Scheduled Temporal Loss Weighting for Neural Operators

Oluwaseun E. Coker[1 [0009−0000−2016−3622]], He Wang[1,2 [0000−0002−2281−5679]], Amirul Khan[1 [0000−0002−7521−5458]], and Peter K. Jimack[1 [0000−0001−9463−7595]]

[1] University of Leeds, Leeds, LS2 9JT, UK.
[2] University College London, London, WC1E 6BT, UK.

Abstract. Neural operators offer promise for efficient solutions to time-dependent partial differential equations, but face challenges in long-term prediction due to complex dynamics, gradient accumulation, and error propagation. To address these limitations, we propose a novel curriculum learning strategy, temporal weighted loss. This method mitigates overfitting to early dynamics by dynamically adjusting the weights applied to the loss across the temporal sequence, prioritising initial time steps during early training. This approach enhances model generalisation and prediction accuracy for extended time horizons, demonstrating improved performance compared to baseline curriculum learning techniques.

Keywords: PDEs · Curriculum Learning · Neural Operator.

1 Introduction

Neural operators offer a promising, computationally efficient alternative to traditional numerical methods for solving time-dependent partial differential equations (PDEs), particularly in fluid flow modelling [7]. By learning discretisation-independent mappings between functional spaces, they enable fast inference at varying resolutions, providing significant speed-ups for tasks like design optimisation [14]. Deep autoregressive neural operators have become a popular choice for modelling PDE time evolution. These **constant-timestep models** predict the next state from the previous one at a fixed interval. Unlike recurrent neural networks (RNNs), they lack internal hidden states, bypassing vanishing gradient problems and making them easier to train [17]. They are also less susceptible to **exposure bias**—the compounding error caused by discrepancies between training ground truths and testing predictions.

Despite these advantages, long-term autoregressive prediction remains challenging due to multiscale complexity, involving sophisticated spatial and temporal scales and unsteady behaviours; gradient accumulation, where "long-rollout" training can trigger numerical instability and out-of-memory (OOM) errors [12]; and exponential error propagation, where small per-step inaccuracies can compound, causing predicted trajectories to diverge from the ground truth [4, 15]. To mitigate these model-agnostic issues, research has focused on improving training robustness through noise injection or refinement processes [10, 20]. We focus on

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_15

curriculum learning, which prioritises "easy" samples before introducing complexity [24]. In RNNs, this often involves transitioning from **teacher forcing** to **autoregressive rollout**. However, this paradigm remains under-investigated for neural operators. Our investigation into existing curricula for operators such as FNO [9] and U-NO [19] suggests that they can overfit to early dynamics, resulting in poor performance on later dynamics.

We propose **Scheduled Temporal Loss Weighting (STLW)**, a curriculum-based approach that performs only autoregressive rollout training, avoiding sub-optimal teacher forcing. Unlike methods that neglect long-term gradients to avoid early-stage instability, STLW prioritises short-term accumulation without ignoring long-term gradients. We achieve this by assigning a weight to the loss at each timestep, governed by a transition function that dynamically rebalances them as training progresses. STLW initially assigns lower weights to later timesteps, reducing their loss contribution while maintaining their influence on gradient accumulation. Weights increase via a smooth schedule, converging to a uniform weighting. This facilitates a stable optimisation landscape, yielding higher accuracy and superior generalisation. Our key contributions include the **consolidation** of curriculum learning strategies for autoregressive neural operators, the **development** of STLW to bridge the gap between easy one-rollout and hard full-rollout training, the **benchmarking** of STLW against existing strategies across multiple PDE datasets, and the demonstration of the superior accuracy and robustness provided by the STLW framework.

2 Problem Statement

We formulate the learning problem by considering explicit time-stepping for a general class of time-dependent PDEs over the domain $(t, x) \in (0, T] \times D$. The governing equations, subject to boundary conditions B, are defined as:

$$\begin{aligned} \partial_t u(t,x) &= \mathcal{F}^\dagger\left(u(t,x)\right) \qquad (t,x) \in [0,T] \times D, \\ u(0,x) &= u_0(x) \qquad x \in D, \\ u(t,x) &= B \qquad (t,x) \in (0,T] \times \partial D, \end{aligned} \tag{1}$$

where $u : [0, T] \times D \to \mathbb{R}^d$ is the solution function, $u_0(x)$ is the initial condition at $t = 0$ and , and $\mathcal{F}^\dagger$ denotes a spatial differential operator. Our objective is to learn a forward-explicit autoregressive neural operator, $\mathcal{F}_\theta$ with parameters $\theta \in \mathbb{R}^p$, that maps an input function $u(t)$ to the predicted solution at the subsequent timestep:

$$\tilde{u}\left(t + \Delta t\right) = \mathcal{F}_\theta\left(u(t)\right) \quad \theta \in \mathbb{R}^p. \tag{2}$$

The learning problem seeks to identify the optimal parameters θ^* by minimising the discrepancy between the ground truth u and the prediction $\tilde{u}$. Training is performed using a finite set of time-evolution trajectories generated by a high-accuracy numerical scheme. Each trajectory is represented as an evenly spaced temporal sequence with a uniform spatial resolution. To assess performance, we use an empirical train-test split and evaluate the model on unseen test samples.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_15

During inference, the trained model $\mathcal{F}_{\theta^*}$ is applied recursively to generate future solutions at a constant timestep Δt. We can therefore reformulate the model from Equation 2 to define a sequence of T autoregressive rollouts, yielding the solutions $\{\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_T\}$:

$$\tilde{u}_t = \mathcal{F}_{\theta^*}(\tilde{u}_{t-1}), \quad 1 \leq t \leq T. \tag{3}$$

3 Related Work

Recent efforts to enhance PDE solver stability and accuracy are broadly categorised into architectural modifications, training paradigms, and inductive biases [25, 3, 18]. Strategies include integrating modules like PINNs for physics-based loss [25], using RNNs for temporal dynamics [16], and employing denoising or adversarial training to combat error accumulation [11, 20]. Other methods utilise temporal bundling to reduce forward passes [3], latent evolution [26], or inductive biases like attention mechanisms and U-Net architectures to resolve multi-scale dynamics [18, 8]. Our work aligns with the second category, focusing on enhancing training paradigms without altering underlying models.

