

ST-PCNN: Spatio-Temporal Physics-Coupled Neural Networks for Dynamics Forecasting

Yu Huang, James Li, Min Shi, Hanqi Zhuang, Xingquan Zhu *Senior Member, IEEE*,
 Laurent Chérubin, James VanZwieten, and Yufei Tang *Member, IEEE*

Abstract—Ocean current, fluid mechanics, and many other spatio-temporal physical dynamical systems are essential components of the universe. One key characteristic of such systems is that certain physics laws – represented as ordinary/partial differential equations (ODEs/PDEs) – largely dominate the whole process, irrespective of time or location. Physics-informed learning has recently emerged to learn physics for accurate prediction, but they often lack a mechanism to leverage localized spatial and temporal correlation or rely on hard-coded physics parameters. In this paper, we advocate a physics-coupled neural network model to learn parameters governing the physics of the system, and further couple the learned physics to assist the learning of recurring dynamics. A spatio-temporal physics-coupled neural network (ST-PCNN) model is proposed to achieve three goals: (1) learning the underlying physics parameters, (2) transition of local information between spatio-temporal regions, and (3) forecasting future values for the dynamical system. The physics-coupled learning ensures that the proposed model can be tremendously improved by using learned physics parameters, and can achieve good long-range forecasting (e.g., more than 30-steps). Experiments, using simulated and field-collected ocean current data, validate that ST-PCNN outperforms existing physics-informed models.

Index Terms—Spatio-temporal, Physics-informed, Neural networks, Dynamics modeling, Spatial datasets, Ocean current

1 INTRODUCTION

Spatio-temporal modeling is essential in many scientific fields ranging from studies in biology [1], [2], information flow in social networks [3], sensor network communications [4], traffic predictions [5], [6], climate and environment forecasts [7], [8], to recent COVID-19 spread modeling [9]. These applications rely on accurate predictions of spatio-temporal structured data reflecting the real-world phenomena. In all mentioned cases, the major challenge is to infer, model, and predict the underlying causes, which generate the perceived data stream and propagate the involved causal dynamics through graphs and distributed sensor meshes. A stunning characteristic of such dynamical systems is that the widely distributed members (or sensors) share striking *homogeneity* and *heterogeneity*. The former is driven by the physics laws governing the systems, whereas the latter is impacted by the localized factor in spatial and temporal regions.

Take information propagation mechanisms of ocean current in Fig. 1 as an example, where each node denotes a geographic location observing ocean current. Three types of

This work was supported in part by the U.S. National Academy of Sciences Gulf Research Program through Grant No. xxx and the U.S. National Science Foundation through Grant Nos. OAC-2017597 and IIS-1763452.

- Y. Huang, Y. Tang, X. Zhu, J. VanZwieten, and H. Zhuang are with the Department of Computer & Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA.
E-mail: {yhwang2018, zhuang, xzhu3, jvanzwi, tangy}@fau.edu
- L. Chérubin is with the Harbor Branch Oceanographic Institute, Florida Atlantic University, Fort Pierce, FL 34946, USA. E-mail: lcherubin@fau.edu
- J. Li was with the Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA. E-mail: jli494@gatech.edu
- M. Shi is with the Department of Genetics, Washington University School of Medicine, St. Louis, MO 63110, USA. E-mail: mins@wustl.edu

Manuscript received xxx xx, 2020; revised xxx xx, 2021.

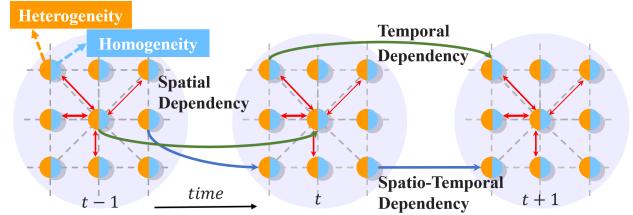


Fig. 1: Three types of dependencies in spatio-temporal modeling. A node is mixed with heterogeneity (orange) and homogeneity (blue) information, propagating between neighbors. At any time point t , the status of a center node is influenced by its previous time point ($t-1$) and its neighbors (red-arrows with various levels of strengths/weights).

dependencies exist in spatio-temporal modeling: 1) Spatial dependence: a node in the mesh concurrently affects, and be affected by, its neighbors; 2) Temporal dependence: a node status depends on its previous status and affects its future status; and 3) Spatio-temporal dependence: a node even directly influence its neighbor nodes across the time.

Deep learning methods, such as graph neural networks (GNNs), have been applied to spatio-temporal modeling. Existing methods take temporal information into account – e.g. ARIMA [10], or integrate complex spatial dependencies into temporal models – e.g. ConvLSTM [11] and ST-3DNet [12]. Most recently, researchers utilize graph convolution methods, such as DCRNN [13] and STGCN [6], to model spatial correlations in spatio-temporal structured data. Instead of modeling the spatio and temporal correlations separately, STSGCM [14] and STG2Seq [15] try to simultaneously capture the localized spatio-temporal correlations. It is worth mentioning that STSGCM deploys multiple modules

on each time period to capture the heterogeneity, which is computationally intensive.

One major issue of existing deep learning methods is that they seldom include prior knowledge of the underlying physics, i.e. taking the “homogeneity” into consideration. A key property, that all spatio-temporal processes have in common, is that some generally underlying principles will apply irrespective of time or location when observing natural processes. As a result, the same predictable patterns individually modified by local spatial and temporal influences are observable repeatedly at different spatial locations in time. Recently, physics informed learning has emerged to incorporate physics into the deep learning [16], [17], [18]. PDE-Net [19] and ODE-Net [20] are proposed to train neural networks that simultaneously approximate the simulations and conforms to the PDEs representing the physical knowledge of systems. However, these methods suffer two major shortcomings: (1) they are often restricted to 1D temporal sequence or to a regular grid where constraints on the learnable filters can be easily defined; and (2) there is no good solution to combine homogeneity and heterogeneity for effective prediction. In summary, three research challenges are identified as follows:

Learning physics: Physics is one of the fundamental pillars describing how the real-world behaves. Although physics informed learning has [16], [17], [18] taken physics into consideration, such methods are only applicable when specific/general equations are explicitly given and none of them consider the spatio-temporal cases. In reality, it is not always possible to describe all rules governing real-world data. We need to have a learning mechanism to automatically discover the physics underneath the data observations.

Coupling physics and spatio-temporal information: Homogeneity and heterogeneity are two key characteristics of dynamical systems, but governed by different modules. We need to have a new way to enable the learning of physics, and use the inferred physics to further guide the spatio-temporal learning with robust prediction, regardless of physical plausibility.

Long-term forecasting: For dynamical systems, long-term forecasting allows proactive controls and early planning. However, most existing models only work on a one-step-ahead short-term forecasting [21], [22]. An intuitive way to achieve long-term prediction is to recursively reuse previous-step predictions as input for the next-step prediction. Inevitably, such a mechanism leads to prediction errors that are probable to accumulate over time, and the results obtained may be modest.

In this paper, we propose a spatio-temporal physics-coupled neural networks (ST-PCNN) model to capture spatio-temporal correlations, and heterogeneity and its inherited homogeneity in spatially distributed manner. ST-PCNN is a three-network architecture, consisting of a forecasting net (FN), a transition net (TN), and a physics net (PN). ST-PCNN learns predictive neural network (FNs) that are distributively executed at different locations of a grid. Additional information routing transition neural network (TNs) laterally connect the FN. Both FN and TN share their weights respectively, allowing efficient parallel computation and capturing heterogeneity from all spatial locations.

In order to incorporate physics laws with which an effective model is supposed to follow and helps training with less samples and being robust to unseen data, a third network PN is developed to reveal unknown governing physics from pre-given spatio-temporal data and vice versa facilitates the overall model to capture the homogeneity.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 defines the spatio-temporal forecasting problem for the dynamical systems from a machine-learning perspective. The proposed model ST-PCNN is introduced in Section 4, followed by experiments in Section 5 that includes comparative and ablation studies and conclusion in Section 6.

2 RELATED WORK

Physical process modeling is close to the field of spatio-temporal statistical models that are increasingly being used across a wide variety of scientific disciplines to describe and predict spatially explicit processes that evolve over time [23]. [24] advocate the use of physical prior knowledge to develop statistical models, e.g., PDEs related to the observed real-world phenomenon. They mainly consider auto regressive models within a hierarchical Bayesian framework. Another research direction is the use of neural networks (NNs) for enhancing the performance and reducing the complexity of numerical physical process simulation. There are three major approaches: 1) NNs are used in place of a computational demanding component of the simulation process. For examples, [25] uses a random forest to compute particle location and [26] adopts a CNN to approximate part of a numerical PDE scheme. 2) NNs are combined with related physics equations to model the whole physical process. For example, [27] uses physics-informed neural networks (PINNs) to learn nonlinear relations between spatial- and temporal-coordinates with a given PDE. The given physics could be changed for various purposes, such as advection-diffusion equation informed in [28], fluid dynamics in [7], Lagrangian mechanics in [29], and Hamiltonian in SymODEN [30], etc. These methods are only applicable when the specific equations are explicitly given but hard to be generalized to incorporate other types of physics equations. 3) NNs are used to uncover the underlying hidden physics and model the dynamics of complex systems. State-of-the-art work includes discovering the PDEs [16], [19], [31] or ODEs [20] from given observations of the systems.

