POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

M-SHRED: Masked Shallow Recurrent Decoders for Fault-Tolerant Sparse Sensing under Missing Sensor Measurements

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Matteo Romilio Pasqual

Advisor: Prof. Andrea Manzoni

Co-advisor: Dr. Matteo Tomasetto

Academic year: 2025-2026

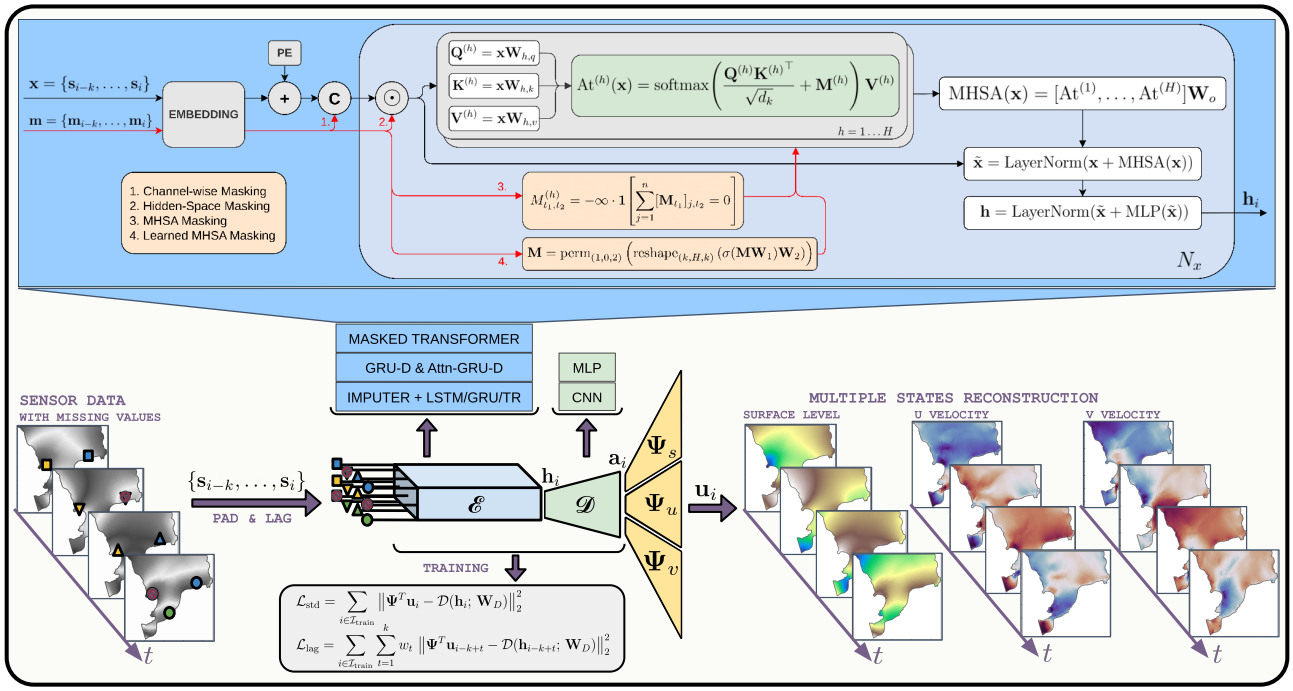
1. Introduction

Real-world dynamical systems evolve in high-dimensional state spaces that cannot be directly observed in their entirety [9]. In practical settings such as fluid dynamics [17], oceanographic monitoring [4], and atmospheric modeling [2], sensors are the principal interface between theory and reality, yet their deployment is inevitably limited by cost, reliability, physical inaccessibility, and operational constraints. Reconstructing a full high-dimensional field from sparse measurements is therefore a fundamental challenge in science and engineering.

Classical techniques based on Proper Orthogonal Decomposition (POD) [7] combined with linear inverse methods provide a mathematically tractable framework, yet they exhibit two major limitations. First, the reconstruction accuracy is inherently constrained by the low-rank approximation induced by the truncation of the POD basis. Second, both the reconstruction accuracy and the conditioning of the associated inverse problem depend critically on the sensor placement strategies. Shallow Decoder Networks (SDNs) [5] partially address the first issue by learning a nonlinear mapping from sparse measurements to the full state field. Nevertheless, all methods share a fundamental structural limitation: they operate on static snapshots, treating each time instant independently and thus ignoring the temporal correlations that are a core feature of dynamical systems.