Operator learning models, specifically the Fourier Neural Operator (FNO) [9] and U-shaped Neural Operator (U-NO) [19], have succeeded in solving various PDEs due to their resolution-independent nature [13, 5]. However, research into their effectiveness for autoregressive time-stepping is limited, as standard architectures often struggle with long-term dynamics in complex PDEs. This study utilizes state-of-the-art neural operators to demonstrate a novel, robust training strategy.

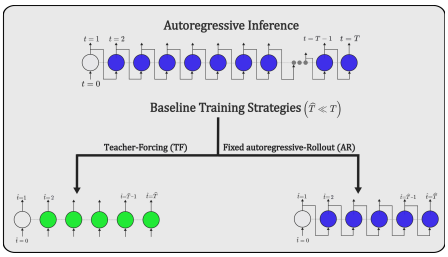
Curriculum Learning (CL) improves convergence by transitioning from "easy" to "hard" samples [1], defined by a curriculum metric and a scheduler. In PDE applications, this typically involves transitioning between teacher forcing and autoregressive rollout [23, 24, 21]. However, teacher forcing can lead to exposure bias and suboptimal long-term performance [6]. Consequently, we avoid teacher forcing entirely, formulating our curriculum approach exclusively within the autoregressive rollout regime to ensure superior stability.

4 Baseline and Curriculum Training Strategies

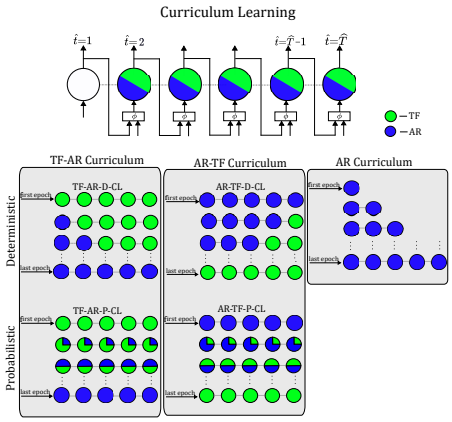
Training neural operators for long-term PDE prediction is computationally demanding due to the risks of *exposure bias* and memory constraints during deep gradient accumulation. Consequently, models often train on shorter subsequences of length $\widehat{T} \ll T$. Standard static strategies include **Teacher Forcing (TF)**, a one-step rollout ($\widehat{T} = 1$) using ground-truth inputs that is efficient but prone to divergence during inference, and **Fixed-length Autoregressive Rollout (AR)**, a multi-step approach ($1 < \widehat{T} \ll T$) that better aligns training with inference (Figure 1a). To improve robustness, Gaussian noise is often added to the inputs in the TF strategy (TF+N).

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_15

Beyond these baselines, curriculum learning (CL) improves convergence by transitioning from "easy" to "hard" tasks (Figure 1b). **Deterministic Curricula (D-CL)** follow fixed schedules; for example, **TF–AR-D-CL** transitions from teacher forcing to autoregressive rollouts, while **AR-CL** incrementally increases the rollout length $\widehat{T}$ without employing teacher forcing. Alternatively, **Probabilistic Curricula (P-CL)** use a probability α_e to determine the training mode for each rollout. In **TF–AR-P-CL**, the model initially samples rollouts as TF, with the probability of sampling an AR rollout increasing throughout training to facilitate a stable transition to complex dynamics.



(a) Baseline training strategies of $\widehat{T}$ timestep: Teacher forcing in green (one-rollout) and autoregressive rollout in blue (multi-rollout) where T represents the total timesteps of the data and $\widehat{T}$ represent the training batch timesteps.



(b) Combinations of Teacher Forcing (TF) and Autoregressive Rollout (AR) across training epochs

Fig. 1: Baseline and curriculum learning strategies. Autoregressive rollout (AR) and Teacher-Forcing (TF) at each timestep are represented as blue and green, respectively.

5 Scheduled Temporal Loss Weighting (STLW)

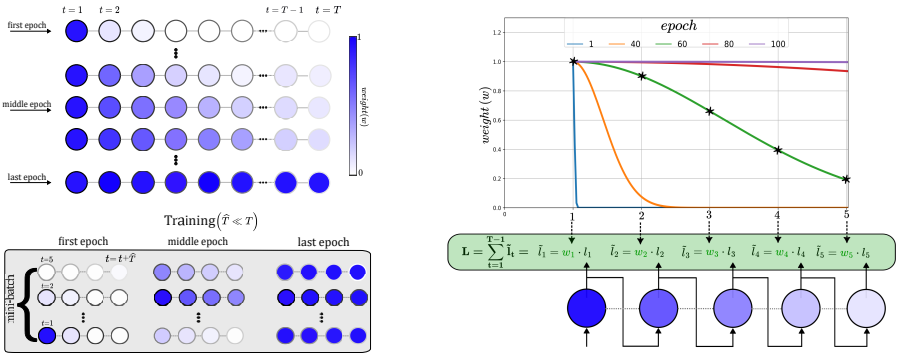
We propose Scheduled Temporal Loss Weighting (STLW), a curriculum learning strategy that treats sequence length as a continuum of difficulty. By assigning dynamic weights to each timestep, STLW enables the model to attend to the entire training batch simultaneously while prioritising "easy" early timesteps to reduce error propagation (Figure 2). For a rollout of length $\widehat{T}$ at epoch e, the total weighted loss $\widehat{L}_t$ is:

$$\widehat{L}_t = \sum_{\hat{t}=1}^{\widehat{T}} w_{e,\,t+\hat{t}} \cdot l(u_{t+\hat{t}},\, \tilde{u}_{t+\hat{t}}), \quad w_{e,t} = \exp\left(-C_e \cdot t^2\right). \tag{4}$$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_15

