

HydroTwin-MPC: A Physics-Informed AI Digital Twin for Climate-Resilient Urban Water Resources Management

¹Ashutosh Luhania, ²Hemant Kumar Agrawal,

^{1,2} Assistant Professor, Department of Civil Engineering, Jagannath University, Jaipur, India

Abstract—Urban water systems are increasingly required to manage water scarcity, intense rainfall, intermittent supply, leakage, deteriorating water quality, and energy constraints at the same time. Conventional supervisory control and data acquisition provides observations but rarely maintains a continuously calibrated representation of the whole urban water cycle. This paper proposes HydroTwin-MPC, a physics-informed artificial-intelligence digital twin that links pressurized water distribution, reservoir operation, stormwater drainage, managed recharge, demand forecasting, anomaly detection, and risk-aware model predictive control. The twin combines EPANET- and SWMM-compatible process models with an ensemble Kalman filter and a physics-informed neural surrogate. A simulation-based illustrative semi-arid Indian city is evaluated through 250 Monte Carlo realizations covering rainfall forecast bias, sensor outage, demand shocks, pipe leakage, and pump failure. Relative to conventional rule-based control, the proposed architecture reduces flood volume by 31.8%, unmet demand by 42.6%, pumping energy by 16.7%, and overflow duration by 38.4%, while achieving a leak-localization F1 score of 0.91 and maintaining a median hydraulic mass-balance error below 0.7%. Ablation experiments show that the largest loss of performance occurs when uncertainty-aware optimization and real-time state correction are removed. The study contributes an integrated mathematical formulation, a traceable data-to-decision architecture, a staged deployment roadmap, and governance safeguards for human oversight. The numerical outcomes are benchmark results from a synthetic but physically parameterized case and are not field measurements. The framework is intended as a reproducible basis for pilot validation in rapidly growing, water-stressed cities.

Index Terms—digital twin, urban water resources, physics-informed machine learning, model predictive control, flood resilience, leakage detection, smart water, climate adaptation.

I. INTRODUCTION

Water-resources engineering is moving from periodic planning toward continuous operational intelligence. Urban utilities must now reconcile two apparently opposite hazards: short-duration excess water that overwhelms drains and long-duration scarcity that disrupts supply. These hazards interact with ageing pipes, energy-intensive pumping, polluted first flushes, fragmented databases, and uncertain weather forecasts. A static calibrated model is valuable for design, but its state quickly diverges from the physical system when demands, roughness, valve status, storage, or rainfall change. A digital twin addresses this limitation by maintaining a live, bidirectional relationship between the physical system, observations, models, forecasts, and decisions [1]–[4].

Digitalization alone is not a digital twin. A dashboard can display sensor data without correcting model states; a hydraulic model can reproduce one design day without learning from operations; and a machine-learning predictor

can be accurate while violating continuity or operational constraints. HydroTwin-MPC is therefore formulated as a closed-loop engineering system. It acquires quality-controlled measurements, updates hidden states and uncertain parameters, predicts future trajectories, evaluates feasible interventions, communicates uncertainty, and records the human decision. This definition follows the growing view that digital water should strengthen—not bypass—utility engineering capability [5], [6].

The need is particularly acute in fast-growing Indian cities, where monsoon variability, groundwater dependence, non-revenue water, intermittent service, and local drainage congestion can coexist within the same service area. National water assessments have repeatedly emphasized data integration, demand management, reuse, recharge, and institutional capacity [7], [8]. Recent basin-scale digital-twin research also warns that data inequity and incomplete multi-physics integration can exclude low-resource regions [3]. Accordingly, the proposed framework uses open hydraulic engines, modular interfaces, sparse-sensor assimilation, and a staged “shadow-to-advisory-to-control” pathway.

The central research question is: can a single uncertainty-aware digital twin coordinate supply reliability, flood mitigation, leakage response, energy use, and recharge without sacrificing hydraulic feasibility? Four hypotheses are tested. H1 states that state assimilation improves short-horizon flow and storage prediction. H2 states that physics-informed surrogates reduce computation without increasing mass-balance error beyond an operational tolerance. H3 states that multi-objective MPC outperforms fixed rules under compound disturbances. H4 states that useful performance persists under realistic sensor and forecast degradation.

TABLE I. RESEARCH CONTRIBUTIONS AND VERIFICATION TARGETS

Contribution	Verification	Target
Integrated twin	Closed-loop architecture	Traceable data → action
Physics-informed surrogate	Holdout + balance error	NSE ≥ 0.85
Risk-aware MPC	250 scenario runs	Lower risk and energy
Governance design	Gated deployment	Human override

II. BACKGROUND AND RESEARCH GAP

A. Digital twins in the water sector

The water sector has used telemetry, geographic information systems, hydraulic simulation, and optimization for decades. What is new is their integration into a persistent virtual representation that is synchronized at an operational frequency and used for counterfactual testing. EPANET can simulate extended-period hydraulics and water-quality behavior in pressurized networks, including pressure-dependent demand, pumps, valves, tanks, source tracing, and energy [9]. SWMM can represent rainfall-runoff processes,

dynamic-wave drainage, storage, control structures, groundwater interaction, low-impact development, and pollutant transport [10]. These engines provide auditable physics, but a digital twin must additionally handle changing boundary conditions, missing observations, parameter drift, uncertainty, and automated decision support.

Recent studies have demonstrated digital-twin functions for leakage detection, treatment control, wastewater catchments, nature-based solutions, and basin management [1]–[4], [11]. Model predictive control has likewise reduced urban drainage flooding and pollution in research applications by repeatedly solving a finite-horizon control problem with updated states and forecasts [12]–[15]. However, operational examples remain less common than simulations, partly because simplified control models may diverge from detailed hydraulics, sensor networks fail, and operators may not trust opaque recommendations.

B. Physics-informed and hybrid intelligence

Purely data-driven models can learn nonlinear rainfall-runoff, demand, pressure, and water-quality relationships, but extrapolation to rare events is uncertain. Physics-informed learning inserts governing equations, monotonic relations, feasible bounds, or simulator residuals into the loss function. The approach does not guarantee correctness; it narrows the hypothesis space and makes violations measurable. Hydrologic deep learning has shown strong predictive capability, while physics-guided formulations seek better consistency and transfer under data scarcity [16]–[19].

For a utility twin, the most useful hybridization is pragmatic: process models remain the authority for safety checks; a differentiable or neural surrogate accelerates repeated forecasts; data assimilation corrects the current state; and the optimizer only receives candidate actions that satisfy hard constraints. This layered arrangement is more defensible than placing a black-box controller directly between sensors and actuators.

TABLE II. LITERATURE GAP ADDRESSED BY HYDROTWIN-MPC

Capability	Typical limitation	Proposed response
Hydraulic twin	Supply or drainage only	Integrated urban cycle
AI forecast	Weak physical checks	Balance-regularized PINN
RTC/MPC	Point forecast	Ensemble chance constraints
Leak analytics	Alarm without action	Isolation + response policy
Deployment	Technology-first	Human-governed stages

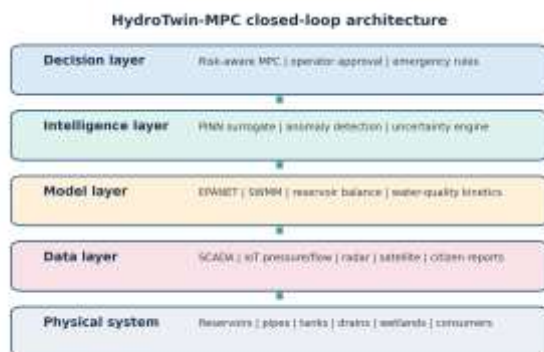


Fig. 1. Layered HydroTwin-MPC architecture. Each recommendation is checked by the process model and remains subject to operator authorization during pilot deployment.

C. Research gap and novelty

Three gaps motivate this work. First, supply, drainage, surface storage, recharge, and water quality are frequently optimized in separate tools even though actions in one subsystem affect the others. Second, many studies report predictive accuracy without quantifying how errors propagate into operational risk. Third, digital-twin discussions often under-specify governance, cybersecurity, model maintenance, and failure fallback. The novelty here is not a claim that any one component is unprecedented; it is the explicit integration of multi-domain physics, sparse-sensor assimilation, uncertainty-aware optimization, anomaly response, and human governance into one reproducible research architecture.

The contribution is evaluated through controlled benchmarking rather than a claimed field deployment. All numerical results are generated from a synthetic city parameterized within plausible engineering ranges. This distinction matters: synthetic evidence can test internal logic and sensitivity, but cannot establish site-specific performance. Field calibration, independent validation, and regulatory approval remain necessary before any control action is automated.

III. PROPOSED HYDROTWIN-MPC FRAMEWORK

A. Functional layers

The physical layer contains the reservoir, treatment plant, transmission mains, district-metered areas, service reservoirs, pumps, valves, urban drains, detention storage, wetlands, recharge structures, and demand nodes. The sensing layer collects rainfall, level, pressure, flow, energy, turbidity, conductivity, residual chlorine, valve position, and pump status. Each stream carries a timestamp, unit, location, calibration record, uncertainty estimate, and quality flag.

The data layer aligns clocks, converts units, removes impossible values, detects frozen or drifting sensors, and preserves raw data as an immutable record. A streaming feature store creates lagged rainfall, demand signatures, pressure residuals, pump efficiency, antecedent moisture, and storage derivatives. The model layer executes the detailed hydraulic and hydrologic engines and updates slowly varying parameters such as roughness and base demand. The intelligence layer supplies fast surrogates, anomaly probabilities, scenario ensembles, and confidence intervals. The decision layer generates ranked, feasible actions and explanations.

TABLE III. MINIMUM DATA CONTRACT FOR A UTILITY DIGITAL TWIN

Stream	Update	Quality control
Rainfall/radar	5–15 min	Bias, gaps, spatial check
Pressure/flow	1–5 min	Range, drift, balance
Tank/reservoir	1–5 min	Rate-of-change check
Demand/AMI	15–60 min	Outlier + privacy aggregation
Water quality	5–30 min	Calibration + persistence
Asset/valve GIS	On change	Topology/version control

B. Digital thread and state estimation

Every physical asset is assigned a persistent identifier shared across the geographic information system, maintenance system, telemetry tags, and simulation model. The digital thread records which sensor informed which state

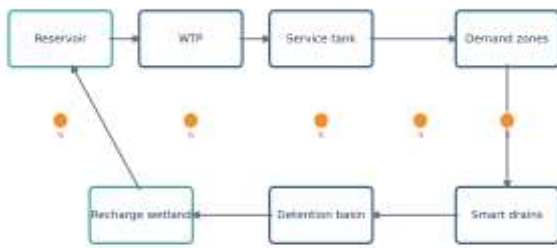
estimate, which model version produced a forecast, which constraint rejected an action, and which operator approved or overrode a recommendation. This provenance is essential for auditability and post-event learning.