3 PROBLEM DEFINITION

A dynamical system is observed from a grid of nodes (e.g., a sensor network), distributed/located at different locations. $s^{(t,i,j)} \in \mathbb{R}$ denotes observed value of a node located at (i, j) at time t . For ease of representation we use $s^{(t,:)}$ to denote value of any node at time t , and $s^{(:,i,j)}$ denotes values observed at (i, j) . Accordingly, each time slice of the grid observation is denoted by $\mathcal{S}^{(t,:)} = \{s^{(t,i,j)}\}_{i,j \in \Omega} \in \mathbb{R}^{\Omega}$ and the time series of each node is defined by $\mathcal{S}^{(:,i,j)} = \{s^{(t,i,j)}\}_{t \in T} \in \mathbb{R}^T$ (Ω is the total nodes and T is the total recorded time steps). $\mathcal{S} \in \mathbb{R}^{\Omega \times T}$ denotes the whole observations and $\hat{\mathcal{S}}$ represents the variable flow in the future.

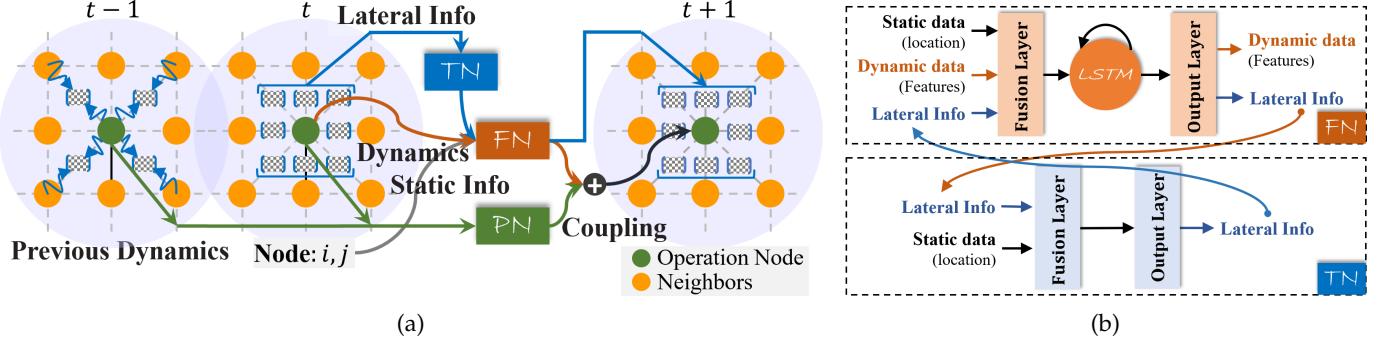


Fig. 2: (a): The lateral connection schema of FN, TN, and PN. Green node denote a center node located at (i, j) , and orange nodes are its neighbors. Three types of information are used to characterize each node: (1) Dynamics $\mathcal{S}^{(t,i,j)}$: an embedding vector that represents status of node at time t , (2) Static Info $\tilde{\mathbf{p}}^{(i,j)}$: an embedding vector that represents node location; and (3) Lateral Info $\mathcal{L}^{(t,i,j)}$: an embedding vector (dashed dot-square set) capturing interaction (lateral info) between each node and its neighbors; (b): The inputs, outputs, and an exemplary topology of FN and TN with recurrent connections.

Assume variable observed from a node is governed by an unknown physics rule, i.e., PDE/ODE¹, which relates a function $u(\mathbf{x}, t)$ with its derivatives, i.e., $\mathcal{D}^a u(\mathbf{x}, t) := \frac{\partial^{a_1}}{\partial x_1^{a_1}} \cdots \frac{\partial^{a_k}}{\partial x_k^{a_k}} u(\mathbf{x}, t)$, where $a = (a_1, \dots, a_k)$ are non-negative integers. Consequently, any of the PDEs can be defined by this notion as:

$$\mathcal{G}(\mathbf{x}, t, u(\mathbf{x}, t), \mathcal{D}^a u(\mathbf{x}, t)) = 0 \quad (1)$$

where \mathcal{G} is a function that relates position $\mathbf{x} = (i, j) \in \Omega$ and time $t \in \mathbb{R}$ with u and its partial derivatives at \mathbf{x} and t . We say u is a solution to the PDE if Eq. (1) holds for every point $\mathbf{x} \in \Omega$ and time step $t \in \mathbb{R}$. In this paper, we observe data points $(\mathbf{t}, \mathbf{x}; \mathcal{S}) = \{t, i, j; s^{(t,i,j)}\}_{i,j \in \Omega}$, where \mathcal{S} are noisy function values at \mathbf{x} , that is, $s^{(t,i,j)} = u(t, i, j) + \xi^{(t,i,j)}$ with noise $\xi^{(t,i,j)}$ impacted by the temporal localized factor. The $u(t, i, j)$ refers to homogeneity while $\xi^{(t,i,j)}$ is the cause of heterogeneity in dynamical systems.

Spatio-Temporal Forecasting Problem: From a machine-learning perspective, the spatio-temporal forecasting problem is to learn a non-linear mapping function f that maps the historical spatio-temporal observations $\{\mathcal{S}^{t_p-\tau}, \mathcal{S}^{t_p-\tau+1}, \dots, \mathcal{S}^{t_p}\}$ into the future predictions $\{\hat{\mathcal{S}}^{t_p+\tau+1}, \hat{\mathcal{S}}^{t_p+\tau+2}, \dots, \hat{\mathcal{S}}^{t_p+\tau'}\}$, where τ denotes the length of observation conditioned on and τ' denotes the prediction horizon. The learning is formulated as a deterministic optimization problem that constitutes both minimizing the data mismatches and estimating the hidden underlying PDE of a physical model by equating derivatives of the neural network approximation.

4 THE PROPOSED FRAMEWORK: ST-PCNN

To better modeling the dynamical systems in terms of considering both of the homogeneity and heterogeneity, a spatio-temporal physics-coupled neural networks (ST-PCNN) model is proposed. ST-PCNN is a tri-network architecture, as shown in Figs. 2a and 2b, consisting of physics network (PN), forecasting network (FN), and transition network (TN). FN receives 1) dynamic data, which is subject

1. Ordinary differential equations form a subclass of partial differential equations, corresponding to functions of a single variable.

to prediction and changes over time, 2) static information, which stays constant and characterizes the location of each FN, and 3) lateral information from neighbors. The output of each FN includes predicted dynamics and additional lateral information that will be interacted with its neighbors. Such interaction, that distinguishes our architecture from others, is conducted through a TN with two-stacked linear layers. TN aims to model the location-sensitive transitions between adjacent FNs and thus enabling local context-dependent spatial information propagation.

4.1 Spatio-Temporal Heterogeneity

In many natural phenomena, data are collected in a distributed manner and exhibit heterogeneous properties: each of these distributed locations present a different view of the natural process at the same time, where each view has its own individual representation of the space and dynamics [32]. Theoretically, each location may contain information that other location do not have access to. Therefore, all local views must be interacted in some way in order to describe the global activity comprehensively and accurately.

How to explicitly encode location information into neural networks is critical in this location-wise forecasting. Inspired by the Transformer [33] that encodes word positions in sentences, we extend the absolute positional encoding to represent grid positions. In particular, let i, j be the desired position in a regular grid, $\tilde{\mathbf{p}}^{(i,j)} \in \mathbb{R}^d$ be its corresponding encoding, and d be the encoding dimension, then the encoding scheme is defined as:

$$\tilde{\mathbf{p}}^{(i,j)} := \begin{cases} [\sin(\omega_k, i), \sin(\omega_k, j)] & \text{if } i, j = 2k \\ [\cos(\omega_k, i), \cos(\omega_k, j)] & \text{if } i, j = 2k + 1 \end{cases} \quad (2)$$

where $\omega_k = \frac{1}{10,000^{2k/d}}, k \in \mathbb{N}_{\leq \lfloor \frac{d}{2} \rfloor}$. The wavelengths form a geometric progression from 2π to $10000 \cdot 2\pi$. The positional embedding as a vector contains pairs of *sines* and *cosines* for each decreasing frequency along the vector dimension, it would allow the model to easily learn to attend by relative positions [33].

As illustrated in Fig. 2, the FN and TN are executed in space simultaneously. At each time t , the TN first encodes

the current operation node's lateral info $\hat{\mathbf{L}}$ and static info $\tilde{\mathbf{p}}$ as follows:

$$\hat{\mathbf{L}}_{enc}^{(t,i,j)} = \text{Relu}([\tilde{\mathbf{p}}^{(i,j)}, \hat{\mathbf{L}}^{(t,i,j)}] \mathcal{W}_T^T + b_T) \quad (3)$$

where $\theta_T = [\mathcal{W}_T, b_T]$ denote the weights and bias of TN. $\hat{\mathbf{L}}^{(t,i,j)}$ is a vector used to characterize interaction between a node at i, j to its neighbors. It is initialized as zeros at first step and continuously updated by Eq. (7) when $t > 0$.