The Shallow Recurrent Decoder (SHRED) [15] addresses this limitation by combining a recurrent temporal encoder, implemented via Long Short-Term Memory (LSTM) [6], with a shallow spatial decoder (SDN). This architecture exploits temporal context to substantially reduce reconstruction error, even when only a small number of randomly placed sensors are available. Subsequent extensions have further generalized this framework: SHRED-ROM [13] adapts the methodology to parametric dynamical systems, while T-SHRED [16] replaces the recurrent encoder with a Transformer to capture long-range temporal dependencies more effectively.

While the theoretical foundations of all SHRED-based models are robust when tested on simulated data where sensor measurements are *(i)* pointwise field evaluations and *(ii)* assumed to be continuously operational throughout the observation period, their deployment in real-world environments raises significant challenges, such as incomplete, unreliable, and asynchronous



measurements [8, 12]. This thesis addresses this limitation through two complementary contributions:

1. We introduce a hierarchy of Transformer-based masking mechanisms that make the encoder explicitly aware of sensor availability. This is accomplished by incorporating a binary missingness mask at multiple levels of the architecture: at the input feature level via channel-wise masking (1.), at intermediate layers through hidden-space masking (2.), or within the multi-head self-attention mechanism through a direct MHSA masking (3.), or a learned MHSA masking (4.).
2. We further propose a multi-step decoding training objective that encourages smoother and more coherent latent trajectories and enhances reconstruction fidelity in the presence of heterogeneous missing-data patterns without increasing the computational complexity at inference time.

2. Background

Let $\mathbf{u}(\mathbf{x}, t, \boldsymbol{\mu}) \in \mathbb{R}^m$ denote the high-dimensional state of a dynamical system, where $\mathbf{x}$ denotes the spatial coordinates, t the temporal variable, and $\boldsymbol{\mu}$ the vector of system parameters. The evolution of the system is governed by a partial differential equation (PDE), that after spatial discretization can be expressed in the following, abstract form:

$$\dot{\mathbf{u}} = \mathcal{N}(\mathbf{u}, \mathbf{x}, t, \boldsymbol{\mu}).$$

The system is monitored via $n \ll m$ sparse sensors, represented by a nonlinear sensing operator $\mathbf{C} : \mathbb{R}^m \to \mathbb{R}^n$, which produce noisy measurements

$$\mathbf{s}_i = \mathbf{C}(\mathbf{u}_i) + \boldsymbol{\eta}_i \in \mathbb{R}^n.$$

The sparse sensing problem is formulated as the task of determining a reconstruction operator $\mathcal{R} : \mathbf{s} \mapsto \hat{\mathbf{u}}$ that reliably recovers the high-dimensional state from sparse observations.

2.1. The SHRED Framework

Rather than mapping a single snapshot $\mathbf{s}_i$ directly to $\mathbf{u}_i$, the SHRED framework exploits temporal context by employing a sliding window of k lagged sensor measurements. The architecture is given by

$$\hat{\mathbf{u}}_i = \mathcal{D}(\mathcal{E}(\{\mathbf{s}_{i-k}, \ldots, \mathbf{s}_i\}; \mathbf{W}_E); \mathbf{W}_D),$$

where $\mathcal{E} : \mathbb{R}^{n \times k} \to \mathbb{R}^h$ is a recurrent encoder (implemented, for example, via an LSTM [6] or Transformer [14] network), and $\mathcal{D} : \mathbb{R}^h \to \mathbb{R}^m$ is a shallow spatial decoder. This architectural design is motivated by the separation-of-variables solution of linear PDEs: indeed, the encoder learns the temporal evolution, while the decoder reconstructs the spatial structure.

For large-scale systems, a compressive training strategy is employed by projecting the high-dimensional state onto an r-dimensional proper orthogonal decomposition (POD) subspace $\mathbf{\Psi} \in \mathbb{R}^{m \times r}$. In this reduced-order setting, the decoder is required to estimate only the r POD coefficients, rather than the full m-dimensional state, thereby substantially reducing the computational complexity of training and inference.

3. Handling Missing Data

In operational settings, sensor measurements are frequently incomplete. To represent this, a binary mask $\mathbf{M}_i \in \{0,1\}^{n \times k}$, which encodes sensor availability over the lag window, is introduced and passed to the encoder alongside the observed data matrix $\mathbf{S}_i$.

The simplest strategy for handling heterogeneous sensor availability is to discard any sensor that is not jointly active over the entire observation period. This approach sacrifices critical information and, in the extreme case where no two sensors share overlapping observation windows, makes the reconstruction impossible.