The decay parameter C_e is defined by $C_e(e) = \exp(-A \cdot e \cdot \frac{T}{E})$, where E is the total epochs and $A = \exp(5.0 \cdot T^{-0.723})$ is an empirical constant obtained by fitting a exponential curve over varying T. This Gaussian-style formulation as seen in Figure 2b ensures a smooth transition where later timesteps begin with low weights and gradually converge to unity ($w_t \approx 1$) by the end of training.

The curriculum metric $\bar{w}_e = (1 + \sum_{t=2}^{T} w_{e,t})/T$ defines the task difficulty, transitioning from $\bar{w}_e \approx 1/T$ to full weighting ($\bar{w}_e = 1$). Unlike discrete curricula that increment the rollout length $\widehat{T}$, STLW includes all timesteps from the onset. This prevents the model from becoming "blind" to later dynamics, mitigating the risk of overfitting to early-stage temporal behaviours while maintaining numerical stability through gradient damping. STLW functions as a continuous generalisation of the discrete AR-Curriculum (see Figure 3). While fixed AR training ($w_{e,t} = 1$) is prone to instability and discrete AR-CL ($\widehat{T}_e < T$) may overfit early dynamics, STLW balances these by applying $w_{e,t} \in (0, 1]$ across the full sequence. This approach ensures the model maintains a global view of the trajectory throughout the optimisation process, yielding superior generalisation across complex PDE benchmarks.



(a) STLW training Strategy (b) Weighting function

Fig. 2: A schematic of the scheduled temporal loss Weighting (STLW) Curriculum. (a) The colour strength ranges from 0 (low) to 1 (high), indicating the weight assigned to each rollout during training. Later rollouts in the early training phase have low strength, whereas these weights gradually increase in magnitude. The weights are applied to the loss of its respective timestep and epoch. (b) For a dataset with a subsequence of length T ($T = 6$), a weight w is assigned to each timestep for $t > 0$, which controls the loss l contributed by each timestep (for example, at epoch = 60, the weight for each timestep is highlighted in black markers). The weighting function (curves), which determines the weight values, changes during training, and at the end of training (epoch = 100), the weighting function sets all weights to 1.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_15

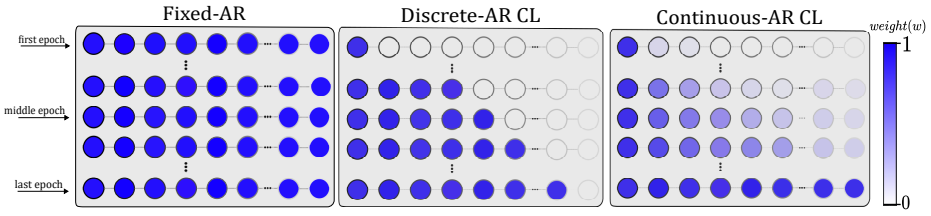


Fig. 3: Schematic comparison between the discrete AR-CL and the Scheduled Temporal Weight Loss (STWL) curriculum (right). STWL acts as a continuous variant, with a weight applied at every timestep, represented by the colour intensity of the bars.

6 STLW Gradient Propagation and Multi-dimensional Scheduling Interpretation

Under the STLW framework, the total weighted loss for a rollout of length T is $\widehat{L} = \sum_{t=1}^{T} w_{e,t} \cdot l(u_t, \tilde{u}_t) = \sum_{t=1}^{T} \hat{l}_t$. The gradient of this weighted loss at each timestep t can be derived by modifying the standard backpropagation formulation [12]:

$$\frac{\partial \hat{l}_t}{\partial \theta} = w_{e,t} \left(\sum_{tt=1}^{t} \left[\frac{\partial l_t}{\partial \tilde{u}_t} \left(\prod_{ttt=1}^{t-tt} \frac{\partial \tilde{u}_{t-ttt+1}}{\partial \tilde{u}_{t-ttt}} \right) \frac{\partial \tilde{u}_{tt}}{\partial \theta} \right] \right) = w_{e,t} \frac{\partial l_t}{\partial \theta}. \tag{5}$$

This reveals that the scheduled weights are directly backpropagated, meaning STLW acts as a scheduled gradient-weighting mechanism. In early training, earlier timesteps receive minimal damping while later steps are heavily damped, mitigating the impact of exploding gradients while the model learns fundamental dynamics.

This mechanism allows for a spatio-temporal interpretation of learning rate scheduling. By substituting the weighted gradient into the standard update rule $\theta_{e+1} = \theta_e - \eta_e \nabla_\theta l$, we obtain an effective learning rate $\widehat{\eta}_{e,t} = \eta_e \cdot w_{e,t}$. This formulation effectively adds a temporal dimension to the standard scheduler, enabling adaptive step sizes across different timesteps t. Despite this mathematical equivalence, we categorise STLW as curriculum learning because the weights w increase over time to introduce complexity, whereas standard learning rates typically decrease to facilitate late-stage convergence.

