An Attention-Based Spatio-Temporal Neural Operator with Uncertainty Quantification for Dynamical Systems

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper we present the Attention-based Spatio-Temporal Neural Operator (ASNO), an operator-learning architecture that decouples temporal evolution from spatial coupling. The design follows an implicit–explicit interpretation of Backward Differentiation Formula (BDF) integration: a time-series Transformer delivers explicit temporal extrapolation while a Nonlocal Attention Operator applies implicit spatial refinement. Epistemic uncertainty is estimated post hoc via a diagonal Linear Laplace Approximation with negligible overhead. Across Lorenz, Darcy, and two-dimensional incompressible Navier–Stokes systems, ASNO attains state-of-the-art or competitive accuracy under comparable parameter budgets, is resolution-agnostic, and maintains stable long-horizon rollouts, enabling uncertainty-aware modeling of high-dimensional fields.

1 Introduction

Learning surrogates for operators governed by ordinary and partial differential equations enables fast, resolution-independent prediction [Lu et al., 2021, Kovachki et al., 2023]. However, representing temporal dynamics and spatial couplings within a single module conflates distinct sources of error and can reduce stability under iterative rollouts [Vaswani et al., 2017, Li et al., 2020]. In scientific settings with variability in initial or boundary conditions, calibrated uncertainty is also necessary for reliable use [Daxberger et al., 2021, Cinquin et al., 2024, Zou et al., 2024].

We pursue a principled separation inspired by implicit–explicit (IMEX) time integration. The explicit component advances the state based on recent history, whereas the implicit component enforces consistency with current forcing and spatial interactions [Ascher et al., 1995]. ASNO instantiates this separation by combining a time-series Transformer for temporal extrapolation with a nonlocal spatial operator; uncertainty is quantified by a diagonal Laplace approximation propagated through first-order sensitivity [Karkaria et al., 2025]. The main contributions are: (i) an IMEX-guided decomposition for spatio-temporal operator learning; (ii) a lightweight uncertainty mechanism providing pixel-wise intervals; (iii) an evaluation across Lorenz, Darcy, and Navier–Stokes with matched budgets; and (iv) implementation details that facilitate reproduction. We also discuss extrapolation risks and detection ideas relevant to scientific deployment [Madras et al., 2019].

2 Method

Backward differentiation formula (BDF) methods provide high-order accuracy and large stability regions for stiff problems Fredebeul [1998] [Wanner and Hairer, 1996]. For the initial-value problem