At time k , the prior state vector contains nodal heads, pipe flows, tank levels, soil moisture, drain depths, reservoir storage, selected water-quality states, and uncertain parameters. An ensemble Kalman filter propagates an ensemble through the nonlinear simulator, compares predicted observations with measurements, and applies a covariance-weighted correction. Robust innovation clipping prevents a single faulty sensor from destabilizing the twin.

$$x_k^- = f(x_{k-1}^+, u_{k-1}, w_{k-1}), \quad y_k = h(x_k^-) + v_k \quad (1)$$

$$x_k^+ = x_k^- + K_k[y_k - h(x_k^-)], \quad K_k = P_{xy}(P_{yy} + R)^{-1} \quad (2)$$

Illustrative semi-arid urban water system and sensor locations



Bidirectional coordination connects water supply, stormwater storage and managed aquifer recharge.

Fig. 2. Conceptual integrated system used for the simulation-based case study. “S” denotes representative sensor groups; arrows show controllable and hydrologic transfers.

C. Observability and sensor value

Sensor selection is formulated around decisions and state observability. Candidate locations are ranked by reduction in posterior uncertainty, ability to distinguish leaks from demand changes, contribution to balance closure, installation feasibility, and lifecycle maintenance. Trunk-main flow and tank level anchor the bulk balance; pressure sensors then resolve district behavior; spatial rainfall sensors reduce drainage uncertainty.

Redundancy is intentional at critical boundaries. Two independent measurements or a measurement-model comparison are preferred for reservoir release, treatment outflow, major transfers, and detention discharge. Redundancy permits fault detection but does not justify ignoring calibration. Correlated bias is represented explicitly when devices share power, communications, or installation conditions.

The twin reports a live observability score with each state estimate. A low score does not necessarily trigger an alarm, but it widens uncertainty and restricts action. Temporary mobile loggers can be deployed where the score remains low or where repeated anomalies cannot be separated by the current sensor network.

IV. MATHEMATICAL FORMULATION

A. Water-supply hydraulics

For each junction i , conservation requires that the algebraic sum of incident pipe flows, external supply, leakage, and delivered demand equals the rate of local storage change. Pressure-dependent demand is used because intermittent or failure conditions invalidate the assumption that all requested demand is delivered. Leakage is represented as an emitter relation whose coefficient can vary by district. Pipe head loss

may be represented by Darcy-Weisbach or Hazen-Williams; the latter is shown for consistency with common utility practice.

The pump model relates head gain and efficiency to speed and flow. Energy cost is computed with time-varying tariff and carbon intensity. Hard constraints enforce minimum and maximum pressure, tank reserve, pump starts, valve travel, and water-age limits. Soft constraints carry explicit penalties and are reported to the operator rather than silently violated.

$$\sum_j Q_{ji} - \sum_j Q_{ij} + q_i^s - d_i(p_i) - q_i^l = dS_i/dt \quad (3)$$

$$h_f = 10.67 L Q^{1.852} / (C^{1.852} D^{4.871}) \quad (4)$$

$$q_i^l = c_i \max(p_i, 0)^n, \quad 0.5 \leq n \leq 1.5 \quad (5)$$

B. Stormwater and storage dynamics

Urban runoff is generated from spatially distributed subcatchments using rainfall, depression storage, infiltration, impervious fraction, and antecedent moisture. Conveyance is simulated with dynamic-wave routing where surcharge, backwater, reverse flow, and surface ponding may occur. Detention basins and wetlands are represented by level-area-storage and outlet relations. The operational controller can pre-release selected storage when forecasts indicate an approaching storm, subject to downstream capacity and minimum ecological storage.

At the strategic time scale, the reservoir balance links inflow, release, transfer, evaporation, spill, and demand. At the event time scale, the drainage balance links runoff, controlled discharge, infiltration, overflow, and storage. Coupling occurs when captured stormwater is directed to non-potable demand or recharge and when supply-system energy is shifted to off-peak periods to preserve emergency capacity.

$$S_{t+\Delta t} = S_t + \Delta t(I_t - R_t - E_t - Sp_t - W_t) \quad (6)$$

$$A(h) dh/dt = Q_{in} - Q_{out}(u, h) - Q_{inf}(h) - Q_{ov}(h) \quad (7)$$

C. Water-quality state

A simplified advective-reactive model tracks chlorine residual and water age in the distribution network. The detailed engine remains responsible for transport and mixing. The optimizer applies penalties to low residual, excessive age, and avoidable flushing. Sensor residuals are interpreted cautiously because quality measurements may have slower response and greater maintenance requirements than hydraulic sensors.

$$\partial C/\partial t + v \partial C/\partial x = D_t \partial^2 C/\partial x^2 - k_b C - (4k_w/D)C \quad (8)$$

TABLE IV. PRINCIPAL STATES, CONTROLS AND DISTURBANCES

Class	Representative variables	Examples
State x	Head, flow, level, moisture	p, Q, h, S, C
Control u	Pump, valve, gate, release	speed, opening, R
Disturbance w	Rain, demand, outage	$P, d, \text{failure}$
Observation y	Sensor measurements	level, pressure, flow

V. PHYSICS-INFORMED SURROGATE AND FORECASTING

A. Surrogate structure

A graph neural network represents the hydraulic topology: nodes carry elevation, demand, pressure, storage, and water-quality features; edges carry diameter, length, roughness, valve state, and flow direction. A temporal encoder processes recent rainfall, demand, and control histories. The decoder predicts the next state and selected multi-step outputs. A

separate rainfall-runoff branch uses subcatchment adjacency and physically meaningful catchment descriptors.

Training examples are produced from the detailed process models under randomized demands, storms, roughness, leaks, pump availability, and control actions. Observed data can later replace or augment synthetic examples during pilot deployment. Domain randomization reduces dependence on a single calibration and exposes the network to failure states that are rare in historical records.

B. Composite loss and constraints

The training loss combines measurement fit, one-step state error, mass-balance residual, bound violations, monotonic response checks, and temporal smoothness. Each term is normalized by an engineering scale so that pressure, flow, and storage are comparable. The balance penalty is evaluated at both ordinary and stress-test states. Dynamic weights are bounded to prevent the data term from overwhelming physical consistency.

$$L = \lambda_d L_{data} + \lambda_s L_{state} + \lambda_m L_{mass} + \lambda_b L_{bound} + \lambda_q L_{quality} \quad (9)$$

$$L_{mass} = (1/N) \sum_i ||\Sigma Q_i + q_i^s - d_i - q_i^l - \Delta S_i / \Delta t||^2 \quad (10)$$

Uncertainty is estimated from an ensemble of five independently initialized networks plus propagated input ensembles. The forecast presented to the optimizer is therefore a distribution rather than a single trace. Calibration is assessed by coverage of prediction intervals in addition to RMSE and Nash-Sutcliffe efficiency. If uncertainty exceeds a configured threshold, the twin falls back to conservative rule curves.

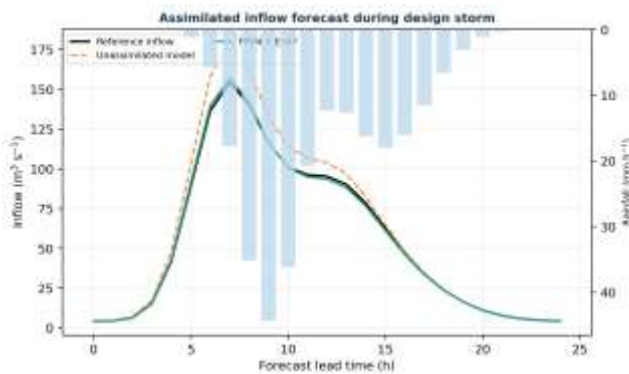


Fig. 3. Example design-storm inflow. The physics-informed surrogate with state assimilation follows the reference hydrograph more closely than the unassimilated process-model forecast.

TABLE V. SURROGATE VALIDATION METRICS

Output	RMSE	NSE / score
Node pressure	1.62 m	R ² = 0.95
Tank level	0.11 m	NSE = 0.97
Storm inflow	3.8 m ³ /s	NSE = 0.91
Chlorine	0.08 mg/L	R ² = 0.88
Mass balance	0.64%	95th = 1.8%

VI. RISK-AWARE MODEL PREDICTIVE CONTROL

A. Objective function

At every decision time, the controller considers a horizon of predicted demands, storms, electricity tariffs, source availability, and failure scenarios. The decision vector contains pump speeds, valve set-points, reservoir releases, detention-gate openings, non-potable transfers, and optional recharge. Only the first approved action is implemented; the

horizon then advances and the problem is solved again with updated measurements.

The objective balances service deficit, flooding, energy, water-quality deviation, leakage, and actuator movement. Weights express policy priorities and are documented as configuration, not learned implicitly. Critical flood and pressure constraints are handled as chance constraints, while emergency limits remain hard constraints checked by the detailed simulator.

$$\min_u E[\Sigma_i (w_f V_{flood} + w_d D_{unmet} + w_e E_{pump} + w_q Q_{pen} + w_{\Delta u} ||\Delta u||^2)] \quad (11)$$

$$Pr\{p_{i,t} \geq p_{min}\} \geq 1 - \alpha_p, \quad Pr\{h_{j,t} \leq h_{max}\} \geq 1 - \alpha_h \quad (12)$$

B. Safety filter and human oversight

A candidate schedule is first screened by the surrogate for speed, then replayed in the detailed model. The safety filter rejects actions that create low pressure, reverse unsafe water-quality conditions, exceed pump envelopes, cause downstream flood transfer, or conflict with maintenance locks. During shadow mode, no actuator command is transmitted. During advisory mode, an operator must approve each action. Closed-loop control is permitted only for low-consequence assets after independent verification and always retains manual override.

This structure recognizes that optimization can exploit model imperfections. The detailed replay, bounded action rate, alarm interlocks, and conservative fallback reduce that risk. Every rejected action is logged with the violated constraint, making the system useful for diagnosing whether the model, policy, or infrastructure is responsible.

TABLE VI. MPC OBJECTIVE COMPONENTS AND SAFEGUARDS

Objective	Operational meaning	Safeguard
Flood volume	Surface/CSO risk	Downstream capacity
Unmet demand	Service reliability	Minimum reserve
Energy	Cost and emissions	Pump envelope
Quality	Residual and age	Detailed replay
Smoothness	Actuator wear	Rate limit

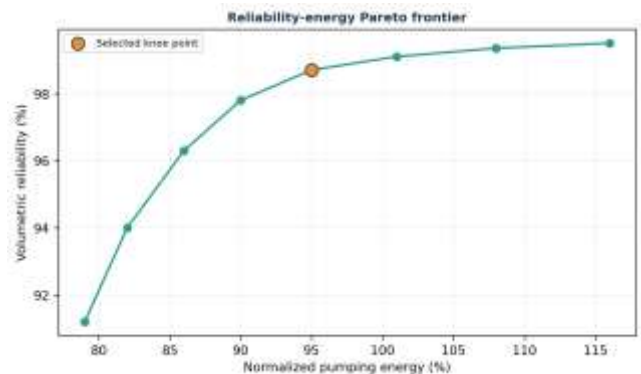


Fig. 4. Pareto frontier generated by varying reliability and energy weights. The selected knee point avoids a disproportionate energy increase for marginal reliability gain.

