Neural State Estimation for a Water Distribution Network: An Ablation Study and Gradient-Based Explainability

Anastasios Arsenos[1,*]

[1]ArtinisLabs SA, Greece – https://artinislabs.com

Abstract. We benchmark four families of neural state estimators for a nonlinear water distribution network (WDN) when only past and current actuation signals are available, i.e. without feeding back any measurement of the network state. Using a simulation of a three-conduit network with five exogenous inputs and ten internal states, we compare a memoryless MLP, a windowed MLP, an LSTM, a GRU, and a regularized Temporal Fusion Transformer (TFT) under an identical chronological split. The GRU with a 20-step input window reaches the best validation RMSE of 7.59×10^{-2} (validation $R^2 = 0.976$) with only 14,410 parameters, outperforming the LSTM, the windowed MLP and the TFT. We then probe the best GRU with permutation importance and gradient × input attribution; both methods agree that the upstream reservoir head u_1 dominates the prediction, in line with the physics of the network. Compact recurrent estimators thus appear to be a strong default for input-driven WDN state estimation, and gradient-based attribution is a cheap, model-agnostic way to add explainability to an already-trained estimator.

1 Introduction

Monitoring and control of water distribution networks requires real-time knowledge of internal hydraulic and quality states — pipe flow rates, nodal pressure heads and disinfectant concentrations [1, 2] — yet direct instrumentation of every conduit and node is rarely economical, motivating platform-level monitoring architectures [3] and fast data-driven surrogates that estimate the full state from a few boundary signals. We focus on the strictest variant of this problem, *input-only state estimation*: from a sliding window $\mathbf{u}_{t-k+1:t}$ the model must emit $\mathbf{x}_t$ without any state-feedback. Our contributions are (i) a controlled ablation across MLP, windowed MLP, LSTM, GRU and a regularised TFT on an identical chronological split; (ii) identification of the GRU at $k = 20$ as the most accurate and parameter-efficient estimator; and (iii) post-hoc explainability of the best GRU via permutation and gradient × input attribution that aligns with the physical role of each boundary signal.

2 Related Work

Early WDN surrogates used shallow networks for friction and demand prediction [5, 6]; deeper recurrent, convolutional and graph models have since been applied to pressure,

*e-mail: a.arsenos@artinislabs.com

demand and chlorine forecasting and to leak localisation [7–10]. Most assume state-feedback or dense instrumentation; the input-only formulation we adopt is closer to classical state observers, including the recent switching-observer soft sensor of [4], and to NARX/Hammerstein–Wiener system identification.

3 Methodology

Problem and dataset.

Let $\mathbf{u}_t \in \mathbb{R}^{n_u}$ and $\mathbf{x}_t \in \mathbb{R}^{n_x}$ be the exogenous inputs and network states at time t. We learn

$$\hat{\mathbf{x}}_t = f_{\boldsymbol{\theta}}(\mathbf{u}_{t-k+1:t}), \quad k \in \{1, 5, 10, 20\}, \tag{1}$$

minimising the MSE on standardised targets, with no state-feedback. Data are generated from a Simulink model of a three-conduit WDN with three reservoirs and chlorine transport (Fig. 1). Inputs switch at random times within [0, 1 200] s sampled uniformly from their admissible ranges, producing a 241-row record at $\Delta t = 5$ s. States are three pipe flow rates (x_1–x_3), the nodal pressure head (x_4) and six chlorine concentrations at the segment boundaries (x_5–x_{10}); inputs are the three reservoir heads (u_1–u_3), the aggregated demand (u_4) and the chlorine concentration injected at reservoir 1 (u_5). We use a chronological 70/15/15 split (rows 0–168 train, 169–204 validation, 205–240 test) and fit a z-score normaliser on the training section only.

Model families.