To build a solid baseline, three classical imputation strategies (mean, median, and last observation carried forward) pre-fill missing entries before passing the sequence to the encoder; however, they do not propagate any uncertainty about the imputed values and introduce systematic biases under high missingness regimes.

In addition to the baseline, specific models for missing data in timeseries such as GRU-D [3] and its attention-augmented variant (attnGRU-D) [18] are evaluated as an advanced baseline. These architectures explicitly model missingness through learnable input and hidden-state decay mechanisms. However, these approaches rely on specific assumptions regarding the underlying missingness distribution, which may not hold in real-world settings.

3.1. Transformer-Based Masking

To address our first problem, four Transformer-based masking strategies are proposed to inject the binary availability mask $\mathbf{M}$ directly into the encoder pipeline:

1. **Channel-wise masking**: the mask is concatenated to the observed input prior to embedding, doubling the input channels from n to $2n$. The model jointly encodes sensor values and availability indicators from the first layer and propagates it through the encoder.
2. **Hidden-space masking**: the mask is projected into the latent space via a learned linear map followed by a sigmoid, producing a $(0,1)$ gating signal $\mathbf{G} = \sigma(\mathbf{M}^\top \mathbf{W}_{\text{mask}})$ that is multiplied element-wise at each hidden representation before each Transformer block.
3. **MHSA masking**: modifying the autoregressive mask used in the classical Transformer decoder, an additive bias $A_{qt} = -\infty$ is applied in the scaled dot-product attention for any timestep t where all sensor channels are missing, enforcing exactly zero attention weights to fully absent positions.
4. **Learned MHSA masking**: Extending the previous approach, a lightweight two-layer MLP maps the embedded mask sequence $\mathbf{M}$ to head-specific attention bias matrices $\mathbf{L}^{(h)}$, enabling the network to learn flexible, soft availability patterns between timesteps that adapt dynamically to each sequence's specific missingness configuration: $\mathbf{L} = \text{ReLU}(\mathbf{M}\mathbf{W}_1)\mathbf{W}_2$.

These strategies are not mutually exclusive. Different masking strategies can be combined since they operate at different encoder levels.

4. Multi-Step Decoding

Our second observation is that, in the standard T-SHRED pipeline, only the final encoder representation $\mathbf{h}_k$ is decoded; all intermediate timestep representations $\mathbf{h}_1, \ldots, \mathbf{h}_{k-1}$ are discarded. This is especially problematic under heterogeneous missing data: the model must heavily rely on the full temporal context to compensate for absent observations, yet receives no explicit training signal to maintain informative intermediate representations. Inspired by the auxiliary classifier mechanism in GoogLeNet [11] and the sequence-to-sequence (seq2seq) training loss [10], we propose a *multi-step decoding* training strategy in which the same decoder $\mathcal{D}$ is applied to a set of timesteps of the encoder output:

$$\mathcal{L}_{\text{lagged}} = \sum_{i \in \mathcal{I}_{\text{train}}} \sum_{t=1}^{k} w_t \left\| \mathbf{a}_{i,t} - \mathcal{D}(\mathbf{h}_{i,t}; \mathbf{W}_D) \right\|_2^2,$$

where $\mathbf{a}_{i,t}$ are the lagged POD coefficients and $w_t = \alpha^{k-t}/Z$ are normalized exponential

Model	MRE
Mask x2 channels (T.3.B)	**1.03**
Mask x2 + MHSA (T.3.E)	1.31
Mask x2 + Learn MHSA (T.3.G)	1.40
Mask MHSA learned (T.3.F)	1.51
Mask hidden space (T.3.C)	1.56
Mask MHSA (T.3.D)	1.59
Mask baseline (T.3.A)	1.88
Median imputation (T.1.A)	1.91
GRU-D encoder (T.2.A)	1.99
Attn GRU-D encoder (T.2.B)	2.30
Last-value imputation (T.1.C)	2.44
Mean imputation (T.1.B)	2.92
Always-available sensors (T.0.A)	6.36

Table 1: Reconstruction performance on the Double Gyre flow benchmark.

