Enabling Uncertainty-Aware Time-Series Forecasting in Federated Learning for Urban Water Dynamics

Golnoosh Abdollahinejad [*1 2] **Patrik Okanovic** [*3] **Denisa constantinescu** [1] **Sergey Shevchik** [2] **Torsten Hoefler** [3] **David Atienza** [1]

Abstract

Urban water forecasting is essential for flood awareness and resilient city operations. Yet, in real-world deployment, sensor data is fragmented due to utility privacy policies and sensor failures. Therefore, city-scale forecasting must operate across fragmented sensor ownership while providing calibrated uncertainty under sensor failures. To address this, we extend the functionality of the transformer forecasting model by introducing i) a quantile regression head for uncertainty calibration on forecasting intervals, and ii) a sensor-disjoint federated training scheme, scaled up to $1{,}000$ sensors across 100 clients without raw-data exchange in between, with remaining robustness to a Zipf-1 ownership skew. The quantile head preserves point-forecast absolute error from 0.0489 to 0.0338. The federated training retains this skill end-to-end. The developed framework is easily adjustable to local privacy policies, provides uncertainty-aware forecasting at its backbone, and applies to city-scale risk management. We map the architecture developed onto the canonical world-model recipe and identify the missing components for closed-loop urban water management.

1. Introduction

Motivation. Urban water utilities operate heterogeneous sensor networks across sewer collectors, retention basins, wastewater treatment plants, and combined sewer overflow outfalls. Their clients must decide in real time whether to close treatment-step inlets before a flood or to issue lake-side beach-closure alarms. Such decisions require a forecasting backbone calibrated so that a controller can threshold on overflow probability rather than rely on a point estimate. It must also be deployable at the urban scale without pooling raw data across clients. In addition, it should remain robust to client skew, where one large utility is often surrounded by many smaller ones with heterogeneous sensor footprints.

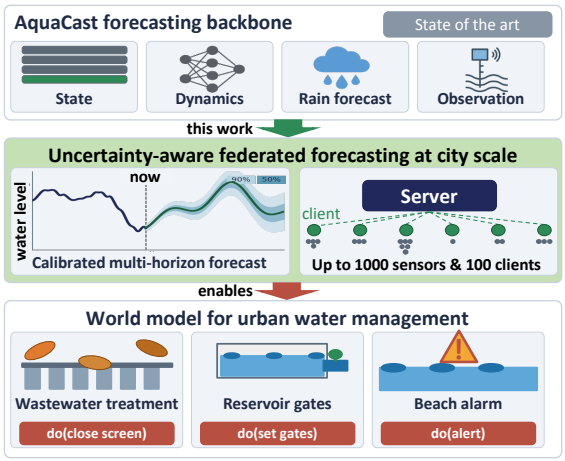

Figure 1. This work (middle box) extends AquaCast with calibrated quantile forecasts and federated training for heterogeneous clients, enabling a world model for urban water management.

Background. Probabilistic forecasting with quantile regression is now standard, supported by specialized tools for uncertainty estimation such as pinball loss (Koenker & Bassett, 1978), Temporal Fusion Transformers (Lim et al., 2021), quantile heads for foundation models (Ansari et al., 2024), and conformal coverage guarantees (Angelopoulos & Bates, 2023). Hydrological forecasting has adopted transformers, LSTMs, and hybrid backbones (Abdollahinejad et al., 2026; Feng et al., 2024). Multi-agent reinforcement learning and model predictive control for urban drainage report a 25.4% reduction in flooding-and-overflow volume *when fed accurate forecasts* (Ding et al., 2025; Tian et al., 2022). Federated learning (FL), via FedAvg (McMahan et al., 2017), averages local model updates without raw data

[1]Embedded System Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland [2]Laboratory for Advanced Materials Processing, Swiss Federal Laboratories for Materials Science and Technology (Empa), Thun, Switzerland [3]Department of Computer Science, ETH Zürich, Zürich, Switzerland. Correspondence to: Golnoosh Abdollahinejad <golnoosh.abdollahinejad@empa.ch>, Patrik Okanovic <patrik.okanovic@inf.ethz.ch>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

exchange, and recent hydrological FL studies demonstrate predictive value across distributed basins (Xie et al., 2024; Kerkez et al., 2016; Bartos et al., 2018). Yet the published water forecasting FL literature evaluates at most 18 clients under structurally homogeneous splits, one to two orders of magnitude below a real urban deployment and blind to the privacy policies that dominate in real-life (Appendix A.2).