Then, FN encodes each view (i.e., static $\tilde{\mathbf{p}}$, dynamics \mathcal{S} , and encoded $\hat{\mathbf{L}}_{enc}$ of each node) using a fusion layer:

$$f^{(t,i,j)} = [\tilde{\mathbf{p}}^{(i,j)}, \mathcal{S}^{(t,i,j)}, \hat{\mathbf{L}}_{enc}^{(t,i,j)}] \mathcal{W}_{fusion}^T + b_{fusion} \quad (4)$$

These features, $f^{(t,i,j)} \in \mathbb{R}^{d_{f_{i,j}}}$, are then fed into an LSTM to model the node-specific interactions over time. The update mechanism of the LSTM cell is defined as:

$$[\mathcal{I}^{(t)}; \mathcal{F}^{(t)}; \tilde{\mathcal{C}}^{(t)}; \mathcal{O}^{(t)}] = \sigma(\mathcal{W} \cdot f^{(t,i,j)} + \mathcal{T} \cdot h^{(t-1)}) \quad (5)$$

$$\mathcal{C}^{(t)} = \tilde{\mathcal{C}}^{(t)} \circ \mathcal{I}^{(t)}; h^{(t)} = \mathcal{O}^{(t)} \circ \mathcal{C}^{(t)} \quad (6)$$

where $\sigma(\cdot)$ applies sigmoid on the input gate $\mathcal{I}^{(t)}$, forget gate $\mathcal{F}^{(t)}$, and output gate $\mathcal{O}^{(t)}$, and $\tanh(\cdot)$ on memory cell $\tilde{\mathcal{C}}^{(t)}$. The parameters are characterized by $\mathcal{W} \in \mathbb{R}^{d_{f_{i,j}} \times d_{h_{i,j}}}$ and $\mathcal{T} \in \mathbb{R}^{d_{h_{i,j}} \times d_{h_{i,j}}}$, where $d_{h_{i,j}}$ is the output dimension. A cell updates its hidden states $h^{(t)}$ based on the previous step $h^{(t-1)}$ and the current input $f^{(t,i,j)}$.

An output layer is stacked at the end of FN to transform the LSTM output into the expected dynamic prediction and additional lateral information as:

$$[\hat{s}^{(t,i,j)}; \hat{\mathbf{L}}^{(t,i,j)}] = \text{Relu}(\mathcal{W}_{out} \cdot f^{(t,i,j)} + b_{out}) \quad (7)$$

where $\hat{s}^{(t,i,j)}$ denotes the prediction of the node dynamics at time step t . The learnable parameters are characterized by $\mathcal{W}^{(t)} \in \mathbb{R}^{d_{f_{i,j}} \times d_{y_{i,j}}}$ and $b_{out} \in \mathbb{R}^{d_{y_{i,j}}}$, where $d_{y_{i,j}}$ denotes the total dimension of the dynamic and the lateral outputs.

4.2 Homogeneity by Underlying Physics

Physicists attempt to model natural phenomena in a principled way through analytic descriptions. Conservation laws, physical principles, or phenomenological behaviors are generally formalized using differential equations, which can best reasoning the homogeneity of observations. Knowledge accumulated for modeling physical processes in well developed fields such as maths or physics could be a useful guideline for dynamics learning [28]. Our proposed ST-PCNN includes a physics-aware module, the physics network PN, to learn underlying hidden physics. Two main approaches are considered:

- **PDE-learning-net:** explicitly estimating the underlying partial differential equation (PDE) from time series assisted by neural network model fitting to PDE solution.
- **ODE-informed-net:** implicitly approximating time dependence with neural network and solving via an ordinary differential equation (ODE) solver.

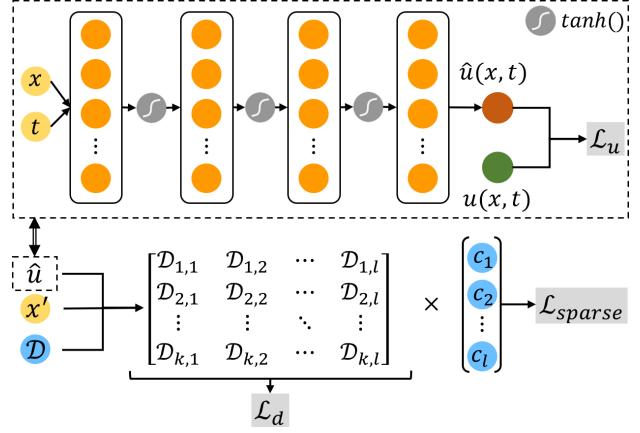


Fig. 3: Diagram of the PDE-learning-net structure and loss calculations.

4.2.1 PDE-learning-net

Partial differential equation (PDE) is an equation which imposes relations between the various partial derivatives of a multivariable function. PDEs are ubiquitous in mathematically-oriented scientific fields, such as physics and engineering. For instance, they are foundations in the modern scientific understanding of sound, heat, diffusion, fluid dynamics, general relativity, quantum mechanics, etc. PDEs for a single variable v can be generally described as a linear combination of functions of v , its derivatives, and the dimensions. More formally, let $\mathcal{D} = \{\mathcal{D}_1, \dots, \mathcal{D}_K\}$ be a dictionary of such terms, we can then generally define PDEs as:

$$\sum_{k=1}^K c_k \mathcal{D}_k(v, u(v), \mathcal{D}^a u(v)) = 0, \forall v \in \Omega_v \quad (8)$$

where $\mathbf{c} = \{c_k\}_{k=1}^K$ is a set of coefficients to be determined. These terms are determined by best estimate of what would be relevant for each use case. For most physical systems, this would typically be limited to second order derivatives².

Suppose we observe $u(x, t)$ such that u is a solution to Eq. (8). It is then possible to approximate u by a neural network $\hat{u} : \Omega \in \mathbb{R}$, where observations at (x, t) becomes the training inputs. To approximate u , the PDE-learning-net [31], as illustrated in the dotted box of Fig. 3, consists of a four-linear-layer stacked, fully-connected network with linear input and output layers as denoted below:

$$\hat{u}(x, t) = \tanh([\mathbf{x}, t] \mathcal{W}_1^T + b_1) \cdots \mathcal{W}_n^T + b_n \quad (9)$$

where the inputs are position x and time t , and yields the prediction value \hat{u} at that position. The $\tanh()$ activation function is employed after each layer except for the output layer.

This approximation is optimized by a combination of multiple loss terms to place emphasis on different parts of the model, defined as:

$$\mathcal{L}(\mathbf{x}, \mathbf{x}'; \mathbf{c}, \theta) = \mathcal{L}_u^{1/2} (1 + \lambda_d \mathcal{L}_d + \lambda_{sparse} \mathcal{L}_{sparse}) \quad (10)$$

2. Example: suppose the 1D wave equation is $au_{tt} - bu_{xx} = g(u_{tt}, u_{xx}) = 0, \forall \mathbf{x} \in \Omega \subset \mathbb{R}^2, \forall t \in \mathbb{R}$. If we define our dictionary as $\mathcal{D}_1 = u_{tt}$, $\mathcal{D}_2 = u_{xx}$, where $a\mathcal{D}_1(u, x, t) - b\mathcal{D}_2(u, x, t) = 0$, the PDE can then be represented by the coefficients $\mathbf{c} = [a, -b]$.

Algorithm 1: PDE-learning-net

Input : Spatio-temporal point (\mathbf{x}, t) ; Dictionary of differential terms $\{\mathcal{D}(\hat{u}, \mathbf{x}, t)\}$

Output: \mathbf{c} (the estimated coefficients of PDE)

Initialize

- | Neural network parameters: θ of \hat{u} , $\theta_c / \|\theta_c\|$;
- | **for** number of epochs **do**
- | $\mathcal{L}_u(\mathbf{x}; \theta) \leftarrow \frac{1}{N} \sum_{i=1}^N (u_i - \hat{u}(\mathbf{x}_i; \theta))^2$;
- | $\mathcal{L}_d \leftarrow \|\mathcal{D}(\hat{u}, \mathbf{x})\mathbf{c}\|_2^2$;
- | $\mathcal{L}_{sparse} \leftarrow \|\mathbf{c}\|_1$;
- | $\mathcal{L} \leftarrow \mathcal{L}_u^{1/2} (1 + \lambda_d \mathcal{L}_d + \lambda_{sparse} \mathcal{L}_{sparse})$;
- | Update θ and $\theta_c \leftarrow \mathcal{L}.\text{backward}()$;
- | $\mathbf{c} \leftarrow \theta_c / \|\theta_c\|$;
- | **end**

where the use of \mathcal{L}_u as a scaling factor for the other losses acts as a regularizing term to maintain a more consistent ratio of losses even when the mean square error (MSE) becomes very large or small. The λ_d and λ_{sparse} coefficients are set such that the individual loss function contributions will have similar magnitude.

