathbf{z}}_t - \mathbf{z}_t\|^2 \right] \right), \quad (15)$$

where  $\lambda$  balances physical and latent consistency. Both terms in Eq. (15) employ the L2 norm (mean squared error), which is a standard choice for continuous physical fields and provides smooth, stable gradients for training masked autoencoder architectures.

### 3.5. Evaluation metrics

To comprehensively evaluate model performance, we employ three complementary metrics commonly used in spatiotemporal forecasting [20,36]: The MSE measures pixel-wise prediction accuracy over all time steps [74]. The Structural Similarity Index Measure (SSIM) quantifies structural similarity of spatial fields, capturing perceptual quality and spatial coherence [75]. The Peak Signal-to-Noise Ratio (PSNR) assesses reconstruction fidelity in decibels, providing a measure of signal-to-noise ratio [76]. These metrics collectively capture pointwise accuracy, structural preservation, and reconstruction quality, which are essential for evaluating spatiotemporal predictions in physical systems.

### 3.6. Ablation study design

We compare P-STMAE against two established RNN-based models to evaluate the effectiveness of transformer-based latent modelling versus recurrent approaches:

1. ConvRAE [20] employs a two-stage approach: a convolutional autoencoder for spatial compression followed by LSTM for temporal modelling in the latent space. This enables a direct comparison of transformer-based versus LSTM-based temporal modelling within the same latent representation framework.

2. ConvLSTM [23] integrates convolutional operations with LSTM cells to model spatiotemporal relations directly in full physical space. This provides a comparison with end-to-end full-space approaches, contrasting with the latent-space methods used by ConvRAE and P-STMAE.

To ensure fair comparison, we train these RNN-based baselines using ground truth inputs during training and autoregressive predictions during inference. Since RNN-based models require regularly sampled inputs, we adapt them to handle irregular time steps using linear interpolation, following common practice in irregular time series modelling [77,78]. This preprocessing step allows the baselines to process the data while enabling comparison with P-STMAE, which handles irregular sampling directly without interpolation.

## 4. Experiments

### 4.1. Overview

#### 4.1.1. Datasets and benchmarking

Following the methodology outlined in Section 3, we evaluate P-STMAE against the baseline models introduced in the Ablation Study Design on three representative datasets. These datasets span synthetic PDE simulations and real-world climate observations, enabling validation of both accuracy under controlled conditions and generalisation to noisy, large-scale data. Specifically, we use two datasets from PDE simulations and one from real-world observations:

- Shallow Water [79]: nonlinear fluid flow, testing robustness to chaotic dynamics.
- Diffusion Reaction from PDEBench [36]: chemical patterns, testing coupled-variable modelling.
- SST [37]: NOAA sea surface temperature data, which are noisy data with long-range dependencies.

Each dataset is split into training, validation, and test sets with ratios of (0.8, 0.1, 0.1). Channel-wise normalisation is applied to transform data values into the range of [0, 1] prior to both training and evaluation. This ensures a consistent dynamic range across heterogeneous physical variables.

We use a shifting window approach to sample input sequences from the original dataset, which contains longer sequences generated from simulations. This approach allows the model to fully leverage all available input sequences, enhancing its adaptability to various prediction scenarios. The input length is  $T_{\text{in}} = 10$ , and the forecasting length is  $T_{\text{out}} = 5$  for all datasets. We emphasise that this choice of forecasting length is made solely for fair comparison with ConvRAE and ConvLSTM baselines and does not reflect a limitation of the proposed model. In P-STMAE, future time steps are treated as masked positions in the latent space and reconstructed in a non-autoregressive manner. As a result, the forecasting horizon can be flexibly adjusted by specifying a different set of masked future time indices, without any change to the model architecture or training procedure. To model the irregular time series, missing steps are sampled with a random mask for each input sequence with a missing ratio of 0.5, except for Section 4.2.5 which uses mixed ratios, consistent between training and evaluation.

We compare P-STMAE with two representative models: ConvRAE, which also relies on latent representations with RNN temporal modelling, and ConvLSTM, which performs full-space sequence learning. This ensures a fair comparison between latent-space transformer, latent-space RNN, and full-space RNN-based approaches. Evaluation metrics include:

- MSE: pixel-wise accuracy,
- SSIM: preservation of structural information,
- PSNR: reconstruction fidelity and noise robustness.

All metrics are computed on the normalised fields. In particular, SSIM and PSNR are evaluated using their standard definitions after channel-wise normalisation, with the dynamic range parameter for PSNR set to  $\text{MAX} = 1$ . This avoids unit dependence on the original physical variables and enables consistent comparison across datasets [20,26,36].

#### 4.1.2. Implementation details

Since both P-STMAE and ConvRAE use a CAE combined with a latent model structure, our goal is to compare their performance in latent space inference using the same spatial autoencoder. To achieve this, we pre-trained an optimal CAE on the training dataset, then froze its parameters and used it with both models in subsequent time series experiments. The latent dimension is set to 128 for all datasets. To maintain consistency between predictions and ground truth during evaluation, a Sigmoid activation function is applied at the output of the CAE decoder, ensuring that all reconstructed fields lie strictly within the normalised range of  $[0, 1]$ .

Both the encoder and decoder transformer blocks possess 2 heads, with the encoder having a depth of 4 and the decoder a depth of 1. The positional embedding settings follow those of the original Transformer architecture [25]. An exception is made for SST data, where the model is expanded to include 8 heads and an encoder depth of 8. We use the RAdam optimiser with a learning rate of  $3 \times 10^{-4}$  for training P-STMAE, and use the Adam optimiser with a learning rate of  $1 \times 10^{-3}$  for training RNN baselines. The batch size is 32. The weighting coefficient of the combined loss is set to  $\lambda = 0.5$  as shown in Eq. (15).

A common concern is that Transformer-based models require extremely large datasets, an observation that primarily arises from applications in natural language processing and natural-image modelling, where data distributions are highly complex and high-entropy [28,80,81].

In contrast, the scientific computing problems studied here are governed by smooth, structured PDE dynamics that lie on low-dimensional manifolds. Together with latent-space compression via the convolutional autoencoder, this reduces temporal modelling complexity. As a result, dataset sizes of  $10^5$ – $10^6$  frames are sufficient in our setting, and the Transformer-based temporal model does not exhibit data inefficiency compared to ConvLSTM.

The architecture is shown in Appendix Table A.1. All convolutions use kernel size  $3 \times 3$  with same padding.

Table 1 presents a comparison of test performance metrics across all three datasets. The results demonstrate that P-STMAE consistently achieves competitive or better performance compared to baseline models. On the Shallow Water dataset, P-STMAE outperforms both ConvRAE and ConvLSTM across all metrics, achieving the lowest MSE ( $6.16 \times 10^{-5}$ ), highest SSIM (0.9538), and highest PSNR (43.90). For the Diffusion-Reaction dataset, P-STMAE achieves the lowest MSE ( $5.99 \times 10^{-5}$ ) while ConvLSTM performs slightly better in SSIM and PSNR. On the real-world SST dataset, P-STMAE delivers the strongest overall performance, outperforming both baselines with the lowest MSE ( $8.02 \times 10^{-5}$ ), highest SSIM (0.9817), and highest PSNR (41.03). These results indicate the robustness and generalisation capability of P-STMAE across diverse spatiotemporal systems, from synthetic PDE simulations to real-world climate observations.

## 4.2. Shallow water test case

### 4.2.1. Dataset description

Shallow Water Equations (SWEs) are a set of hyperbolic PDEs that model the flow beneath a pressure surface in a fluid. This dataset tests the models' ability to capture chaotic nonlinear fluid dynamics under varying physical parameters, which are relevant to geophysical flows such as atmospheric and oceanic dynamics. The system is formulated as

**Table 1**

Test metrics in the full physics space across all three datasets with a missing ratio of 0.5. Metrics for Shallow Water are averaged over all  $h$ ,  $u$ , and  $v$  variables; metrics for Diffusion-Reaction are averaged over both  $u$  and  $v$  variables. Bold values indicate the best performance among the three forecasting models (excluding CAE).

Dataset	Model	MSE	SSIM	PSNR
Shallow Water	CAE	$5.32 \times 10^{-5}$	0.9596	44.64
	P-STMAE	<b><math>6.16 \times 10^{-5}</math></b>	<b>0.9538</b>	<b>43.90</b>
	ConvRAE	$9.86 \times 10^{-5}$	0.9394	42.47
	ConvLSTM	$1.82 \times 10^{-4}$	0.9231	40.72
Diffusion-Reaction	CAE	$3.66 \times 10^{-5}$	0.9887	48.78
	P-STMAE	<b><math>5.99 \times 10^{-5}</math></b>	0.9870	44.20
	ConvRAE	$8.48 \times 10^{-5}$	0.9875	44.41
	ConvLSTM	$6.80 \times 10^{-5}$	<b>0.9928</b>	<b>47.36</b>
SST	CAE	$7.11 \times 10^{-5}$	0.9819	41.53
	P-STMAE	<b><math>8.02 \times 10^{-5}</math></b>	<b>0.9817</b>	<b>41.03</b>
	ConvRAE	$1.03 \times 10^{-4}$	0.9803	40.33
	ConvLSTM	$4.57 \times 10^{-4}$	0.9384	36.84

follows:

$$\frac{\partial h}{\partial t} + \frac{\partial(hu)}{\partial x} + \frac{\partial(hv)}{\partial y} = 0, \quad (16)$$

$$\frac{\partial u}{\partial t} + g \frac{\partial h}{\partial x} + bu = 0, \quad (17)$$

$$\frac{\partial v}{\partial t} + g \frac{\partial h}{\partial y} + bv = 0, \quad (18)$$

where  $h$  is the surface height of water,  $u$  and  $v$  are the orthogonal velocity components averaged in depth,  $g$  is the gravitational acceleration, and  $b$  is the friction coefficient of the fluid. The initial condition is a cylinder bump in the water with a small height  $h$  above the surface average, with a variable radius  $r$ , and zero velocities  $u$  and  $v$ . The boundary conditions are periodic in both directions.