Gap and Scope. No published system jointly delivers (a) quantile-calibrated multi-horizon trained federated forecasts, (b) at the scale of a real urban network ($\sim 1\,000$ sensors across up to 100 clients), and (c) under heterogeneous client splits, where clients differ in the number of sensors they hold. We frame this gap against the canonical world-model recipe (Ha & Schmidhuber, 2018; Hafner et al., 2025) and map our contributions onto it in Figure 1.

Contributions. This work is an extension of earlier developed AquaCast (Abdollahinejad et al., 2026) - a centrally trained precipitation-informed transformer, with three methodological contributions that jointly yield a forecasting backbone suitable for operational deployment:

- **Probabilistic multi forecasting head.** The point-regression head is replaced by a quantile head that produces calibrated predictive intervals over future water levels and discharges, letting downstream controllers condition decisions on exceedance probabilities rather than point estimates, as required for flood-warning, spilling basins, and closing beach interventions.
- **Federated training across distributed clients.** Training is recast in a strictly sensor-disjoint federated setting in which each client owns a private subset of the sensor network and the server does not observe raw measurements. We show that forecasting skill is preserved as the federation grows toward city-scale.
- **Robustness to heterogeneous client ownership.** Across all tested configurations, forecasting skill under a highly skewed client partition matches or exceeds the homogeneous equal-split baseline. This shows that realistic client skew is compatible with the proposed scheme and does not require artificial rebalancing. In practice, this means one dominant utility can be surrounded by many smaller ones without degrading performance.

2. Methodology

2.1. Metrics

Forecasting Skill. Point accuracy is reported via MSE, MAE, and RMSE (lower is better), and goodness of the fit via R^2 (higher is better). We additionally report the Nash–Sutcliffe efficiency (Gupta et al., 2009) (NSE$\in(-\infty, 1]$), which coincides with R^2 in this single-output setting (definition in Appendix A.1).

Probabilistic skill. For each quantile $q \in Q$ the pinball loss is defined as $\ell_q(y, \hat{y}) = \max\big(q(y-\hat{y}),\ (q-1)(y-\hat{y})\big)$. With $Z := \mathbb{E}_{b,t,v}[|y|] + \varepsilon$, we utilize the scaled quantile loss $\mathrm{SQL} = \mathbb{E}_q\big[\mathbb{E}_{b,t,v}[\ell_q(y, \hat{y}^{(q)})]\big]/Z$ and the median-only variant $\mathrm{SQL}_{\text{point}} = \mathbb{E}_{b,t,v}[\ell_{0.5}(y, \hat{y}^{(0.5)})]/Z$, where b, t, v index batch, time step and variable, $\hat{y}^{(q)}$ is the forecasted q-quantile and ε is a small numerical-stability constant. A gap $\mathrm{SQL} < \mathrm{SQL}_{\text{point}}$ shows that the non-median quantiles carry information beyond the median estimate.

Deployment Cost. Per FedAvg round, we report the per-client model size, the per-round communication payload, the client-side training time, and the server-side aggregation time. Together, these expose whether forecasting gains remain practical when clients face limited bandwidth and compute, and when the total number of training rounds is capped, a setting we refer to as the *round budget*.

2.2. Quantile Head

To enable probabilistic forecasting, the original deterministic output head of AquaCast was replaced with a multi-output quantile head that forecasts a set of conditional quantiles instead of a single point estimate. The model jointly estimates the quantile set $Q = \{0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95\}$ for each future time step and target variable. The median forecast ($q = 0.5$) is used as the central forecast, while the remaining quantiles provide calibrated uncertainty intervals around the forecasts. This allows the model to represent predictive uncertainty without assuming a parametric output distribution. Training is performed using the pinball loss, which is the standard objective for quantile regression. Centralized and federated comparisons against the original head are reported in Table 1 and Figure 4 (Section 3).

2.3. Federated Learning Setup

We train AquaCast with FedAvg (McMahan et al., 2023): C clients perform $E{=}1$ local epoch per round on their private sensors, and the server uniformly averages the parameters over $R{=}100$ rounds (Adam, lr 10^{-4}, batch 32; full description in Appendix A.4).

We adopt a *narrow* feature-partitioning scheme. Each client owns a disjoint subset of the N sensors; the encoder input width is $n_{\max}{+}1$, where $n_{\max}$ is the largest per-client subset and the $+1$ slot carries the shared rain forcing. Clients with $n_c{<}n_{\max}$ sensors zero-pad the unused slots, and the loss is masked so padded entries contribute neither to the objective nor to its gradients. The full construction, including the heterogeneous-split layout, is given in Appendix A.4.

We evaluate four configurations at two complexity regimes

(SynthLow, SynthMid (Abdollahinejad et al., 2026)) and two network sizes ($N \in \{100, 1000\}$), denoted **S1–S4** (S1, S2: N=100; S3, S4: N=1000). For each we sweep $C \in \{10, 50, 100\}$, ranging from centralized training to one-sensor-per-client at N=100, and preserving ten sensors per client even at C=100 for N=1000. This separates the effect of the *degree of partitioning* from that of *per-client signal density*. At each round, the server evaluates the global model on every client's held-out split.