The *MLP* is two hidden ReLU layers, applied either to the current input ($k = 1$) or to a flattened window of past inputs. The *recurrent* models are a single GRU or LSTM layer with hidden size 64, followed by LayerNorm and a linear head on the last time-step. The *TFT* reference uses a Variable Selection Network over $\mathbf{u}$, a single-layer LSTM encoder, multi-head self-attention over the window, and gated residual networks; we report the regularised configuration ($d_{\text{hidden}} = 32$, $h = 2$, dropout 0.3, weight decay 10^{-3}), the only TFT variant that did not collapse on this small-data regime.

Training and explainability.

All models use AdamW (lr 10^{-3}, weight decay 10^{-5}, batch size 32), gradient clipping at $\|g\|_2 = 1$, and early stopping on validation MSE (patience 120, max 1 000 epochs) with a fixed seed. For the best model we compute two complementary signals on the test set: (i) *permutation importance*, replacing each input u_i across all 20 window steps with its training mean and recording the RMSE inflation Δ_i; and (ii) *gradient* × *input*, $|\partial \hat{x}_j / \partial u_{i,\tau}| \cdot |u_{i,\tau}|$, averaged over samples and targets to obtain a (k, n_u) saliency map.

4 Experiments and Results

Ablation.

Table 1 and Fig. 2(a) summarise the comparison. The memoryless MLP is a strong floor (val. RMSE = 0.216) but cannot disambiguate state trajectories sharing the same instantaneous inputs. Widening the MLP to a 10-step window improves validation but *worsens* test RMSE (0.430): the flattened representation triples the parameter count and overfits the 169

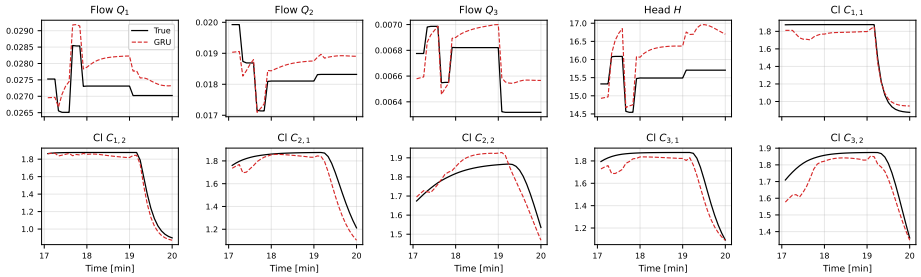
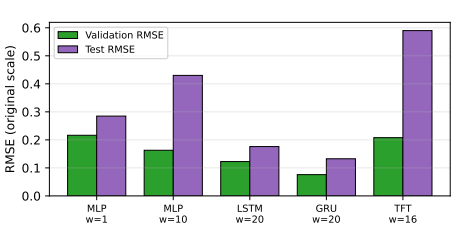


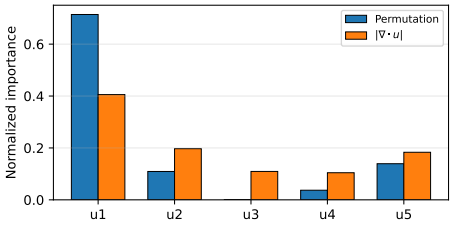
Figure 1. Test-set predictions of the best estimator (GRU with $k = 20$, hidden size 64) against the Simulink ground truth for all ten states.

Table 1. Ablation across model families. RMSE is averaged over the ten states in the original units. Parameter counts include the linear head.

Model	Window	Params	Val. RMSE	Test RMSE
MLP	1	5 194	0.2164	0.2849
MLP	10	15 434	0.1629	0.4301
LSTM ($h=64$)	20	18 954	0.1225	0.1761
GRU ($h=64$)	**20**	**14 410**	**0.0759**	**0.1325**
TFT (reg., $d=32$)	16	50 293	0.2078	0.5899



(a) RMSE per model.



(b) Input attribution of the best GRU.