Model	MRE
Mask x2 + Learn MHSA (T.3.G)	**15.73**
Mask hidden space (T.3.C)	15.73
Mask x2 + MHSA (T.3.E)	15.77
Mask baseline (T.3.A)	15.77
Last-value imputation (T.1.C)	15.78
Mask MHSA (T.3.D)	15.78
Mask MHSA learned (T.3.F)	15.81
Median imputation (T.1.A)	15.90
Mask x2 channels (T.3.B)	15.97
Mean imputation (T.1.B)	16.07
GRU-D encoder (T.2.A)	16.18
Attn GRU-D encoder (T.2.B)	16.21
Always-available sensors (T.0.A)	16.24

Table 2: Reconstruction performance on the Hydrodynamic model of the Southern North Sea.

weights, assigning greater importance to more recent timesteps. The hyperparameters $\alpha \in (0, 1)$ (decay factor) and $p \leq k$ (number of supervised timesteps) control the shape of the weighting vector. At inference time, all auxiliary decoding heads are discarded, reducing the model to the standard SHRED pipeline with no additional computational overhead.

5. Experimental Evaluation

5.1. Double Gyre Flow

The proposed methods are first evaluated on the Double Gyre flow, a synthetic time-dependent two-vortex system with controlled heterogeneous missingness: one sensor experiences random dropouts (Bernoulli, $p = 0.25$), another undergoes a complete failure over a random contiguous interval, and a third remains fully available. All models share the same architecture (single Transformer layer, hidden size 32, 4 attention heads) and training setup. Table 1 reports Test MRE for the main model families.
All proposed Transformer-based masking configurations consistently achieve a Test MRE between 1.03% and 1.88%, yielding a peak relative improvement of approximately **83.8%** over the always-available sensor baseline. Furthermore, these masking strategies substantially outperform both classical imputation pipelines and specialized recurrent architectures like GRU-D, highlighting the clear advantage of explicitly incorporating sensor availability patterns directly into the Transformer encoder architecture.

5.2. Hydrodynamic model of the Southern North Sea

The second benchmark uses a real-world hydrodynamic dataset from the Southern North Sea (MIKE21 framework [4]), with four operational in situ stations (Dover 48.6%, ZeebruggeWielingendok 100%, NWA 75.7%, K141 99.8% availability) and three reconstructed fields: surface elevation (12 POD modes), U velocity (40 POD modes), and V velocity (68 POD modes). Unlike the synthetic benchmark, this dataset presents measurement noise, irregular sampling, communication delays, and real hardware failures.
Under a single-run evaluation (Table 2), the best configuration (T.3.G: channel-wise masking with learned MHSA masking) achieves a Test MRE of 15.73%, a relative improvement of **3.2%** over the always-available Transformer baseline (16.24%). The reduced margin reflects the higher intrinsic difficulty of the real-world task and the noise inherent in the measurements.

Cumulative error analysis. The advantage of the best masking configuration becomes increasingly pronounced at longer prediction horizons. At the final test timestep (step 3207), T.3.G achieves cumulative MRE reductions of **19.3%** for surface elevation, **3.3%** for U velocity, and **13.7%** for V velocity relative to the always-available baseline. This indicates that the proposed modifications are especially beneficial in long-horizon prediction regimes, where error accumulation is most severe.

Hyperparameter Tuning. A Bayesian hyperparameter search via Optuna [1] over 18 432 configurations confirms the robustness of these gains. The optimal configuration (T.3.G, encoder depth 4, hidden size 32, CNN decoder, dropout 0.1, lag 48) achieves a **2.9%** relative improvement in validation MRE and a consistent **3.3%** improvement in test MRE. A key practical finding is that no masking strategy universally dominates: attention-level and channel-level masking can be combined or deployed independently, with both configurations consistently among the top performers.

5.3. Multi-Step Decoding Results

Multi-step decoding is assessed exclusively on the most challenging scenario of the Southern North Sea dataset, using paired comparisons across independent runs and a grid search over the temporal weighting hyperparameters $\alpha \in \{0.2, \ldots, 0.8\}$ and $p \in \{4, \ldots, 20\}$. Across all paired runs, lagged training consistently achieves lower or comparable test MRE than the single-step baseline. With the best configuration ($\alpha = 0.7$, $p = 16$), the relative improvement reaches **1.89%** in Test MRE, with the gain maintained under heterogeneous missing-data conditions (T.3.G) confirming that the two contributions are mutually reinforcing.

Trajectory Smoothness. Beyond reconstruction accuracy, multi-step decoding induces substantially smoother latent trajectories: the mean total variation of the encoder representations decreases from $\mu_{\text{baseline}} = 70.55 \pm 11.84$ to $\mu_{\text{lagged}} = 33.59 \pm 7.25$, supporting the hypothesis that auxiliary supervision over intermediate timesteps encourages the encoder to maintain informative and temporally coherent latent representations throughout the lookback window.