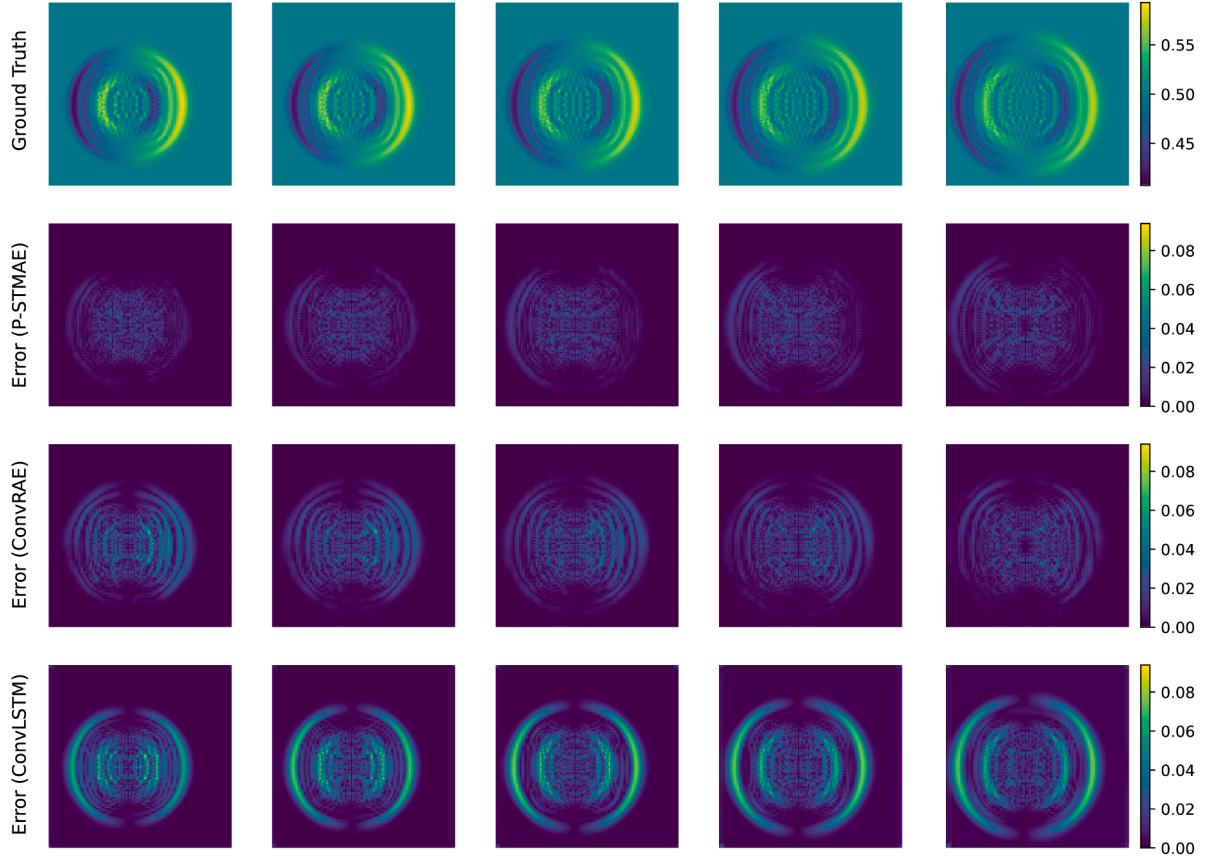
Our numerical simulation is performed on a grid of size  $128 \times 128$  with three variables  $[h, u, v]$ . We sample snapshots from the simulation at fixed intervals when generating each data sequence. The fixed time step is set to  $\Delta t = 10^{-4}$ , space step  $\Delta x = 10^{-2}$ , and gravity  $g = 1.0$ . Other parameters, including the fluid friction  $b$  and the centre, radius, and height of the cylinder bump, and the snapshot gap, are randomised to produce different initial conditions of fluid dynamics. The detailed ranges of simulation parameters are shown in Appendix Table B.1. We generate 600 sequences, with each sequence having 200 spatio-temporal frames.

### 4.2.2. CAE reconstruction

First, we train a convolutional autoencoder on the shallow water dataset. The CAE can capture the spatial features of input physical fields and reconstruct them with high fidelity. It achieves an MSE of  $5.32 \times 10^{-5}$ , SSIM of 0.9596, and PSNR of 44.64 on the test set (see Appendix Fig. B.1). These strong reconstruction metrics indicate that the CAE provides reliable latent representations for subsequent sequence modelling.

### 4.2.3. Model comparison

The validation results for the shallow water dataset demonstrate the efficacy of the P-STMAE model compared to the baseline models. In terms of full-space MSE, the P-STMAE outperforms both ConvRAE and ConvLSTM. The full-space MSE curve for P-STMAE lies much closer to that of the CAE's performance, indicated by the dotted line (see Appendix Fig. B.2). This proximity suggests that the time series masked autoencoder is effective at exploiting the semantic information hidden in latent representations, making P-STMAE's performance nearly match the reconstruction capability of the CAE. Analysing the latent MSE curves, we observe a consistent pattern whereby the P-STMAE surpasses the ConvRAE in terms of reducing prediction error within the latent space. The performance indicates an ability to handle complex nonlinear spatiotemporal patterns, which the baseline models struggle



**Fig. 3.** Ground truth (top) and error maps of P-STMAE, ConvRAE, and ConvLSTM for forecasting the variable  $u$  in the shallow water dataset with a sampling dilation of 3. Columns represent successive forecasting steps. Among the models, P-STMAE yields the smallest errors, indicating predictive accuracy.

with. This confirms that transformer-based latent inference is more effective than RNN-based alternatives in the presence of chaotic fluid dynamics.

Despite its performance, the P-STMAE exhibits a relatively slower convergence speed compared to the RNN-based models. This behaviour can be attributed to the complex self-attention mechanisms within the transformer architecture, which require a slower learning process to adequately tune weights to capture overall time series dependencies.

The test performance metrics for the shallow water dataset (see Table 1) further substantiate the superior performance of the P-STMAE model over the baseline models. Apart from the lowest full-space MSE, the P-STMAE demonstrates the highest SSIM, reflecting its ability to maintain structural integrity and perceptual quality of the reconstructed physical fields. In terms of PSNR, the P-STMAE also scores the highest, confirming its effectiveness in minimising noise and enhancing the clarity of predictions compared to the other models. Fig. 3 provides a detailed visual comparison of prediction errors across all three models. The error maps reveal that P-STMAE consistently produces the smallest prediction errors across successive forecasting steps, with error magnitudes lower than both ConvRAE and ConvLSTM. The spatial distribution of errors for P-STMAE is more uniform and concentrated in regions with higher physical complexity, while the baseline models exhibit larger and more widespread error patterns in areas with strong nonlinear dynamics (see Appendix Fig. B.3).

#### 4.2.4. Ablation study on loss weighting coefficient

We conduct an ablation study on the shallow water dataset to investigate the sensitivity of P-STMAE to the weighting coefficient  $\lambda$  in the combined loss function (Eq. 15). The ablation study results demonstrate that model performance is not sensitive to  $\lambda$  within the range [0.2, 1.0] (see Appendix Table B.2 and Fig. B.4). Based on these findings,

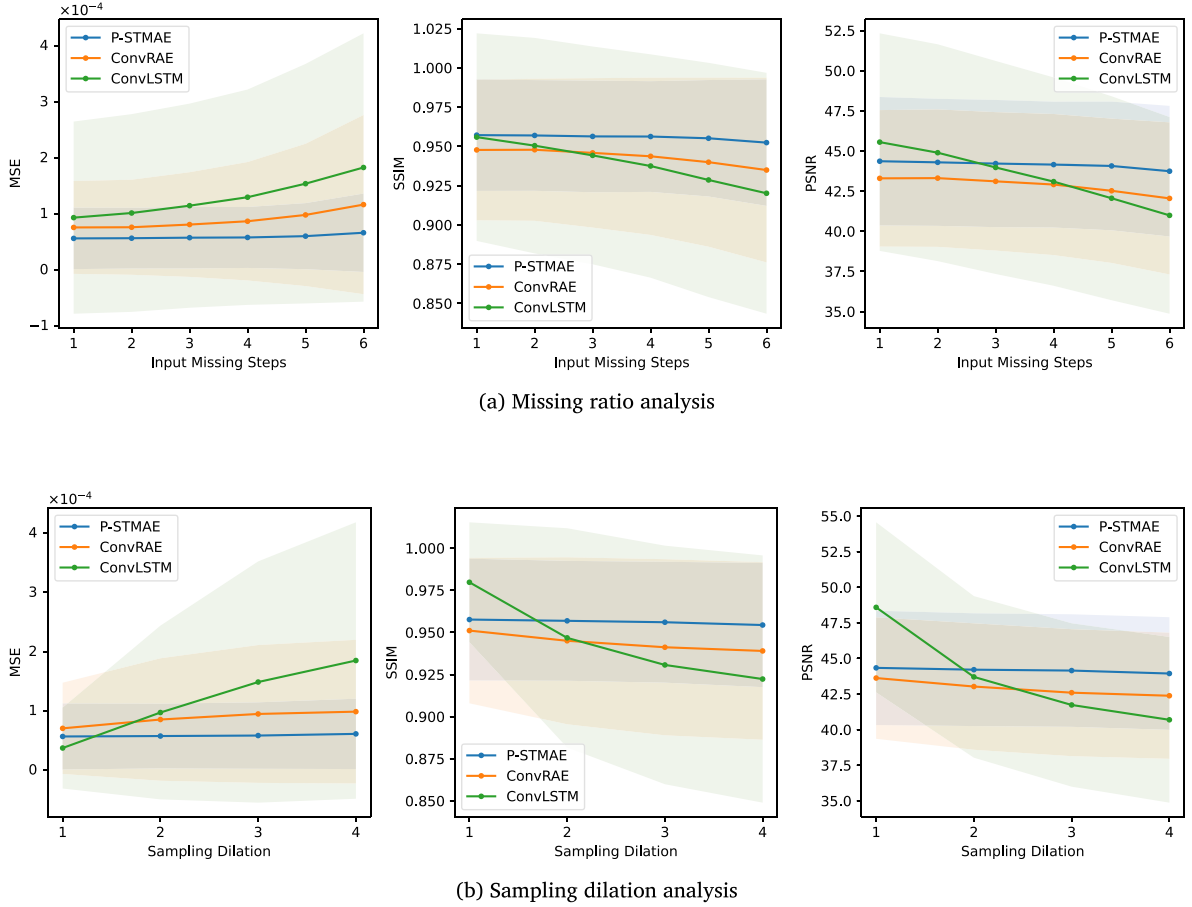
we fix  $\lambda = 0.5$  for all experiments for simplicity, as this value provides a balanced trade-off between physical and latent space consistency while maintaining robust performance across different  $\lambda$  values.