2.4. Towards a World Model for Water Management

World models learn a compact predictive representation of environment dynamics, typically comprising (i) a latent state z_t, (ii) a transition model $p(z_{t+1} \mid z_t, a_t, u_t)$ under actions a_t and exogenous variables u_t, (iii) an observation head, (iv) a rollout/planning mechanism (Ha & Schmidhuber, 2018; Rezazadeh et al., 2025; Wang, 2024). AquaCast already supplies the dynamics, observation, and exogenous-driver components through its Transformer-based sequence modeling conditioned on precipitation, and the quantile head adds the calibrated-uncertainty component needed for probabilistic rollouts (Caetano et al., 2025; Tang & Matteson, 2021). Its internal representations serve as implicit state embeddings that can be formalized into a latent state, while its input structure naturally admits control variables (pump operations, reservoir releases) for intervention-aware dynamics. The missing pieces–an action interface, latent rollout, and reward signal–are discussed in Section 4.

3. Results

Quantile Head – Centralized. Table 1 compares the original AquaCast head with the proposed quantile head on real and synthetic datasets. On the real dataset, the quantile head improves probabilistic forecasting, achieving $\mathrm{SQL} = 0.1743$ compared to $\mathrm{SQL}_{\text{point}} = 0.2669$. The gap $\mathrm{SQL} < \mathrm{SQL}_{\text{point}}$ confirms that multiple quantiles provide meaningful uncertainty beyond the median forecast. This is accompanied by reduced point accuracy, with MSE increasing from 0.1925 to 0.2943 and R^2 decreasing from 0.6669 to 0.5373, reflecting the shift from mean-optimized to distribution-aware training. For operational water management, calibrated uncertainty is often more valuable than small point-wise error gains.

On the synthesized datasets, the quantile head improves both probabilistic and deterministic performance. For Lorenz Synth (Figure 4) and MeteoSwiss (Figure 6 in A.3), MSE decreases, NSE improves, and SQL remains lower than $\mathrm{SQL}_{\text{point}}$. This is expected since synthetic datasets are larger and smoother, while the real Lausanne dataset contains more noise, fewer sensors, and imperfect rainfall forecasts. Overall, the quantile head consistently improves uncertainty estimation and can also improve point forecasting when sufficient data coverage is available.

Table 1. Performance comparison of AquaCast with the original point head versus the quantile head, on the Lausanne real dataset (4 sensors) and on two synthesized datasets (S1, S2; 100 sensors); both history and future are 96 time steps in normalized scale.

Variant	MSE↓	MAE↓	RMSE↓	R^2 ↑	NSE↑	SQL↓	$\mathrm{SQL}_{\mathrm{p}}$↓
orig. head (real)	0.1925	0.2436	0.4374	0.6970	0.6669	–	–
quant. head (real)	0.2943	0.2702	0.5425	0.5373	0.5126	0.1743	0.2669
orig. head - S1	0.0489	0.0983	0.2211	0.9421	0.9373	–	–
quant. head - S1	0.0338	0.0548	0.1839	0.9599	0.9565	0.0308	0.0525
orig. head - S2	0.0347	0.1225	0.1864	0.9676	0.9630	–	–
quant. head - S2	0.0197	0.0828	0.1406	0.9815	0.9786	0.0223	0.0365

Scaling Federated Clients. For each of S1–S4, we sweep $C \in \{10, 50, 100\}$ with the quantile-head variant of AquaCast and report the best server-side per-partition MSE over R=100 rounds; the band shows dispersion across partitions. As C grows, each client observes a smaller subnetwork, yet the federated runs remain within a narrow MSE band of 0.18–0.31 for $C \geq 10$ and barely degrade from C=10 to C=100 (Figure 2). The N=1000 configurations (S3, S4) show visibly tighter bands than the N=100 ones because each partition retains many sensors. AquaCast thus retains the bulk of its centralized forecasting skill even when each client observes only a small fraction of the network.

Client Heterogeneity. Fixing C=10, we contrast the homogeneous split (each client owns N/C sensors) with a Zipf-1 partition in which sensor counts are highly skewed; the heterogeneous case uses the zero-pad-and-mask layout of Section 2.3. The heterogeneous setup matches or exceeds the homogeneous one on every configuration (Figure 5). We interpret this as the dominant client observing a near-centralized slice of the network dynamics and anchoring the global model, whereas averaging over many small, equally limited clients lacks such an anchor.