7 Experiments

To validate the **AR-STLW-CL** strategy, we evaluate it against baseline and established curriculum learning (CL) methods across four diverse PDE datasets. Our analysis addresses several research questions, including baseline stability (TF, TF+N, AR), STLW performance relative to existing CL, the impact of training rollout length $\widehat{T}$, generalisation to extended inference horizons, and sensitivity to both random initialisations and transition function parameterisation.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_15

Evaluated strategies are categorised into three groups: **Baseline (Static)** methods, comprising Teacher Forcing (TF), TF with Gaussian noise (TF+N), and Fixed Autoregressive (AR) rollouts; **Deterministic Curricula (D-CL)**, which utilise fixed schedules like TF–AR-D-CL (transitioning from TF to AR) and AR-CL (incrementally increasing $\widehat{T}$); and **Probabilistic Curricula (P-CL)**, such as TF–AR-P-CL, which use an evolving probability α_e to shift training modes over time.

Experiments are conducted on four 1D time-dependent PDEs: **Advection (Av)**, to assess stability in linear transport [22]; two regimes of **Burgers' Equations**, including **Viscous (vB)** and **Inviscid with forcing (iB)** to test transitions between smooth and sharp features [22, 3]; and the **Korteweg–De Vries (KdV)** equation, a dispersive, non-dissipative model that poses a significant challenge over long sequences of 640 timesteps [2].

8 Results

Table 1 presents the mean normalised Root Mean Squared Error (nRMSE) across four PDE datasets, evaluated over nine distinct training strategies. Each row reflects the curriculum parameter configuration that yielded the lowest nRMSE during our exploratory investigation. For each dataset, the best-performing strategy is highlighted in **bold**, and the second-best is underlined.

8.1 Baseline Performance Analysis

We first evaluate the static baseline strategies: TF, TF+N, and AR. Our results indicate that the AR strategy significantly outperforms TF in three out of the four datasets (Av, vB, and KdV). In contrast, TF outperforms AR on the iB dataset. To provide a rigorous comparison, we use the best-performing baseline for each dataset as the reference point to calculate the relative improvement (*rel. impro.*) of the curriculum strategies.

8.2 Comparative Effectiveness of Curriculum Learning

The empirical results demonstrate that curriculum learning strategies consistently enhance model performance. Our proposed **AR-STLW-CL** method achieved the lowest error (and consequently the greatest relative improvement) on three of the four datasets: vB, iB, and KdV, with improvements of **62.5%**, **12.0%**, and **40.3%**, respectively. The AR-CL strategy ranked second in three cases (Av, iB, and KdV), achieving relative improvements of 16.8%, 11.4%, and 36.9%, respectively.

Among the deterministic strategies, TF–AR-D-CL and TF+N–AR-D-CL showed strong performance on the viscous Burgers' (vB) dataset, ranking as the joint best and second-best strategies. Conversely, probabilistic approaches (AR–TF-P-CL and TF–AR-P-CL) generally underperformed. Notably, AR–TF-P-CL (decreasing complexity) often led to performance degradation, failing to

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_15

surpass the best baseline strategies in several instances. This confirms that increasing the task difficulty over time—rather than decreasing it—is critical for training stability in autoregressive neural operators.

	Strategy	Best Curriculum	nRMSE		
		$\alpha_{start} \to \alpha_{end}$	$mean^{\pm s.d}$ ↓	rel. impro. ↑	last 10% ↓
Av	TF	0.0	$0.0375^{\pm 1.76e-2}$	–	0.0883
	TF+N	0.0	$0.0165^{\pm 5.68e-3}$	–	0.0446
	AR	1.0	$0.00353^{\pm 8.70e-5}$	–	0.00536
	AR-CL	1.0	$\underline{0.00293}^{\pm \underline{8.98e-5}}$	16.8%	<u>0.00450</u>
	TF–AR-D-CL	0.0 ↗ 1.0	$\mathbf{0.00272}^{\pm 9.58e-5}$	**22.6%**	**0.00423**
	TF+N–AR-D-CL	0.0 ↗ 1.0	$0.00296^{\pm 1.43e-4}$	15.9%	0.00455
	AR–TF-P-CL	1.0 ↘ 0.0	$0.00538^{\pm 1.18e-3}$	−52.4%	0.01460
	TF–AR-P-CL	0.0 ↗ 1.0	$0.00294^{\pm 9.48e-5}$	16.5%	0.00473
	AR-STLW-CL	1.0, $A_c = 5.0$	$0.00303^{\pm \mathbf{0.0}}$	14.1%	0.00485
vB	TF	0.0	$0.0632^{\pm 4.23e-2}$	–	0.1040
	TF+N	0.0	$0.0125^{\pm 2.39e-2}$	–	0.0154
	AR	1.0	$0.00749^{\pm 1.15e-3}$	–	0.00771
	AR-CL	1.0	$0.00376^{\pm \mathbf{1.30e-4}}$	49.7%	0.00623
	TF–AR-D-CL	0 ↗ 1	$\mathbf{0.00280}^{\pm 4.83e-4}$	**62.5%**	**0.00363**
	TF+N–AR-D-CL	0.0 ↗ 1.0	$\underline{0.00291}^{\pm 3.59e-4}$	61.1%	<u>0.00422</u>
	AR–TF-P-CL	1.0 ↘ 0.0	$0.0112^{\pm 8.29e-4}$	−44.9%	0.01427
	TF–AR-P-CL	0.0 ↗ 1.0	$0.00413^{\pm \underline{2.13e-4}}$	56.3%	0.00501
	AR-STLW-CL	1.0, $A_c = 5.0$	$\mathbf{0.00280}^{\pm 4.66e-4}$	**62.5%**	0.00566
iB	TF	0.0	$0.274^{\pm 3.9e-3}$	–	0.625
	TF+N	0.0	$0.278^{\pm 8.1e-3}$	–	0.628
	AR	1.0	$0.306^{\pm 8.7e-3}$	–	0.683
	AR-CL	1.0	$\underline{0.243}^{\pm 3.4e-3}$	11.4%	<u>0.559</u>
	TF–AR-D-CL	0.0 ↗ 1.0	$0.252^{\pm 4.4e-3}$	8.2%	0.571
	TF+N–AR-D-CL	0.0 ↗ 1.0	$0.250^{\pm 3.1e-3}$	8.7%	0.573
	AR–TF-P-CL	1.0 ↘ 0.0	$0.291^{\pm 1.4e-2}$	−6.1%	0.660
	TF–AR-P-CL	0.0 ↗ 1.0	$0.266^{\pm \mathbf{1.7e-3}}$	3.0%	0.603
	AR-STLW-CL	1.0, $A_c = 5.0$	$\mathbf{0.241}^{\pm \underline{2.8e-3}}$	**12.0%**	**0.558**
KdV	TF	0.0	$0.541^{\pm 9.50e-2}$	–	0.896
	TF+N	0.0	$0.347^{\pm 1.01e-1}$	–	0.623
	AR	1.0	$0.189^{\pm 5.29e-3}$	–	0.311
	AR-CL	1.0	$\underline{0.119}^{\pm 5.24e-3}$	36.9%	<u>0.226</u>
	TF–AR-D-CL	0.0 ↗ 1.0	$0.159^{\pm \mathbf{4.13e-3}}$	15.9%	0.290
	TF+N–AR-D-CL	0.0 ↗ 1.0	$0.145^{\pm \underline{4.74e-3}}$	23.3%	0.267
	AR–TF-P-CL	1.0 ↘ 0.0	$0.226^{\pm 2.0e-2}$	−19.1%	0.411
	TF–AR-P-CL	0.0 ↗ 1.0	$0.189^{\pm 7.29e-3}$	0%	0.342
	AR-STLW-CL	1.0, $A_c = 5.0$	$\mathbf{0.113}^{\pm 5.02e-3}$	**40.3%**	**0.214**