#### 4.2.5. Missing ratio analysis

We analyse model performance under varying amounts of missing data to assess generalisation capability as the number of missing steps increases. To evaluate their performance, we train all models with missing steps ranging from 1 to 6 with a fixed input sequence length of 10 and evaluate them under corresponding settings. During training, we maintain a fixed missing ratio within each batch and randomise this ratio between different batches to leverage parallelism and accelerate training.

Fig. 4 presents the results of these robustness experiments. Panel (a) shows performance comparison under varying numbers of missing steps, while panel (b) shows test performance comparison regarding sampling dilations. For missing ratio analysis (Fig. 4(a)), P-STMAE consistently outperforms the baseline models in MSE across all missing step conditions. The error curves demonstrate that P-STMAE maintains consistently low prediction errors even as the number of missing steps increases from 1 to 6, with the MSE curve remaining relatively flat. In contrast, both ConvRAE and ConvLSTM show progressively increasing errors as missing steps increase, with ConvLSTM exhibiting sharp degradation. This indicates that the transformer-based architecture of P-STMAE, which leverages attention mechanisms and contextual encoding, maintains lower prediction errors when reconstructing sequences with varying levels of missing data.

ConvLSTM performs well at lower missing steps (1 and 2) and achieves higher PSNR values than P-STMAE in these conditions. As the number of missing steps increases, its performance deteriorates sharply, revealing sensitivity to temporal disruptions and difficulty in handling



**Fig. 4.** Robustness analysis of P-STMAE on the shallow water dataset. **(a)** Performance comparison under varying numbers of missing steps in the input sequence with a length of 10. Each model is trained and evaluated with missing steps ranging from 1 to 6. P-STMAE demonstrates consistent performance and robustness, while the RNN-based models, especially ConvLSTM, show higher sensitivity to increasing missing steps. **(b)** Test performance comparison regarding the sampling dilations of data sequences. All models are separately trained on the shallow water dataset of different dilations.

irregular sequences due to its end-to-end prediction in full-space. ConvRAE exhibits a similar trend, with decent performance at lower missing steps but a noticeable decline as missing data increases.

The experiment demonstrates that P-STMAE maintains high performance across different missing ratios. In contrast, the RNN-based models are more effective with limited missing data but struggle as gaps increase, highlighting the advantage of transformer-based attention mechanisms for handling irregular sampling.

#### 4.2.6. Nonlinear robustness analysis

We explore nonlinear robustness by introducing dilations into the sampling window with a fixed missing ratio of 0.5. Dilation increases the time gaps between consecutive data points by a fixed factor, testing whether models can generalise when temporal dynamics become more irregular and chaotic.

Mathematically, given an original time series sequence:

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_T\}, \quad (19)$$

where  $\mathbf{x}_t$  represents the value at time step  $t$ , and a dilated sequence with dilation factor  $d$  is defined as:

$$\mathbf{X}_d = \{\mathbf{x}_1, \mathbf{x}_{1+d}, \mathbf{x}_{1+2d}, \dots, \mathbf{x}_{1+kd}\}, \quad (20)$$

where  $k$  is the largest integer satisfying  $1 + kd \leq T$ . This means that every  $d$ -th element is sampled, increasing the time gap between consecutive points.

This operation simulates irregular, nonlinear time steps by introducing structured sparsity into the sequence. The dilation parameter  $d$  controls the degree of this gap expansion, allowing us to introduce more variability in the time structure. As  $d$  increases, the sequence becomes less regular and more nonlinear, challenging the models to generalise and capture complex, evolving temporal dynamics.

Panel (b) of Fig. 4 shows the test performance results over different dilations. The performance curves reveal distinct patterns: P-STMAE exhibits stable performance across different dilations, with MSE values remaining consistently low even as dilation increases, demonstrating robustness when faced with increasing nonlinearities. In contrast, ConvLSTM shows sensitivity to dilation changes, performing well at minimal dilation but deteriorating rapidly as dilation increases, with error values rising. ConvRAE shows moderate sensitivity, with performance declining gradually but remaining better than ConvLSTM at higher dilations. This suggests that ConvLSTM struggles to generalise in complex scenarios where temporal dependencies become harder to capture due to high nonlinearities, likely because it operates directly in full physics space where temporal relations are chaotic on sparse features.

Both P-STMAE and ConvRAE show resilience to dilation, suggesting that latent-space architectures are better suited for handling chaotic data. By processing spatial information into dense features via CAEs, they can manage complex, nonlinear patterns without being directly impacted by irregularities in the full physics space. The performance gap between P-STMAE and ConvRAE further reveals the transformer-based

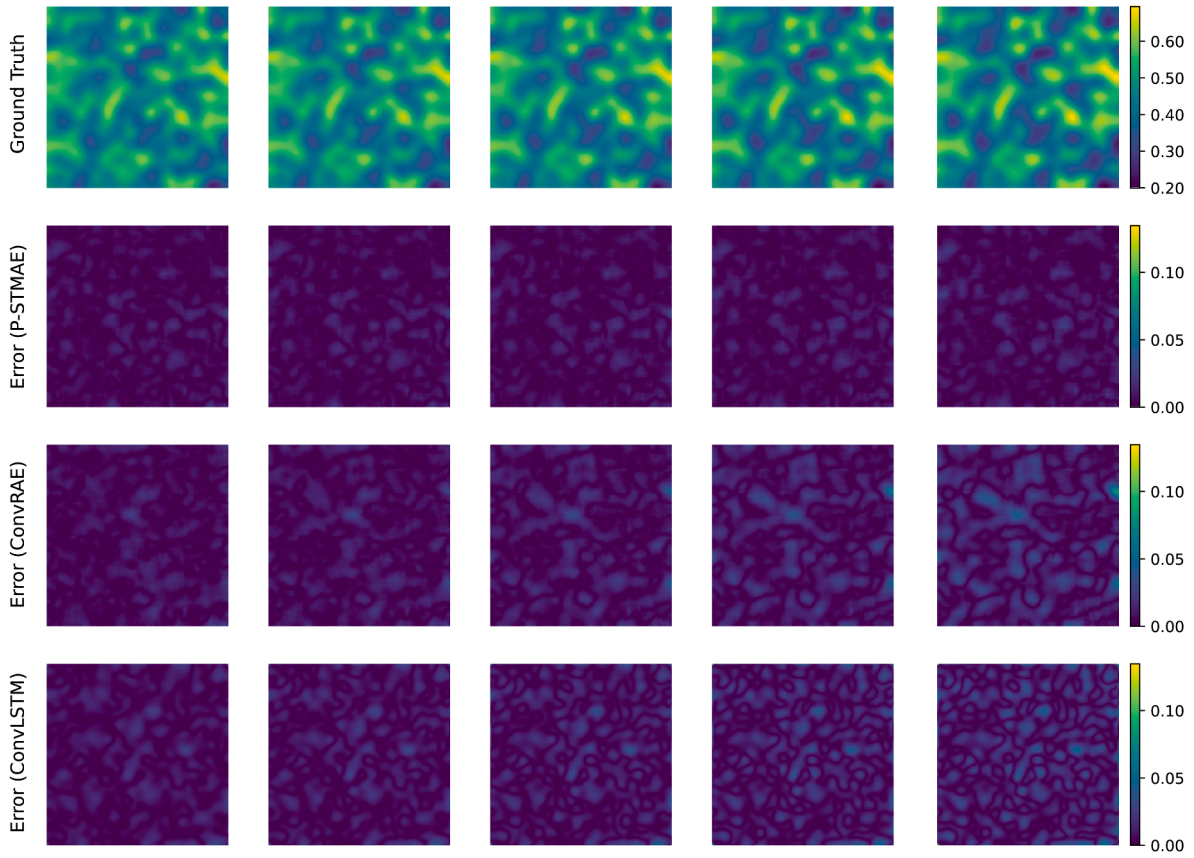


Fig. 5. Ground truth (top) and error maps of P-STMAE, ConvRAE, and ConvLSTM for forecasting the variable  $u$  in the diffusion-reaction dataset with a sampling dilation of 5. Columns represent successive forecasting steps. The results show that P-STMAE consistently achieves lower errors than the baselines, confirming its advantage.