Deployment Metrics Evaluation. Beyond forecasting skill, we report per-round per-client communication volume (uplink + downlink, in MB) and wall-clock time decomposed into client-side training and server-side aggregation (Figure 3). Per-round payload drops rapidly with C: at N=1000, the model shrinks from $\approx$ 1.3 GB at C=1 to $\approx$ 24 MB at C=100 ($\sim$55×; derivation in Appendix A.5), because the encoder width `enc_in`=N/C+1 and the associated per-sensor embedding parameters scale with the per-client subnetwork. Client training time per round drops with less data, while server time grows since the server now runs per-partition evaluation on C test splits in addition to averaging C parameter vectors. This is the central deployment trade-off: scaling across clients lowers per-client cost at the cost of heavier server work.

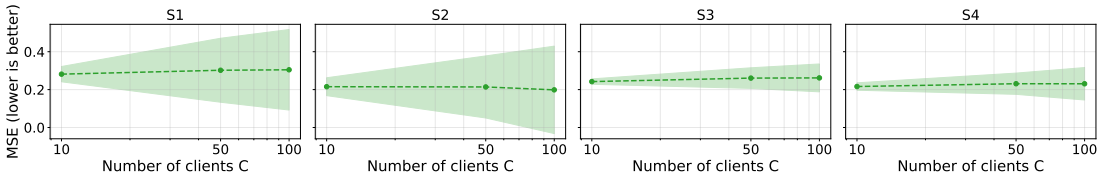

Figure 2. Server-side per-partition MSE (quantile head) versus number of federated clients C on S1-S4.

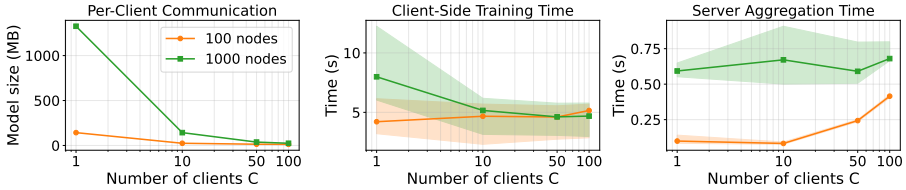

Figure 3. Deployment metrics versus number of clients C at N=100 and N=1000.

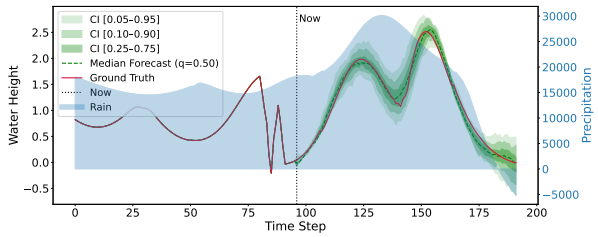

Figure 4. Sample 96-step forecast on the S2 dataset; shaded band shows the Q7 quantile envelope around median forecast.

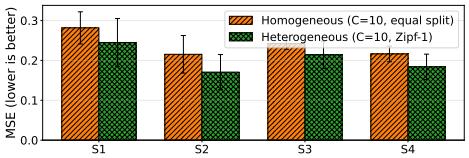

Figure 5. Homogeneous versus heterogeneous Zipf-1 client partitioning at C=10 across S1–S4.

4. Discussion

The three federated experiments jointly show that strict sensor-disjoint training is a practical substrate for urban water forecasting at scale, not a costly compromise. Forecasting skill is preserved up to C=100 clients with no raw-data exchange (Figure 2), and degrades most gracefully in the N=1000 regime where each client still retains a sizeable subnetwork. Realistic client skew does not penalize this behavior: under a Zipf-1 split, the federated model matches or exceeds the homogeneous baseline, with MSE reductions of 0.03–0.05 across S1–S4 (Figure 5), which we attribute to a dominant client anchoring aggregation (Section 3). The deployment metrics (Figure 3) make the design space explicit: increasing C shrinks per-client model size, bandwidth, and local compute at the cost of heavier server-side aggregation, so C is chosen jointly from privacy boundaries and server capacity rather than as a modeling hyperparameter. Combined with the quantile head, which adds calibrated predictive intervals at a small additional cost, this yields a forecasting backbone that real utilities can plausibly deploy: privacy-preserving by construction, uncertainty-aware by design, and robust to the ownership skew that prior FL hydrology benchmarks do not test.

