Operator Learning for Smoothing and Forecasting

Andrew Stuart

Computing and Mathematical Sciences
California Institute of Technology

US-DoD Vannevar Bush Faculty Fellow

⋆ ⋆ ⋆

15^{th} **AIMS Conference**
Athens, July 6^{th}–10^{th} **2026**

June 2nd 2026

Collaborators and Reference

Collaborators

- Edo Calvello (Caltech $\rightarrow$ Lawrence Berkeley Lab)
- Elizabeth Carlson (Oregon State University)
- Nik Kovachki (Nvidia $\rightarrow$ NYU)
- Michael Manta (Stanford)

Reference:

E. Calvello, E. Carlson, N. Kovachki, M. Manta, and AMS
Operator Learning for Smoothing and Forecasting.
`arXiv:2603.20359`

Overview

Talk Outline

Data-Driven Data Assimilation

Well-Posedness

Approximation Theory

Numerical Experiments

Closing

Data-Driven Data Assimilation

Data Assimilation Text: Reich and Cotter [18] (2015)

Data Assimilation Text: Law, AMS and Zygalakis [15] (2015)

Data Assimilation (DA)

Continuous Time

$$\text{Observed:} \quad \frac{dp}{dt} = f(p, q), \quad t \in \mathbb{R}^+$$
$$\text{Unobserved:} \quad \frac{dq}{dt} = g(p, q), \quad t \in \mathbb{R}^+$$

What Can We Recover, Given $p|_{[0,T]}$?

1. **Filtering:** Recover $q|_t$ from observed $p|_{[0,t]}$, $t \in [0, T]$.
2. **Smoothing:** Recover $q|_{[0,T]}$ from observed $p|_{[0,T]}$.
3. **Forecasting:** Predict $p|_{[T,T+\tau]}$ from observed $p|_{[0,T]}$.
4. **No Model:** f, g (and dimension of q) are not known.

This Talk

- Pure data-driven data assimilation is being widely adopted:
    1. physical contexts (e.g. weather and climate) [14, 5, 17, 2, 1, 12]
    2. time-series forecasting (e.g. commercial applications) [21, 10, 8, 16, 3]
- Analysis to underpin this is needed.
- We will show:
    1. A principled approach to data-driven smoothing.
    2. A principled approach to data-driven forecasting.
    3. Existence theory for relevant operators on C^k spaces.
    4. Approximation theory for relevant operators on C^k spaces.
    5. Implementation with transformer neural operator.
    6. Trajectory accuracy (smoothing).
    7. Statistical accuracy (forecasting, learned delay equation).

Talk Outline

Data-Driven Data Assimilation

Well-Posedness

Approximation Theory

Numerical Experiments

Closing

Well-Posedness

Observability in Control: Hermann and Krener [11] (1977)

Data Assimilation (DA)

Continuous Time

$$\text{Observed:} \quad \frac{dp}{dt} = f(p, q), \quad t \in \mathbb{R}^+$$
$$\text{Unobserved:} \quad \frac{dq}{dt} = g(p, q), \quad t \in \mathbb{R}^+$$

Initial Conditions

1. $p(0) = p_0$.
2. $q(0) = q_0$.

Observe $p|_{[0,T]}$; **smoothing** and **forecasting**?

Existence Of $\Psi_S^\dagger : p\big|_{t\in[0,T]} \mapsto q\big|_{t\in[0,T]}$

Key Idea: Smoothing

Find, from $p\big|_{t\in[0,T]}$, $\quad (t^\star, q^\star) : q(t^\star) = q^\star$

Then $q\big|_{t\in[0,T]}$ **is given by**

$$\dot{q} = g(p, q), \quad q(t^\star) = q^\star.$$

Existence Of $\Psi_F^\dagger : p\big|_{t\in[0,T]} \mapsto p\big|_{t\in[T,T+\tau]}$

Key Idea: Forecasting

$$\begin{aligned} &\textbf{Find, from} \quad p\big|_{t\in[0,T]}, \quad (t^\star, q^\star) : q(t^\star) = q^\star \\ &\textbf{Then} \quad p\big|_{t\in[T,T+\tau]} \quad \textbf{is given by} \end{aligned}$$

$$\begin{aligned} \dot{p} &= f(p,q), \quad p(T) = \textbf{from observation}, \\ \dot{q} &= g(p,q), \quad q(T) = \textbf{from smoothing}. \end{aligned}$$