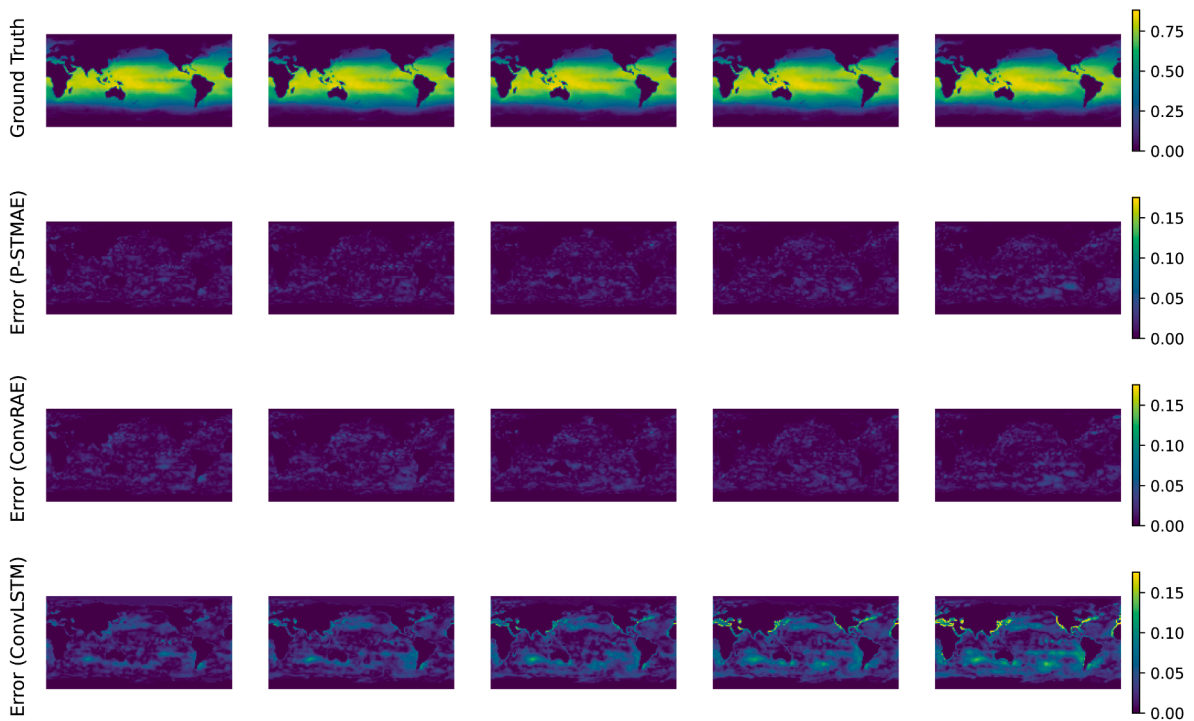


Fig. 6. Error maps of P-STMAE, ConvRAE, and ConvLSTM predictions for the forecasting steps in the SST dataset.

capability in capturing irregular temporal dependencies compared to RNNs.

#### 4.3. Diffusion reaction test case

##### 4.3.1. Dataset description

The 2D diffusion-reaction equations are commonly employed to model phenomena where diffusion and reaction processes interact in a spatial domain, such as biological pattern formation. Compared to the shallow water case, this case emphasises coupled nonlinear interactions and pattern formation. The system consists of two nonlinearly coupled variables, the activator  $u$  and the inhibitor  $v$ . The equations governing their evolution are given by [36]:

$$\frac{\partial u}{\partial t} = \alpha_u \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) + F_u(u, v), \quad (21)$$

$$\frac{\partial v}{\partial t} = \alpha_v \left( \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} \right) + F_v(u, v). \quad (22)$$

Here,  $\alpha_u$  and  $\alpha_v$  denote the diffusion coefficients for the activator and inhibitor. The reaction functions  $F_u$  and  $F_v$  follow the FitzHugh-Nagumo model:

$$F_u(u, v) = u - u^3 - c - v, \quad (23)$$

$$F_v(u, v) = u - v, \quad (24)$$

where  $c$  is a constant parameter that affects the reaction kinetics. The domain for the simulation extends over  $x, y \in [-1, 1]$  with time  $t \in (0, 5]$ . The no-flow Neumann boundary conditions ensure that the flux of both  $u$  and  $v$  across the boundaries remains zero.

The training dataset, available from the PDEBench [36] project, is discretised into  $128 \times 128$  spatial grid points and 100 temporal steps, with 10,000 sample sequences.

The 2D diffusion-reaction dataset poses a challenge due to the nonlinear coupling between the activator and inhibitor, and its applicability to real-world problems such as biological pattern formation.

##### 4.3.2. Model comparison

For the diffusion-reaction dataset (see Table 1), P-STMAE achieves the lowest MSE among the baselines, indicating numerical accuracy in minimising pointwise error. However, P-STMAE slightly underperforms in SSIM and PSNR compared to ConvLSTM, suggesting a trade-off between pixel-wise accuracy and higher-order spatial consistency (see Appendix Fig. C.1). This may arise because ConvLSTM operates directly in the full physics space, potentially better preserving structural patterns and perceptual quality for complex coupled-variable systems, while P-STMAE's latent-space compression may introduce subtle spatial distortions despite pointwise accuracy.

Fig. 5 provides a visual comparison of prediction errors across all three models. The error maps demonstrate that P-STMAE consistently achieves lower prediction errors than both baseline models across successive forecasting steps. The spatial error patterns reveal that P-STMAE maintains higher accuracy in regions with complex pattern formations, where the activator and inhibitor variables exhibit strong coupling. In contrast, ConvRAE and ConvLSTM show larger error magnitudes in areas where the reaction dynamics create intricate spatial structures (see Appendix Fig. C.2).

#### 4.4. NOAA sea surface temperature test case

##### 4.4.1. Dataset description

The SST [37] dataset provides a long-term climate record of weekly sea surface temperature observations spanning the period from 1981 to 2018. These data are collected from multiple sources, including satellites, ships, buoys, and Argo floats, and are then interpolated to produce a continuous global grid of temperature data. The spatial resolution of the dataset is  $360 \times 180$ , corresponding to a global grid where each unit covers a 1-degree area of latitude and longitude.

The dataset consists of 1914 snapshots, each representing the global distribution of sea surface temperatures at weekly intervals. These observations are stored as a single temperature variable. The temporal and spatial continuity of the dataset makes it valuable for studying long-term climate trends, oceanic processes, and their influence on weather and marine ecosystems.

The SST dataset is widely used for analysing climate variability, detecting anomalies such as El Niño and La Niña, and studying oceanic heat content changes. This dataset presents a complex spatiotemporal problem, as ocean temperatures are influenced by long-term climatic patterns, oceanic currents, and seasonal variations.

##### 4.4.2. Model comparison

The performance on the SST dataset (Table 1, missing ratio 0.5) demonstrates P-STMAE's strong generalisation to real-world climate data. P-STMAE delivers the strongest overall performance, achieving the lowest MSE ( $8.02 \times 10^{-5}$ ), highest SSIM (0.9817), and highest PSNR (41.03) compared with the baselines. The transformer architecture enables P-STMAE to better handle irregularities and capture long-term dependencies in the SST dataset, leading to more accurate temporal predictions. The performance of ConvRAE closely follows P-STMAE, while ConvLSTM falls behind, suggesting that latent-space compression is crucial for modelling real-world systems, and reduces the impact of noise and chaotic patterns that ConvLSTM struggles to capture at global scale.

Fig. 6 presents error maps comparing all three models across forecasting steps. The global error patterns reveal that P-STMAE achieves the smallest prediction errors across most oceanic regions, with strong performance in areas with complex temperature gradients such as ocean fronts and upwelling zones. The error distribution shows that P-STMAE maintains consistent accuracy across different latitudinal bands and oceanic basins, demonstrating its ability to capture both large-scale climate patterns and regional temperature variations. In contrast, ConvLSTM exhibits larger and more spatially widespread errors in regions with strong seasonal variations and complex current systems, highlighting the challenges of full-space modelling for global-scale climate data (see Appendix Fig. D.1).

## 5. Conclusion and future work

In this paper, we introduced the P-STMAE, a novel model designed to address irregular time series prediction in high-dimensional dynamical systems. By integrating CAEs with transformer-based masked autoencoders, P-STMAE employs placeholder-based attention to handle missing data and irregular time steps directly without preprocessing Table B.2. Experiments across synthetic PDE benchmarks and real-world SST data demonstrated its robustness, computational efficiency, and better accuracy compared to traditional RNN-based approaches Fig. D.1.

A key advantage of P-STMAE is its latent-space masked training, which enables efficient processing of high-dimensional data. By operating in a compressed representation, the model reduces computational cost while maintaining spatiotemporal pattern learning. This efficiency is beneficial for large-scale systems where computational resources are constrained.

Despite its promising performance, several limitations highlight areas for future research. The quadratic complexity of the transformer's global self-attention poses challenges for processing very long sequences. Exploring alternatives like local or sparse attention mechanisms could enhance scalability. To efficiently handle relative time embedding in irregular time series, future work could consider advanced positional embedding techniques, such as Attention with Linear Biases (ALiBi) [82] and Rotary Position Embedding (RoPE) [83], which can better capture relative temporal relationships without explicit positional encodings. Additionally, the reliance on the convolutional autoencoder may introduce a bottleneck, potentially limiting reconstruction fidelity. Future work could investigate advanced physical field encoding techniques, such as VAEs or Vision Transformers, to overcome this

limitation. Furthermore, the observed trade-off between minimizing point-wise prediction errors and preserving structural fidelity in spatiotemporal data remains an open challenge. Future research could focus on multi-objective optimisation strategies that balance numerical accuracy with the preservation of global structures. Finally, while the encouraging results on SST provide an initial demonstration of real-world applicability, broader validation on diverse real-world datasets will be essential to fully establish the model's generalizability. In summary, this work advances irregular time series forecasting for high-dimensional dynamical systems. P-STMAE offers a purely data-driven, adaptable, and computationally efficient solution, positioning it as a promising tool for scientific and industrial applications requiring accurate prediction of complex spatiotemporal systems.

### CRedit authorship contribution statement

**Kewei Zhu:** Writing – original draft, Software, Investigation, Formal analysis, Data curation; **Yanze Xin:** Writing – original draft, Methodology, Investigation, Formal analysis, Data curation; **Jinwei Hu:** Writing – review & editing, Validation, Methodology; **Xiaoyuan Cheng:** Writing – review & editing, Validation, Investigation; **Yiming Yang:** Writing – review & editing, Validation, Investigation; **Sibo Cheng:** Writing – review & editing, Validation, Investigation, Funding acquisition, Conceptualization.