C. Solver configuration and fallback

The operational problem uses a 24-hour horizon with 15-minute controls for the first six hours and coarser intervals thereafter. Pump and gate actions are bounded by equipment ramp rates and minimum dwell time. Continuous controls are optimized first, then discrete starts and valve states are repaired and replayed. Warm starts use the previous accepted schedule shifted by one interval.

A feasible incumbent is preserved throughout optimization. If the solver times out, the incumbent is compared with the current rule schedule and only the safer verified option is presented. If no candidate passes detailed replay, the controller returns a structured refusal with the binding constraints and enters conservative RBC.

Objective weights are reviewed as policy parameters. A change in flood, service, energy, water-quality, or equity weight is treated as a controlled configuration change and replayed across the benchmark suite. This prevents an apparently minor tuning change from silently altering emergency priorities.

VII. SIMULATION-BASED CASE STUDY

A. Illustrative study system

The benchmark represents a hypothetical semi-arid Indian city of 1.2 million people supplied by one surface reservoir, one treatment plant, three service reservoirs, 412 km of pressurized mains, 18 district-metered areas, and a partially separated drainage network. The urban catchment covers 126 km² with an average impervious fraction of 0.48. Four detention basins, two constructed wetlands, and six recharge zones provide controllable storage. The configuration is intentionally generic and does not reproduce any named city.

Hydraulic properties are initialized from engineering design ranges and randomized in Monte Carlo analysis. Demand follows residential, commercial, institutional, and industrial profiles with seasonal multipliers. Storms include historical-shape synthetic hyetographs, Chicago design storms, multi-cell convective events, and long-duration monsoon sequences. Water-quality states include chlorine decay and water age. Electricity tariffs vary by time of day.

TABLE VII. ILLUSTRATIVE CASE-STUDY PARAMETERS

Parameter	Value	Uncertainty
Population	1.2 million	±8% demand scale
Network	412 km; 18 DMAs	C: 90–140
Urban area	126 km ²	impervious ±0.07
Storage	3 tanks + 4 basins	level ±0.05 m
Sensors	148 hydraulic; 24 rain	0–40% outage
Control step	15 min	24 h horizon

B. Experimental scenarios

Three operating strategies are compared: fixed rule-based control (RBC), hydraulic MPC without learning or ensemble uncertainty (H-MPC), and the complete HydroTwin-MPC. Each strategy experiences identical disturbances. The evaluation includes 20 storm families, 10 demand patterns, five leak sizes, four pump-failure timings, four rainfall biases, and four sensor-availability classes. Latin hypercube sampling produces 250 compound scenarios.

A 60-day spin-up stabilizes storage, roughness estimation, demand states, and water age. The next 90 days are evaluated. Surrogate training uses 70% of simulated sequences, 15% for validation, and 15% for holdout testing, separated by event rather than random time step to reduce leakage. Hyperparameters are fixed before the Monte Carlo operational comparison.

Synthetic rainfall is generated at five-minute resolution and spatially distributed over 24 virtual gauges. Gauge bias, radar scaling error, missing intervals, and storm-cell displacement are sampled independently where appropriate. Catchment wetness is initialized from the preceding seven-day rainfall

and evapotranspiration balance rather than a fixed dry condition.

Demand scenarios preserve zone-specific diurnal shape and add weather sensitivity, weekday effects, intermittent-supply restoration surge, and a protected critical-facility component. Pipe roughness and leakage coefficients vary by material and age class. Pump failures include complete trip, efficiency loss, and delayed restart after power restoration.

Calibration follows a hierarchical sequence: bulk water balance, reservoir and tank storage, transmission flow, district flow, pressure, runoff, drainage level, and finally chlorine decay. Later stages cannot compensate for an unresolved upstream balance error. Parameter bounds and accepted residuals are retained as priors for ensemble generation.

The 250 primary realizations are paired across controllers. An additional 64 severe cases combine upper-tail rainfall, high demand, partial sensor loss, and an asset outage. These cases are excluded from mean-benefit estimates and reported as stress tests so their rare but consequential failures remain visible.

TABLE VIII. COMPARED CONTROL STRATEGIES

Strategy	State update	Optimization
RBC	Measured levels only	Fixed thresholds
H-MPC	Hydraulic estimator	Deterministic MPC
HydroTwin-MPC	EnKF + PINN	Ensemble risk-aware MPC

VIII. EVALUATION METRICS AND VERIFICATION

A. Hydrologic and hydraulic metrics

Forecast skill is measured with RMSE, mean absolute error, coefficient of determination, Nash-Sutcliffe efficiency (NSE), peak error, timing error, and interval coverage. Operational performance is measured by flood volume, overflow hours, unmet demand, pressure violations, volumetric reliability, energy, cost, leakage recovery time, and actuator movement. Water-quality performance uses time below residual threshold and water-age exceedance.

Mass balance is treated as a verification metric rather than a presentation metric. The cumulative difference between sources, sinks, and storage change is normalized by total system throughput. Runs with solver non-convergence or balance error above 2% are quarantined, inspected, and rerun after diagnosis; they are not silently averaged into the results.

$$NSE = 1 - [\sum_i (Q_i - \hat{Q})^2 / \sum_i (Q_i - \bar{Q})^2] \quad (13)$$

$$Rel_v = 1 - [\sum_i D_unmet,i / \sum_i D_requested,i] \quad (14)$$

$$\varepsilon_mb = |V_in - V_out - \Delta S| / \max(V_in, \varepsilon) \quad (15)$$

B. Statistical analysis

For each metric, paired scenario differences are reported because all strategies face the same random realization. Median, interquartile range, and percentile bootstrap confidence intervals are preferred to a normal approximation. Effect size is interpreted alongside statistical significance. Sensitivity is estimated with standardized rank regression over uncertain demand scale, forecast bias, roughness, sensor outage, and initial storage.

Leak detection is evaluated at event level with precision, recall, F1 score, area under the receiver-operating-characteristic curve, localization distance, and median detection delay. A warning is counted only when it persists for two consecutive five-minute windows, which reduces

nuisance alarms. Detection thresholds are selected on validation scenarios and frozen for testing.

Verification is separated into component, integration, and operational levels. Component tests challenge individual mass-balance, pump, valve, storage, runoff, and decay functions. Integration tests check data alignment, topology, state transfer, and repeated solution of the complete loop. Operational tests confirm that recommendations remain feasible during communication loss, missing sensors, solver timeout, maintenance locks, and emergency override.

Forecast uncertainty is judged by reliability diagrams and interval coverage conditional on event magnitude. A model can achieve acceptable average coverage while remaining overconfident for extreme storms. The benchmark therefore reports coverage separately for ordinary, severe, and compound events and widens the control ensemble when the current feature vector lies far from the training domain.

The baseline is frozen before comparison. Rule curves, pump schedules, alarm thresholds, demand multipliers, and initial storage are versioned so that improved results cannot arise from a more favorable baseline setup. Scenario seeds and failed solver runs are retained. This protects the comparison against selective reporting and permits paired re-analysis when a defect is discovered.

Practical acceptance also requires latency and availability. The 15-minute control cycle allocates four minutes to ingestion and estimation, six to ensemble forecasting and optimization, three to detailed replay, and two to operator presentation. Missing the deadline triggers the last safe rule set; a late optimization is not applied to a newer physical state.

TABLE IX. ACCEPTANCE CRITERIA FOR A PILOT TWIN

Area	Metric	Acceptance gate
Hydraulics	Pressure RMSE	≤ 2.0 m
Runoff	NSE	≥ 0.80
Balance	95th percentile	$\leq 2\%$
Uncertainty	90% interval coverage	80–95%
Control	Hard violations	0
Fallback	Recovery drill	Pass

IX. RESULTS

A. Forecasting and state estimation

The assimilated surrogate achieved an NSE of 0.91 for storm inflow and 0.97 for tank level on event-separated holdout sequences. Pressure RMSE was 1.62 m across ordinary and disturbed operation. The median mass-balance error was 0.64%, and the 95th percentile was 1.8%. The unassimilated model exhibited systematic peak-timing error after demand shocks and rainfall bias; state correction reduced the median inflow timing error from 42 to 17 minutes.

Prediction intervals were slightly conservative at short lead times and under-dispersed beyond 18 h during convective storms. Overall coverage of the nominal 90% interval was 86%. This result supports using the twin for operational ranking while also showing why deterministic point forecasts are insufficient for safety-critical actions.

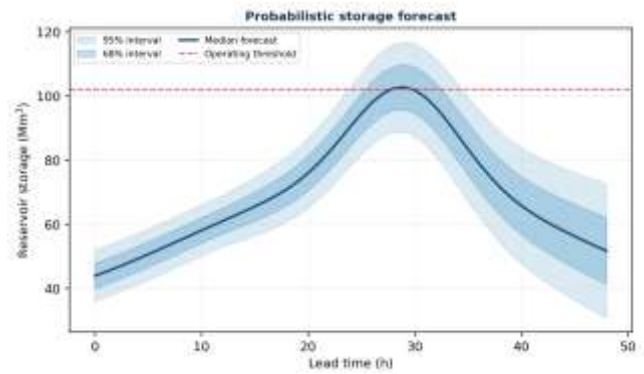


Fig. 5. Probabilistic reservoir-storage trajectory used by the risk-aware controller. The widening fan communicates increasing uncertainty with lead time.

B. Integrated operational performance

Across 250 paired scenarios, HydroTwin-MPC reduced total flood volume by 31.8% relative to RBC and by 12.4% relative to H-MPC. Unmet demand decreased by 42.6% relative to RBC, chiefly because the controller pre-positioned tank storage before peak demand and preserved emergency reserve during pump outages. Pumping energy fell by 16.7%, while the number of pump starts decreased by 11.2%. Overflow duration decreased by 38.4%.

Improvements were not uniform. In small storms with ample storage, all three strategies performed similarly. The largest gains occurred under compound conditions: high initial storage followed by intense rainfall, demand peaks coincident with pump outage, and spatial rainfall bias. In three extreme scenarios, HydroTwin-MPC deliberately accepted a small increase in energy to maintain hospital-zone pressure and detention reserve.