$$\dot{X}(t) = F\big(t, X(t)\big), \qquad X(t_0) = X_0, \tag{1}$$

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

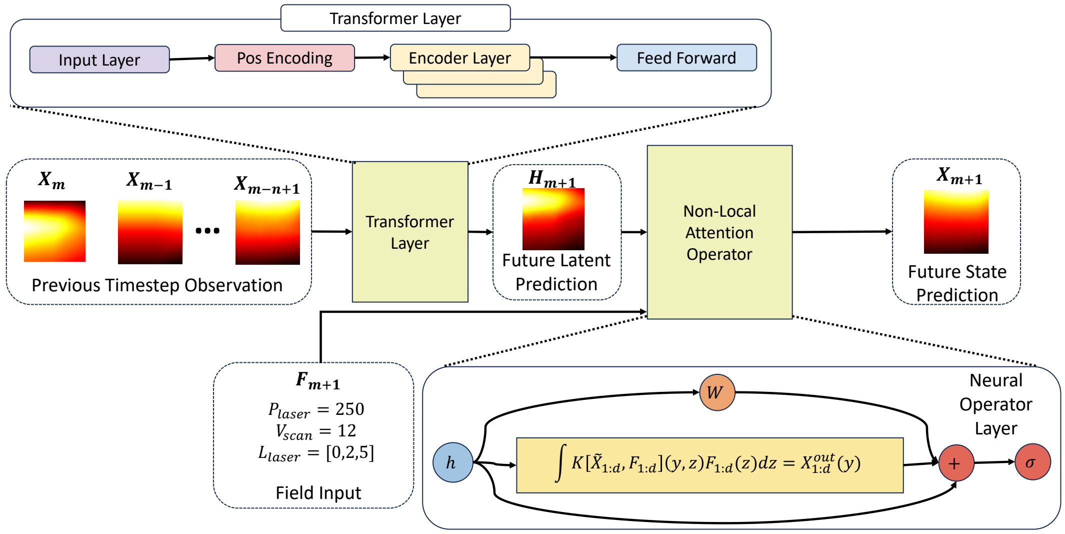

Figure 1: ASNO architecture. A temporal Transformer consumes the last p fields $X_{m-p+1:m}$ to produce the explicit extrapolate $\tilde{X}_{m+1}$; a Non-Local Attention Operator refines this state using conditioning channels F_{m+1} to yield the final prediction X_{m+1}.

a p-th order BDF step of size Δt balances a linear combination of past states against a residual involving $F(t_{m+1}, X_{m+1})$. Rearranging yields an IMEX view

$$X_{m+1} \;=\; \underbrace{\Big(-\textstyle\sum_{k=1}^{p} \alpha_k X_{m+1-k}\Big)}_{\tilde{X}_{m+1}\ \text{(explicit extrapolate)}} \;+\; \Delta t\, \beta\, F\big((m+1)\Delta t,\; X_{m+1}\big)\,, \tag{2}$$

which motivates an architectural split: first construct the explicit extrapolate $\tilde{X}_{m+1}$ from a short temporal history; then compute an implicit correction that encodes spatial coupling and consistency with the driving operator [Ascher et al., 1995]. The resulting ASNO architecture is summarized in Figure 1.

Temporal extrapolation. Let $X(t) \in \mathbb{R}^{N\times d}$ denote a field discretized into N spatial tokens and d channels. The temporal path (Figure 1, left) uses a time-series Transformer encoder $\mathcal{T}_{\theta_T}$ that processes the last p states along time for each spatial token. Inputs are linearly embedded and augmented with temporal positional encodings; multi-head self-attention aggregates information across lags, followed by a position-wise feed-forward block with residual connections and layer normalization. The output is the explicit extrapolate

$$\tilde{X}_{m+1} \;=\; \mathcal{T}_{\theta_T}\big(X_m, \ldots, X_{m-p+1}\big) \in \mathbb{R}^{N\times d}, \tag{3}$$

which plays the explicit role in (2). Isolating temporal memory in a dedicated path reduces competition with spatial modeling and mitigates accumulation error during recursive rollouts [Vaswani et al., 2017, Zerveas et al., 2021, Lim et al., 2021, Zhou et al., 2021].

Spatial refinement. The spatial path (Figure 1, center/right) applies a Nonlocal Attention Operator $\mathcal{S}_{\theta_S}$ to $\tilde{X}_{m+1}$, optionally conditioned on auxiliary channels F_{m+1} (e.g., boundary indicators or source terms). Tokens attend over space to capture long-range interactions and boundary influence; cross-attention incorporates known forcings at t_{m+1}. A residual stack of attention and feed-forward layers yields the refined update

$$X^{\text{out}}_{m+1} \;=\; \mathcal{S}_{\theta_S}\Big(\tilde{X}_{m+1}\Big) \;=\; \mathcal{S}_{\theta_S}\Big(\mathcal{T}_{\theta_T}(X_m, \ldots, X_{m-p+1})\Big), \tag{4}$$

which assigns temporal extrapolation and spatial coupling to distinct, composable modules. This assignment improves interpretability and empirically stabilizes long-horizon forecasts in advection–diffusion and elliptic regimes [You et al., 2022, Yu et al., 2024, Li et al., 2020].