### Data availability

The code of this study is available at <https://github.com/RyanXinOne/PSTMAE>.

### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Detailed Visualization of Model Predictions

### A.1. Model architecture

**Table A.1**

Network structures of CAE encoder (left) and decoder (right) used in the P-STMAE. All of the convolutions and transpose convolutions use the kernel size of  $3 \times 3$  and the same padding. The input dimension and channel number can vary depending on the dataset.

Layer Type	Output Shape	Activation
Input	(128, 128, 3)	
Conv 2D	(128, 128, 8)	GELU
Conv 2D	(64, 64, 16)	GELU
Conv 2D	(32, 32, 32)	GELU
Conv 2D	(16, 16, 64)	GELU
Conv 2D	(8, 8, 128)	GELU
Linear	(128)	
Input	(128)	
Linear	(8, 8, 128)	GELU
TransConv 2D	(16, 16, 64)	GELU
TransConv 2D	(32, 32, 32)	GELU
TransConv 2D	(64, 64, 16)	GELU
TransConv 2D	(128, 128, 8)	GELU
Conv 2D	(128, 128, 3)	Sigmoid

## Appendix B. Shallow Water Test Case

**Table B.1**

Ranges of computational parameters used in the shallow water equation simulations. All parameters are uniformly sampled.

Parameter	Symbol	Min Value	Max Value
Initial bump centre (x)	$p_x$	54.00	74.00
Initial bump centre (y)	$p_y$	54.00	74.00
Bump height	$h$	0.05	0.20
Bump radius	$r$	8.94	12.65
Friction coefficient	$b$	0.02	2.00
Snapshot interval (steps)	–	60.00	100.00

**Table B.2**

Quantitative ablation study on  $\lambda$ .

$\lambda$	MSE	SSIM	PSNR
0.01	$5.85 \times 10^{-5}$	0.9276	41.02
0.02	$5.25 \times 10^{-5}$	0.9317	41.52
0.05	$4.72 \times 10^{-5}$	0.9264	40.82
0.10	$5.20 \times 10^{-5}$	0.9315	41.60
0.20	$6.10 \times 10^{-5}$	0.9177	39.56
0.50	$8.00 \times 10^{-5}$	0.9356	42.07
0.60	$2.17 \times 10^{-4}$	0.9037	35.46
0.70	$2.20 \times 10^{-4}$	0.9017	35.38
1.00	$2.18 \times 10^{-4}$	0.9037	35.43

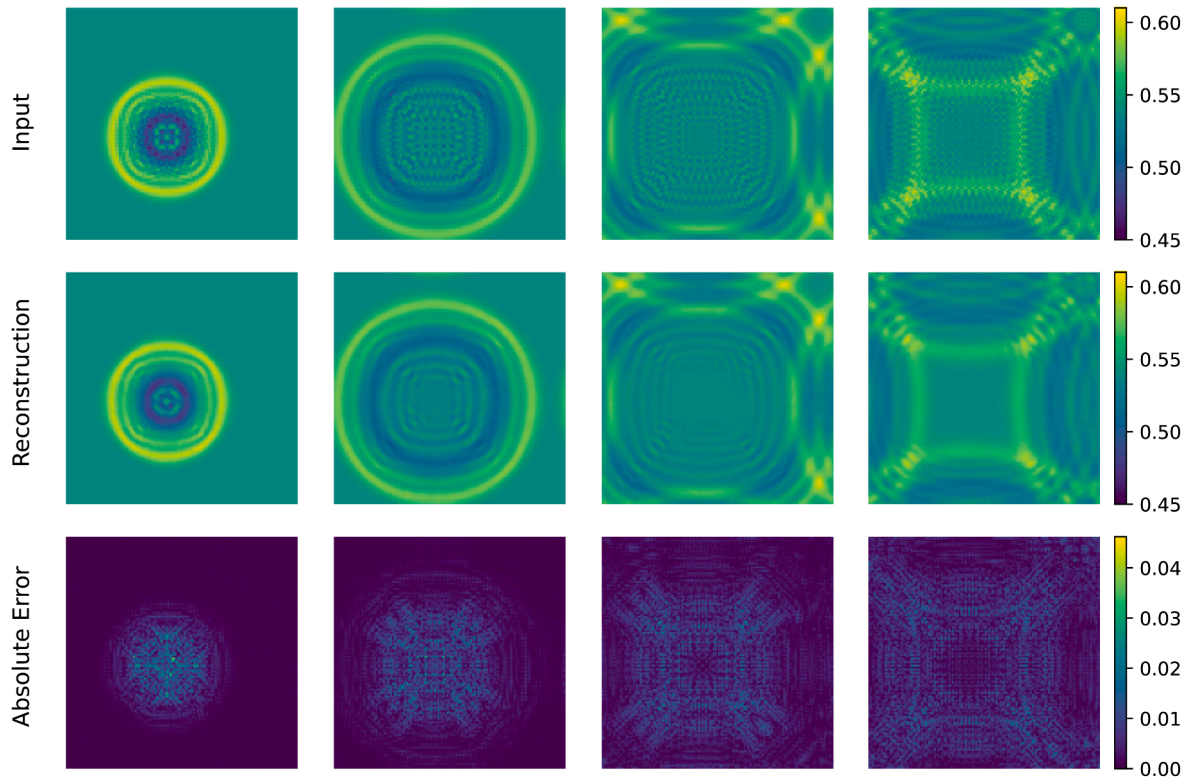


Fig. B.1. Visualisation of the trained CAE model performance on the variable  $h$  of shallow water samples.

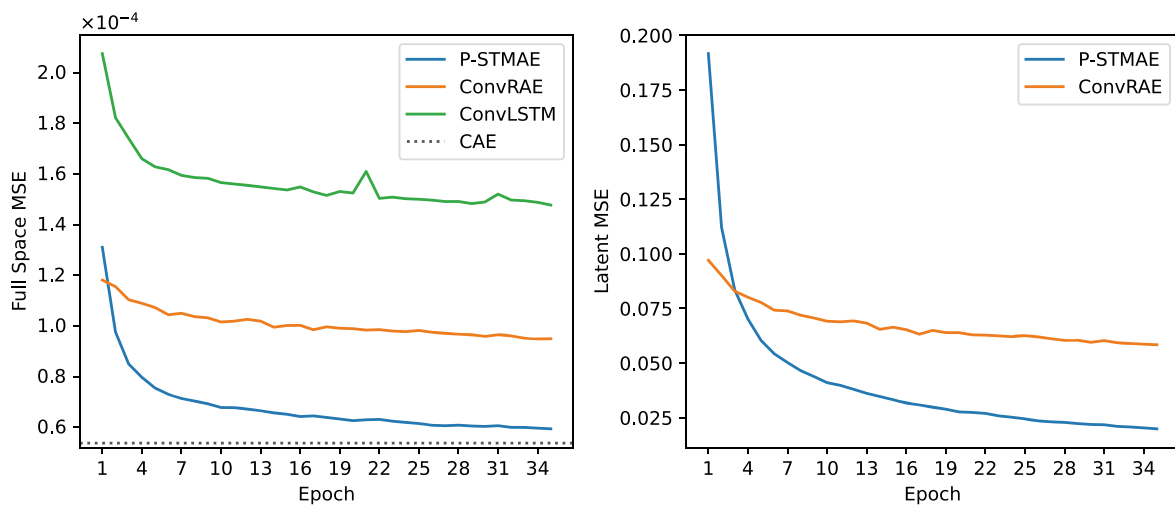
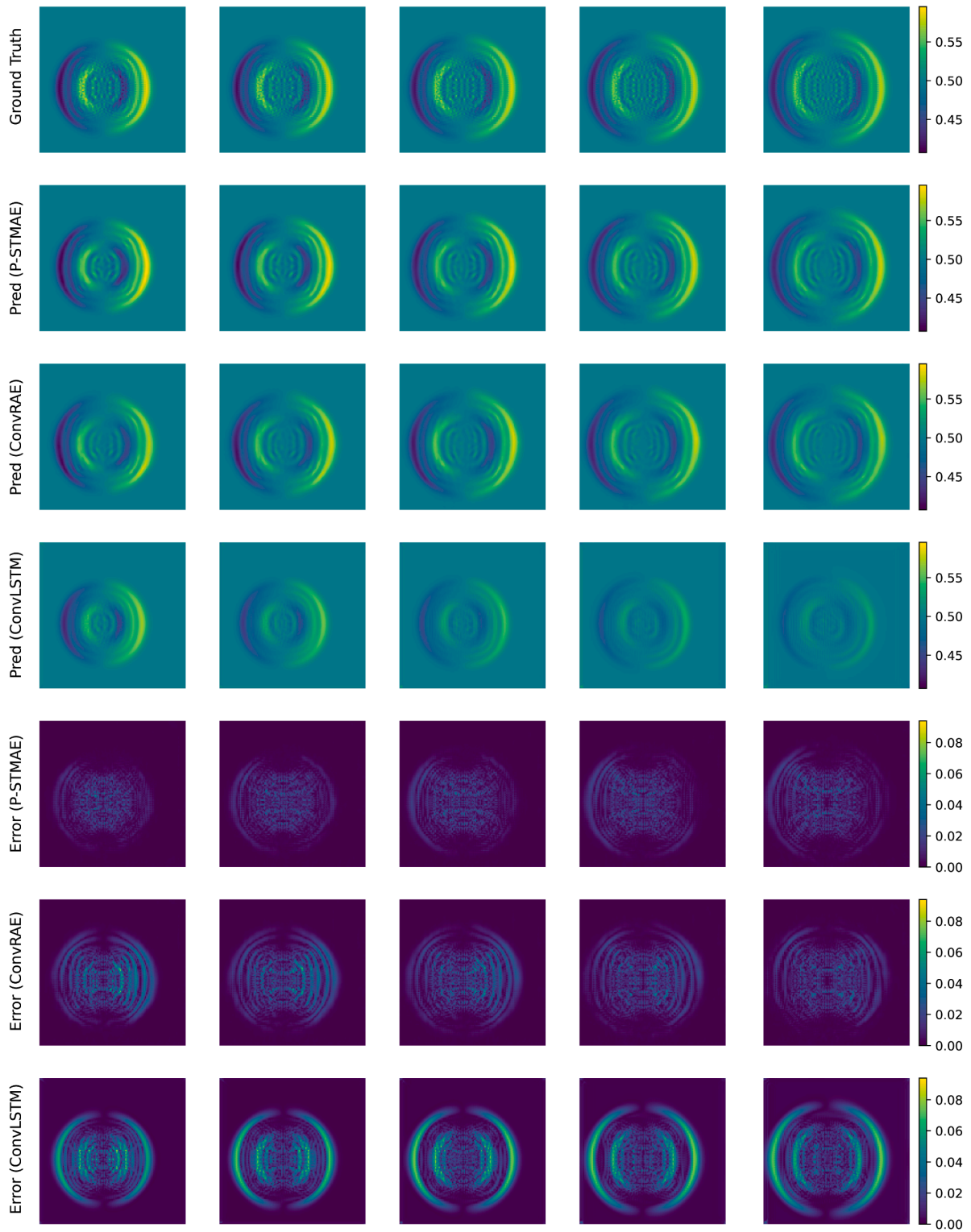