Observability-Rank Condition

Observability

$$\textbf{Find, from} \quad p|_{t\in[0,T]}, \quad (t^\star, q^\star) : q(t^\star) = q^\star$$

Lie derivatives of f along (f, g)

For $(\mathfrak{p}, \mathfrak{q}) \in \mathbb{R}^{d_p+d_q}$, define $F^{(n)}(\mathfrak{p}, \mathfrak{q}) \in \mathbb{R}^{nd_p}$ by $f^{(0)} = f$ and

$$F^{(n)}(\mathfrak{p}, \mathfrak{q}) = \left(f^{(0)}(\mathfrak{p}, \mathfrak{q})^\top, f^{(1)}(\mathfrak{p}, \mathfrak{q})^\top, \dots, f^{(n-1)}(\mathfrak{p}, \mathfrak{q})^\top \right)^\top,$$

$$f^{(k)}(\mathfrak{p}, \mathfrak{q}) = \partial_p f^{(k-1)}(\mathfrak{p}, \mathfrak{q}) \cdot f(\mathfrak{p}, \mathfrak{q}) + \partial_q f^{(k-1)}(\mathfrak{p}, \mathfrak{q}) \cdot g(\mathfrak{p}, \mathfrak{q})$$

Observability-Rank Condition

Use derivatives of p to define equation for q

Let

$$\left(P^{(n)}(p)\right)(t) := \left(\partial_t^1 p(t)^\top, \ldots, \partial_t^n p(t)^\top\right)^\top.$$

Then for all $L\colon \mathbb{R}^{nd_p} \to \mathbb{R}^{d_q}$, any p, q from model satisfies

$$LF^{(n)}\left(p(t), q(t)\right) = L\Big(\left(P^{(n)}(p)\right)(t)\Big). \tag{1}$$

Observability-Rank Condition special case of Hermann and Krener [11] (1977)

The observability-rank condition holds at $(\mathfrak{p}, \mathfrak{q}) \in \mathbb{R}^{d_p+d_q}$, if $\exists n \in \mathbb{N}$ and $L\colon \mathbb{R}^{nd_p} \to \mathbb{R}^{d_q}$ such that the rank of the matrix $\mathrm{D}_q LF^{(n)}(\mathfrak{p}, \mathfrak{q})$ is d_q.

Invertibility of $LF^{(n)}(\mathfrak{p}, \bullet)$

If condition satisfied at $(\mathfrak{p}, \mathfrak{q})$, can invert $LF^{(n)}(\mathfrak{p}, \bullet)$ locally.

Smoothing and Forecasting Operators Exist

Calvello, Carlson, Kovachki, Manta, AMS. [6](2026)

Theorem: Existence of $\Psi_S^\dagger : p\big|_{t\in[0,T]} \mapsto q\big|_{t\in[0,T]}$

- Let observability-rank condition hold at some point $(\mathfrak{p}, \mathfrak{q})$.
- Consider trajectories with (p_0, q_0) in neighbourhood of $(\mathfrak{p}, \mathfrak{q})$.
- Then $\Psi_S^\dagger : C^k([0, T]; \mathbb{R}^{d_p}) \to C^k([0, T]; \mathbb{R}^{d_q})$ exists.
- $\Psi_S^\dagger$ is **continuous**.

Theorem: Existence of $\Psi_F^\dagger : p\big|_{t\in[0,T]} \mapsto p\big|_{t\in[T,T+\tau]}$

- Let observability-rank condition hold at some point $(\mathfrak{p}, \mathfrak{q})$.
- Consider trajectories with (p_0, q_0) in neighbourhood of $(\mathfrak{p}, \mathfrak{q})$.
- Then $\Psi_F^\dagger : C^k([0, T]; \mathbb{R}^{d_p}) \to C^k([T, T+\tau]; \mathbb{R}^{d_p})$ exists.
- $\Psi_F^\dagger$ is **continuous**.