Limitations and Future Work. Three limitations frame the next steps. First, the federated layer relies on FedAvg with uniform averaging and assumes honest-but-curious clients: we provide no formal privacy bound on shared parameter updates, do not couple aggregation with differential privacy or secure aggregation, and do not evaluate robustness to Byzantine or adversarial participants; all standard concerns for production deployment across competing utilities. Second, the evaluation is synthetic-dominant: real data is restricted to the four-sensor Lausanne subnetwork, while the $C\in\{1, 10, 50, 100\}$ sweeps at $N\in\{100, 1000\}$ use MeteoSwiss-synthesized and Lorenz trajectories. No multi-utility real federation is exercised, and geographic generalization across catchments with different hydrological regimes remains untested. Third, the action interface, latent rollout, and reward signal of Figure 1's bottom tier are still absent, and empirical interval coverage is not yet validated (e.g., via conformal calibration), both of which are required to reach closed-loop, federated decision support.

5. Conclusion

We extended AquaCast with a pinball-loss quantile head that produces calibrated predictive intervals across three datasets, with strict point-error improvements on the synthetic benchmarks and only a small trade-off on real Lausanne data. A sensor-disjoint federated training scheme preserves forecast skill as the federation grows to 100 cooperating clients on networks of up to 1,000 sensors, while the per-client communication payload contracts by approximately 55× at the highest client count. Robustness to realistic client skew is further

established: a Zipf-1 partition consistently matches or improves on the homogeneous equal-split baseline. Within the canonical world-model recipe, the present contributions populate the uncertainty and deployable-training components; the remaining action interface, latent rollout, and reward signal define the natural next steps toward closed-loop, federated decision support for urban water infrastructure.

Impact Statement

This paper presents work aimed at advancing the field of Machine Learning. The application domain of this work, urban water management, carries clear public-safety benefits, since calibrated, federated forecasts can support flood preparedness, basin-spill avoidance, and lake-side beach-closure decisions across clients that cannot share raw sensor data.

References

Abdollahinejad, G., Baghersalimi, S., Constantinescu, D.-A., Shevchik, S., and Atienza, D. Aquacast: Urban water dynamics forecasting with precipitation-informed multi-input transformer. *Journal of Hydrology*, 670:135177, 2026. doi: 10.1016/j.jhydrol.2026.135177. URL https://www.sciencedirect.com/science/article/pii/S002216942600274X.

Angelopoulos, A. N. and Bates, S. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101. URL https://www.nowpublishers.com/article/Details/MAL-101.

Ansari, A. F., Stella, L., Turkmen, C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S. S., Pineda Arango, S., Kapoor, S., Zschiegner, J., Maddix, D. C., Mahoney, M. W., Torkkola, K., Wilson, A. G., Bohlke-Schneider, M., and Wang, Y. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. doi: 10.48550/arXiv.2403.07815. URL https://arxiv.org/abs/2403.07815.

Bartos, M., Wong, B., and Kerkez, B. Open storm: A complete framework for sensing and control of urban watersheds. *Environmental Science: Water Research & Technology*, 4(3):346–358, 2018. doi: 10.1039/C7EW00374A.

Caetano, R., Oliveira, J., and Ramos, P. Transformer-based models for probabilistic time series forecasting with explanatory variables. *Mathematics*, 2025. URL https://www.mdpi.com/2227-7390/13/5/814.

Chen, W., Nguyen, K. A., and Lin, B.-S. Rethinking evaluation metrics in hydrological deep learning: Insights from torrent flow velocity prediction. *Sustainability*, 2025. doi: 10.3390/su17198658.

Ding, W., Tian, W., Liao, Z., Wang, T., and Wang, X. Pollution-based integrated real-time control for urban drainage systems: A multi-agent deep reinforcement learning approach. *npj Clean Water*, 2025. doi: 10.1038/s41545-025-00512-z. URL https://www.nature.com/articles/s41545-025-00512-z.

Duc, L. and Sawada, Y. A signal-processing-based interpretation of the nash–sutcliffe efficiency. *Hydrology and Earth System Sciences*, 2023. doi: 10.5194/hess-27-1827-2023.

Farooq, M. S., Tehseen, R., Qureshi, J. N., Omer, U., Yaqoob, R., Tanweer, H. A., and Atal, Z. Ffm: Flood forecasting model using federated learning. *IEEE Access*, 11:24472–24483, 2023. doi: 10.1109/ACCESS.2023.3252896.

Feng, Z.-k., Zhang, J.-s., and Niu, W.-j. A state-of-the-art review of long short-term memory models with applications in hydrology and water resources. *Applied Soft Computing*, 167:112352, 2024. doi: 10.1016/j.asoc.2024.112352. URL https://www.sciencedirect.com/science/article/abs/pii/S1568494624011268.

Gupta, H., Kling, H., Yilmaz, K., and Martinez, G. Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377:80–91, 2009. doi: 10.1016/j.jhydrol.2009.08.003.

Ha, D. and Schmidhuber, J. World models. *arXiv preprint arXiv:1803.10122*, 2018. doi: 10.48550/arXiv.1803.10122. URL https://arxiv.org/abs/1803.10122.