Fig. B.2. Validation MSEs on the shallow water dataset with the sampling dilation of 3. **Left:** Full space MSE curves of the P-STMAE, ConvRAE, and ConvLSTM. The dotted line shows the performance of the trained CAE. **Right:** Latent space MSE curves of the P-STMAE and ConvRAE. Note that the ConvLSTM does not use latent representations.



**Fig. B.3.** Model predictions and error maps for the forecasting steps of the variable  $u$  in the shallow water dataset with a sampling dilation of 3. The first row shows the ground truth input, rows 2–4 show the model-predicted outputs of the three models, and rows 5–7 show the pixel-level absolute differences between the two.

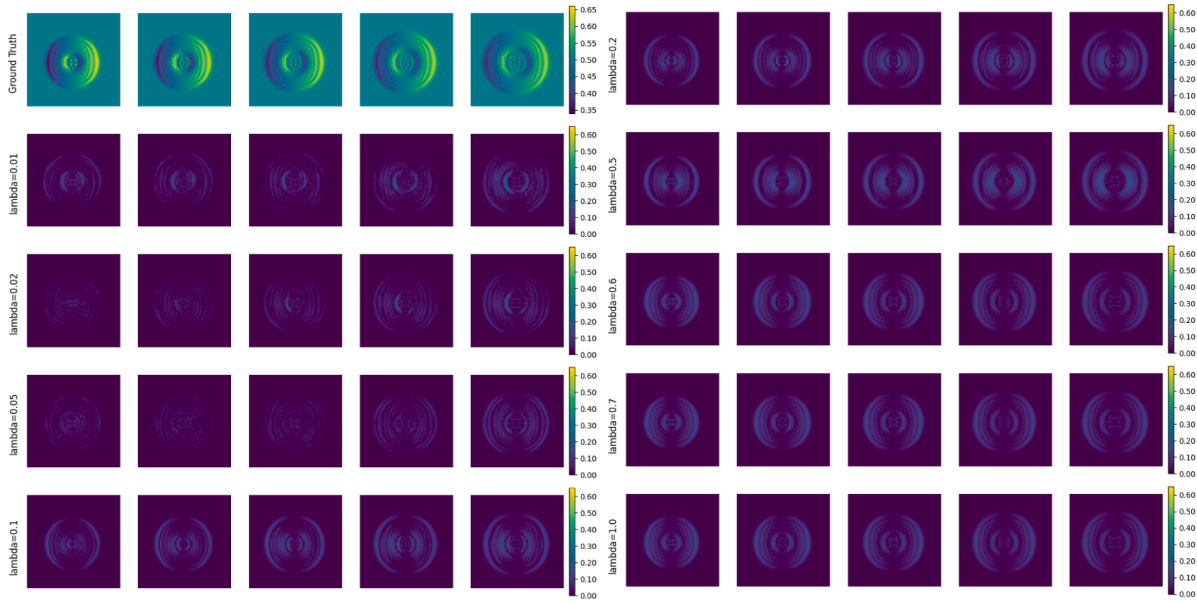


Fig. B.4. Model error maps for different values of  $\lambda$  across forecasting steps in the shallow water dataset.

Appendix C. Diffusion Reaction Test Case

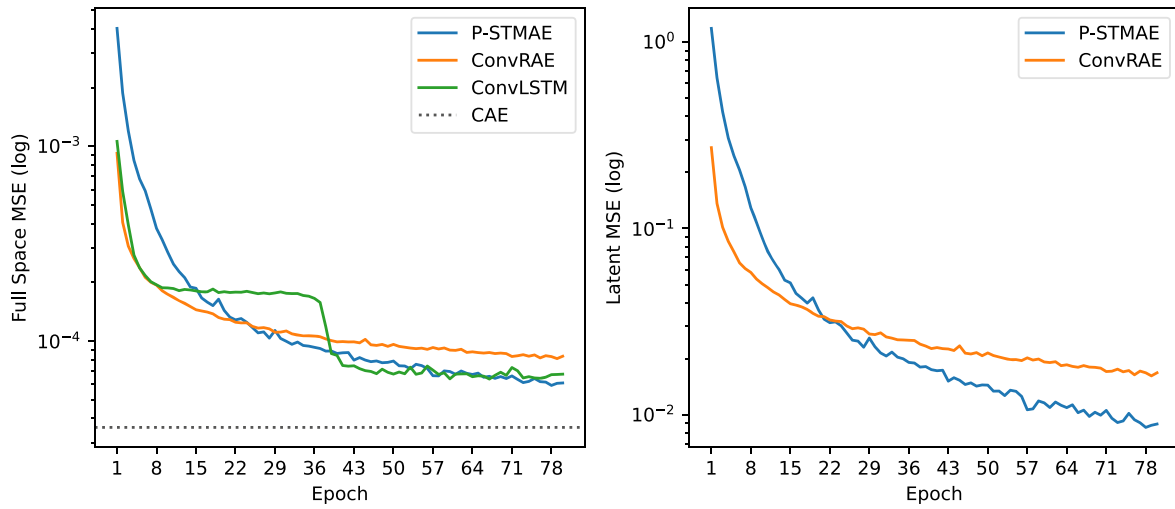


Fig. C.1. Validation MSEs on the diffusion reaction dataset with the sampling dilation of 5.

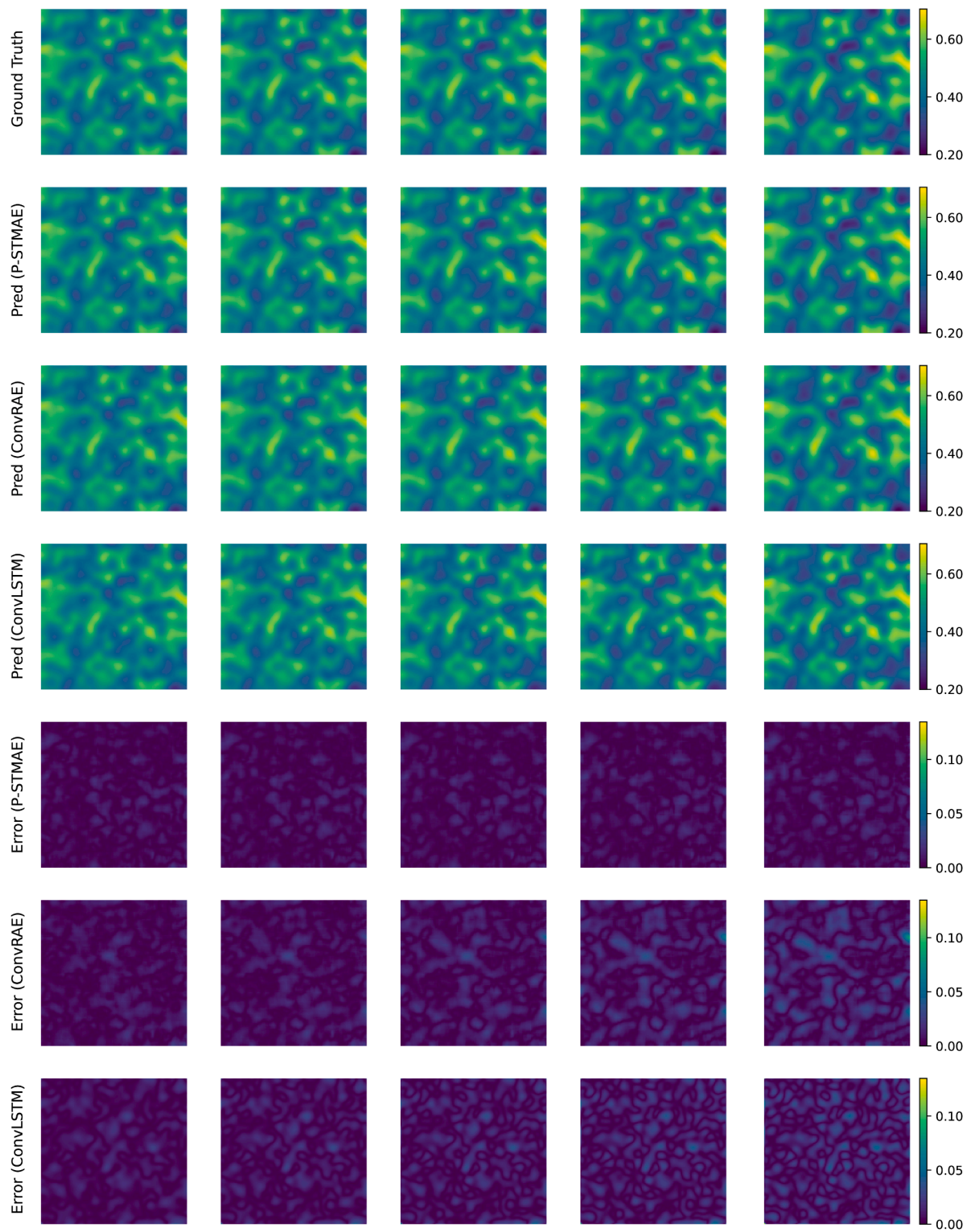


Fig. C.2. Model predictions and error maps for the forecasting steps of the variable  $u$  in the diffusion reaction dataset with a sampling dilation of 5.