Talk Outline

Data-Driven Data Assimilation

Well-Posedness

Approximation Theory

Numerical Experiments

Closing

Approximation Theory

Neural Operators: Kovachki, Lanthaler and AMS [13](2024)

Attention Mechanism: Calvello, Kovachki, Levine and AMS [7](2025)

Operator Learning

Approximating By Neural Operator (NO) Kovachki, Lanthaler, AMS. [13](2025)

$$\begin{aligned} \text{Parameter Space:} \quad & \Theta \subseteq \mathbb{R}^p \\ \text{Function Class:} \quad & \Psi : \mathcal{X} \times \Theta \to \mathcal{Y} \\ \text{Function Approximation:} \quad & \Psi(\cdot\,; \theta^\star) \approx \Psi^\dagger \end{aligned}$$

Theorem (Transformer Neural Operator (TNO) is Universal)

Calvello, Kovachki, Levine, AMS. [7](2025)

- Let $\Psi^\dagger$ be a continuous operator.
- Let $\mathcal{K} \subset \mathcal{X}\big(D; \mathbb{R}^{d_i}\big)$ be compact.

Then, for any $\varepsilon > 0$, parameter θ^* may be chosen so that

$$\sup_{i \in \mathcal{K}} \|\Psi^\dagger(i) - \Psi(i; \theta^*)\|_{\mathcal{Y}} \leq \varepsilon.$$

Universal Approximation for $\Psi_S^\dagger : p|_{t\in[0,T]} \mapsto q|_{t\in[0,T]}$

Theorem (Smoothing) Calvello, Kovachki, Levine, AMS. [6](2026)

- Let observability-rank condition hold at some point $(\mathfrak{p}, \mathfrak{q})$.
- Let U denote the invertibility neighbourhood of $(\mathfrak{p}, \mathfrak{q})$.
- Let $\mathcal{K}$ denote a compact subset of trajectories starting in U.
- Let $\mathcal{K}_p$ denote projection onto p of those trajectories.

For any $\varepsilon > 0$ there exists a TNO $\Psi(\bullet;\theta)$ such that

$$\sup_{p\in\mathcal{K}_p} \|\Psi_S^\dagger(p) - \Psi(p;\theta)\|_{C^k} < \varepsilon.$$

Universal Approximation for $\Psi_F^\dagger : p|_{t\in[0,T]} \mapsto p|_{t\in[T,T+\tau]}$

Theorem (Forecasting) Calvello, Kovachki, Levine, AMS. [6](2026)

- Let observability-rank condition hold at some point $(\mathfrak{p},\mathfrak{q})$.
- Let U denote the invertibility neighbourhood of $(\mathfrak{p},\mathfrak{q})$.
- Let $\mathcal{K}$ denote a compact subset of trajectories starting in U.
- Let $\mathcal{K}_p$ denote projection onto p of those trajectories.

For any $\varepsilon > 0$ there exists a TNO $\Psi(\bullet;\theta)$ such that

$$\sup_{p\in\mathcal{K}_p} \|\Psi_F^\dagger(p) - \Psi(p;\theta)\|_{C^k} < \varepsilon.$$

Talk Outline

Data-Driven Data Assimilation

Well-Posedness

Approximation Theory

Numerical Experiments

Closing

Numerical Experiments

Calvello, Carlson, Kovachki, Manta and AMS [6](2025)