Table 1: **Accuracy**: Mean Normalised Root Mean Squared Error (nRMSE) using FNO. **Bold** and <u>underline</u> denote the best and second-best results, respectively.

8.3 Long-Term Accuracy and Stability

To assess long-term predictive performance, we computed the nRMSE over the final 10% of the temporal horizon (the "Last 10%" column in Table 1). Excluding the probabilistic variants, curriculum learning consistently enhanced accuracy at later timesteps. **AR-STLW-CL** demonstrated the lowest terminal error for two datasets (iB and KdV), while TF–AR-D-CL achieved the lowest error for Av and vB.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_15

We also evaluated the stability of these strategies by examining the standard deviation across five independent runs per configuration. Our findings reveal that non-probabilistic curriculum strategies exhibit smaller standard deviations than static baselines, indicating greater robustness to weight initialisation. Specifically, AR-STLW-CL demonstrated the highest stability for the Av dataset, while AR-CL, TF–AR-P-CL, and TF–AR-D-CL showed the smallest variations for vB, iB, and KdV, respectively.

8.4 Error Propagation Analysis

Figure 4 illustrates the accumulation of error over time for the three top-performing curricula for each dataset. These trajectories clearly demonstrate the benefit of curriculum-based training across all physical regimes. While **AR-STLW-CL** generally achieves the lowest overall nRMSE, the plots indicate that its performance advantage over the top discrete curricula remains consistent throughout the rollout, effectively "flattening" the error growth curve compared to the best baseline.

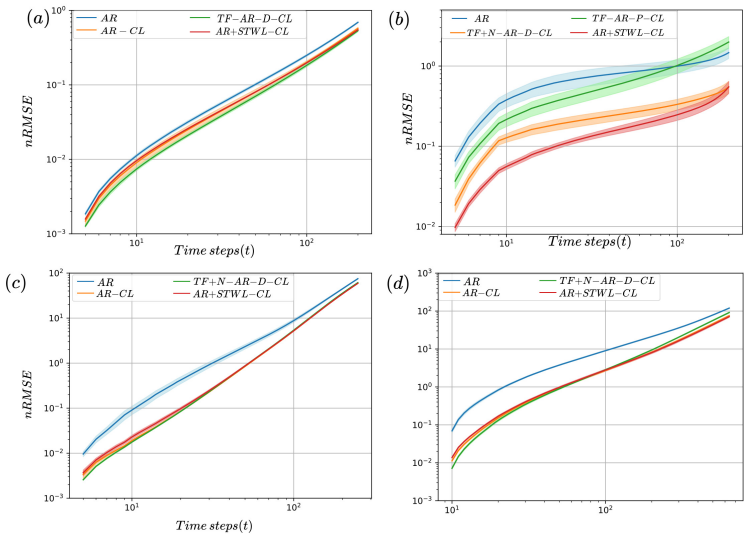


Fig. 4: Error propagation over time for the best baseline and the top two curriculum strategies for: (a) Advection (Av), (b) Viscous Burgers (vB), (c) Inviscid Burgers (iB), and (d) Korteweg–De Vries (KdV).

8.5 Impact of Training Rollout Length

As established in Section 5, AR-STLW-CL can be conceptualised as a continuous generalisation of the discrete AR-CL strategy. While discrete AR-CL progressively increases the number of training rollouts $\widehat{T}$, our continuous approach

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_15

applies temporal weights across the entire sequence length. We investigate the impact of varying the training rollout length $\widehat{T} \in \{2, 4, 8, 16, 32\}$ on the performance of the fixed-AR, discrete AR-CL, and continuous AR-STLW-CL strategies.