## Appendix D. NOAA Sea Surface Temperature Test Case

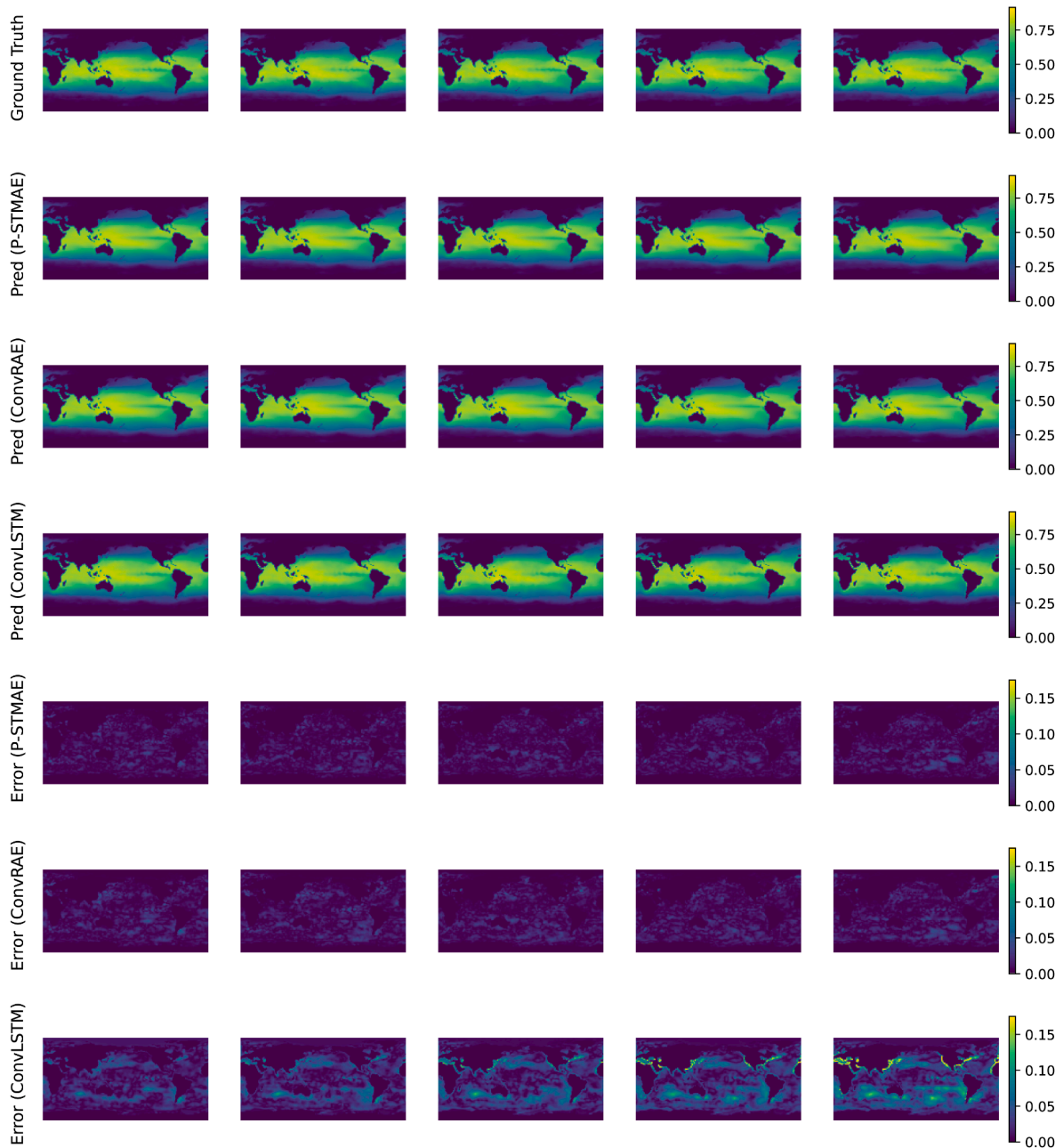


Fig. D.1. Model predictions and error maps for the forecasting steps in the SST dataset.

## References

- [1] K. Fukami, R. Maulik, N. Ramachandra, K. Fukagata, K. Taira, Global field reconstruction from sparse sensors with voronoi tessellation-assisted deep learning, *Nat. Mach. Intell.* 3 (11) (2021) 945–951.
- [2] L.A. Caffarelli, P.E. Souganidis, Rates of convergence for the homogenization of fully nonlinear uniformly elliptic pde in random media, *Invent. Math.* 180 (2) (2010) 301–360.
- [3] M. Brin, G. Stuck, *Introduction to Dynamical Systems*, Cambridge university press, 2002.
- [4] S.L. Brunton, B.W. Brunton, J.L. Proctor, E. Kaiser, J.N. Kutz, Chaos as an intermittently forced linear system, *Nat. Commun.* 8 (1) (2017) 19.
- [5] S. Siami-Namini, N. Tavakoli, A.S. Namin, The performance of LSTM and BiLSTM in forecasting time series, in: 2019IEEE International Conference on Big Data (Big Data), IEEE, 2019, pp. 3285–3292.
- [6] P.B. Weerakody, K.W. Wong, G. Wang, W. Ela, A review of irregular time series data handling with gated recurrent neural networks, *Neurocomputing* 441 (2021) 161–178.
- [7] B. Johnson, S.B. Munch, An empirical dynamic modeling framework for missing or irregular samples, *Ecol. Modell.* 468 (2022) 109948.
- [8] S. Cheng, C. Liu, Y. Guo, R. Arcucci, Efficient deep data assimilation with sparse observations and time-varying sensors, *J. Comput. Phys.* 496 (2024) 112581.
- [9] E. Afrifa-Yamoah, U.A. Mueller, S.M. Taylor, A.J. Fisher, Missing data imputation of high-resolution temporal climate time series data, *Meteorol. Appl.* 27 (1) (2020) e1873.
- [10] H. Ahn, K. Sun, K.P. Kim, et al., Comparison of missing data imputation methods in time series forecasting, *Comput. Mater. Continua* 70 (1) (2022) 767–779.
- [11] B.K. Nelson, Time series analysis using autoregressive integrated moving average (ARIMA) models, *Acad. Emerg. Med.* 5 (7) (1998) 739–744.