Objective and rollout training. Given samples $\mathcal{D} = \{(X_{m-p+1:m}, X_{m+1})\}$, parameters $\theta = (\theta_T, \theta_S)$ are trained by regularized empirical risk minimization,

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(m)\in\mathcal{D}} \left\| X_{m+1}^{\text{out}}(\theta) - X_{m+1} \right\|_2^2 + \lambda \|\theta\|_2^2. \tag{5}$$

Teacher forcing supervises single steps. For multi-step stability, a short unroll replaces the single-step loss by a sum over q future steps and can employ scheduled sampling. Inputs and targets are standardized per channel; reported errors are de-standardized.

3 Uncertainty quantification

Calibrated uncertainty is needed to quantify confidence in field predictions, detect extrapolation, and support downstream decisions in scientific modeling. Epistemic uncertainty is estimated post hoc via a Laplace approximation around the trained parameters. Let $R(\theta)$ denote the regularized risk in (5); the maximum a posteriori estimate θ_{MAP} minimizes $R(\theta)$. A local quadratic approximation yields

$$p(\theta \mid \mathcal{D}) \approx \mathcal{N}\big(\theta_{\text{MAP}}, \Sigma\big), \qquad \Sigma = H^{-1}, \qquad H = \nabla_\theta^2 R(\theta)\big|_{\theta_{\text{MAP}}}. \tag{6}$$

To scale, we replace H by a diagonal generalized Gauss–Newton surrogate formed from averages of Jacobian outer products plus weight decay; the diagonal is accumulated over mini-batches or estimated with Hutchinson probes using Jacobian–vector and vector–Jacobian products [Daxberger et al., 2021, Ritter et al., 2018, Eschenhagen et al., 2023, George et al., 2018, Schraudolph, 2002, Amari, 1998].

Predictive uncertainty follows from first-order propagation at θ_{MAP}. Writing $J_m = \partial X_{m+1}^{\text{out}}/\partial\theta$ evaluated at θ_{MAP},

$$\mu_{m+1} = X_{m+1}^{\text{out}}(\theta_{\text{MAP}}), \qquad \operatorname{Cov}\big[X_{m+1}^{\text{out}}\big] \approx J_m \, \Sigma \, J_m^\top. \tag{7}$$

Pixel-wise $(1-\alpha)$ credible intervals are $\mu_{m+1} \pm z_{1-\alpha/2}\sigma$, where σ^2 is the corresponding diagonal element of (7); a scalar temperature $\tau > 0$ can rescale Σ on validation to improve empirical calibration. For reporting, we use prediction interval coverage probability (PICP) and mean prediction interval width (MPIW):

$$\text{PICP} = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}\big\{\, y_j \in [\mu_j - z_{1-\alpha/2}\sigma_j,\, \mu_j + z_{1-\alpha/2}\sigma_j] \,\big\}, \qquad \text{MPIW} = \frac{2\, z_{1-\alpha/2}}{M} \sum_{j=1}^{M} \sigma_j. \tag{8}$$

Gaussian negative log-likelihood and CRPS are computed in standard closed forms; related interval-construction and calibration perspectives appear in [Nikulchev and Chervyakov, 2023, Xue et al., 2024]. Libraries such as NeuralUQ support broader UQ workflows for neural operators [Zou et al., 2024].

4 Benchmarks and results

We summarize datasets, training, metrics, and results in a single narrative for coherence. Lorenz isolates temporal extrapolation under chaotic dynamics. Trajectories are integrated by fourth–order Runge–Kutta with step 0.01; models observe five past states and predict the next, for the system

$$\dot{x} = \sigma(y - x), \quad \dot{y} = x(\rho - z) - y, \quad \dot{z} = xy - \beta z, \qquad (\sigma, \rho, \beta) = (10,\, 28,\, 8/3). \tag{9}$$

[Lorenz, 1963] Darcy isolates nonlocal spatial coupling on two-dimensional grids with heterogeneous permeability and Dirichlet boundaries; the strong form is

$$-\nabla \cdot \big(a(x)\nabla u(x)\big) = f(x) \text{ in } \Omega, \qquad u(x) = g(x) \text{ on } \partial\Omega. \tag{10}$$