\mathcal{L}_u is the loss between observed data points u and values calculated by the neural network at these points \hat{u} :

$$\mathcal{L}_u(\mathbf{x}; \theta) = \frac{1}{N} \sum_{i=1}^N (u_i - \hat{u}(\mathbf{x}_i; \theta))^2 \quad (11)$$

where θ is the neural network parameters. This is the primary regression loss term for the neural network.

\mathcal{L}_d is a differentiation loss used to measure the error of the estimated PDE coefficients. Here, a dictionary of L differential terms $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_L\}$ is used. These functions are evaluated at K points $\mathbf{x}' = \{\mathbf{x}_i\}_{i=1}^K$, sampled from Ω , resulting in a $\mathbb{R}^{K \times L}$ matrix with entries:

$$\mathcal{D}(\hat{u}, \mathbf{x}', \theta)_{k,l} := \mathcal{D}_l(\mathbf{x}'_k, \hat{u}(\mathbf{x}'_k; \theta), \mathcal{D}^\alpha \hat{u}(\mathbf{x}'_k)) \quad (12)$$

By definition of the PDE, we require $\mathcal{D}(u, \mathbf{x}')\mathbf{c} = 0$, thus \mathbf{c} must lie within the null space of $\mathcal{D}(u, \mathbf{x}')$. Equivalently, \mathbf{c} is a singular vector of $\mathcal{D}(u, \mathbf{x}')$, with associated singular value 0. If $\|\hat{u} - u\|_{m,2} < \epsilon$ and assuming some regularity conditions on \mathcal{D} , we can have $\|\mathcal{D}(u, \mathbf{x}') - \mathcal{D}(\hat{u}, \mathbf{x}')\|_{m,2} < \epsilon C$ for some constant C that depends on \mathcal{D} . Therefore, the singular vector of $\mathcal{D}(\hat{u}, \mathbf{x}')$, associated with its smallest singular value, is an approximation of \mathbf{c} .

The loss term \mathcal{L}_d is then defined along with the constraint $\|\mathbf{c}\| = 1$ to avoid \mathbf{c} being minimized to zero:

$$\mathcal{L}_d(\mathbf{x}'; \theta, \mathbf{c}) = \|\mathcal{D}(\hat{u}(\mathbf{x}'; \theta), \mathbf{x}')\mathbf{c}\|_2^2 \quad (13)$$

The contribution of \mathcal{L}_d is twofold: minimizing $\mathcal{L}_d(\mathbf{x}', \theta, \mathbf{c})$ over \mathbf{c} recovers the PDE and minimizing $\mathcal{L}_d(\mathbf{x}', \theta, \mathbf{c})$ over \hat{u} additionally that encourages fitting to a solution of the PDE, thus preventing over-fitting to noise.

The loss term \mathcal{L}_{sparse} is further introduced:

$$\mathcal{L}_{sparse}(\mathbf{c}) = \|\mathbf{c}\|_1 \quad (14)$$

to impose the assumption that natural systems are inherently simple and are thus dependent on a few terms. The PDE-learning-net training is presented in Algorithm 1.

4.2.2 ODE-informed-net

Ordinary differential equation (ODE) is a differential equation containing one or more functions of one independent variable and the derivatives of those functions. The term ordinary is used in contrast with the term partial differential equation which may be with respect to more than one independent variables. ODE-informed-net, in this paper, is proposed to implicitly approximate time-dependence with neural network and solving via an ODE solver. The model structure is shown in Fig. 4.

ODE-informed-net training is performed with the DOPRI5 method [34] in the ODE solver with an absolute and relative tolerance of $1e^{-3}$. The DOPRI5 method is well established as an ODE solver and allows for adaptive steps, thus the model can then perform network evaluations with a dynamic and arbitrary number of times. This can be tuned in and outside of training by changing tolerances before network evaluation.

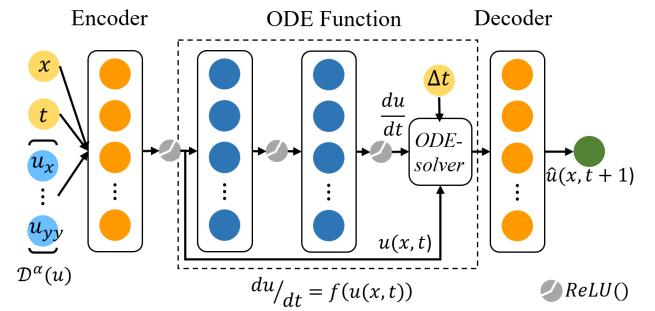


Fig. 4: Diagram of the ODE-informed-net model structure with ODE-solver integrated. The neural network consists of an Encoder u_E , an ODE-function u_O , and a Decoder u_D .

Model inputs are a set of derivatives of the target function, the current state, and the previous state. This is then encoded into the latent space, h_1 , as:

$$h_1 = \text{Relu}([\{\mathcal{D}^\alpha u(\mathbf{x}, t)\}] \mathcal{W}_{encoder}^T + b_{encoder}) \quad (15)$$

The model interprets this as the initial state for the ODE-solver³ DOPRI5 [34]. $g(u, \frac{du}{dt}, \Delta t)$, where Δt is the ODE solver time step and $\frac{du}{dt}$ is approximated by a neural network, in this case, a two-linear-layer stacked fully-connected network. The following process is looped within the ODE solver:

$$h'_1 = \text{Relu}(h_1 \mathcal{W}_{ode}^T + b_{ode}) \quad (16)$$

$$h_2 = g(h_1, h'_1, \Delta t) \quad (17)$$

where h'_1 is the time derivative of the latent space state, h_2 is the calculated latent space state at $t + \Delta t$, and t is the current time of the solver. If this time is not yet $t + 1$, t and h_1 are updated for the next iteration. Once this converges to expected tolerance, the final h_2 is decoded as the model's estimate for the true state at $t + 1$ as:

$$\hat{u}(\mathbf{x}, t + 1) = \mathcal{W}_{decoder} \cdot h_2 + b_{decoder} \quad (18)$$

3. <https://docs.scipy.org/doc/scipy/reference/generated/scipy.integrate.ode.html>

Algorithm 2: ODE-informed-net

Input : Observations: $u(\mathbf{x}, t), u(\mathbf{x}, t - 1)$
Output: Prediction: $\hat{u}(\mathbf{x}, t + 1)$

Initialize

- Neural network parameters $\theta: [\theta_E, \theta_O, \theta_D]$ of Encoder \hat{u}_E , ODE Function \hat{u}_O and Decoder \hat{u}_D ;
- for** number of epochs **do**

 - for** \mathbf{x} in Ω **do**

 - Perform finite difference for $\{\mathcal{D}^\alpha u(\mathbf{x}, t)\}$;
 - $u_E \leftarrow \hat{u}_E(u(\mathbf{x}, t), u(\mathbf{x}, t - 1), \{\mathcal{D}^\alpha u(\mathbf{x}, t)\}; \theta_E)$;
 - $u_O \leftarrow \int_0^1 \hat{u}_O(u(\mathbf{x}, \tau); \theta_O) d\tau + u_E$;
 - $\hat{u}(\mathbf{x}, t + 1) \leftarrow \hat{u}_D(u_O; \theta_D)$;

 - end**
 - $\mathcal{L} \leftarrow \|u(\mathbf{x}, t + 1) - \hat{u}(\mathbf{x}, t + 1)\|_2 + \alpha \|\theta_E, \theta_O, \theta_D\|_2$;
 - Update $\theta_E, \theta_O, \theta_D \leftarrow \mathcal{L}.\text{backward}()$;

- end**

Here, the ODE-informed-net unites a sub-network acting as an ODE function \hat{u}_O . By leveraging the multiple evaluations of this sub-network, the ODE-informed-net can effectively act as a deeper network while using fewer parameters and being more stable in training. When used to predict time-series, the ODE function solves for future states from some initial state at $u(t_0)$ as:

$$\frac{du}{dt} = f(u; \theta); \quad u(t_0) = (u_1(t_0), \dots, u_d(t_0))^T \quad (19)$$

where the united ODE function \hat{u}_O fits to the time derivative of u , mimicking the behavior of the physical systems. In more complex systems, if it is assumed that the system can be described with PDEs, then time gradients can be estimated as a function of the local state, previous state, and a set of spatial derivatives. These derivatives can be estimated via finite difference and be more formally defined as the set $\{\mathcal{D}^\alpha u\}$; $\alpha = 0, 1, \dots, m$, where m is the highest order of derivative with a typical value of 2.

Our implementation performs this calculation at individual points in space and predicts the state of that point in the next time step. By leveraging ODE integration tools, any future values can be solved with arbitrary accuracy. When approximating time series data, the ODE-informed-net is trained using pairs of data at sequential time steps to predict values for the next time step. The data is sampled at successive time points $t - 1$ and t at some spatial point \mathbf{x} and their corresponding values $u(\mathbf{x}, t - 1), u(\mathbf{x}, t)$, where $u = (u_1, u_2, \dots, u_d)^T$ is a d -dimension vector representing data values at time t and point \mathbf{x} . Finally, an estimate of $u(\mathbf{x}, t + 1)$ is calculated at each point using the input states as the initial state for the ODE integration.