Hafner, D., Pašukonis, J., Ba, J., and Lillicrap, T. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2. URL https://www.nature.com/articles/s41586-025-08744-2.

Kerkez, B., Gruden, C., Lewis, M., Montestruque, L., Quigley, M., Wong, B., Bedig, A., Kertesz, R., Braun, T., Cadwalader, O., Poresky, A., and Pak, C. Smarter stormwater systems. *Environmental Science & Technology*, 50(14):7267–7273, 2016. doi: 10.1021/acs.est.5b05870.

Knoben, W., Freer, J., and Woods, R. Technical note: Inherent benchmark or not? comparing nash–sutcliffe and

kling–gupta efficiency scores. *Hydrology and Earth System Sciences*, 2019. doi: 10.5194/hess-23-4323-2019.

Koenker, R. and Bassett, G. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. doi: 10.2307/1913643.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764, 2021. doi: 10.1016/j.ijforecast.2021.03.012.

Lin, Y., Wang, D., Jiang, T., and Kang, A. Assessing objective functions in streamflow prediction model training based on the naïve method. *Water*, 2024. doi: 10.3390/w16050777.

Mahir, S. H., Tashrif, M. T. A., Karim, M. A., Kundu, D., Rahman, A., Hamza, M. A., Farid, F. A., Miah, A. S. M., and Mansor, S. Advanced hydro-informatic modeling through feedforward neural network, federated learning, and explainable ai for enhancing flood prediction. *IEEE Open Journal of the Computer Society*, 6:726–738, 2025. doi: 10.1109/ojcs.2025.3556424.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and Agüera y Arcas, B. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pp. 1273–1282. PMLR, 2017. doi: 10.48550/arXiv.1602.05629. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data, 2023. URL https://arxiv.org/abs/1602.05629.

Raza, M. O., Umar, A., Rasheed, J., Asuroglu, T., and Alsubai, S. Federated transfer learning for distributed drought stage prediction. *Discover Artificial Intelligence*, 5(1):55, 2025. doi: 10.1007/s44163-025-00288-8. URL https://link.springer.com/article/10.1007/s44163-025-00288-8.

Rezazadeh, F., Chergui, H., Debbah, M., and Song, H. Agentic world modeling for 6g: Near-real-time generative state-space reasoning. *arXiv preprint arXiv:2511.02748*, 2025. URL https://arxiv.org/pdf/2511.02748.

Tang, B. and Matteson, D. Probabilistic transformer for time series analysis. In *NeurIPS*, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c68bd9055776bf38d8fc43c0ed283678-Paper.pdf.

Tian, W., Liao, Z., Zhang, Z., Wu, H., and Xin, K. Combined sewer overflow and flooding mitigation through a reliable real-time control based on multi-reinforcement learning and model predictive control. *Water Resources Research*, 58(7):e2021WR030703, 2022. doi: 10.1029/2021WR030703.

Wang, Y. e. a. Timexer: Empowering transformers for time series forecasting with exogenous variables. In *NeurIPS*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/0113ef4642264adc2e6924a3cbbdf532-Paper-Conference.pdf.

Wang, Y.-Q., Wang, H., Wang, W.-Z., Yang, H.-L., ji Chen, J., Fan, Y.-X., Yin, W.-X., qiang Lv, J., Luo, X.-Q., Zhou, X., and Wang, A.-J. Federated machine learning enables risk management and privacy protection in water quality. *Environmental science & technology*, 2025. doi: 10.1021/acs.est.5c00369.

Xia, Y., Feng, H., Ge, Q., Rodrigues, J., Gadekallu, T. R., and Fang, K. Federated learning based water streamflow forecasting via multi-sensor data fusion. *Information Fusion*, 129:104020, 2026. doi: 10.1016/j.inffus.2025.104020. URL https://www.sciencedirect.com/science/article/abs/pii/S1566253525010826.

Xie, Z., Zhang, X., Wang, Y., Jie, X., Fang, W., and Cai, Y. Improving adaptive runoff forecasts in data-scarce watersheds through personalized federated learning. pp. 180–198, 2024. doi: 10.1007/978-3-031-78183-4_12.

A. Appendix

A.1. Nash–Sutcliffe Efficiency

NSE is dimensionless and ranges in $(-\infty, 1]$, with 1 a perfect fit and 0 matching the observed mean (Gupta et al., 2009). For the single-output forecasting setting considered here, NSE coincides with R^2:

$$R^2 = \mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

where y_i, $\hat{y}_i$, and $\bar{y}$ are the observed value, the forecast, and the mean of the observations. NSE is sensitive to observed-variance regimes; we therefore report it alongside MSE, MAE, RMSE, and AUC.