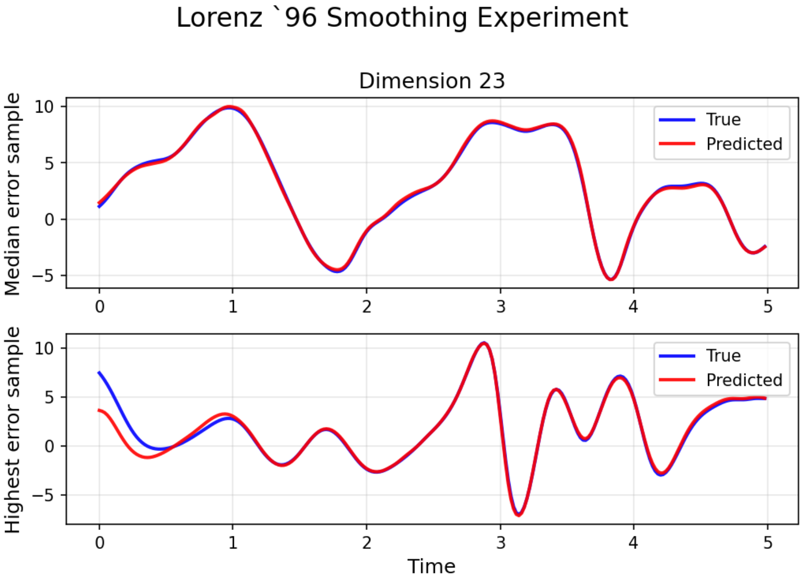
The Lorenz '96 Model

Lorenz '96 Model

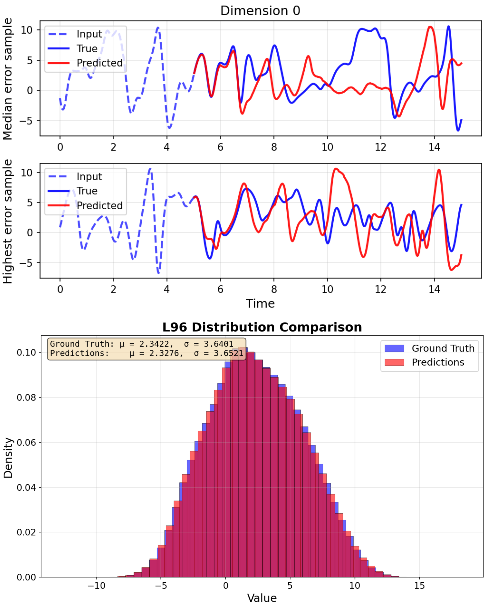
$$\frac{dv_i}{dt} = \big(v_{i+1} - v_{i-2}\big)v_{i-1} - v_i + F, \quad i = 1, 2, \dots, 40$$

Category	**Smoothing**	**Forecasting**
Parameters	$F = 8$	
Observed	$(u_1, \dots, u_{21}, u_{23}, u_{25}, \dots, u_{39})$ $\in C([0,5]; \mathbb{R}^{30})$	$(u_1, u_2, u_3, u_5, u_6, u_7, u_9, \dots)$ $\in C([0,5]; \mathbb{R}^{30})$
Unobserved	$(u_{22}, u_{24}, \dots, u_{40})$ $\in C([0,5]; \mathbb{R}^{10})$	$(u_1, u_2, u_3, u_5, u_6, u_7, u_9, \dots)$ $\in C([5, 5.2]; \mathbb{R}^{30})$
Obs. time	$\Delta t = 0.02$	

Lorenz ‘96: Smoothing



Lorenz '96: Long-horizon Forecasting



- Long-horizon forecast obtained via composition;
- Trajectory prediction diverges, but statistics match.

Closing

Conclusions and References

Conclusions

- Data-driven data assimilation is being widely adopted:
  1. physical contexts (e.g. weather and climate) [14, 5, 17, 2, 1, 12]
  2. time-series forecasting (e.g. commercial applications) [21, 10, 8, 16, 3]
- Analysis to underpin this is needed.
- We have shown:
  1. A principled approach to data-driven smoothing.
  2. A principled approach to data-driven forecasting.
  3. Existence theory for relevant maps.
  4. Approximation theory for relevant maps.
  5. Implementation with transformer neural operator.
  6. Trajectory accuracy and statistical accuracy.
- Many open problems:
  1. Discrete time analog.
  2. Effect of noise.
  3. Connection to Takens embedding.
  4. Theory for PDE (infinite dimensional) models.
  5. Extending theory to global well-posedness.

References I

[1] M. Alexe, E. Boucher, P. Lean, E. Pinnington, P. Laloyaux, A. McNally, S. Lang, M. Chantry, C. Burrows, M. Chrust, F. Pinault, E. Villeneuve, N. Bormann, and S. Healy.
Graphdop: towards skilful data-driven medium-range weather forecasts learnt and initialised directly from observations.
arXiv preprint arXiv:2412.15687, 2024.

[2] A. Allen, S. Markou, W. Tebbutt, J. Requeima, W. P. Bruinsma, T. R. Andersson, M. Herzog, N. D. Lane, M. Chantry, J. S. Hosking, and R. E. Turner.
End-to-end data-driven weather prediction.
Nature, 641(8065):1172–1179, May 2025.

[3] A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. P. Arango, S. Kapoor, et al.
Chronos: learning the language of time series.
arXiv preprint arXiv:2403.07815, 2024.

[4] E. Bach, R. Baptista, E. Calvello, B. Chen, and A. Stuart.
Learning enhanced ensemble filters.
Journal of Computational Physics, page 114550, 2026.