Figure 5 presents the nRMSE as a function of $\widehat{T}$. Our observations indicate that **AR-STLW-CL** consistently achieves lower errors and greater stability across all rollout lengths tested. Notably:

- **Advection (Av):** AR-STLW-CL remains stable even at low $\widehat{T}$ values where other methods fluctuate.
- **Viscous Burgers (vB):** While errors generally reach a minimum at $\widehat{T} = 8$, AR-STLW-CL consistently maintains a lower error profile across the entire range.
- **iB and KdV:** Whereas the error for the fixed-AR strategy begins to diverge as $\widehat{T}$ increases beyond 8, both AR-CL and AR-STLW-CL maintain a downward trend, with AR-STLW-CL achieving the most significant reduction in error.

In summary, AR-STLW-CL demonstrates enhanced scalability and stability across different training rollouts, highlighting its effectiveness in mitigating gradient-related instabilities associated with long-horizon training.

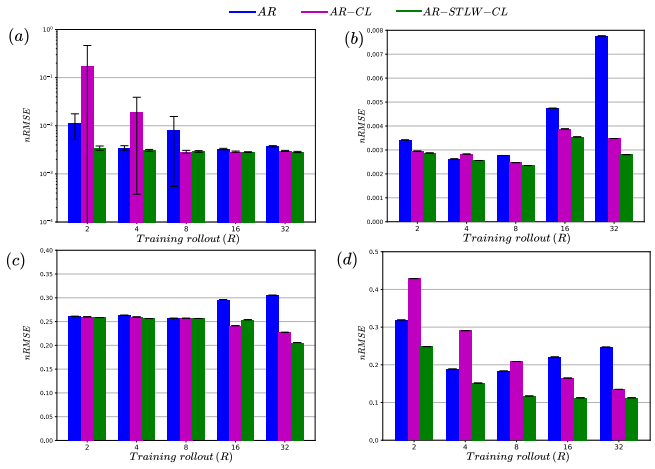


Fig. 5: Effect of increasing training rollout length $\widehat{T}$ on the nRMSE of the best baseline and the top two curriculum strategies across the four PDE datasets.

8.6 Inference Rollout Analysis

To evaluate the robustness of **AR-STLW-CL** for long-term forecasting, we analyse its performance across varying inference horizons (T). Table 2 details the

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_15

mean nRMSE and standard deviation for the Advection (Av) and Korteweg–De Vries (KdV) datasets.

In this analysis, each reported value corresponds to a model trained with a specific output size (S), defined as the number of timesteps predicted per forward pass. For the Av dataset, we evaluate the configurations $(T, S) = (40, 5)$ and $(200, 1)$. For the KdV dataset, we examine increasingly long horizons: $(T, S) \in \{(128, 5), (320, 2), (640, 1)\}$.

The results indicate that AR-STLW-CL consistently outperforms alternative strategies across nearly all temporal horizons. Notably, the performance gap between our curriculum and the baselines becomes more pronounced as the rollout length T increases. While the deterministic curriculum (TF–AR-D-CL) remains competitive at shorter horizons (e.g., $T = 40$ for Av), it exhibits significant instability or divergence in long-term regimes ($T = 200$).

Strikingly, at the longest horizon for KdV ($T = 640$), most baseline and discrete-curriculum strategies either fail to converge or produce errors several orders of magnitude higher than those of our approach. In contrast, AR-STLW-CL maintains superior stability and a significantly lower error profile. This confirms that our continuous weighting strategy effectively mitigates the compounding error and gradient instabilities typically associated with extended autoregressive rollouts.

	Strategy	**Inference Rollout T**		
		40		**200**
Av	TF	$0.0375^{\pm 1.76e-2}$		$0.486^{\pm 4.53e-2}$
	TF+N	$0.0165^{\pm 5.68e-3}$		NaN
	AR	$0.00353^{\pm 8.70e-5}$		$859.6^{\pm 1913.4}$
	AR-CL	$\underline{0.00293}^{\pm 8.98e-5}$		$\underline{0.0551}^{\pm 1.01e-1}$
	TF–AR-D-CL	$\mathbf{0.00272}^{\pm 9.58e-5}$		$10.9^{\pm 23.1}$
	TF+N–AR-D-CL	$0.00296^{\pm 1.43e-4}$		$1.057^{\pm 1.93}$
	AR–TF-P-CL	$0.00538^{\pm 1.18e-3}$		$0.196^{\pm 1.12e-1}$
	TF–AR-P-CL	$0.00294^{\pm 9.48e-5}$		$29.9^{\pm 65.7}$
	AR-STLW-CL	$0.00303^{\pm 0.0}$		$\mathbf{0.0135}^{\pm 8.72e-3}$
	Strategy	**Inference Rollout T**		
		128	**320**	**640**
KdV	TF	$0.541^{\pm 9.50e-2}$	–	–
	TF+N	$0.347^{\pm 1.01e-1}$	$0.408^{\pm 1.14e-2}$	$0.810^{\pm 1.37e-1}$
	AR	$0.189^{\pm 5.29e-3}$	$0.298^{\pm 6.62e-2}$	$0.518^{\pm 1.94e-1}$
	AR-CL	$\underline{0.119}^{\pm 5.29e-3}$	$\underline{0.196}^{\pm 6.61e-3}$	$2.49e5^{\pm 5.58e5}$
	TF–AR-D-CL	$0.159^{\pm 4.13e-3}$	$0.244^{\pm 1.42e-2}$	$\underline{0.366}^{\pm 1.29e-2}$
	TF+N–AR-D-CL	$0.145^{\pm 4.74e-3}$	–	–
	AR–TF-P-CL	$0.226^{\pm 2.00e-2}$	–	–
	TF–AR-P-CL	$0.189^{\pm 7.29e-3}$	–	–
	AR-STLW-CL	$\mathbf{0.113}^{\pm 5.02e-3}$	$\mathbf{0.170}^{\pm 1.09e-2}$	$\mathbf{0.307}^{\pm \mathbf{2.85e-2}}$

Table 2: **Inference Rollout Performance**: nRMSE across varying temporal horizons T. **Bold** and underline denote the best and second-best results. "NaN" and extremely high values indicate model divergence denoted with a dash).