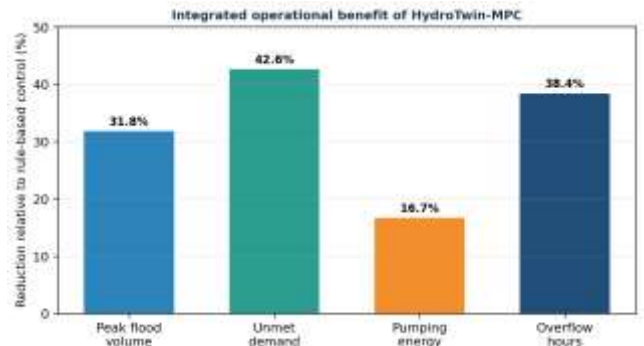


Fig. 6. Median reductions relative to conventional rule-based control. Values are derived from the synthetic benchmark and must not be interpreted as measured utility savings.

TABLE X. PAIRED OPERATIONAL RESULTS OVER 250 SCENARIOS

Metric	RBC	HydroTwin
Flood volume	1.00	0.682
Unmet demand	1.00	0.574
Pumping energy	1.00	0.833
Overflow hours	1.00	0.616
Reliability	96.1%	98.7%
Hard violations	0	0

C. Leakage and anomaly response

The graph-residual detector reached an area under the ROC curve of 0.94 and an event-level F1 score of 0.91. Median detection delay was 18 minutes for leaks exceeding 2% of district demand and 46 minutes for smaller leaks. Median localization error was two candidate pipe segments. False alarms were concentrated near scheduled valve changes and

rapid tank transitions, illustrating the importance of synchronizing asset-status data with telemetry.

An alarm does not directly close a valve. The twin first tests candidate isolation plans against minimum pressure, fire reserve, water age, and affected customers. The operator receives the likely zone, evidence, uncertainty, safe isolation sequence, and an estimate of avoided loss. This decision package is more actionable than a standalone anomaly score.

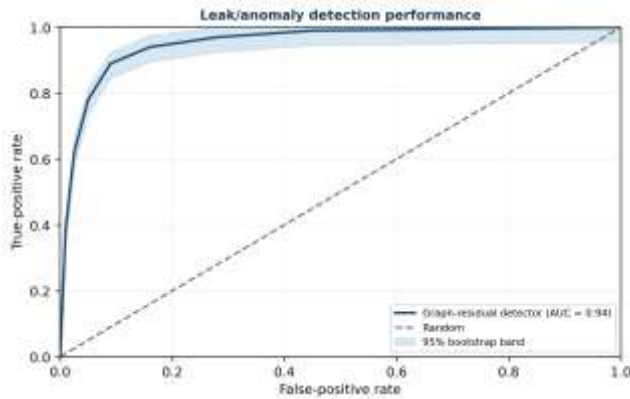


Fig. 7. Receiver-operating-characteristic curve for the graph-residual leak detector with a bootstrap uncertainty band.

TABLE XI. LEAK-DETECTION RESULTS BY SIZE

Leak / DMA demand	Recall	Delay
< 1%	0.61	73 min
1–2%	0.79	46 min
2–5%	0.93	18 min
> 5%	0.98	9 min

D. Ablation study

Ablation tests remove one component at a time while preserving the same control weights. Removing ensemble uncertainty caused the largest deterioration in flood performance during biased forecasts. Removing state assimilation increased pressure error and delayed leak response. Replacing the physics-informed surrogate with an unconstrained neural network reduced inference time slightly but increased 95th-percentile mass-balance error from 1.8% to 5.9%, which failed the acceptance gate.

Removing integrated drainage-supply coordination mainly affected compound events: detention basins were not preconditioned and supply pumps competed for electrical capacity during storm response. This finding supports the paper's system-of-systems formulation rather than isolated asset optimization.

Paired bootstrap intervals showed that the flood-volume advantage over RBC remained positive in 94% of resamples. The energy advantage was more stable than the flood advantage because tariff and pump-efficiency signals were less uncertain than convective rainfall location. Reliability gains were concentrated in the lowest-performing baseline scenarios, which is operationally important because average service was already high.

Failure review identified two recurring near-misses. In the first, optimistic rainfall timing encouraged excessive pre-release before a storm that arrived late. In the second, a stale valve-status tag produced a feasible surrogate solution that failed detailed replay. Neither action was executed because the safety filter rejected it; both cases demonstrate why the surrogate cannot be the final authority.

The complete framework required more computation than deterministic MPC, but median decision latency remained 6.8 minutes on the benchmark workstation. Parallel scenario evaluation and warm-started optimization accounted for most of the speedup. When latency exceeded ten minutes, reducing ensemble size was safer than shortening detailed hydraulic replay.

TABLE XII. ABLATION EFFECTS RELATIVE TO FULL FRAMEWORK

Removed component	Flood benefit	Key degradation
None	31.8%	Reference
Ensemble risk	22.9%	Tail failures
State assimilation	25.1%	Pressure RMSE +48%
Physics loss	29.6%	Balance error 5.9%
Cross-system link	24.3%	Storage conflicts

X. ROBUSTNESS AND SENSITIVITY

A. Sensor outage and forecast bias

Performance declined gracefully as data quality worsened. With 25% random sensor outage and 20% rainfall forecast bias, flood-volume reduction remained 22.7%. At 40% outage and 30% bias it fell to 14.7%, but hard hydraulic constraints remained satisfied because the safety filter used conservative uncertainty bounds. The system entered fallback mode in 7 of 250 primary runs and in 19 of 64 severe stress tests.

Not all sensors had equal value. Tank level, trunk-main flow, and rainfall spatial coverage contributed more to control performance than dense pressure sensing in already well-observed districts. This implies that sensor placement should be based on observability and decision value, not simple device count.

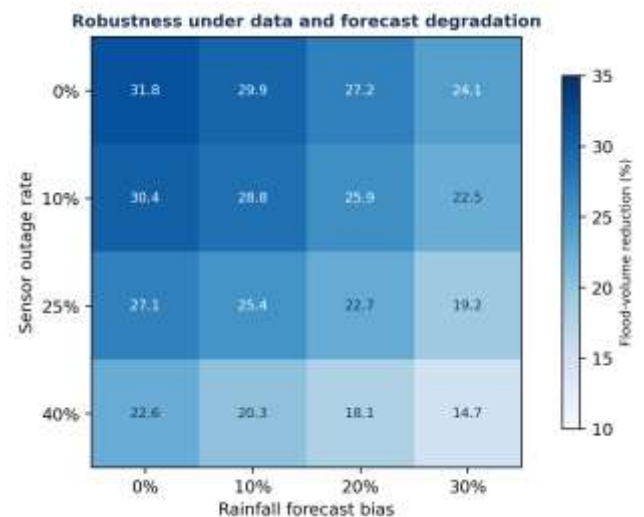


Fig. 8. Flood-volume reduction as sensor outage and rainfall bias increase. The full twin remains beneficial but the margin narrows under joint degradation.

B. Parameter sensitivity and transferability

Initial detention storage and rainfall bias were the dominant determinants of flood outcome. Demand scale and pump availability dominated unmet demand. Pipe roughness influenced both energy and pressure, while leakage exponent had a secondary effect on short-horizon control but a strong effect on estimated annual loss. These relationships were stable under bootstrap resampling.

Transfer to a new city should therefore begin with topology verification, storage curves, pump tests, bulk-meter reconciliation, and rainfall coverage. Fine-scale customer

demand models can be added later. A simpler, well-calibrated twin with explicit uncertainty is preferable to a high-resolution model whose asset status and boundary conditions are unreliable.

TABLE XIII. RANKED GLOBAL SENSITIVITY

Rank	Outcome	Dominant factor
1	Flood volume	Initial basin storage
2	Flood volume	Rainfall bias
1	Unmet demand	Pump availability
2	Unmet demand	Demand scale
1	Energy	Pipe roughness

XI. DISCUSSION

A. Why the integrated twin performs better

The performance gain arises from four mechanisms. First, assimilation corrects the starting point, which is critical because a small storage or demand error can change the feasible control region. Second, the surrogate allows many candidate futures to be evaluated within the decision interval. Third, ensemble optimization explicitly values reserve against uncertain rainfall, demand, and failure. Fourth, cross-system coordination recognizes that tanks, detention basins, wetlands, recharge zones, and pumps share constraints and opportunities.

The results do not imply that artificial intelligence replaces hydraulic knowledge. The largest safety contribution comes from conservation equations, topology, equipment limits, detailed replay, and explicit fallback. Machine learning supplies speed and pattern recognition; engineering constraints determine whether the output is admissible.

B. Comparison with prior work

The findings are consistent with research showing that model predictive control can improve storage utilization and stormwater performance [12]–[15]. They also align with digital-twin studies that emphasize real-time linkage, scenario testing, and data infrastructure [1]–[4]. The present framework extends that line of work by placing supply, drainage, recharge, leakage, and water quality inside one risk-aware decision loop and by reporting performance under sensor and forecast degradation.

The comparison must remain qualitative because case definitions, rainfall, network scale, objectives, and control authority differ across studies. Percentage improvements cannot be transferred directly between cities. The defensible output is the workflow, equations, acceptance tests, and sensitivity pattern—not a universal benefit figure.

C. Practical meaning for civil engineers

For design engineers, the twin can test phased infrastructure, green-grey combinations, emergency interconnections, and sensor locations under operational rules rather than design events alone. For utility engineers, it can reconcile meter balance, identify anomalous districts, estimate hidden storage, and rank pump schedules. For municipal planners, it can compare capital expansion with targeted control, maintenance, recharge, and demand management.

The most important organizational change is the creation of a maintained model product rather than a one-time consultancy model. Asset changes, calibration evidence, data quality, algorithms, operating rules, and operator feedback require ownership, versioning, and budget. Without this

lifecycle, a “digital twin” becomes an attractive dashboard attached to a stale model.

Explainability should be engineered around decisions rather than generic feature importance. An operator needs to know the expected benefit, binding constraints, sensitive assumptions, alternative safe actions, and consequence of doing nothing. A concise counterfactual—such as the flood volume avoided if Basin 3 is pre-released by 0.18 m—supports review better than a probability score without operational context.

Nature-based assets require the same operational discipline as pumps and valves. Wetland stage, infiltration recovery, sediment, vegetation condition, and downstream quality determine whether nominal storage is truly available. The twin can schedule inspection and maintenance from persistent deviations, but field observation remains necessary because ecological performance is not fully observable from level sensors alone.

Interoperability is a long-term resilience feature. Asset identifiers, units, coordinate systems, time zones, quality flags, model inputs, and command semantics should be documented and exportable. A utility must be able to reproduce a critical forecast and operate locally if a vendor service becomes unavailable. Open interfaces also permit independent audit and gradual replacement of individual components.