[5] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian.
Accurate medium-range global weather forecasting with 3d neural networks.
Nature, 619(7970):533–538, Jul 2023.

References II

[6] E. Calvello, E. Carlson, N. Kovachki, M. N. Manta, and A. M. Stuart.
Operator learning for smoothing and forecasting.
arXiv preprint arXiv:2603.20359, 2026.

[7] E. Calvello, N. B. Kovachki, M. E. Levine, and A. M. Stuart.
Continuum attention for neural operators.
Journal of Machine Learning Research, 26(300):1–52, 2025.

[8] C. Challu, K. G. Olivares, B. N. Oreshkin, F. G. Ramirez, M. M. Canseco, and A. Dubrawski.
Nhits: Neural hierarchical interpolation for time series forecasting.
In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 6989–6997, 2023.

[9] D. Donoho.
Data science at the singularity.
Harvard Data Science Review, 6(1), 2024.

[10] C. Eisenach, Y. Patel, and D. Madeka.
Mqtransformer: multi-horizon forecasts with context dependent and feedback-aware attention.
arXiv preprint arXiv:2009.14799, 2020.

References III

[11] R. Hermann and A. Krener.
Nonlinear controllability and observability.
IEEE Transactions on automatic control, 22(5):728–740, 1977.

[12] J. Kossaifi, N. Kovachki, M. Mardani, D. Leibovici, S. Ravuri, I. Shokar, E. Calvello, M. S. Abbas, P. Harrington, A. Subramaniam, N. Brenowitz, B. Bonev, W. Byeon, K. Kreis, D. Durran, A. Vahdat, M. Pritchard, and J. Kautz.
Demystifying data-driven probabilistic medium-range weather forecasting.
arXiv preprint arXiv:2601.18111, 2026.

[13] N. B. Kovachki, S. Lanthaler, and A. M. Stuart.
Operator learning: algorithms and analysis.
Handbook of Numerical Analysis, 25:419–467, 2024.

[14] T. Kurth, S. Subramanian, P. Harrington, J. Pathak, M. Mardani, D. Hall, A. Miele, K. Kashinath, and A. Anandkumar.
Fourcastnet: accelerating global high-resolution weather forecasting using adaptive Fourier neural operators.
In *Proceedings of the Platform for Advanced Scientific Computing Conference*, PASC '23, New York, NY, USA, 2023. Association for Computing Machinery.

References IV

[15] K. Law, A. Stuart, and K. Zygalakis.
Data Assimilation.
Texts in Applied Mathematics 62. Springer, 2015.

[16] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long.
itransformer: inverted transformers are effective for time series forecasting.
arXiv preprint arXiv:2310.06625, 2023.

[17] I. Price, A. Sanchez-Gonzalez, F. Alet, T. R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia, R. Lam, and M. Willson.
Probabilistic weather forecasting with machine learning.
Nature, 637(8044):84–90, Jan 2025.

[18] S. Reich and C. Cotter.
Probabilistic Forecasting and Bayesian Data Assimilation.
Cambridge University Press, New York, 2015.

[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer.
High-resolution image synthesis with latent diffusion models.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

References V

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin.
Attention is all you need.
In I. Guyon *et al*, editor, *NEURIPS*, 2017.

[21] R. Wen, K. Torkkola, B. Narayanaswamy, and D. Madeka.
A multi-horizon quantile recurrent forecaster.
arXiv preprint arXiv:1711.11053, 2017.

Attention Key Component of LLMs, CLIP, Stable Diffusion

- Attention is all you need: Vaswani et al [20] (2017)
- $235,508$ Google Scholar citations (March 9th 2026).
- Nonlinear correlation map: Donoho [9] (2024).

Attention Map $\mathcal{O}(N^2)$ to evaluate

$$\mathsf{A}^N : F^N \to F^N, \qquad \vartheta := (Q, K, V) \quad \text{to be learned}$$

$$p(k|j;u) \propto \exp\Big(\big\langle Qu(j), Ku(k)\big\rangle_{\mathbb{R}^d}\Big), \qquad u \in F^N \quad p \in \mathcal{P}(D^N)$$

$$\mathsf{A}^N(u)(j) = V\,\mathbb{E}_{k\sim p(k|j;u)}[u(k)].$$

positional encoding is used Rombach et al [19] (2022)

Cross Attention (Diffusion Based Generative Modelling)

Attention is all you need: Vaswani et al [20] (2017),

Cross Attention in diffusion based conditional image generation: Rombach et al [19] (2022)

Discrete Cross Attention

- Define $D^N = \{1, \cdots, N\}$.
- Let $F^N = \{f : D^N \to \mathbb{R}^c\}$.