Interpretation. NSE is commonly interpreted relative to a mean-observation benchmark: values close to 1 indicate very good performance, values greater than 0 indicate that the model performs better than forecasting the observed mean, and values below 0 indicate worse-than-mean performance. Its main strengths are that it is dimensionless, making it comparable across sites and scales, and that it emphasizes the reproduction of variability, including peaks and troughs, rather than only average levels (Gupta et al., 2009; Duc & Sawada, 2023; Chen et al., 2025). At the same time, NSE has important limitations: it is sensitive to the variance of the observations, may appear favorable in highly variable regimes despite sizable errors, and may appear unfavorable in low-variance regimes despite small absolute errors (Chen et al., 2025; Gupta et al., 2009). Optimizing only for NSE can underestimate flow variability and introduce calibration bias. At the same time, fixed thresholds such as "NSE > 0.5 is acceptable" can be misleading if RMSE, MAE, bias, and application context are ignored (Gupta et al., 2009; Lin et al., 2024; Chen et al., 2025; Knoben et al., 2019). We therefore use NSE as one metric among several, rather than relying on it alone for model selection (Chen et al., 2025; Duc & Sawada, 2023).

A.2. Prior FL Hydrology Work

Table 2. Forecasting-oriented prior work and benchmark configurations considered in this study. NR: not reported.

Study (Year)	Setting	FL?	# Clients	Metrics	Key result
FedMSF (2026) (Xia et al., 2026)	Streamflow forecasting	Yes	NR	RMSE, MAE, NSE	RMSE 0.4322, MAE 0.2358, and NSE 0.8371 on distributed hydrologic data.
Raza et al. (2025) (Raza et al., 2025)	Drought-stage prediction	Yes	NR	Task-specific prediction metrics	Demonstrates federated transfer learning for distributed drought-stage prediction under decentralized data constraints.
Mahir et al. (2025) (Mahir et al., 2025)	Flood prediction	Yes	NR	R^2, MAE, RMSE, PBIAS	Best FNN under FL: $R^2 = 0.96$, MAE $= 0.02$, PBIAS $= 0.4185$, with lower RMSE than alternative deep models.
Siva et al. (2024) (Farooq et al., 2023)	Flood forecasting	Yes	18	Accuracy	Federated FFNN model over 18 client stations achieves $\approx 84\%$ accuracy for 5-day flood alarms; CNN2D variants exceed 90% accuracy.
AWSFA (2025) (Wang et al., 2025)	WWTP water-quality forecasting	Yes	6	MAPE, MSE, RMSE, MAE, adj.-R^2	Global AWSFA model: MAPE $\approx 4.93\%$ (no disturbance), rising to 7.5–10% under 10–50% feature disturbance.
FedHydroDSW (2024) (Xie et al., 2024)	Runoff forecasting (CAMELS)	Yes	NR	NSE, RAE, RMSE	Personalized FL yields on average 30.1% higher NSE and 26.6% lower RMSE than standalone and conventional FL across multiple data-scarce basins.
AquaCast (2026) (Abdollahinejad et al., 2026)	Synthetic MeteoSwiss (SynthLow), urban wastewater / drainage	No	100	MSE, MAE, RMSE, R^2, AUC	1d: 0.0489, 0.0983, 0.1962, 0.9374, 0.9004 2d: 0.0502, 0.0997, 0.1940, 0.9313, 0.8974 5d: 0.0915, 0.1354, 0.2624, 0.8659, 0.8876 7d: 0.1964, 0.2290, 0.4252, 0.7093, 0.8922
This: AquaCast + quantile head, centralized (Lausanne real, 4 sensors)	Real Lausanne sensor network	No	1	MSE, MAE, RMSE, R^2, NSE, SQL, $\mathrm{SQL}_{\mathrm{point}}$	MSE 0.2943, MAE 0.2702, RMSE 0.5425, R^2 0.5373, NSE 0.5126, SQL 0.1743, $\mathrm{SQL}_{\mathrm{point}}$ 0.2669 (Table 1).
This: AquaCast + quantile head, centralized (synthetic S1/S2, 100 sensors)	MeteoSwiss + Lorenz synthesized	No	1	MSE, MAE, RMSE, R^2, NSE, SQL, $\mathrm{SQL}_{\mathrm{point}}$	S1 (MeteoSwiss): MSE 0.0338, NSE 0.9565, SQL 0.0308. S2 (Lorenz): MSE 0.0197, NSE 0.9786, SQL 0.0223. Quantile head improves *both* point and probabilistic skill over the deterministic head (Table 1).
This: AquaCast + quantile head, federated (S1–S4, $N \in \{100, 1000\}$)	Public synthetic benchmark, sensor-disjoint partitioning	Yes	$C \in \{1, 10, 50, 100\}$	Per-partition MSE; bandwidth and wall-clock time	Best per-partition MSE stays in 0.18–0.31 for $C \geq 10$ across S1–S4; Zipf-1 ownership skew matches or improves on the homogeneous split; per-client payload contracts $\sim 55\times$ from $C{=}1$ to $C{=}100$ at $N{=}1000$ (Figures 2–3).