Sensitivity analysis. The analysis reveals that reliable gains are concentrated in the region $\alpha \in [0.7, 0.8]$, $p \in [12, 20]$, corresponding to weight vectors that apply a moderate-to-shallow exponential decay over a sufficiently long temporal horizon. Poorly chosen hyperparameters (low α combined with short p) can yield mildly negative improvements, underscoring the importance of appropriate tuning.

6. Conclusions

This thesis addressed a central limitation of the SHRED framework in real-world deployments: its inability to handle incomplete, heterogeneous, and noisy sensor data. Two complementary contributions were proposed and validated on both a controlled synthetic benchmark and a real-world operational dataset.

The first contribution introduces a hierarchy of Transformer-based masking strategies that make the encoder explicitly aware of sensor availability. On the Double Gyre flow benchmark, the best configuration achieves an 83.8% relative improvement in Test MRE over the always-available sensor baseline, with all masking approaches consistently outperforming classical imputation and GRU-D-based methods. On the Southern North Sea hydrodynamic dataset, the best configuration (channel-wise masking combined with learned MHSA masking) achieves a 3.2% aggregate improvement and substantially larger reductions in cumulative error at long prediction horizons, confirmed by a systematic Bayesian hyperparameter search over 18 432 configurations.

The second contribution introduces a multi-step decoding training strategy that enforces structured latent representations by applying the shared decoder to all intermediate encoder timesteps during training and discarding the auxiliary heads at inference. This yields a 1.89% improvement in Test MRE, smoother latent trajectories (total variation halved), and sustained robustness under heterogeneous missing-data patterns, confirming the complementarity of the two proposed modifications.

Both contributions incur no additional computational overhead at inference time, offering a practical and principled methodology for deploying SHRED-based models in the heterogeneous sensing environments that characterize real-world applications. Future directions include extension to parametric settings for parameter estimation and uncertainty propagation; latent-space forecasting exploiting the smoother embeddings produced by multi-step decoding; and the replacement of the spatial decoder with geometry-aware architectures such as graph neural networks to improve generalization across varying mesh structures and resolutions.

References

[1] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework, 2019.

[2] I. Bey et al. Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. *Journal of Geophysical Research*, 106(D19):23073–23095, 2001.

[3] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

[4] DHI. Hydrodynamic model of the southern north sea: Mike21 model setup, outputs and observation data, March 2025.

[5] N. B. Erichson, L. Mathelin, Z. Yao, S. L. Brunton, M. W. Mahoney, and J. N. Kutz. Shallow neural networks for fluid flow reconstruction with limited sensors. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2238):20200097, 06 2020.

[6] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[7] J. N. Kutz. Data-driven modeling & scientific computation: Methods for complex systems & big data. *Journal of Statistical Software*, 67, 01 2015.

[8] R. Mitra, S. F. McGough, T. Chakraborti, C. Holmes, R. Copping, N. Hagenbuch, S. Biedermann, J. Noonan, B. Lehmann, A. Shenvi, et al. Learning from data with structured missingness. *Nature Machine Intelligence*, 5(1):13–23, 2023.

[9] A. Quarteroni, A. Manzoni, and F. Negri. *Reduced Basis Methods for Partial Differential Equations: An Introduction*, volume 92. Springer, 2015.

[10] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[11] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[12] Q. Tan et al. A review of irregular time series data handling with gated recurrent neural networks. *Neurocomputing*, 441:161–173, 2021.

[13] M. Tomasetto, J. P. Williams, F. Braghin, A. Manzoni, and J. N. Kutz. Reduced order modeling with shallow recurrent decoder networks. *Nature Communications*, 16(1):10260, 2025.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[15] J. P. Williams, O. Zahn, and J. N. Kutz. Sensing with shallow recurrent decoder networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 480(2298):20240054, 09 2024.

[16] A. Yermakov, D. Zoro, M. L. Gao, and J. N. Kutz. T-shred: Symbolic regression for regularization and model discovery with transformer shallow recurrent decoders, 2025.

[17] J. Yu and J. S. Hesthaven. Flowfield reconstruction method using artificial neural network. *AIAA Journal*, 57(2):482–498, 2019.

[18] X. Zhang, X. Liang, A. Zhiyuli, S. Zhang, R. Xu, and B. Wu. At-lstm: An attention-based lstm model for financial time series prediction. *IOP Conference Series: Materials Science and Engineering*, 569(5):052037, jul 2019.