Regression loss \mathcal{L} is defined for ODE-informed-net as a function of the MSE and the network parameters θ :

$$\mathcal{L}(\theta, \alpha) = \|u - \hat{u}\|_2 + \alpha \|\theta\|_2 \quad (20)$$

where \hat{u} denotes the prediction. The ODE-informed-net training is presented in Algorithm 2.

4.3 Coupling Hetero- and Homogeneity

Based on above analysis, we describe the ST-PCNN to model the heterogeneous properties of spatio-temporal data and to reveal the homogeneous physics from raw data.

Algorithm 3: ST-PCNN Training

Input : Dynamics: $\mathcal{S} \in \mathbb{R}^{T \times H \times W}$; Grid: $\Omega \in \mathbb{R}^{H \times W}$;
 $\mathcal{P}\mathcal{N} \leftarrow \text{ODE-informed-net/PDE-learning-net}$;

Initialize

- Neural network parameters: $\theta_T, \theta_F, \theta_C$;
- Static info: $\tilde{\mathbf{p}} \leftarrow \text{Positional-Encoding}(d, H, W)$;
- Lateral info: $\mathbf{L} \leftarrow \mathbf{0}$;
- for** number of epochs **do**

 - for** t in T **do**

 - for** i, j in Ω **do**

 - $\hat{\mathbf{L}}_{enc}^{(t,i,j)} \leftarrow \mathcal{T}\mathcal{N}(\tilde{\mathbf{p}}^{(i,j)}, \mathbf{L}^{(t,i,j)}, \theta_T)$;
 - $[\hat{\mathcal{S}}_{he}^{(t+1,i,j)}, \hat{\mathbf{L}}^{(t+1,i,j)}] \leftarrow$
 - $\mathcal{F}\mathcal{N}(\tilde{\mathbf{p}}^{(i,j)}, \mathcal{S}^{(t,i,j)}, \hat{\mathbf{L}}_{enc}^{(t,i,j)}, \theta_F)$;
 - $\hat{\mathcal{S}}_{ho}^{(t+1,i,j)} \leftarrow \mathcal{P}\mathcal{N}(\tilde{\mathbf{p}}^{(i,j)}, \hat{\mathcal{S}}^{(t-1,i,j)}, \mathcal{S}^{(t,i,j)})$;
 - $\hat{\mathcal{S}}^{(t+1,i,j)} = \text{Coupling}(\hat{\mathcal{S}}_{he}^{(t+1,i,j)}, \hat{\mathcal{S}}_{ho}^{(t+1,i,j)}, \theta_C)$;

 - end**
 - Update $\mathbf{L}^{(t+1,:)} \leftarrow \hat{\mathbf{L}}^{(t+1,:)}$;

 - $\mathcal{L} \leftarrow \frac{1}{n} \|\mathcal{S}^{(t,:)} - \hat{\mathcal{S}}^{(t,:)}\|_1 + \frac{1}{n} \|\mathcal{S}^{(t,:)} - \hat{\mathcal{S}}^{(t,:)}\|_2$;
 - Update $\theta_T, \theta_F, \theta_C \leftarrow \mathcal{L}.\text{backward}()$

Here, a stacking coupling mechanism is proposed to integrate the obtained physics into the spatio-temporal learning.

As shown in Fig. 2a, at each time step t and location i, j , the FN produces the initial prediction $\hat{\mathcal{S}}_{he}^{(t+1,i,j)}$ based on the current observation $\mathcal{S}^{(t,i,j)}$, the hidden states $h^{(t-1,i,j)}$ from previous-step (within LSTM), and the lateral info from its neighbors produced by TN. Here the ‘heterogeneous’ initial prediction leverages its own specific local attributes only. Then, regarding the integration of differential equations, the previous-step initial prediction $\hat{\mathcal{S}}^{(t-1,i,j)}$ and the current observation $\mathcal{S}^{(t,i,j)}$ are fed into the learnt physics PDEs/ODEs from PN to derive the numerical solution $\hat{\mathcal{S}}_{ho}^{(t+1,i,j)}$, which is the ‘homogeneous’ part of the dynamics regularized by governing physics. Finally, a coupling layer with parameters $\theta_C = [\mathcal{W}_C, b_C]$ in Eq. (21), is used to produce final prediction $\hat{\mathcal{S}}^{(t+1,i,j)}$ by synthesizing $\hat{\mathcal{S}}_{he}^{(t+1,i,j)}$ and $\hat{\mathcal{S}}_{ho}^{(t+1,i,j)}$.

$$\hat{\mathcal{S}}^{(t+1,i,j)} = \text{Relu}([\hat{\mathcal{S}}_{he}^{(t+1,i,j)}, \hat{\mathcal{S}}_{ho}^{(t+1,i,j)}] \mathcal{W}_C^T + b_C) \quad (21)$$

ST-PCNN training is presented in Algorithm 3 with supervised loss including sum of l_1 -norm and l_2 -norm loss.

5 EXPERIMENTS AND COMPARATIVE STUDY

5.1 Datasets

In this section, we evaluate ST-PCNN on a synthetic dataset and a real-world ocean current dataset, with the data statistics summarized in Table 1 and details introduced as below.

Reflected Wave Simulation Data As illustrated in Fig. 5, single-waves are propagating outwards, where waves are reflected at borders such that wave fronts become interactive. The following 2D wave equation was used for reflected wave data generation:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) \quad (22)$$

TABLE 1: The synthetic and real-world data statistics.

Data Sets	# of Time Points	Grid Size	Sampling Rate
Reflected Wave	8,000	16×16	0.1s
LC of GoM	1,810	29×36	12h

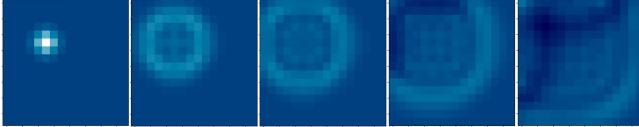


Fig. 5: Exemplary circular wave with reflecting borders. Plots from left to right denote temporal evolving of the circular wave (propagate from center to boundaries). After the wave reaches the border, reflecting effects are generated through boundary conditions.

PDE solutions were solved numerically using an explicit central difference approach:

$$\frac{\partial^2 u}{\partial b^2} = \frac{u(b+h) - 2u(b) + u(b-h)}{h^2} = u_{bb} \quad (23)$$

where b stands for a variable of function u , and h is the approximation step size. In the case of calculating simulated wave data, we apply Eq. (23) to Eq. (22) to obtain:

$$c^2(u_{xx} + u_{yy}) = \frac{u(x, y, t + \Delta t) - 2u(x, y, t) + u(x, y, t - \Delta t)}{\Delta_t^2} \quad (24)$$

which can be solved for $u(x, y, t + \Delta t)$ to obtain an equation for determining state of the field at the next time step $t + \Delta t$ at each point.

Both the boundary conditions (when $x < 0$ or $x > fieldwidth$, analogously for y) and initial condition (in time step 0) are treated as zero. The following variable choices were met: $\Delta_t = 0.1$, $\Delta_x = \Delta_y = 1$ and $c = 3.0$. The field was initialized using a Gaussian distribution:

$$u(x, y, 0) = a \exp \left(- \left(\frac{(x - s_x)^2}{2\sigma_x^2} + \frac{(y - s_y)^2}{2\sigma_y^2} \right) \right) \quad (25)$$

with amplitude factor $a = 0.34$, wave width in x and y directions $\sigma_x^2 = \sigma_y^2 = 0.5$, and s_x, s_y being the starting point or center of the circular wave.

Gulf of Mexico (GoM) Loop Current Data As illustrated in Fig. 6, the sensor array are placed in the GoM region, covered from $89^\circ W$ to $85^\circ W$, and $25^\circ N$ to $27^\circ N$ with $30-50$ km horizontal resolution, where the Loop Current (LC) extended northward and, more importantly, where eddy shedding events occurred most often [35]. This sensor array consisted of 25 pressure-recording inverted echo sounders (PIES), 9 full-depth tall moorings with temperature, conductivity and velocity measurements, and 7 near bottom current meter moorings deployed under the LC region. The dataset contains velocity data gathered from June 2009 to June 2011. Since the sampling frequency from multiple sensors varied from minutes to hours, we processed the dataset with a fourth order Butterworth filter and sub-sampled at 12-hour intervals, leading to a total of 1,810 records (905 days).