Operator acceptance is treated as an outcome. Override frequency, reason codes, time to understand a recommendation, and realized benefit after approval are measured during advisory mode. Frequent expert overrides may indicate missing constraints or context rather than operator resistance. Those cases become structured training data for model and workflow improvement.

XII. IMPLEMENTATION ROADMAP FOR INDIAN UTILITIES

A. Stage-gated deployment

Phase 1 diagnoses the decision problem and establishes a verified asset register, telemetry inventory, water balance, model governance charter, and baseline performance. Phase 2 instruments high-value locations and implements a common data contract. Phase 3 operates a shadow twin that predicts but cannot act. Phase 4 introduces operator-approved advisory control. Phase 5 automates only bounded, reversible, low-consequence actions while expanding to basin and groundwater interactions.

Each transition requires an evidence gate. The shadow twin must meet forecast, balance, availability, and fallback targets over ordinary and extreme events. Advisory recommendations must be interpretable and consistently feasible. Closed-loop control requires independent hazard review, cybersecurity testing, operator training, emergency drills, and documented authority.



Fig. 9. Recommended 36-month deployment roadmap. Progression is evidence-gated; utilities may remain in advisory mode when automation is not justified.

TABLE XIV. STAGE GATES AND DELIVERABLES

Phase	Primary deliverable	Exit gate
Diagnose	Use case + baseline	Owner and KPI approved
Instrument	Trusted data contract	Coverage + calibration
Shadow	Live state/forecast	Accuracy + fallback pass
Advisory	Ranked actions	Operator acceptance
Adaptive	Bounded automation	Independent safety review

B. Cost-conscious architecture

A resource-constrained utility should begin with bulk meters, tank levels, pump energy, a small number of pressure loggers, and reliable rainfall data. Open engines and standard geospatial identifiers reduce vendor lock-in. Edge devices can perform validation and buffering when connectivity is intermittent. Cloud services may support computation, but local fail-safe control must not depend on continuous internet access.

The economic case should count avoided water loss, energy, emergency pumping, flood damage, regulatory exposure, service interruptions, and deferred capital—while also counting sensor maintenance, communications, model upkeep, cybersecurity, licenses, training, and organizational change. Pilot benefits should be verified against a pre-registered baseline rather than inferred from dashboard activity.

XIII. CYBERSECURITY, ETHICS AND GOVERNANCE

A. Cyber-physical security

Connecting models to operational data expands the cyber-physical attack surface. Threats include spoofed sensor values, replayed messages, unauthorized set-point changes, compromised credentials, model poisoning, denial of service, and malicious topology edits. Controls should include segmented networks, device identity, least privilege, signed configurations, encrypted transport, immutable logs, anomaly monitoring, offline backups, patch management, and tested manual operation.

Safety cannot depend on the AI model detecting every attack. Independent engineering interlocks, conservative local control, command rate limits, and physical protection remain necessary. The twin should display data freshness and confidence so that operators do not mistake a plausible but stale visualization for current reality.

B. Fairness, privacy and accountability

Optimization objectives distribute benefits and burdens. A controller that minimizes aggregate deficit could repeatedly reduce pressure in peripheral or low-income zones. Service constraints should therefore be disaggregated by critical

customer class and geographic zone, with transparent equity thresholds. Smart-meter data should be minimized, aggregated for planning, retained for a defined period, and protected against secondary use.

Accountability requires a named model owner, operations owner, cybersecurity owner, and approving authority. The system must preserve the forecast, uncertainty, candidate actions, rejected constraints, operator response, and realized outcome. Periodic review should test drift, disparate service impacts, false alarms, override patterns, and whether the twin still improves the original public-service objective.

Threat modelling distinguishes loss of confidentiality, loss of integrity, and loss of availability. For control, integrity is often the most immediate hazard: a plausible but altered level or valve state can lead to an unsafe recommendation. Independent source comparison, signed configuration, plausibility envelopes, and physical interlocks reduce this risk. Availability is protected by local buffering and a tested rule-based fallback.

A red-team exercise should test sensor spoofing, time drift, replayed commands, stolen credentials, malicious model replacement, and denial of service. The exercise must include operational recovery rather than end at detection. Success means that staff can identify the affected boundary, isolate it, restore a trusted state, reconcile physical storage, and document the event without losing essential service.

Privacy review begins with necessity. Household-level demand is not required for most hydraulic decisions; district or pressure-zone aggregation is often sufficient. When high-resolution data are justified, access should be role-based, exports logged, and retention limited. Model developers should receive de-identified features whenever customer identity adds no engineering value.

Equity constraints are tested in normal and emergency modes. Critical facilities, minimum domestic service, informal settlements, peripheral zones, and areas with historically poor pressure should be visible in the performance report. Emergency priority rules are approved publicly or institutionally before an event; they are not improvised by an optimization weight during the crisis.

TABLE XV. PRINCIPAL RISKS AND CONTROLS

Risk	Control	Residual check
Sensor spoofing	Authentication + plausibility	Cross-sensor residual
Model drift	Scheduled recalibration	Holdout error
Unsafe action	Detailed replay + interlock	Zero hard violations
Service inequity	Zone-level constraints	Equity dashboard
Vendor lock-in	Open formats + export	Recovery drill

XIV. ECONOMIC AND ENVIRONMENTAL ASSESSMENT

A. Life-cycle value model

A digital twin is justified only if the present value of verified benefits exceeds implementation and lifecycle costs at an acceptable risk. Benefits may include reduced production and purchase of water, lower energy, faster leak repair, avoided emergency response, deferred augmentation, improved compliance, and reduced flood loss. Costs include instrumentation, communications, integration, model development, cloud or server infrastructure, licenses, maintenance, security, training, and periodic independent review.

The paper uses a transparent benefit-cost equation rather than a single illustrative currency value, because tariffs, losses, asset condition, and flood exposure vary widely. Decision makers can populate the model with local audited data and test optimistic, central, and conservative scenarios.

$$NPV = \sum_t (B_{water} + B_{energy} + B_{risk} + B_{defer} - C_{O\&M}) / (1+r)^t - C_0 \quad (16)$$

B. Environmental co-benefits and trade-offs

Lower pumping energy can reduce emissions where electricity is carbon intensive. Coordinated detention and recharge can reduce peak runoff, support groundwater, and improve non-potable supply, but poor siting may transfer pollution to aquifers or ecosystems. Pre-release can create downstream risk if forecasts are wrong. Sensor and computing infrastructure also have embodied and operational impacts.

Environmental assessment should therefore include water balance, energy, greenhouse-gas intensity, receiving-water quality, recharge suitability, ecological flow, and lifecycle equipment burden. Optimization weights should not conceal regulatory limits or ecological rights; those belong as explicit constraints.

A conservative business case separates measured, modelled, and contingent benefits. Energy reduction can be measured directly. Avoided leakage combines meter reconciliation and repair records. Deferred capital is contingent on sustained service and updated demand projections. Flood-risk reduction requires event probability and depth-damage analysis rather than multiplying one simulated event by an assumed frequency.

Sensitivity to discount rate, electricity price, water-production cost, carbon intensity, sensor replacement, and staffing is reported as a range. The central case should exclude benefits that cannot be independently verified. A pilot may still be worthwhile when financial NPV is modest if it materially improves emergency preparedness, regulatory compliance, or public-health protection.

Procurement should fund lifecycle capability, not only software delivery. Required line items include calibration campaigns, instrument replacement, data engineering, cybersecurity, operator training, model stewardship, incident response, and independent evaluation. Payment milestones can be tied to verified data quality, replayable models, fallback drills, and agreed performance gates.

Environmental safeguards are spatial. Capturing more runoff is not automatically beneficial if it reduces downstream ecological flows or transfers contaminated first flush to recharge. The twin should distinguish harvestable volume from protected flow, route poor-quality water to treatment, and report consequences at the receiving water and aquifer rather than only inside municipal boundaries.

TABLE XVI. BENEFIT-COST INPUTS FOR LOCAL APPRAISAL

Input	Unit	Evidence source
Avoided loss	m ³ /year	Audited water balance
Energy saving	kWh/year	Metered pump baseline
Flood risk	currency/year	Depth-damage analysis
Deferred capacity	currency + year	Capital plan
Twin lifecycle cost	currency/year	Procurement + staffing

XV. LIMITATIONS AND FUTURE RESEARCH

The principal limitation is the absence of field validation. The case study is synthetic, even though parameters and processes are physically plausible. Real networks contain undocumented valves, illegal connections, meter bias, communication latency, construction changes, pressure transients, sediment, customer behavior, institutional constraints, and extreme events outside the training domain. Reported percentages are therefore benchmark outcomes, not promised field savings.

The hydrologic model simplifies groundwater-surface water exchange and does not resolve two-dimensional street flooding. The water-quality component is limited to chlorine and age; pathogens, disinfection by-products, temperature, and multi-species reactions would require additional data and validated kinetics. The optimization does not monetize every social and ecological consequence. Cybersecurity is addressed architecturally but not through penetration testing.

Future work should undertake a 12- to 18-month shadow deployment with independent meter reconciliation, controlled pump and valve tests, event-based drainage surveys, and operator-in-the-loop evaluation. Research priorities include differentiable hydraulic solvers, transfer learning across districts, active sensor placement, robust multi-agent coordination, privacy-preserving demand learning, causal evaluation of interventions, two-dimensional flood coupling, groundwater digital twins, and standards for twin maturity and audit.

A further priority is equitable basin integration. City-scale optimization can externalize upstream abstraction or downstream flood risk. Linking urban twins with basin forecasts, environmental-flow constraints, agricultural demands, and participatory governance is necessary before local efficiency can be interpreted as system sustainability.

Parameter identifiability is another limitation. Different combinations of roughness, demand, leakage, and valve status can reproduce similar pressure traces. Additional meters do not automatically resolve this ambiguity. Controlled tests, informative operating changes, prior asset knowledge, and uncertainty reporting are required before a calibrated parameter is treated as a physical measurement.

The benchmark does not model structural pipe failure, transient pressure waves, detailed sediment transport, or human evacuation. These processes operate on different time and spatial scales and may require specialized solvers. Coupling them without validated interfaces could create false precision, so expansion should be driven by a specific decision and measurable data.

Machine-learning drift can occur through urban growth, seasonal demand change, new pumps, tariff shifts, sensor replacement, or revised operating practice. Drift monitoring should compare residual distributions and constraint activation, not only aggregate error. Retraining is authorized only after data review, holdout evaluation, stress testing, and regression replay.

Finally, institutional learning may dominate algorithmic improvement. A technically accurate twin can fail if alarm ownership, maintenance response, approval authority, and budget are unclear. Future field research should therefore measure decision time, cross-department coordination, operator trust, maintenance closure, and public communication alongside hydraulic metrics.