Cross Attention Map $\mathcal{O}(NM)$ to evaluate

$$C^N : F^N \times F^M \to F^N, \qquad \vartheta = (Q, K, V) \quad \text{to be learned}$$

$$q(k|j; u, w) \propto \exp\Big(\big\langle Qu(j), Kw(k)\big\rangle_{\mathbb{R}^d}\Big), \qquad (u, w) \in F^N \times F^M \quad q \in \mathcal{P}(D^N)$$

$$C^N(u, w)(j) = V\,\mathbb{E}_{k\sim q(k|j;u,w)}[w(k)].$$

Attention On Banach Space

Calvello, Kovachki, Levine and AMS 2024 [7]

Continuum Attention

- Define $D \subseteq \mathbb{R}^m$, an open set.
- Let $F = \{f : D \to \mathbb{R}^c\}$.

Attention Map (On Functions)

$$
\begin{aligned}
\mathsf{A} : F \to F, &\qquad \vartheta := (Q, K, V) \quad \text{to be learned}\\
p(y|x;u) \propto \exp\Big(\big\langle Qu(x), Ku(y)\big\rangle_{\mathbb{R}^d}\Big), &\qquad u \in F \quad p \in \mathcal{P}(D)\\
\mathsf{A}(u)(x) = V\,\mathbb{E}_{y\sim p(y|x;u)}[u(y)]. &
\end{aligned}
$$

Attention Map On Spaces of Probability Measures

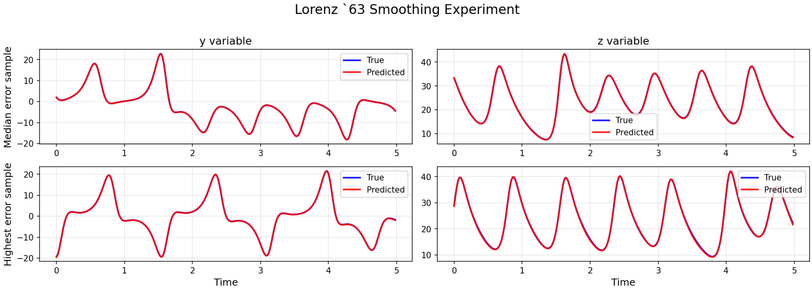
Bach, Baptista, Calvello, Chen and S 2024 [4]

Attention Map (On Probability Measures)

$$
\begin{aligned}
\mathsf{A}: \mathcal{P}(\mathbb{R}^c) &\to \mathcal{P}(\mathbb{R}^c), \qquad \vartheta := (Q, K, V) \quad \text{to be learned} \\
\mathfrak{A}(\cdot\,; p): \mathbb{R}^c &\to \mathbb{R}^c, \qquad p \in \mathcal{P}(\mathbb{R}^c), \\
\mathsf{A}(p) &= \mathfrak{A}(\cdot\,; p)_\sharp p. \\
&\star\star\star\star\star\star \\
\rho(u; s) &= \frac{1}{Z(s)} \exp\big(\langle Qs, Ku\rangle_{\mathbb{R}^d}\big) p(u), \\
\mathfrak{A}(s; p) &= V\,\mathbb{E}_{u\sim\rho(\cdot;s)}[u], \qquad u \in \mathbb{R}^c.
\end{aligned}
$$

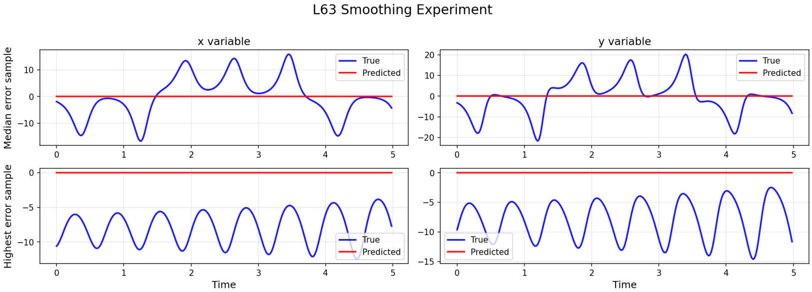
Smoothing in Lorenz '63; Observability

(y, z) from x: observability-rank condition **is satisfied**.