Incompressible Navier–Stokes probes coupled advection–diffusion with nonlocal constraints on the two-dimensional torus; the vorticity–streamfunction system is

$$\partial_t \omega + J(\psi, \omega) = \nu\, \Delta\omega, \qquad \Delta\psi = \omega, \qquad J(\psi, \omega) = \partial_x\psi\, \partial_y\omega - \partial_y\psi\, \partial_x\omega. \tag{11}$$

Table 1: Unified benchmark performance across Lorenz, Darcy, and Navier–Stokes. Best loss per system is marked with †.

Model	Lorenz			Darcy			Navier–Stokes		
	Params	Time	Loss	Params	GPU	Loss	Params	GPU	Loss
ASNO	258K	1.55s	0.00079†	760K	181MB	0.0368†	4.66M	880MB	0.0213†
Transolver	396K	1.46s	0.00083	811K	422MB	0.0428	4.14M	911MB	0.0234
DeepONet	266K	1.74s	0.00175	6.23M	2146MB	0.0537	5.10M	3100MB	0.0921
Transformer	258K	1.18s	0.00182	1.62M	173MB	0.0559	5.19M	961MB	0.0967
FNO	–	–	–	900K	214MB	0.0768	4.10M	846MB	0.1186
U-Net	–	–	–	821K	123MB	0.1150	5.02M	991MB	0.1940
GNOT	401K	1.99s	0.00219	760K	208MB	0.0516	5.25M	1024MB	0.0322
Linear+NAO	306K	1.29s	0.00529	720K	165MB	0.0547	4.05M	791MB	0.0328

Training uses Adam or AdamW with initial learning rate in $[10^{-4}, 10^{-3}]$, cosine decay with warmup, early stopping, batch sizes chosen to saturate device memory, gradient clipping, matched parameter budgets, standardized inputs, and de-standardized outputs. Single-step losses are computed under teacher forcing; long-horizon stability is assessed by autoregressive iteration. Deterministic accuracy uses mean-squared error (MSE) and fieldwise L^2 norm for predicted $\hat{X}$ and truth X:

$$\mathrm{MSE} = \frac{1}{Nd}\sum_{i=1}^{N}\sum_{c=1}^{d}(\hat{X}_{i,c} - X_{i,c})^2, \qquad \| \hat{X} - X \|_2 = \Big(\sum_{i=1}^{N}\sum_{c=1}^{d}(\hat{X}_{i,c} - X_{i,c})^2\Big)^{1/2}. \tag{12}$$

Uncertainty quality at ninety five percent nominal is summarized by PICP and MPIW using pixelwise means μ_j and standard deviations σ_j.

Table 2: Uncertainty metrics for a representative Darcy test case.

Metric	Value
PICP (coverage %)	94.00 %
MPIW	0.3046

Table 1 indicates systematic gains from separating temporal extrapolation and spatial refinement. On Lorenz, the temporal pathway stabilizes five-step memory and achieves the lowest loss with fewer parameters; on Darcy, nonlocal refinement reduces bias under boundary-induced long-range correlations and attains the best loss with comparable sizes and lower memory; on Navier–Stokes, the split design mitigates rollout drift and preserves coherent structures, consistent with reduced single-step error. Uncertainty estimates are well-calibrated in practice: for Darcy, Table 2 reports coverage near nominal with moderate interval width (PICP 94.00%, MPIW 0.3046). Removing the spatial operator increases error on Darcy and Navier–Stokes; a purely spatial variant conditioned only on the latest frame lacks temporal memory and becomes unstable in autoregression; disabling uncertainty preserves means but worsens calibration (higher Gaussian NLL, worse CRPS), indicating that the Laplace layer provides useful reliability at low cost [You et al., 2022, Li et al., 2020, Vaswani et al., 2017].