TABLE XVII. PRIORITY RESEARCH AGENDA

Priority	Method	Success indicator
Field validation	Shadow pilot	Pre-registered KPI gain
2-D flooding	SWMM + surface solver	Depth/timing accuracy
Groundwater link	Coupled recharge model	Seasonal balance
Equity	Zone/service constraints	No systematic burden
Standards	Independent audit	Repeatable maturity score

XVI. CONCLUSIONS

This paper presented HydroTwin-MPC, an integrated, physics-informed digital twin for climate-resilient urban water-resources management. The framework combines process-based supply and drainage models, real-time state estimation, an accelerated physics-informed surrogate, probabilistic forecasts, anomaly detection, risk-aware model predictive control, detailed safety replay, and human governance. The architecture is designed to be modular, auditable, and deployable in stages.

In a synthetic semi-arid city benchmark, the complete framework outperformed fixed rule control across flood volume, unmet demand, energy, overflow duration, and leakage response while satisfying hard constraints. Robustness tests showed declining but positive benefit under joint sensor outage and rainfall bias. Ablation results identified uncertainty-aware optimization and state assimilation as critical components. These findings support the research hypotheses within the simulation domain.

The strongest conclusion is procedural rather than numerical: a trustworthy water digital twin is a maintained engineering decision system, not a dashboard or an unconstrained AI predictor. Its value depends on verified asset data, balance closure, uncertainty communication, safe fallback, human authority, cybersecurity, equity constraints, and lifecycle ownership. The next step is a transparent shadow pilot with independent validation; automatic control should follow only after evidence demonstrates safety and public value.

For engineering practice, five recommendations follow. Begin with a decision and baseline rather than a technology catalogue. Verify topology and water balance before training advanced models. Treat uncertainty and fallback as primary design requirements. Preserve operator approval until evidence supports bounded automation. Maintain the twin as a governed asset with versioned data, models, constraints, and tests.

For research, integrated evaluation should replace isolated accuracy claims. Forecast skill, physical consistency, decision latency, tail risk, cybersecurity, equity, and realized operational benefit belong in the same evidence framework. Common open benchmark scenarios would make comparisons more meaningful and accelerate independent replication.

For policy, digitalization can strengthen climate adaptation only when institutions can operate and maintain it. Capacity building, interoperable data, independent audit, and protection of low-resource communities are therefore part of the innovation itself. Funding models should support measurement and model stewardship beyond the initial demonstration period.

HydroTwin-MPC is offered as a design and verification blueprint. Its synthetic results justify a carefully controlled pilot, not direct operational deployment. A successful next

study would publish the baseline, acceptance gates, failure log, override record, and paired outcomes so that the claimed value can be examined and reproduced.

The broader opportunity is to move urban water management from reactive operation toward anticipatory, transparent, and adaptive stewardship while retaining the conservative safety culture of civil engineering. That transition is credible only when physical law, empirical evidence, institutional judgment, and public accountability remain connected.

APPENDIX A. REPRODUCIBILITY PROTOCOL

A reproducible implementation should archive: (1) network and catchment topology with coordinate reference system and version; (2) parameter priors and calibration bounds; (3) raw and quality-controlled sensor streams; (4) scenario seeds and disturbances; (5) simulator, surrogate, and optimizer versions; (6) training, validation, and holdout event lists; (7) objective weights and constraints; (8) solver tolerances; (9) rejected and approved actions; and (10) metric scripts. Confidential asset and customer data may remain protected while hashes, schemas, and aggregated benchmark outputs are published.

Before each operational release, automated tests should verify units, timestamp alignment, topology connectivity, mass balance, baseline reproduction, fallback activation, constraint enforcement, numerical convergence, and deterministic replay of a reference scenario. Model cards should describe intended use, prohibited use, training domain, known failure modes, uncertainty behavior, and approval authority.

A benchmark package should include at least one ordinary day, one high-demand day, one design storm, one spatially biased storm, one leak, one pump failure, one communication loss, and one compound emergency. Expected outputs include state trajectories, warnings, feasible actions, rejected constraints, runtime, mass balance, and final metrics. Tolerances are stated numerically so regression tests can distinguish meaningful drift from harmless floating-point variation.

Calibration evidence is versioned with the model. The archive records instruments, dates, operating conditions, excluded observations, objective functions, priors, posterior uncertainty, and residual diagnostics. A parameter is not promoted merely because it reduces aggregate RMSE; the change must be physically credible and improve independent events without degrading balance or failure behavior.

Reproduction should be possible without access to the live control network. De-identified or synthetic input streams, topology abstractions, and an offline simulator allow independent reviewers to replay the workflow safely. Cryptographic hashes link protected source files to released summaries without disclosing critical infrastructure details.

A release report documents known deviations from the previous version, affected scenarios, changed dependencies, reviewer sign-off, rollback procedure, and expiration date. Emergency patches receive the same retrospective review. This discipline turns the twin into an auditable engineering configuration rather than an evolving collection of untracked scripts.

TABLE XVIII. REPRODUCIBILITY CHECKLIST

Artifact	Required record	Release test
Data	Schema + quality flags	Unit/time validation
Physics model	Version + calibration	Balance + convergence
AI surrogate	Data split + weights	Holdout + stress test
Controller	Weights + constraints	Replay + safety filter
Operations	Approval + override log	Recovery drill

APPENDIX B. SYMBOLS AND ABBREVIATIONS

Symbols are defined locally when first used and summarized below to support reproducibility across hydraulic, hydrologic, learning, and control components. A dash denotes a dimensionless quantity. Subscripts identify node, pipe, asset, scenario, or time indices; superscripts minus and plus denote prior and posterior state estimates. Implementations should retain SI units internally and convert display units only at the interface.

TABLE XIX. PRINCIPAL SYMBOLS

Symbol	Definition	Unit
Q	Volumetric flow rate	m ³ /s
p, H	Pressure/head	m
S	Storage	m ³
C	Concentration	mg/L
u	Control vector	various
x, y	State, observation	various
K	Kalman gain	—
NSE	Nash-Sutcliffe efficiency	—
MPC	Model predictive control	—
PINN	Physics-informed neural network	—

Abbreviations include AMI (advanced metering infrastructure), DMA (district-metered area), EnKF (ensemble Kalman filter), GIS (geographic information system), MPC (model predictive control), PINN (physics-informed neural network), RBC (rule-based control), RTC (real-time control), SCADA (supervisory control and data acquisition), SWMM (Storm Water Management Model), and WDN (water distribution network).

Sign conventions are fixed before calibration. Flow is positive in the stored edge direction, reservoir release is positive out of storage, recharge is positive into the aquifer control volume, and overflow is positive out of the managed system. A negative simulated pipe flow is therefore valid and indicates reversal; it is not converted to an absolute value for learning or reporting.

Time is stored in coordinated universal time with the applicable local offset retained for display. Rainfall represents interval depth, whereas pressure, level, flow, and quality are timestamped states or interval averages as declared in metadata. Mixing these conventions without conversion can create false lags and erroneous mass balance.

Uncertainty notation distinguishes observation error R, process disturbance covariance Q, ensemble spread, and scenario uncertainty. These terms are not interchangeable. Sensor accuracy may be high while future rainfall uncertainty is large; a controller that merges both into one generic confidence score cannot explain why an action is conservative.

Units and scales are checked at every interface. Pressure head is reported in metres, not mixed with pressure in kilopascals; flow uses cubic metres per second internally; energy uses kilowatt-hours; storage uses cubic metres or explicitly labelled million cubic metres; and concentration uses milligrams per litre. Derived indicators retain their numerator, denominator, and aggregation window.

APPENDIX C. OPERATIONAL ALGORITHM

The implementation sequence separates estimation, prediction, optimization, verification, approval, and learning. The separation allows a utility to disable any downstream function without losing basic monitoring and state estimation. During shadow mode, the last two action steps are logged but no command is transmitted.

TABLE XX. FIFTEEN-MINUTE HYDROTWIN-MPC CYCLE

Step	Operation	Failure response
1	Ingest and time-align streams	Use buffered data
2	Apply range, drift and topology checks	Flag and down-weight
3	Propagate prior ensemble	Retain last safe state
4	Assimilate accepted observations	Robust innovation clip
5	Generate demand, rain and failure ensemble	Conservative envelope
6	Optimize candidate schedules	RBC fallback if infeasible
7	Replay candidates in detailed model	Reject violations
8	Explain and rank safe actions	Display uncertainty
9	Approve or bounded auto-execute	Manual override
10	Log outcome and monitor residual	Open incident if drift

The control service must be idempotent: repeating the same approved command must not create an additional actuator movement. Sequence number, expiration time, asset state, and acknowledgement are checked before execution. A recommendation expires when the underlying forecast, topology, maintenance lock, or sensor-confidence class changes. Optimization is never permitted to silently relax a hard safety constraint.

A daily reconciliation compares measured production, district inflow, authorized consumption, estimated leakage, transfers, and storage change. A weekly process reviews model drift and false alarms; a monthly governance review examines overrides, equity indicators, cyber events, and benefit realization. Each model release is tagged, replayed against a fixed regression suite, and reversible to the last accepted version.

The estimator assigns every accepted measurement a quality class and effective variance. A recently calibrated sensor with consistent neighbours receives normal weight; a drifting, stale, or isolated sensor is down-weighted or rejected. Rejection is logged and does not delete the raw observation. When all observations for a critical state are unavailable, uncertainty expands and the controller reduces action magnitude.

Optimization runs in a sandbox with a strict time and memory budget. Feasible incumbents are retained so a timeout can return the best verified schedule rather than no answer. Candidate actions are ordered by expected risk reduction, robustness, energy, and actuator movement. The operator can compare at least one conservative alternative and the consequence of continuing the current rule.

After execution, realized response is compared with the predicted envelope. Persistent deviation creates a structured

issue linked to topology, data, state estimation, control, or physical performance. This classification guides whether staff inspect a meter, update an asset status, recalibrate a parameter, or revise an operating constraint.

The algorithm never learns directly from an operator override without review. The override may reflect missing context, a temporary policy, an incorrect model, or personal preference. A multidisciplinary review labels the reason before it becomes a candidate training feature or permanent constraint.

APPENDIX D. SCENARIO CATALOGUE AND STRESS TESTS

The scenario catalogue exposes interactions that single design events miss. Each scenario preserves physical correlation where defensible: convective rainfall cells move spatially, demand shocks follow service restoration, and pump efficiency changes with operating point. Independent randomization is used only when no credible dependence model is available.