Smoothing in Lorenz '63; Failure of Observability

(x, y) from z: observability-rank condition **is not satisfied**.



The Kuramoto-Sivashinsky Equation

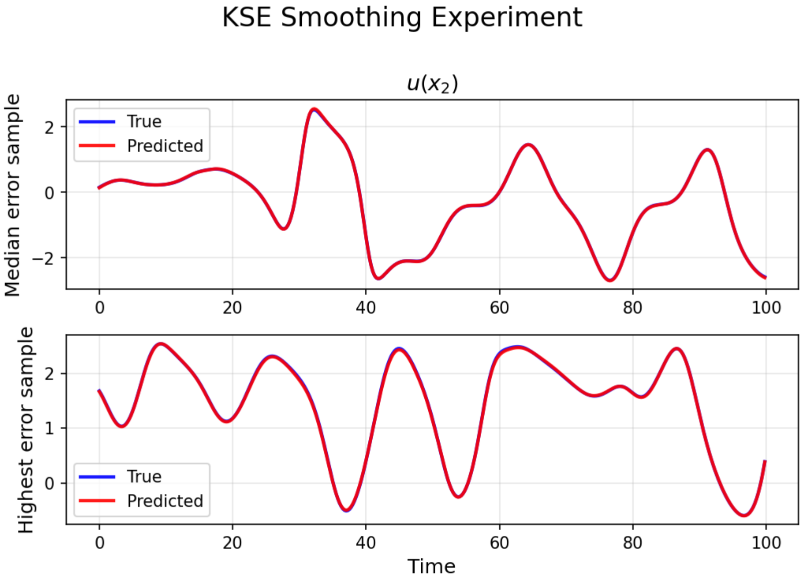
1D Kuramoto-Sivashinsky (KS) Equation

$$\frac{\partial u}{\partial t} + \frac{\partial^4 u}{\partial x^4} + \frac{\partial^2 u}{\partial x^2} + u\frac{\partial u}{\partial x} = 0,$$
$$u : \mathbb{S}_L \times \mathbb{R}^+ \to \mathbb{R}.$$

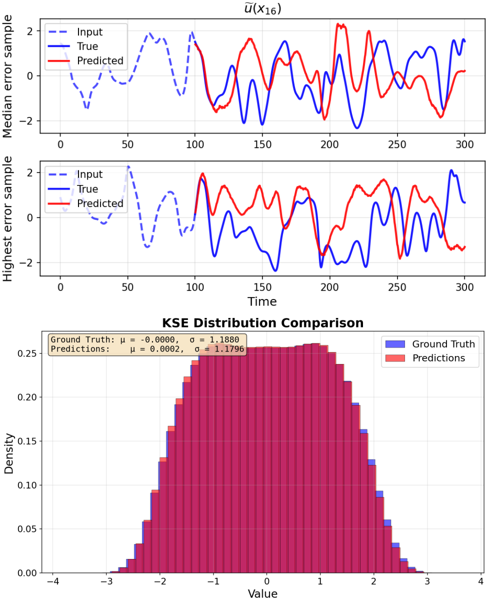
u: solution; $\widetilde{u}$: projection onto first 64 Fourier modes of u.

Category	**Smoothing**	**Forecasting**
Parameters	$L = 32\pi$	
Spatial Grid	$x_j = jL/128, \quad j = 1, \dots, 128$	
Observed	$\left(\widetilde{u}(x_1), \dots, \widetilde{u}(x_{128})\right) \in C([0,100]; \mathbb{R}^{128})$	
Unobserved	$\left(u(x_1), \dots, u(x_{128})\right) \in C([0,100]; \mathbb{R}^{128})$	$\left(\widetilde{u}(x_1), \dots, \widetilde{u}(x_{128})\right) \in C([100,102]; \mathbb{R}^{128})$
Obs. time	$\Delta t = 0.25$	

KS: Smoothing



KS: Long-horizon Forecasting



- Long-horizon forecast obtained via composition;
- Trajectory prediction diverges, but statistics match.