- [12] V. Gómez, A. Maravall, Estimation, prediction, and interpolation for nonstationary series with the Kalman filter, *J. Am. Stat. Assoc.* 89 (426) (1994) 611–624.
- [13] V.I. Kontopoulou, A.D. Panagopoulos, I. Kakkos, G.K. Matsopoulos, A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks, *Future Internet* 15 (8) (2023) 255.
- [14] S. Siami-Namini, N. Tavakoli, A.S. Namin, A comparison of ARIMA and LSTM in forecasting time series, in: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Ieee, 2018, pp. 1394–1401.
- [15] A. Carrassi, M. Bocquet, L. Bertino, G. Evensen, Data assimilation in the geosciences: an overview of methods, issues, and perspectives, *Wiley Interdiscip. Rev. Clim. Change* 9 (5) (2018) e535.
- [16] M.I. Ribeiro, Kalman and extended kalman filters: concept, derivation and properties, *Inst. Syst. Rob.* 43 (46) (2004) 3736–3741.
- [17] E.A. Wan, R. Van Der Merwe, The unscented kalman filter for nonlinear estimation, in: Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373), Ieee, 2000, pp. 153–158.
- [18] S. Cheng, C. Quilodrán-Casas, S. Ouala, A. Farchi, C. Liu, P. Tandeo, R. Fablet, D. Lucor, B. Iooss, J. Brajard, et al., Machine learning with data assimilation and uncertainty quantification for dynamical systems: a review, *IEEE/CAA J. Autom. Sin.* 10 (6) (2023) 1361–1387.
- [19] T. Gleiter, T. Janjić, N. Chen, Ensemble Kalman filter based data assimilation for tropical waves in the MJO skeleton model, *Q. J. R. Meteorol. Soc.* 148 (743) (2022) 1035–1056.
- [20] F.J. Gonzalez, M. Balajewicz, Deep convolutional recurrent autoencoders for learning low-dimensional feature dynamics of fluid systems, (2018).
- [21] S. Hochreiter, The vanishing gradient problem during learning recurrent neural nets and problem solutions, *Int. J. Uncertainty Fuzziness Knowl. Based Syst.* 6 (02) (1998) 107–116.
- [22] B. Chang, M. Chen, E. Haber, E.H. Chi, AntisymmetricRNN: A dynamical system view on recurrent neural networks, (2019).
- [23] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, W.-c. Woo, Convolutional LSTM network: a machine learning approach for precipitation nowcasting, *Adv. Neural Inf. Process. Syst.* 28 (2015).
- [24] V. Iakovlev, H. Lähdesmäki, Modeling Randomly Observed Spatiotemporal Dynamical Systems, (2024).
- [25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Adv. Neural Inf. Process. Syst.* 30 (2017), 6000–6010.
- [26] N. Geneva, N. Zabarar, Transformers for modeling physical systems, *Neural Netw.* 146 (2022) 272–289.
- [27] C. Feichtenhofer, Y. Li, K. He, et al., Masked autoencoders as spatiotemporal learners, *Adv. Neural Inf. Process. Syst.* 35 (2022) 35946–35958.
- [28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, *OpenAI blog* 1 (8) (2019) 9.
- [29] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, (2018).
- [30] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.
- [31] Z. Li, Z. Rao, L. Pan, P. Wang, Z. Xu, Ti-mae: Self-supervised masked time series autoencoders, (2023).
- [32] G. Karlsson, Detecting Anomalies in Imbalanced Financial Data with a Transformer Autoencoder, 2024,
- [33] H. Patel, R. Qiu, A. Irwin, S. Sadiq, S. Wang, EMIT-Event-Based Masked Auto Encoding for Irregular Time Series, (2024).
- [34] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, L. Sun, Transformers in time series: A survey, (2022).
- [35] C. Zhu, J. Fu, D. Xiao, J. Wang, Nonlinear model order reduction of engineering turbulence using data-assisted neural networks, *Comput. Phys. Commun.* 309 (2025) 109501.
- [36] M. Takamoto, T. Praditia, R. Leiteritz, D. MacKinlay, F. Alesiani, D. Pflüger, M. Niepert, Pdebench: an extensive benchmark for scientific machine learning, *Adv. Neural Inf. Process. Syst.* 35 (2022) 1596–1611.
- [37] B. Huang, C. Liu, V. Banzon, E. Freeman, G. Graham, B. Hankins, T. Smith, H.-M. Zhang, Improvements of the daily optimum interpolation sea surface temperature (DOISST) version 2.1, *J. Clim.* 34 (8) (2021) 2923–2939.
- [38] R.J. LeVeque, Finite volume methods for hyperbolic problems, 31, Cambridge university press, 2002.
- [39] S. Cheng, M. Bocquet, W. Ding, T.S. Finn, R. Fu, J. Fu, Y. Guo, E. Johnson, S. Li, C. Liu, et al., Machine learning for modelling unstructured grid data in computational physics: a review, *Inf. Fusion* 114 (2025) 103255.
- [40] Q. Liu, J. Ye, H. Liang, L. Sun, B. Du, TS-MAE: a masked autoencoder for time series representation learning, *Inf. Sci.* 690 (2025) 121576.
- [41] Y.-Y. Xu, J. Luo, D. Pan, W. Lu, T. Liu, G. Yuan, M. Zhong, Q. Li, H. Gong, A latent-coupled neural network for multiphysics long-term forecasting in reactor transients using sparse observations, *Eng. Appl. Artif. Intell.* 162 (2025) 112496.
- [42] S. Riva, A. Missaglia, C. Introini, I.C. Bang, A. Cammi, A Comparison of Parametric Dynamic Mode Decomposition Algorithms for Thermal-Hydraulics Applications, (2025).
- [43] P. Benner, S. Gugercin, K. Willcox, A survey of projection-based model reduction methods for parametric dynamical systems, *SIAM Rev.* 57 (4) (2015) 483–531.
- [44] M.J. Asher, B.F.W. Croke, A.J. Jakeman, L.J.M. Peeters, A review of surrogate models and their application to groundwater modeling, *Water Resour. Res.* 51 (8) (2015) 5957–5973.
- [45] R. Fu, D. Xiao, A.G. Buchan, X. Lin, Y. Feng, G. Dong, A parametric nonlinear non-intrusive reduce-order model using deep transfer learning, *Comput. Methods Appl. Mech. Eng.* 438 (2025) 117807.
- [46] X. Pan, D. Xiao, Domain decomposition for physics-data combined neural network based parametric reduced order modelling, *J. Comput. Phys.* 519 (2024) 113452.
- [47] M. Abbaszadeh, A. Khodadadian, M. Parvizi, M. Dehghan, D. Xiao, A reduced-order least squares-support vector regression and isogeometric collocation method to simulate Cahn-Hilliard-Navier-Stokes equation, *J. Comput. Phys.* 523 (2025) 113650.
- [48] K. Fukami, K. Taira, Grasping extreme aerodynamics on a low-dimensional manifold, *Nat. Commun.* 14 (1) (2023) 6480.
- [49] J. Guo, D. Xiao, Nonlinear Model Reduction by Probabilistic Manifold Decomposition, (2025).
- [50] A. Maćkiewicz, W. Ratajczak, Principal components analysis (PCA), *Comput. Geosci.* 19 (3) (1993) 303–342.
- [51] B. Karasözen, S. Yıldız, M. Uzunca, Intrusive and data-driven reduced order modelling of the rotating thermal shallow water equation, *Appl. Math. Comput.* 421 (2022) 126924.
- [52] P. Li, Y. Pei, J. Li, A comprehensive survey on design and application of autoencoder in deep learning, *Appl. Soft. Comput.* 138 (2023) 110176.
- [53] J. Wu, D. Xiao, M. Luo, Deep-learning assisted reduced order model for high-dimensional flow prediction from sparse data, *Phys. Fluids* 35 (10) (2023) 103115.
- [54] L.R. Medsker, L. Jain, et al., Recurrent neural networks, *Des. Appl.* 5 (64–67) (2001) 2.
- [55] Z.C. Lipton, J. Berkowitz, C. Elkan, A critical review of recurrent neural networks for sequence learning, (2015).
- [56] M. Sundermeyer, R. Schlüter, H. Ney, Lstm neural networks for language modeling, in: *Interspeech*, 2012, 2012, pp. 194–197.
- [57] H. Gong, Q. Li, Accelerating long-term xenon dynamics prediction: a reduced-order hybrid recurrent neural network with intrinsic physics, *Nucl. Sci. Eng.* 200 (2025) 1–21.
- [58] S. Cheng, J. Chen, C. Anastasiou, P. Angeli, O.K. Matar, Y.-K. Guo, C.C. Pain, R. Arcucci, Generalised latent assimilation in heterogeneous reduced spaces with machine learning surrogate models, *J. Sci. Comput.* 94 (1) (2023) 11.
- [59] M. Lechner, R. Hasani, Learning long-term dependencies in irregularly-sampled time series, (2020).
- [60] Z. Yang, Z. Dai, R. Salakhutdinov, W.W. Cohen, Breaking the softmax bottleneck: A high-rank RNN language model, (2017).
- [61] J. Fan, Q. Li, J. Hou, X. Feng, H. Karimian, S. Lin, A spatiotemporal prediction framework for air pollution based on deep RNN, *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* 4 (2017) 15–22.
- [62] P.M.T. Broersen, R. Bos, Estimating time-series models from irregularly spaced data, *IEEE Trans. Instrum. Meas.* 55 (4) (2006) 1124–1131.
- [63] M. Lepot, J.-B. Aubin, F.H. Clemens, Interpolation in time series: an introductive overview of existing methods, their performance criteria and uncertainty assessment, *Water* 9 (10) (2017) 796.
- [64] V. Fortuin, D. Baranchuk, G. Rätsch, S. Mandt, GP-VAE: deep probabilistic time series imputation, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 1651–1661.
- [65] Z. Guo, Y. Wan, H. Ye, A data imputation method for multivariate time series based on generative adversarial network, *Neurocomputing* 360 (2019) 185–197.
- [66] Y. Luo, X. Cai, Y. Zhang, J. Xu, et al., Multivariate time series imputation with generative adversarial networks, *Adv. Neural Inf. Process. Syst.* 31 (2018), 15944–15954.
- [67] C. Fang, C. Wang, Time series data imputation: A survey on deep learning approaches, (2020).
- [68] J. Yoon, J. Jordon, M. Schaar, Gain: missing data imputation using generative adversarial nets, in: International Conference on Machine Learning, PMLR, 2018, pp. 5689–5698.
- [69] Y. Zhuang, S. Cheng, K. Duraisamy, Spatially-aware diffusion models with cross-attention for global field reconstruction with sparse observations, *Comput. Methods Appl. Mech. Eng.* 435 (2025) 117623.
- [70] S. Riva, C. Introini, E. Zio, A. Cammi, Data-driven reduced order modelling with malfunctioning sensors recovery applied to the molten salt reactor case, *EPJ Nuclear Sciences & Technologies* 11 (2025) 55.
- [71] P. Gao, T. Ma, H. Li, Z. Lin, J. Dai, Y. Qiao, Convmae: Masked convolution meets masked autoencoders, (2022).
- [72] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, C. Eickhoff, A transformer-based framework for multivariate time series representation learning, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 2114–2124.
- [73] J. Fu, D. Xiao, R. Fu, C. Li, C. Zhu, R. Arcucci, I.M. Navon, Physics-data combined machine learning for parametric reduced-order modelling of nonlinear dynamical systems in small-data regimes, *Comput. Methods Appl. Mech. Eng.* 404 (2023) 115771.
- [74] G. Mahalakshmi, S. Sridevi, S. Rajaram, A survey on forecasting of time series data, in: 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), IIEEE, 2016, pp. 1–8.
- [75] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, *IEEE Trans. Image Process.* 13 (4) (2004) 600–612.
- [76] Q. Huynh-Thu, M. Ghanbari, The accuracy of PSNR in predicting video quality for different video scenes and frame rates, *Telecommun. Syst.* 49 (2012) 35–48.
- [77] J. Yoon, W.R. Zame, M. van der Schaar, Estimating missing data in temporal data streams using multi-directional recurrent neural networks, *IEEE Trans. Biomed. Eng.* 66 (5) (2018) 1477–1490.
- [78] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu, Recurrent neural networks for multivariate time series with missing values, *Sci. Rep.* 8 (1) (2018) 6085.
- [79] S. Cheng, J.-P. Argaud, B. Iooss, D. Lucor, A. Ponçot, Background error covariance iterative updating with invariant observation measures for data assimilation, *Stochastic Environ. Res. Risk Assess.* 33 (11) (2019) 2033–2051.

- [80] J. Kaplan, S. McCandlish, T. Henighan, T.B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, (2020).
- [81] A. Dosovitskiy, An image is worth 16x16 words: Transformers for image recognition at scale, (2020).
- [82] O. Press, N.A. Smith, M. Lewis, Train short, test long: Attention with linear biases enables input length extrapolation, (2021).
- [83] J. Su, Y. Lu, S. Pan, B. Wen, Y. Liu, RoFormer: Enhanced transformer with rotary position embedding, (2021).