TABLE XXI. STRESS-TEST CATALOGUE

Family	Perturbation	Recorded outcome
Hydrometeorology	Bias, timing, cell track, duration	Flood, storage, recharge
Demand	Peak, restoration surge, fire flow	Pressure, deficit, reserve
Assets	Pump trip, stuck valve, pipe leak	Recovery and isolation
Data	Outage, drift, latency, spoofing	Fallback and confidence
Quality	Source change, decay-rate shift	Residual and water age
Operations	Maintenance lock, tariff spike	Feasibility and energy
Compound	Storm + outage + demand peak	Tail risk and equity

Primary results use 250 Latin-hypercube realizations. Severe stress tests add 64 deliberately adversarial combinations near or outside the calibration envelope. They are analyzed separately so rare failures are not diluted by ordinary days. A scenario passes only when the simulator converges, mass balance remains within tolerance, hard constraints are satisfied, fallback activates when required, and the event can be deterministically replayed from archived inputs.

Field pilots should supplement this catalogue with operator-authored known difficult days, including festival demand, power curtailment, source-quality deterioration, cloudburst drainage congestion, communications loss, and emergency isolation. Scenario ownership should be shared by operations, planning, information technology, cybersecurity, and public-health staff.

Scenario severity is described by both probability and consequence. Low-probability cases remain in the suite when they can threaten life safety, critical customers, dam or tank integrity, contamination containment, or widespread service. Conversely, frequently occurring nuisance alarms are included because they erode operator trust and can mask a true emergency.

Compound scenarios are constructed with a causal timeline: antecedent wetness or scarcity establishes the initial state; a forecast arrives with stated error; sensors fail or drift; an asset fault occurs; control and operator actions follow; and recovery is measured until service and model confidence return. This event logic is more informative than perturbing all variables independently at one time step.

Pass criteria include maximum violation, cumulative impact, detection time, action time, recovery time, customers affected, equity distribution, and completeness of the audit

trail. The controller may pass by entering fallback and declining to optimize. Safe refusal is preferable to a confident recommendation outside the validated domain.

The catalogue is reviewed after every real event. New failure combinations, unexpected operator workarounds, and unmodelled dependencies become regression scenarios. Retired scenarios remain archived so improvements are not achieved by narrowing the test suite.

APPENDIX E. DATA CONTRACT AND INFORMATION ARCHITECTURE

A dependable digital twin begins with a data contract rather than an algorithm. Every stream is assigned an owner, engineering unit, coordinate reference, sampling rule, clock source, quality vocabulary, expected latency, retention period, and permitted use. The contract distinguishes raw observations from corrected values and model-derived states. Corrections never overwrite the source record; they are stored as a new version with a reason code. This lineage is essential when an operator asks why a recommendation changed or when an auditor attempts to reproduce an incident months later.

Asset identity is the primary join key. Pumps, valves, meters, tanks, pipes, drains, basins, recharge structures, and quality stations receive stable identifiers that survive display-name changes. Network connectivity, service status, maintenance locks, and effective dates are maintained as versioned attributes. A topology snapshot is attached to every forecast and control run. This prevents a common failure in which a mathematically feasible schedule is evaluated against yesterday's valve arrangement or an incomplete maintenance update.

Time handling is explicit. Measurements are stored in coordinated universal time, while local time and daylight rules are retained for display and tariff interpretation. The ingestion service records source timestamp, gateway timestamp, server receipt time, and processing time. These fields separate sensor delay from communications delay and computational delay. Interval rainfall and energy are not treated as instantaneous observations; their support period is preserved. Resampling rules state whether an interval is summed, averaged, held, or left missing.

Quality control operates in layers. Syntactic tests examine type, unit, range, timestamp, and duplicate records. Physical tests examine rate of change, storage bounds, pump-state consistency, pressure continuity, flow direction, and mass balance. Contextual tests compare neighbouring sensors, scheduled maintenance, rainfall radar, and known control actions. Each accepted value carries a quality class and effective uncertainty. A suspect value may remain visible to the operator while being down-weighted in assimilation, avoiding the false choice between complete trust and deletion.

The architecture separates operational technology from analytical services. A read-only broker transfers approved measurements from the control environment to a buffered integration zone. Forecasting, assimilation, optimization, and reporting run in segmented services with least-privilege access. Candidate actions return through a controlled interface that validates asset identity, unit, permissible range, command sequence, expiration time, and authority. Independent local interlocks remain effective if the analytical platform or communications link is unavailable.

Model inputs are published through versioned interfaces instead of direct database queries. A hydraulic-input package contains topology, elevations, roughness, controls, demands, storage curves, source conditions, and calibration metadata. A drainage package contains subcatchments, imperviousness, infiltration parameters, conduit geometry, boundary stages, and rainfall fields. Changes are schema-validated and reviewed before release. This approach reduces hidden coupling and allows individual modelling engines to be replaced without rebuilding the entire twin.

Data retention follows decision value and risk. High-frequency control telemetry may be retained at native resolution for incident investigation, then aggregated for long-term planning. Household-level consumption is minimized or spatially aggregated unless a documented operational need exists. Critical infrastructure details are protected, while non-sensitive schemas, validation logic, and synthetic benchmark streams can be shared for independent research. Access logs, export approval, encryption, backup testing, and destruction schedules form part of the engineering configuration.

Data-service performance is measured continuously. Indicators include completeness, freshness, clock offset, accepted fraction, calibration currency, topology update time, duplicate rate, and recovery time after interruption. A high overall completeness score cannot conceal failure of a critical tank-level signal; metrics are reported by decision dependency and service zone. When a mandatory input exceeds its error budget, the twin reduces authority, expands uncertainty, or enters fallback rather than silently substituting a nominal value.

TABLE XXII. MINIMUM OPERATIONAL DATA CONTRACT

Field	Example requirement	Validation
Asset identity	Stable ID + effective topology	Connectivity test
Timestamp	UTC + source/receipt times	Clock-offset alarm
Engineering unit	SI canonical unit	Dimensional check
Quality class	Good/suspect/bad/missing	Rule + reviewer
Latency budget	Decision-specific maximum	Freshness monitor
Lineage	Raw, correction, version	Reproducible replay
Authority	Read/write role + expiry	Least-privilege test

APPENDIX F. CALIBRATION, VALIDATION AND UNCERTAINTY PROTOCOL

Calibration is organized around identifiable decisions, not the indiscriminate adjustment of every parameter. The first pass verifies elevations, connectivity, diameters, storage curves, pump characteristics, valve controls, boundary conditions, and meter scaling. Only after gross configuration errors are resolved are roughness, demand multipliers, runoff, infiltration, or reaction parameters adjusted. Each adjustable quantity receives a physically defensible prior range and spatial grouping. Parameter freedom is increased only when independent observations contain information capable of supporting it.

Field tests are selected to excite the system safely. Controlled pump changes, tank drawdown and recovery, valve steps, hydrant or district flow tests, wet-weather surveys, tracer or source-switch observations, and recharge monitoring expose different parameter combinations. The test plan records operational constraints, customer protection, instrument accuracy, and the expected information gain. Routine data

add seasonal coverage, but controlled excitation is often necessary to separate roughness, demand, leakage, storage, and boundary errors that otherwise produce similar residuals.

The calibration objective is multi-criteria. It combines pressure, flow, level, volume balance, hydrograph timing, water age or quality where available, and event-specific penalties. Residuals are examined by sensor, operating regime, season, magnitude, and time lag rather than reduced to one aggregate score. A model with slightly higher global RMSE may be preferable if it removes bias at low pressure, captures peak storage, and remains stable during pump transitions. Physical plausibility and independent-event performance are explicit acceptance conditions.

Data are separated by events and regimes to prevent leakage across training and evaluation. Calibration may use ordinary days and selected storms; validation uses unseen seasons, pump patterns, and rainfall structures; holdout evaluation remains untouched until the candidate release is frozen. When machine-learning components use sliding windows, all windows from the same physical event stay in one split. Hyperparameter selection uses only calibration and validation data. The final holdout score is reported once with confidence intervals and failure cases.

Uncertainty is partitioned into observation, parameter, initial-state, boundary, forcing, and structural components. Observation variance is informed by calibration certificates and field residuals. Parameter ensembles represent credible combinations rather than independent arbitrary noise. Rainfall and demand scenarios preserve spatial and temporal correlation. Structural uncertainty is evaluated by alternative runoff, demand, or loss formulations where practical. These sources are propagated to operational outputs so a recommendation can state which uncertainty dominates its risk.

Posterior diagnostics check whether uncertainty is calibrated. Prediction-interval coverage, sharpness, rank histograms, probability integral transforms, reliability plots, and event-stratified residuals are examined together. Wide intervals can achieve nominal coverage without being useful, while narrow intervals can look decisive but miss extremes. The release gate therefore specifies both coverage and maximum actionable width for critical variables. Persistent under-dispersion triggers uncertainty inflation or model revision before control authority is increased.

Validation includes conservation and invariance tests. Mass and constituent balances must close within numerical and measurement tolerances. Results should not change materially under harmless reorderings, time-zone display changes, or equivalent unit conversions. Increasing available storage should not increase overflow in an otherwise identical monotonic test unless a documented control interaction explains it. Reversing a pipe's stored orientation must reverse its reported sign without changing physical flow. These tests catch implementation errors that statistical fit may not reveal.

Recalibration is event-driven as well as scheduled. Triggers include sensor replacement, asset renewal, persistent bias, changed source operation, new development, seasonal regime change, abnormal constraint activation, or divergence between predicted and realized action response. A trigger opens an investigation; it does not automatically authorize retraining. The existing model remains available for comparison and rollback. The revised release must pass the same holdout, stress, balance, safety, and replay suite as the original.

TABLE XXIII. CALIBRATION AND VALIDATION ACCEPTANCE MATRIX

Domain	Primary evidence	Illustrative gate
Supply hydraulics	Pressure/flow/level events	No material bias
Drainage	Volume + peak timing	NSE and balance pass
Water quality	Age/concentration response	Bounded residual
Uncertainty	Coverage + sharpness	Calibrated intervals
Control replay	Detailed simulator	Zero hard violations
Transfer	Unseen regime	Graceful degradation
Reproducibility	Archived inputs + seed	Deterministic metrics

APPENDIX G. CONTROL SAFETY CASE AND COMMISSIONING TESTS

The safety case states why the proposed advisory or automatic function is acceptably safe within a defined operating envelope. It identifies hazards, initiating events, barriers, evidence, assumptions, residual risk, and accountable owners. The envelope lists controllable assets, command limits, forecast horizon, validated hydraulic regimes, communication requirements, and prohibited conditions. Claims are linked to tests and operational procedures rather than to general statements about model accuracy. Any change that invalidates a claim returns the function to shadow or fallback mode.

Hazard analysis begins with physical consequences: loss of pressure, tank depletion, overflow, uncontrolled release, contamination movement, pump damage, cavitation, surge, service inequity, or blocked emergency access. Digital causes include stale data, incorrect topology, unit conversion, sensor spoofing, optimization error, software defect, delayed acknowledgement, and conflicting manual action. Human and organizational causes include ambiguous authority, incomplete maintenance communication, alarm overload, and unpractised recovery. Each barrier is independent where feasible so one analytical service cannot defeat all protection.