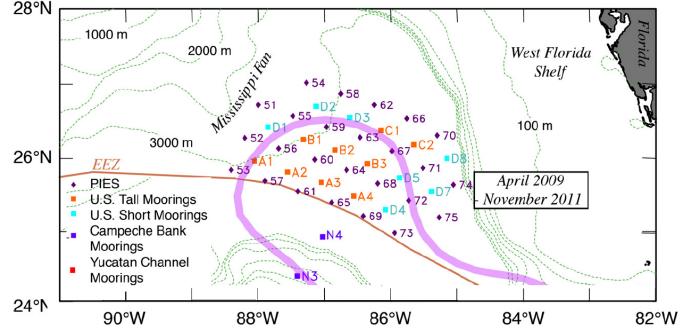


Fig. 6: Locations of moorings and Pressure-Recording Inverted Echo Sounders (PIES) deployed in the U.S. and Mexican sectors in the eastern Gulf of Mexico [35].

5.2 Baselines

- **FC-LSTM** [36] has been proven powerful for handling temporal correlation. The LSTM network here consists of 256 hidden units.
- **ConvLSTM** [11] has convolutional structures in LSTM cell to capture spatiotemporal correlations. The model consists of a 3-layer ConvLSTM network with 128-64-64 hidden states and 3×3 kernels.
- **PredRNN** [37] memorizes both spatial appearances and temporal variations in a unified memory pool, which consists of two ST-LSTM layers with 128 hidden states each and 3×3 convolution filters.
- **CDNN** [28] is a recently developed physics-informed network, which has a convolutional-deconvolutional module with a warping mechanism to produce an interpretable latent state – the motion field⁴ – advecting ocean dynamics.

5.3 Model Details and Implementation

Forecasting Network: The FN consists of a fully-connected layer, followed by an LSTM layer with 256 hidden units, and another fully-connected layer.

Transition Network: The lateral output dimension is set to 8, analogously to the dimension of lateral input.

Physics Network: Two types of PN are adopted: 1) *PDE-learning-net* is a fully-connected network of width 50 with four hidden layers with ReLU activation. 2) *ODE-informed-net* is of width 50 with integrated portion consisting of two fully-connected layers with ReLU activation.

All models are trained with ADAM optimizer with the sum of l_1 -norm and l_2 -norm loss and 0.01 learning rate. The batch size is set to 16. Learning rate decay and scheduled sampling are activated once the model does not improve in 20 epochs (in term of validation loss). The implementation was based on the Pytorch equipped with NVIDIA Geforce GTX 1080Ti and Titan Xp GPU with 32GB memory.

5.4 Physics Learning Analysis

We validate the hypothesis that the PDE-learning-net and ODE-informed-net are able to uncover the underlying hid-

4. The motion field equation $\frac{\partial I}{\partial t} + (w \nabla) I = D \nabla^2 I$ describes the transport of quantity I through advection and diffusion. This equation describes a large family of physical processes (e.g., fluid dynamics, heat conduction, etc.)

TABLE 2: MSE of recovered data by learnt physics.

Models	Single-step Modeling	
	Wave Simulation	GoM Loop Current
PDE-learning-net	$3.60(\pm 0.48) \times 10^{-3}$	$1.36(\pm 1.51) \times 10^{-4}$
ODE-informed-net	$6.39(\pm 1.35) \times 10^{-8}$	$2.86(\pm 1.66) \times 10^{-4}$

TABLE 3: MSE ($\times 10^{-3}$) of recovered reflected wave data.

Estimated PDE+noise%	\mathbf{c}^*	$\mathbf{c}^* + 5\%$	$\mathbf{c}^* + 10\%$
Single-step Modeling	$3.60(\pm 0.48)$	$3.65(\pm 0.48)$	$3.81(\pm 0.50)$

den physics from raw data, and thus, are able to assist the spatio-temporal networks to capture the dynamics of the natural phenomenon. Both models assessed in this section are trained against values at individual points, i.e., trained over a single sequence of synthetic reflected wave data.

The ODE-informed-net model efficacy is demonstrated via a time-series derived from a closed-loop evaluation initialized with a pair of time steps. Qualitative assessments show that the model is capable of accurately recreating the physical behavior of the ground truth data and multi-steps ahead in the future, as shown in Table 2 and Fig. 7. This suggests that for a simple physics example, an ODE approximation can accurately recreate data well with a relatively shallow network. However, computational overhead from finite difference operations required for the implementation of the network offsets some of this benefit.

Both the true and predicted PDEs are represented as a normalized vector of coefficients \mathbf{c} for the following dictionary of differential terms: $\{u_{tt}, u_{xx}, u_{yy}, u_t, u_x, u_y, u\}$. Here, the wave propagating with a speed of $c = 3.0$ with normalized coefficients $\mathbf{c} = [0.0783, -0.7049, -0.7049, 0, 0, 0, 0]$ is set as a ground truth. PDE-learning-net predicts a normalized vector of $\mathbf{c}^* = [-0.086, 0.7274, 0.677, -0.0309, 0.0637, -0.0098, 0.0119]$. An error between the two is calculated by metric⁵, producing an error of approximately 5.74×10^{-2} . According to Table 3, the PDE-learning-net is able to reveal the optimal pde expression from data. To further evaluate the accuracy of the estimated PDE, it is compared to the synthetic data set, as shown in Table 2, which is performed by solving the estimated PDE numerically to directly compare the solutions. Although the estimated PDE produces a similar result, a damping term was incorrectly introduced, resulting in fading in multi-steps modeling, as shown in Fig. 7. This PDE estimation is valuable over a traditional neural network as this method can scale over the space-time dimensions and has flexibility through choice of numerical solver.

5.5 Spatio-Temporal Modeling Analysis

We compare our model with several baselines evaluated with a mean square error (MSE) metric. The complete visualization result of the synthetic reflected-wave forecasting tasks is shown in Fig. 8, where the top-half shows the no-physics-informed approaches while the bottom-half

5. The PDE estimation error is calculated by $err(\mathbf{c}, \mathbf{c}^*) = (1 - |\mathbf{c} \cdot \mathbf{c}^*| / (\|\mathbf{c}\| \|\mathbf{c}^*\|))^{1/2}$, which is always non-negative, and is 0 iff. \mathbf{c} and \mathbf{c}^* are co-linear.

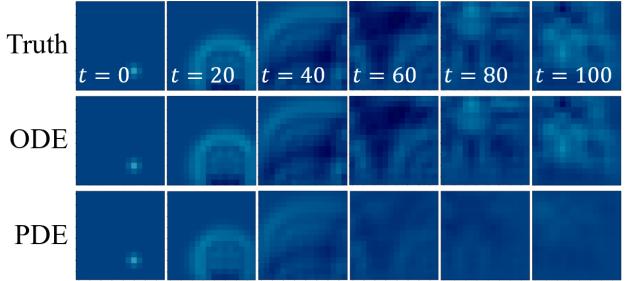


Fig. 7: Numerical solutions of the estimated PDE by PDE-learning-net and approximations by ODE-informed-net. Each column is the visualization at different time steps.

shows the physics-informed approaches. Table 4 (top-half) shows the quantitative comparison of different no-physics-informed approaches for the synthetic reflected-wave forecasting tasks. In most cases, our ST-PCNN consistently outperforms other baseline methods. Although ConvLSTM and FC-LSTM yield the best single-step forecast, they seem easier to over-fit and lead to earlier fading prediction compared to Pred-RNN and ST-PCNN. Considering that the closed-loop performance is more challenging than just single-step forecasting, which requires both intrinsic model stability and the maintenance of plausible ongoing dynamics, our ST-PCNN that distributively execute prediction in space shows the best generalization performance.

Table 4 (bottom-half) presents the performance of physics-informed approaches on reflected-wave data. ST-PCNN[▲] with wave equation (Eq. (22)) informed is regard as a reference here, proving that if the governing physics is known, our model could perfectly capture the spatio-temporal dynamics even in the long-run. The CDNN, that discretizing the solution of the motion field (advection-diffusion equation) with a warping scheme to inform forecasting, is not able to capture the complex long-term dynamics that strongly deviate from the truth after 40 steps closed-loop prediction (shown in Fig. 8). It indicates that pre-given knowledge (e.g., PDE of ‘general form’) may not always be beneficial in physical process modeling, since it neglects the ‘heterogeneity’. In the contrary, the MSE score of ST-PCNN, either informed by predicted ODE or PDE, is better than any of the baselines with/without physics inform. It indicates that ST-PCNN can capture both ‘homogeneity’ (underlying physics) and ‘heterogeneity’ (localized information) in modeling evolution and dynamics of natural phenomena.

5.6 GoM Loop Current Prediction

So far, the synthetic wave data only consider the strict regularly distributed grid, where distances between single measurement point are identical and enriched in physical property. Such situation may not suitable in many real-word applications. We adopt the LC data to validate ST-PCNN’s performance in handling irregularly and widely distributed sensor grid. As shown in Table 5 and Fig. 9, ST-PCNN outperforms other no-physics-informed approaches, not only reaches the lowest single-step forecast error but also yields the best multi-step forecast performance. Among physics-informed approaches, the DCNN achieves the lowest short-term forecast MSE while the predicted ODE-informed ST-

TABLE 4: The MSE \pm Std of closed-loop prediction of *Reflected Wave Simulation Data*.