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_15

8.7 STLW as a Multi-Dimensional Learning Rate Scheduler

The **AR-STLW-CL** curriculum can be conceptualised as a **two-dimensional learning rate scheduler**. Unlike conventional schedulers that modulate the learning rate η_e solely as a function of the epoch e, STLW introduces a temporal weighting function $w_{e,t}$ that varies across both the epoch and the prediction timestep t. This allows the model to selectively prioritise different stages of the temporal rollout throughout training. To isolate the impact of this mechanism, we evaluate variants with both fixed and variable learning rates, comparing STLW against standard AR baselines and discrete AR-CL.

As illustrated in Figure 6 for the viscous Burgers' equation, the inclusion of our curriculum leads to a significant reduction in terminal test loss compared to fixed-η baselines, proving that scheduled weighting independently improves convergence. Furthermore, the qualitative similarity between the fixed-η STLW variants and standard variable-η approaches suggests that temporal loss weighting acts as an implicit proxy for learning rate decay, providing a comparable regularising effect. Notably, in the left figure, STLW exhibits a smoother loss trajectory than the more erratic discrete transitions of AR-CL, suggesting that continuous weighting offers a more stable and fluid optimisation strategy.

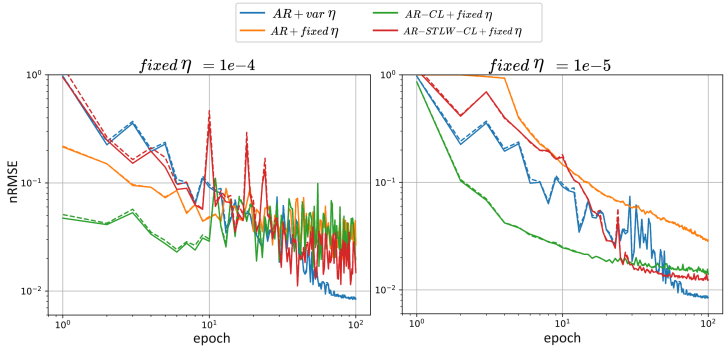


Fig. 6: Training dynamics under fixed and variable learning rates (η) across different curricula. Dashed and solid lines indicate training and validation loss, respectively.

9 Discussion

This section synthesises our empirical findings to address the six core research questions regarding scheduled temporal loss weighting.

– **Q1 Baseline Strategies:** Among non-curriculum baselines, adding Gaussian noise (TF+N) consistently improved Teacher Forcing performance. However, Autoregressive (AR) training proved superior, as it forces the model

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_15

to learn error correction during rollout. This is particularly effective for the well-behaved physical structures of the PDEs investigated here [23].

- **Q2 Curriculum Effectiveness:** Consistent with [1, 24], curriculum learning (CL) significantly enhanced accuracy and stability. **AR-STLW-CL** outperformed curricula that transition toward TF, which we find to be suboptimal for non-chaotic systems. This suggests that for such systems, the optimal curriculum should refine AR tasks rather than incorporate TF.
- **Q3 Training Rollout Length:** Increasing $\widehat{T}$ does not yield monotonic improvements; for the AR baseline, performance often plateaus or degrades beyond $\widehat{T} = 8$ due to error accumulation. **AR-STLW-CL** better manages the difficulty of longer rollouts, though researchers must balance marginal gains against the linear increase in computational and memory costs.
- **Q4 Inference Robustness:** STLW-trained models generalise better to extended inference horizons. While standard AR models often diverge or produce "NaN" values during extended testing in Av and KdV, **AR-STLW-CL** maintains physical plausibility and stable error profiles, effectively mitigating compounding errors.
- **Q5 Stochastic Robustness:** Across multiple independent runs, curriculum strategies exhibited lower variance than the AR baseline. This suggests CL smoothes the optimisation landscape, making models less sensitive to weight initialisation and more reliable for deployment.
- **Q6 Parameter Sensitivity:** A key advantage of **AR-STLW-CL** is its robustness to the curvature parameter A. The stability provided by the "soft" continuous weighting transition reduces the need for extensive hyperparameter tuning compared to the "hard" transitions of discrete curricula or static AR.

10 Conclusion

This work investigated the challenges of training neural operators for long-term temporal PDE forecasting, focusing on instabilities in autoregressive rollouts. We introduced **Scheduled Temporal Loss Weighting (AR-STLW-CL)**, a novel curriculum learning framework that treats training as a continuous transition from short-term to long-term temporal dependencies.