5 Conclusion

ASNO is an IMEX/BDF-inspired operator that separates temporal extrapolation (Transformer) from spatial coupling and loads (neural operator with NAO). On Lorenz, Darcy, and Navier–Stokes, it outperforms baselines in accuracy, rollout stability, and zero-shot generalization. The split improves interpretability and enables real-time decisions; future work targets transfer across systems and broader foundational modeling.

Reproducibility

The supplementary material details architecture hyperparameters, optimizer settings, data-generation scripts, ablation tables, and calibration procedures. Code will be released upon acceptance.

References

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

Uri M Ascher, Steven J Ruuth, and Brian TR Wetton. Implicit-explicit methods for time-dependent partial differential equations. *SIAM Journal on Numerical Analysis*, 32(3):797–823, 1995.

Tristan Cinquin, Marvin Pförtner, Vincent Fortuin, Philipp Hennig, and Robert Bamler. Fsp-laplace: Function-space priors for the laplace approximation in bayesian deep learning. *arXiv preprint arXiv:2407.13711*, 2024.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in Neural Information Processing Systems*, 34:20089–20103, 2021.

Runa Eschenhagen, Alexander Immer, Richard Turner, Frank Schneider, and Philipp Hennig. Kronecker-factored approximate curvature for modern neural network architectures. *Advances in Neural Information Processing Systems*, 36:33624–33655, 2023.

Christoph Fredebeul. A-bdf: a generalization of the backward differentiation formulae. *SIAM journal on numerical analysis*, 35(5):1917–1938, 1998.

Thomas George, César Laurent, Xavier Bouthillier, Nicolas Ballas, and Pascal Vincent. Fast approximate natural gradient descent in a kronecker factored eigenbasis. *Advances in neural information processing systems*, 31, 2018.

Vispi Nevile Karkaria, Doksoo Lee, Yi-Ping Chen, Yue Yu, and Wei Chen. Asno: An interpretable attention-based spatio-temporal neural operator for robust scientific machine learning. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Bryan Lim, Sercan Ö Arık, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021.

Edward N Lorenz. Deterministic nonperiodic flow. *Journal of the atmospheric sciences*, 20(2): 130–141, 1963.

Lu Lu, Pengzhan Jin, and George E Karniadakis. Learning operators with deep neural networks. *Nature Machine Intelligence*, 3(3):218–229, 2021.

David Madras, James Atwood, and Alexander D'Amour. Detecting extrapolation with local ensembles. In *International Conference on Learning Representations*, 2019.

Evgeny Nikulchev and Alexander Chervyakov. Prediction intervals: A geometric view. *Symmetry*, 15(4):781, 2023.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *6th international conference on learning representations, ICLR 2018-conference track proceedings*, volume 6. International Conference on Representation Learning, 2018.

Nicol N Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural computation*, 14(7):1723–1738, 2002.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Gerhard Wanner and Ernst Hairer. *Solving ordinary differential equations II*, volume 375. Springer Berlin Heidelberg New York, 1996.

Long Xue, Kai Zhou, and Xiaoge Zhang. Continuous optimization for construction of neural network-based prediction intervals. *Knowledge-Based Systems*, 293:111669, 2024.

Huaiqian You, Yue Yu, Marta D'Elia, Tian Gao, and Stewart Silling. Nonlocal kernel network (nkn): A stable and resolution-independent deep neural network. *Journal of Computational Physics*, 469: 111536, 2022.

Yue Yu, Ning Liu, Fei Lu, Tian Gao, Siavash Jafarzadeh, and Stewart Silling. Nonlocal attention operator: Materializing hidden knowledge towards interpretable physics discovery. *arXiv preprint arXiv:2408.07307*, 2024.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2114–2124, 2021.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

Zongren Zou, Xuhui Meng, Apostolos F Psaros, and George E Karniadakis. Neuraluq: A comprehensive library for uncertainty quantification in neural differential equations and operators. *SIAM Review*, 66(1):161–190, 2024.