Models	Physics	Magnitude	Multi-step Forecasting with #-steps of teacher-forcing				Single-step Forecasting
			10-steps tf	20-steps tf	30-steps tf	40-steps tf	
FC-LSTM	—	$\times 10^{-3}$	21.89 (± 8.83)	21.24 (± 8.09)	20.14 (± 5.39)	16.62 (± 6.57)	$5.01 (\pm 1.39) \times 10^{-5}$
ConvLSTM	—	$\times 10^{-3}$	10.07 (± 5.00)	10.02 (± 5.53)	8.42 (± 4.24)	7.33 (± 3.47)	$3.80 (\pm 2.16) \times 10^{-5}$
PredRNN	—	$\times 10^{-3}$	8.81 (± 0.78)	9.08 (± 0.92)	8.88 (± 0.58)	8.04 (± 1.08)	$3.09 (\pm 1.98) \times 10^{-4}$
ST-PCNN	—	$\times 10^{-3}$	10.74 (± 0.48)	8.04 (± 0.37)	6.27 (± 0.21)	5.38 (± 0.15)	$2.25 (\pm 0.13) \times 10^{-4}$
ST-PCNN [▲]	Wave Equ	$\times 10^{-6}$	6.41(± 0.11)	5.91(± 0.06)	5.09(± 0.18)	3.40(± 0.21)	$1.76(\pm 0.003) \times 10^{-10}$
CDNN	Motion field	$\times 10^{-3}$	7.21(± 0.64)	7.19(± 1.06)	7.35(± 1.07)	6.39(± 0.06)	$5.59(\pm 0.74) \times 10^{-4}$
ST-PCNN	Predicted ODE	$\times 10^{-4}$	3.43(± 1.11)	4.00(± 1.22)	4.56(± 1.35)	5.58(± 1.66)	$2.27(\pm 0.88) \times 10^{-6}$
ST-PCNN	Predicted PDE	$\times 10^{-4}$	3.01 (± 0.37)	3.25 (± 0.34)	3.46 (± 0.39)	3.88 (± 0.33)	9.11 (± 0.27) $\times 10^{-8}$

* All the models are trained (validated) on 57 (7) sequences of 40 steps and tested on 16 new sequences of 80 steps.

TABLE 5: The MSE \pm Std of closed-loop prediction of *Gulf of Mexico Loop Current Observation*.

Models	# params	Physics	Multi-step Forecasting Horizon (Magnitude: $\times 10^{-2}$)			Single-step Forecasting
			5-steps	10-steps	15-steps	
FC-LSTM	# 3.81m	—	4.46(± 2.38)	5.39(± 3.16)	6.49(± 3.63)	$4.81 (\pm 2.73) \times 10^{-2}$
ConvLSTM	# 1.33m	—	1.15(± 0.60)	3.11(± 1.67)	6.65(± 2.34)	$4.49 (\pm 1.40) \times 10^{-4}$
PredRNN	# 6.41m	—	1.76(± 0.76)	4.68(± 1.97)	6.77(± 2.71)	$8.39 (\pm 6.21) \times 10^{-4}$
ST-PCNN	# 0.53m	—	0.61 (± 0.20)	1.84 (± 0.52)	4.68 (± 1.02)	3.35 (± 0.03) $\times 10^{-4}$
CDNN	# 6.46m	Motion field	0.14 (± 0.02)	1.38(± 0.39)	4.32(± 1.34)	$1.21 (\pm 0.32) \times 10^{-3}$
ST-PCNN	# 0.53m	Predicted ODE	0.38(± 0.05)	1.33 (± 0.43)	3.32 (± 1.81)	1.91 (± 0.81) $\times 10^{-4}$
ST-PCNN	# 0.53m	Predicted PDE	0.40(± 0.08)	1.49(± 0.49)	3.35(± 0.81)	2.29(± 0.73) $\times 10^{-4}$

* All the models are trained (validated) on 1,466 (163) steps and tested on the final 181 steps.

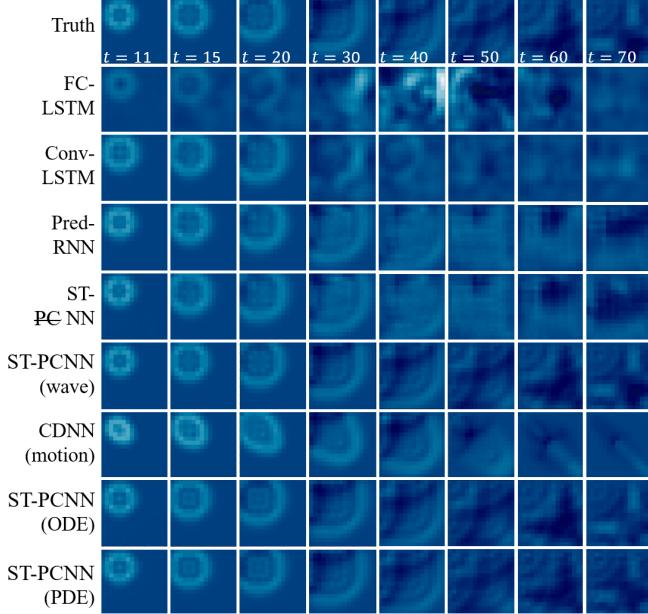


Fig. 8: Visualization of reflected wave closed-loop forecasting with 10-steps teacher-forcing. Each column is the prediction at different time steps (i.e., $t = 11, t = 15, \dots, t = 70$) by baselines and proposed ST-PCNN model.

PCNN reaches the best single-step and lasting forecast performance.

6 CONCLUSIONS

In this paper, we proposed a ST-PCNN model for accurate spatio-temporal forecasting. We argued that real-world

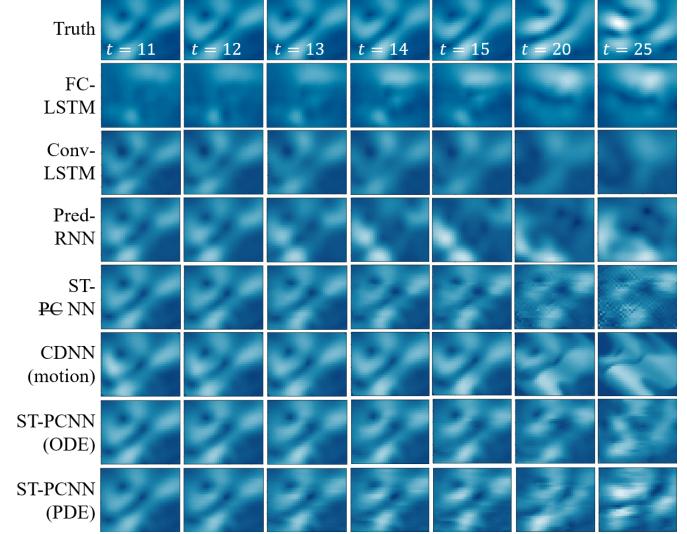


Fig. 9: Visualization of GoM loop current multi-step closed-loop forecast with 10-steps teacher-forcing.

dynamical systems are often challenged by heterogeneity and homogeneity. By coupling three networks to learn underlying physics, enable transition of node interaction, and forecast the future, ST-PCNN shows superior performance to baseline models. The key contribution of the paper, compared to existing research in the field, is three-fold: 1) A novel ST-PCNN framework capturing complex localized spatial-temporal correlations; 2) A 2D PDE-learning-net and ODE-informed-net as the physics nets to learn hidden physics; and 3) The physics-coupled neural network

(PCNN) for long-term forecasting using only limited observations. Further, ST-PCNN is a general framework suitable for many other physical processes modeling.