Empirical evaluation across four physical systems; Advection, Viscous Burgers', Inviscid Burgers', and Korteweg–De Vries; demonstrates that AR-STLW-CL consistently outperforms standard Teacher Forcing and discrete curricula, achieving nRMSE improvements of up to **62.5%** over the best baselines with negligible computational overhead. By providing a more smooth and continuous training through weighted loss, our approach enables more precise and reliable learning of complex physics. Future research could extend this framework to higher-dimensional chaotic dynamics and integrate physics-informed constraints to enhance conservation properties.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_15

References

1. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48 (2009)
2. Brandstetter, J., Welling, M., Worrall, D.E.: Lie point symmetry data augmentation for neural pde solvers. In: International Conference on Machine Learning. pp. 2241–2256. PMLR (2022)
3. Brandstetter, J., Worrall, D., Welling, M.: Message passing neural pde solvers. arXiv preprint arXiv:2202.03376 (2022)
4. Gonzalez, F., Demoulin, F.X., Bernard, S.: Towards long-term predictions of turbulence using neural operators. arXiv preprint arXiv:2307.13517 (2023)
5. Gupta, J.K., Brandstetter, J.: Towards multi-spatiotemporal-scale generalized pde modeling. arXiv preprint arXiv:2209.15616 (2022)
6. Hess, F., Monfared, Z., Brenner, M., Durstewitz, D.: Generalized teacher forcing for learning chaotic dynamics. arXiv preprint arXiv:2306.04406 (2023)
7. Kovachki, N.B., Lanthaler, S., Stuart, A.M.: Operator learning: Algorithms and analysis. Handbook of Numerical Analysis **25**, 419–467 (2024)
8. Li, Z., Peng, W., Yuan, Z., Wang, J.: Long-term predictions of turbulence by implicit u-net enhanced fourier neural operator. Physics of Fluids **35**(7) (2023)
9. Li, Z., Kovachki, N.B., Azizzadenesheli, K., liu, B., Bhattacharya, K., Stuart, A., Anandkumar, A.: Fourier neural operator for parametric partial differential equations. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=c8P9NQVtmnO
10. Lippe, P., Veeling, B., Perdikaris, P., Turner, R., Brandstetter, J.: Pde-refiner: Achieving accurate long rollouts with neural pde solvers. Advances in Neural Information Processing Systems **36** (2024)
11. Lippe, P., Veeling, B.S., Perdikaris, P., Turner, R.E., Brandstetter, J.: Pde-refiner: Achieving accurate long rollouts with neural pde solvers. arXiv preprint arXiv:2308.05732 (2023)
12. List, B., Chen, L.W., Bali, K., Thuerey, N.: Differentiability in unrolled training of neural physics simulators on transient dynamics. Computer Methods in Applied Mechanics and Engineering **433**, 117441 (2025). https://doi.org/https://doi.org/10.1016/j.cma.2024.117441, https://www.sciencedirect.com/science/article/pii/S0045782524006960
13. Lu, L., Meng, X., Cai, S., Mao, Z., Goswami, S., Zhang, Z., Karniadakis, G.E.: A comprehensive and fair comparison of two neural operators (with practical extensions) based on fair data. Computer Methods in Applied Mechanics and Engineering **393**, 114778 (2022)
14. Lu, L., Pestourie, R., Johnson, S.G., Romano, G.: Multifidelity deep neural operators for efficient learning of partial differential equations with application to fast inverse design of nanoscale heat transport. arXiv preprint arXiv:2204.06684 (2022)
15. McCabe, M., Harrington, P., Subramanian, S., Brown, J.: Towards stability of autoregressive neural operators. arXiv preprint arXiv:2306.10619 (2023)
16. Michałowska, K., Goswami, S., Karniadakis, G.E., Riemer-Sørensen, S.: Neural operator learning for long-time integration in dynamical systems with recurrent neural networks. In: 2024 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2024)
17. Pascanu, R., Mikolov, T., Bengio, Y.: On the difficulty of training recurrent neural networks. In: International conference on machine learning. pp. 1310–1318. Pmlr (2013)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_15

18. Peng, W., Yuan, Z., Wang, J.: Attention-enhanced neural network models for turbulence simulation. Physics of Fluids **34**(2) (2022)
19. Rahman, M.A., Ross, Z.E., Azizzadenesheli, K.: U-NO: U-shaped neural operators. Transactions on Machine Learning Research (2023), https://openreview.net/forum?id=j3oQF9coJd
20. Sanchez-Gonzalez, A., Stachenfeld, K., Fielding, D., Kochkov, D., Cranmer, M., Pfaff, T., Godwin, J., Cui, C., Ho, S., Battaglia, P.: Learning general-purpose cnn-based simulators for astrophysical turbulence. In: ICLR 2021 SimDL Workshop (2021)
21. Takamoto, M., Alesiani, F., Niepert, M.: Learning neural pde solvers with parameter-guided channel attention. In: International Conference on Machine Learning. pp. 33448–33467. PMLR (2023)
22. Takamoto, M., Praditia, T., Leiteritz, R., MacKinlay, D., Alesiani, F., Pflüger, D., Niepert, M.: Pdebench: An extensive benchmark for scientific machine learning. Advances in Neural Information Processing Systems **35**, 1596–1611 (2022)
23. Teutsch, P., Mäder, P.: Flipped classroom: effective teaching for time series forecasting. arXiv preprint arXiv:2210.08959 (2022)
24. Vlachas, P.R., Koumoutsakos, P.: Learning on predictions: Fusing training and autoregressive inference for long-term spatiotemporal forecasts. Physica D: Nonlinear Phenomena **470**, 134371 (2024)
25. Wang, S., Perdikaris, P.: Long-time integration of parametric evolution equations with physics-informed deeponets. Journal of Computational Physics **475**, 111855 (2023)
26. Wu, T., Maruyama, T., Leskovec, J.: Learning to accelerate partial differential equations via latent global evolution. Advances in Neural Information Processing Systems **35**, 2240–2253 (2022)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29924-6_15