Figure 2. (a) Validation and test RMSE for the five configurations of Table 1. (b) Normalised input attribution for the best GRU on the test set; permutation is the RMSE inflation when each u_i is replaced by its training mean, and $|\nabla \cdot u|$ is the gradient × input averaged over targets, samples and window positions.

training rows. Recurrent encoders correct this: both LSTM and GRU at $k = 20$ improve validation by another 30%–40% over the memoryless MLP while keeping a comparable parameter budget. The GRU is uniformly the best, with 14.4 k parameters reaching a validation R^2 of 0.976. The regularised TFT, even at 50 k parameters, only matches the memoryless MLP on validation, indicating that attention-heavy architectures need substantially more samples than the 169 available here. The negative test R^2 of the best GRU (−0.49) is dominated by the nodal head x_4, which accounts for $\approx$ 95% of the test SSE and is driven by an unseen excursion of the demand u_4 outside the training envelope; on the remaining nine states the test RMSE is within 40% of the validation RMSE, suggesting that the dynamics were learned and the failure mode is extrapolation rather than overfitting.

Explainability.

Fig. 2(b) compares the normalised per-input permutation and gradient × input scores of the best GRU on the test set. Both attributions place u_1 (the head of reservoir 1, hydraulically the dominant supplier and the source of the chlorinated stream) at the top. Permutation is sharper, attributing $\approx 71\%$ of the RMSE inflation to u_1 alone, while gradient × input is more diffuse ($\approx 41\%$) and gives non-trivial mass to every remaining input. The two signals agree on the ordering $u_1 \succ u_5 \succ u_2$ and on the near-irrelevance of u_3 in this dataset, consistent with the reservoir-3 head sitting at the downstream end of the network and acting only as a weak boundary condition. The full (k, n_u) saliency further reveals that $\approx 92\%$ of u_1's attribution comes from the most recent three window steps, while u_5 keeps a longer effective tail (only 47% from the last three steps), which is in line with the slower advective transport of chlorine relative to pressure.

5 Conclusion and Future Work

We presented an input-only neural state-estimation benchmark for a three-conduit WDN under a strict chronological evaluation protocol. A compact GRU outperforms an LSTM of comparable size, a windowed MLP and a regularised TFT, and remains amenable to lightweight gradient- and permutation-based explainability. The resulting attribution is physically plausible: the upstream reservoir head dominates, the chlorine inlet exhibits a longer effective memory than the heads, and the downstream boundary head is essentially redundant on this dataset. Future work includes multi-seed training with confidence intervals, a physics-informed loss tying x_4 to the conservation laws of the simulator to mitigate the extrapolation gap, and a scale-up to a larger EPANET network with graph-aware encoders.

References

[1] L.A. Rossman, *EPANET 2 Users Manual* (U.S. Environmental Protection Agency, Cincinnati, OH, 2000).
[2] P.F. Boulos, K.E. Lansey, B.W. Karney, *Comprehensive Water Distribution Systems Analysis Handbook for Engineers and Planners*, 2nd edn. (MWH Soft, Pasadena, 2014).
[3] D.G. Fragkoulis, F.N. Koumboulis, A.N. Menexis, M.P. Tzamtzi, N.D. Kouvakas, in *Proc. of the 3rd International Conference on Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary Applications (FAIEMA 25)* (2025) pp. 18–19.
[4] F.N. Koumboulis, D.G. Fragkoulis, N.D. Kouvakas, A. Feidopiasti, Sensors **23**, 2114 (2023).
[5] S. Lingireddy, D.A. Wood, Journal of Hydraulic Engineering **126**, 47 (2000).
[6] J. Zhang, M. Zhang, Water **10**, 1577 (2018).
[7] L. Romero-Ben, V. Puig, D. Rotondo, C. Ocampo-Martínez, Water Resources Research **56**, e2019WR026731 (2020).
[8] N. Mücke, S. Bohte, C. Oosterlee, Journal of Hydroinformatics **25**, 1 (2023).
[9] G. Hajgató, B. Gyires-Tóth, G. Paál, Engineering Applications of Artificial Intelligence **104**, 104336 (2021).
[10] Z. Li, H. Liu, C. Zhang, T. Fu, Water Research **217**, 118419 (2022).