REFERENCES

- [1] Y. Li, W. Zheng, L. Wang, Y. Zong, and Z. Cui, "From regional to global brain: A novel hierarchical spatial-temporal neural network model for eeg emotion recognition," *IEEE Transactions on Affective Computing*, 2019.
- [2] D. Seebacher, J. Häuäler, M. Hundt, M. Stein, H. Müller, U. Engelke, and D. Keim, "Visual analysis of spatio-temporal event predictions: Investigating the spread dynamics of invasive species," *IEEE Transactions on Big Data*, 2018.
- [3] Y. Li, Z. Zhu, D. Kong, M. Xu, and Y. Zhao, "Learning heterogeneous spatial-temporal representation for bike-sharing demand prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 1004–1011.
- [4] J. Bian, H. Xiong, Y. Fu, and S. K. Das, "Cswa: Aggregation-free spatial-temporal community sensing," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [5] H. Nguyen, W. Liu, and F. Chen, "Discovering congestion propagation patterns in spatio-temporal traffic data," *IEEE Transactions on Big Data*, vol. 3, no. 2, pp. 169–180, 2016.
- [6] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," *arXiv preprint arXiv:1709.04875*, 2017.
- [7] S. Seo and Y. Liu, "Differentiable physics-informed graph networks," *arXiv preprint arXiv:1902.02950*, 2019.
- [8] J. Y. Zhu, C. Sun, and V. O. Li, "An extended spatio-temporal granger causality model for air quality estimation with heterogeneous urban big data," *IEEE Transactions on Big Data*, vol. 3, no. 3, pp. 307–319, 2017.
- [9] V. La Gatta, V. Moscato, M. Postiglione, and G. Sperli, "An epidemiological neural network exploiting dynamic graph structured data applied to the covid-19 outbreak," *IEEE Transactions on Big Data*, 2020.
- [10] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results," *Journal of transportation engineering*, vol. 129, no. 6, pp. 664–672, 2003.
- [11] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.
- [12] S. Guo, Y. Lin, S. Li, Z. Chen, and H. Wan, "Deep spatial-temporal 3d convolutional neural networks for traffic data forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3913–3926, 2019.
- [13] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," *arXiv preprint arXiv:1707.01926*, 2017.
- [14] C. Song, Y. Lin, S. Guo, and H. Wan, "Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 914–921.
- [15] L. Bai, L. Yao, S. Kanhere, X. Wang, Q. Sheng *et al.*, "Stg2seq: Spatial-temporal graph to sequence model for multi-step passenger demand forecasting," *arXiv preprint arXiv:1905.10069*, 2019.
- [16] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Data-driven discovery of partial differential equations," *Science Advances*, vol. 3, no. 4, p. e1602614, 2017.
- [17] J. Berg and K. Nyström, "Data-driven discovery of pdes in complex datasets," *Journal of Computational Physics*, vol. 384, pp. 239–252, 2019.
- [18] H. Xu, H. Chang, and D. Zhang, "Dl-pde: Deep-learning based data-driven discovery of partial differential equations from discrete and noisy data," *arXiv preprint arXiv:1908.04463*, 2019.
- [19] Z. Long, Y. Lu, X. Ma, and B. Dong, "Pde-net: Learning pdes from data," in *International Conference on Machine Learning*, 2018, pp. 3208–3216.
- [20] P. Hu, W. Yang, Y. Zhu, and L. Hong, "Revealing hidden dynamics from time-series data by odenet," *arXiv preprint arXiv:2005.04849*, 2020.
- [21] C. Zhang and P. Patras, "Long-term mobile traffic forecasting using deep spatio-temporal neural networks," in *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2018, pp. 231–240.
- [22] K. Wang, C. Xu, Y. Zhang, S. Guo, and A. Y. Zomaya, "Robust big data analytics for electricity price forecasting in the smart grid," *IEEE Transactions on Big Data*, vol. 5, no. 1, pp. 34–45, 2017.
- [23] C. K. Wikle, "Modern perspectives on statistics for spatio-temporal data," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 7, no. 1, pp. 86–98, 2015.
- [24] N. Cressie and C. K. Wikle, *Statistics for spatio-temporal data*. John Wiley & Sons, 2015.
- [25] L. Ladický, S. Jeong, B. Solenthaler, M. Pollefeys, and M. Gross, "Data-driven fluid simulations using regression forests," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 6, pp. 1–9, 2015.
- [26] J. Tompson, K. Schlachter, P. Sprechmann, and K. Perlin, "Accelerating eulerian fluid simulation with convolutional networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3424–3433.
- [27] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations," *arXiv preprint arXiv:1711.10561*, 2017.
- [28] E. de Bezenac, A. Pajot, and P. Gallinari, "Deep learning for physical processes: Incorporating prior scientific knowledge," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2019, no. 12, p. 124009, 2019.
- [29] M. Lutter, C. Ritter, and J. Peters, "Deep lagrangian networks: Using physics as model prior for deep learning," *arXiv preprint arXiv:1907.04490*, 2019.
- [30] Y. D. Zhong, B. Dey, and A. Chakraborty, "Symplectic odenet: Learning hamiltonian dynamics with control," *arXiv preprint arXiv:1909.12077*, 2019.
- [31] A. Hasan, J. M. Pereira, R. Ravier, S. Farsiu, and V. Tarokh, "Learning partial differential equations from data using neural networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3962–3966.
- [32] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [34] G. Wanner and E. Hairer, *Solving ordinary differential equations II*. Springer Berlin Heidelberg, 1996.
- [35] P. Hamilton, A. Lugo-Fernández, and J. Sheinbaum, "A loop current experiment: Field and remote measurements," *Dynamics of Atmospheres and Oceans*, vol. 76, pp. 156–173, 2016.
- [36] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," 1999.
- [37] Y. Wang, M. Long, J. Wang, Z. Gao, and S. Y. Philip, "Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms," in *Advances in Neural Information Processing Systems*, 2017, pp. 879–888.



Yu Huang received the B.S. degree and M.S. degree in Aeronautics and Astronautics Engineering from Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2015 and 2018, respectively.

He is currently working toward the Ph.D. degree in Electrical Engineering at Florida Atlantic University, Boca Raton, FL, USA. His research interests include prognostics and health management, machine learning and its applications in energy systems and oceanography.



James Li received the B.S. degree and M.S. degree in Aerospace Engineering from the Georgia Institute of Technology, Atlanta, Georgia, in 2018 and 2019, respectively.

He gained experiences with a demonstrated history of working in both academia and industry, such as GE Aviation and Lockheed Martin. His research interests include fluid dynamics, data science/machine learning, and combustion.



Laurent Chérubin received the B.S. degree in Fluid Mechanics from the University of Bordeaux I, France, and the M.S. and Ph.D. degrees in Physical and Coastal Oceanography from the University of Méditerranée, Marseille, France. He is currently an Associate Research Professor at HarborBranch Oceanographic Institute, Florida Atlantic University.

Dr. Chérubin is a physical oceanographer specialized in the understanding of ocean dynamics, which is the study of why the water moves

the way it moves. His research has focused on dynamics of motions associated with instabilities in coastal currents and eddies, using both analytical and numerical models in the quasi-geostrophic, shallow-water formalisms, and in realistic models. This research provides a deep understanding of the environmental forces that affect ocean ecosystems at multiple levels of the trophic chain. Both observational analysis and numerical modeling involving hydrodynamic (the Regional Oceanic Modeling System - ROMS) and biophysical models (Connectivity Modeling System - CMS and Ichthyop) are used to study how environmental drivers shape the oceanic ecosystems.



Min Shi (S'15) received the M.S. degree from the School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, China, and the Ph.D. degree from the Department of Computer & Electrical Engineering and Computer Science, Florida Atlantic University, FL, USA, in 2020.

He is currently a postdoctoral researcher at Washington University School of Medicine in St. Louis, Missouri. His research interests include data mining, machine learning, social networks, and service computing.



Hanqi Zhuang (SM'10) received the B.S. degree in Electrical Engineering from Shanghai University, Shanghai, China, and the M.S. and Ph.D. degrees in Electrical Engineering from Florida Atlantic University, Boca Raton, FL, USA. He is the Chair of the Computer and Electrical Engineering and Computer Science Department, Florida Atlantic University.

His research interests include acoustic signal processing, deep neural networks, and their applications. He is currently working on research projects ranging from marine animal detection and classification from acoustic signals to modeling and prediction of the Loop Current System in the Gulf of Mexico.



James VanZwieten received the B.S., M.S., and Ph.D. degrees in Ocean Engineering from Florida Atlantic University, Boca Raton, in 2001, 2003, and 2007, respectively. He is currently an Associate Research Professor with the Department of Civil, Environmental, and Geomatics Engineering at Florida Atlantic University. His research interests include modeling and control of marine vehicles, in-stream hydrokinetic energy production, ocean thermal energy conversion, and sea water air conditioning. Previous experience includes working as an Assistant Research Professor at the Southeast National Marine Renewable Energy Center operated by Florida Atlantic University.

Dr. VanZwieten is a member of the American Society of Civil Engineers (ASCE) Marine Renewable Energy Committee and Chair of its In-stream Hydrokinetic Subcommittee.



Xingquan Zhu (SM'12) received the Ph.D. degree in Computer Science from Fudan University, Shanghai, China. He is a Full Professor with the Department of Computer and Electrical Engineering & Computer Science, Florida Atlantic University, Boca Raton, FL, USA. His research interests include data mining, machine learning, multimedia computing, and bioinformatics. Since 2000, he has authored or co-authored over 260 refereed journal and conference papers in these areas, including three Best Paper Awards and one Best Student Paper Award. Dr. Zhu is an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING from 2008 to 2012, and from 2014 to date. Since 2017, he has been an Associate Editor of the ACM TRANSACTIONS ON KNOWLEDGE DISCOVERY FROM DATA.



Yufei Tang (M'16) received the Ph.D. degree in Electrical Engineering from the University of Rhode Island, Kingston, RI, USA, in 2016. He is currently an Assistant Professor with the Department of Computer and Electrical Engineering & Computer Science at Florida Atlantic University, Boca Raton, FL, USA. His research interests include machine learning, big data analytics, and sustainability for energy and environment.

Dr. Tang is an Early-Career Research Fellow of the National Academies Gulf Research Program (2019). He has also received several other awards, including the Steve Bouley and Rhonda Wilson Graduate Fellowship Award (2016), the Chinese Government Award for Outstanding Student Abroad (2016), the IEEE PESGM Graduate Student Poster Contest, Second Prize (2015), and the IEEE International Conference on Communications (ICC) Best Paper Award (2014).