A.3. Extra Quantile Comparison

Due to lack of space we moved extra visualization here in Figure 6.

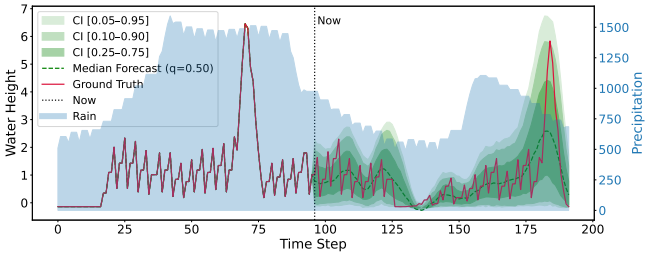

Figure 6. Sample 96-step forecast on the S1 dataset; shaded band shows the Q7 quantile envelope around median forecast.

A.4. Federated Learning Setup – Full Description

FedAvg Update. At each round t the server broadcasts a global parameter vector θ_t; each client c performs E local optimizer steps on its private data to produce $\theta_{t+1}^{(c)}$, and the server aggregates $\theta_{t+1} = \frac{1}{C}\sum_{c=1}^{C}\theta_{t+1}^{(c)}$ via uniform averaging. All clients and the server share a single architecture, and no intermediate activations are exchanged. We use R=100 rounds, E=1 local epoch per round, Adam with learning rate 10^{-4}, and batch size 32, held constant across all configurations.

Narrow Partitioning with Per-Client Encoder Width. With N sensors and C clients, each client owns a disjoint subset of sensors; let $n_{\max}$ denote the largest such subset. The global encoder has input width $\texttt{enc_in} = n_{\max} + 1$: the $n_{\max}$ owned-sensor slots plus one slot for the shared rain forcing, which is broadcast to every client. In a homogeneous split, each client owns N/C sensors, giving $\texttt{enc_in} = N/C + 1$. In a heterogeneous split, a client owning $n_c < n_{\max}$ sensors places its real values at slots $0, \ldots, n_c{-}1$, zero-pads the remaining $n_{\max} - n_c$ slots, and the prediction loss is masked to the client's owned outputs so padded slots contribute neither to the loss nor to its gradients. This isolates feature partitioning from model capacity and lets communication and compute scale with $n_{\max}$ rather than N, while preserving a like-for-like comparison to C=N (one sensor per client). The cost of the strictly sensor-disjoint layout is that the model cannot exploit raw cross-client correlations during training; Section 3 shows that this cost is small in practice.

Server-Side Per-Partition Evaluation. At every round the server evaluates the global model independently on each client's held-out test split, applying the same input zero-padding and output masking as during training. The primary forecasting score is the mean of the per-partition MSEs and reflects the deployed regime in which each client queries the model only on its own subnetwork. The full per-partition array is retained so that cross-partition dispersion can be reported without additional inference passes.

A.5. Federated scaling: additional details

Per-Round Bandwidth Derivation. The per-round per-client communication payload (uplink + downlink) is twice the model size in MB. Because we adopt the narrow partitioning of Section A.4, encoder input width is $\texttt{enc_in} = n_{\max} + 1$, so per-sensor embedding parameters, and the corresponding share of the model, scale with the largest per-client subnetwork rather than with N. In the homogeneous N=1000 sweep, this yields exact model sizes of 1326 MB at C=1, dropping to roughly 24 MB at C=100 — a 55× reduction in per-client bandwidth. Compute behaves analogously: each client backpropagates through a model whose input tensor shrinks with C, so client time drops while server-side cost grows linearly in C from two sources, parameter averaging across C clients and per-partition evaluation over C test splits, the latter dominating.

Why Heterogeneity Helps. The robustness of forecasting skill under Zipf-1 skew (Section 3) admits a simple interpretation. Under a skewed partition, the dominant client owns a large share of the sensors and therefore observes a slice of the network dynamics that is closer to the centralized regime than any single client under the homogeneous split. FedAvg aggregation inherits much of that client's near-centralized signal, and the global model is effectively anchored by it, with smaller clients contributing local corrections rather than dragging the global model toward an equally limited consensus. In the homogeneous split, this anchor is absent, since all clients are equivalent and equally narrow.