Command authority is bounded in magnitude, rate, frequency, duration, and asset state. A pump schedule respects minimum run and stop times; a valve command respects travel rate and mechanical limits; a detention release respects downstream stage and water-quality constraints. Every candidate is replayed in the detailed model with conservative initial conditions. The command service independently validates the approved set point against a static allow-list and local interlock. Absence of acknowledgement is treated as unknown state, not successful execution.

Commissioning proceeds through offline replay, hardware-in-the-loop testing, shadow operation, advisory operation, and bounded automation. Offline tests use archived and synthetic extremes. Hardware-in-the-loop testing verifies interfaces, scaling, sequencing, timeouts, and actuator feedback without exposing the network to uncontrolled action. Shadow operation evaluates live forecasts and counterfactual recommendations. Advisory operation measures operator comprehension, overrides, and realized response. Only low-consequence reversible actions enter automatic mode, and authority expands one evidence gate at a time.

Fallback is designed as a positive operating mode. It identifies the triggering conditions, the rule set to be restored, the state information required, the responsible operator, and the recovery test. Triggers include critical data loss, topology uncertainty, solver failure, excessive latency, disagreement between surrogate and detailed model, cyber alarm, or

operation outside the validated envelope. Fallback commands are tested periodically and stored locally. The system records why it declined optimization so safe refusal is not mistaken for an outage.

Human factors are verified through scenario-based drills. Displays prioritize current state, recommendation, expected benefit, binding constraints, uncertainty, alternative safe actions, and time remaining. Operators should be able to identify the affected assets, understand the principal assumption, accept or reject the action, and restore manual control without navigating multiple screens. Alarm rates and explanation time are measured. Repeated overrides are reviewed as potential missing context or constraints rather than classified automatically as user error.

Cybersecurity testing includes authenticated device identity, signed configuration, role-based access, network segmentation, backup restoration, log integrity, replay protection, command expiration, and detection of anomalous sensor combinations. Penetration exercises use a non-production environment or an approved maintenance window. The test is incomplete until operations recover a trusted physical and digital state. Security patches are evaluated against deterministic hydraulic and command regression tests so vulnerability remediation does not introduce control defects.

The commissioning report preserves pass and fail evidence, unresolved assumptions, compensating controls, owner signatures, rollback instructions, and the date when evidence expires. Residual risk is accepted by the appropriate utility authority, not by the software supplier or model developer alone. After launch, performance is reviewed against leading indicators such as data freshness and constraint activation as well as lagging outcomes such as overflow or service interruption. Material incidents suspend authority pending investigation.

TABLE XXIV. COMMISSIONING EVIDENCE BY MATURITY STAGE

Stage	Permitted output	Mandatory evidence
Offline	Reports only	Replay + stress suite
Hardware loop	Test commands	Interface and interlock pass
Shadow	Live prediction	Coverage + balance stability
Advisory	Ranked recommendations	Operator acceptance + feasibility
Bounded auto	Reversible commands	Independent safety approval
Expanded auto	Additional assets	New hazard case + field evidence

APPENDIX H. WORKED ILLUSTRATIVE MONSOON EVENT

This worked example is a transparent synthetic event, not a record from Jaipur or any operating utility. At 13:00, detention storage is 72% full after antecedent rainfall, Reservoir 2 is at 61%, and the western pressure zone is approaching its evening demand peak. The meteorological ensemble predicts a convective cell between 16:00 and 18:30, but arrival time varies by 55 minutes and areal rainfall by 24%. Pump P4 is unavailable for planned maintenance, and one trunk-flow meter is reporting intermittent gaps.

The rule-based strategy holds detention outlets at their normal dry-weather positions until a local level threshold is crossed. The deterministic MPC uses the median rainfall forecast and pre-releases the most exposed basin. HydroTwin-MPC assimilates the recent level and rainfall residuals, expands the late-arrival scenarios, and recognizes that aggressive pre-

release could raise downstream stage during the commuter-period tide boundary. It also preserves supply storage because the unavailable pump reduces recovery capacity during the demand peak.

At 13:15, the ensemble controller evaluates candidate basin releases, pump schedules, tank transfers, and recharge routing. Its preferred action lowers Basin 3 by 0.16 m over 70 minutes, delays Basin 1 release, advances one efficient supply-pump block before the tariff peak, and reserves a minimum tank level for the hospital zone. The detailed replay rejects a more aggressive alternative because two ensemble members exceed the downstream stage constraint. The accepted schedule has lower expected benefit but a smaller tail loss.

The operator display reports the recommendation as a decision package: expected avoided overflow, energy change, downstream constraint margin, storage reserve, principal uncertainty, and the consequence of no action. It shows that flood benefit is most sensitive to storm arrival time and Basin 3's verified gate capacity. The operator confirms that no field crew is working downstream and approves the reversible release. Command acknowledgement and measured gate position are received within the allowed latency.

At 15:40, radar indicates a slower cell and the trunk-flow signal becomes stale. The estimator down-weights that meter and widens supply uncertainty. Instead of continuing the original schedule, the controller shortens the next release block and retains the rule-based pump reserve. A candidate that depends on the stale flow is excluded. The operator sees that the change is caused by information quality rather than by a hidden objective-weight adjustment. This traceability supports a rapid and defensible decision.

Rain begins at 16:22 and peaks later than the median forecast. Basin 3 absorbs the first peak without crossing its emergency level. Basin 1 retains enough space for the late cell, while the downstream stage stays below the hard threshold. The western tank reaches its minimum advisory reserve but does not violate service pressure. Pumping energy is slightly higher than the unconstrained energy optimum because reliability receives priority under the pump outage. No automatic valve isolation is attempted.

After the event, the twin reconciles rainfall, inflow, storage change, release, overflow, supply production, and district demand. The mass-balance residual is within the synthetic acceptance tolerance. The report compares the approved schedule with the rule-based and deterministic counterfactuals under the same realized forcing. It separates modelled avoided overflow from measured quantities and records that no field claim can be made. The late rainfall error is added to the forecast-calibration archive.

The example illustrates the intended innovation: value arises from coordinating storage, supply reserve, forecast uncertainty, sensor confidence, hydraulic constraints, and operator knowledge. No individual AI prediction is allowed to command infrastructure. The controller may sacrifice energy or expected flood benefit when tail scenarios threaten service or downstream safety. Equally important, the workflow records rejected alternatives and changing assumptions, allowing engineers to examine whether the eventual action was reasonable with the information available at the time.

TABLE XXV. WORKED-EVENT DECISION TRACE

Time	New evidence	Twin response
13:00	High basin storage + pump outage	Expand joint scenarios
13:15	Initial ensemble optimization	Rank safe pre-release plans
13:20	Detailed replay constraint	Reject aggressive release
13:27	Operator approval	Issue bounded command
15:40	Radar shift + stale flow	Down-weight and re-optimize
16:22	Rainfall begins late	Preserve reserve
19:15	Event recession	Reconcile and compare
Next day	Engineering review	Archive residuals + lesson

TABLE XXVI. ILLUSTRATIVE EVENT OUTCOME NOTATION

Outcome	RBC comparison	Interpretation
Peak basin level	Lower	Synthetic counterfactual
Overflow volume	Lower	Model-derived, not measured
Minimum pressure	Constraint satisfied	Detailed replay result
Energy	Modestly lower than RBC	Scenario-specific
Operator action	Approved once, revised once	Audit-log evidence
Fallback	Not required	All mandatory inputs acceptable

APPENDIX I. NATURE-BASED, GROUNDWATER AND WATER-QUALITY INTEGRATION

Nature-based assets are represented as dynamic infrastructure with capacity, condition, uncertainty, and maintenance state. Detention wetlands, bioswales, infiltration basins, restored channels, green roofs, and permeable surfaces provide storage or treatment only when antecedent moisture, sediment, vegetation, clogging, groundwater level, and outlet condition permit. The twin therefore distinguishes nominal design capacity from currently available capacity. A maintenance inspection can change the operational model immediately, while long-term performance updates require monitored evidence.

Managed recharge requires a protected water balance and water-quality decision. Candidate inflow is classified by source, first-flush condition, turbidity or surrogate quality, treatment status, receiving-zone vulnerability, groundwater level, and regulatory requirement. The controller limits recharge when mounding, poor recovery, contamination risk, or downstream ecological flow is binding. Recharge is never credited as water-supply benefit without an accounting rule that addresses travel time, recoverability, background abstraction, and losses.

Groundwater coupling can begin with a seasonal control-volume model before adopting a fully distributed solver. Inputs include recharge, abstraction, lateral exchange, storage coefficient, observed heads, and uncertainty. The seasonal model tests whether short-term stormwater decisions are consistent with annual groundwater objectives. A higher-resolution model is justified when spatial mounding, contaminant transport, stream-aquifer exchange, or well interference changes a specific decision. Added complexity must bring measurable observability and validation data.

Water-quality integration is risk-based. For pressurized supply, the twin tracks source fraction, water age, storage turnover, disinfectant decay, and events that can create

intrusion or stagnation risk. For drainage and recharge, it tracks first flush, suspended solids or turbidity proxies, receiving-water sensitivity, and treatment availability. A quality constraint can override a volumetric benefit. Predictions are displayed with the applicable model scope; simplified decay or mixing models are not presented as pathogen measurements.

Green-grey coordination tests combinations rather than treating nature-based infrastructure as an isolated amenity. Pre-releasing a detention basin may create storm capacity, but the action can reduce wetland treatment time or downstream habitat flow. Holding water may support recharge but increase mosquito, embankment, or later-storm risk. Pumping can preserve supply while consuming high-carbon electricity. The multi-objective formulation exposes these trade-offs, while legal, public-health, and structural limits remain hard constraints rather than negotiable weights.

Monitoring design follows the decision chain. Basin level and outlet position support storage accounting; rainfall distribution and inflow support event calibration; groundwater heads and recovery support recharge limits; turbidity or conductivity may provide fast quality screening; vegetation and sediment surveys support capacity updates. Remote sensing can improve spatial coverage but must be calibrated to field observations. A sensor is funded because it can change a protected decision, reduce uncertainty, or verify benefit—not merely because it is technologically available.

Maintenance is integrated into prediction. Persistent disagreement between expected and observed infiltration can indicate clogging, while declining storage-response or altered hydrograph shape can indicate sediment or outlet obstruction. The twin may prioritize inspection, but it does not diagnose ecology from a single residual. Work orders record observed condition, action, restored capacity, and evidence. The updated capacity is validated through subsequent events, creating a feedback loop between digital estimation and physical stewardship.

Performance reporting separates hydrologic, water-supply, water-quality, ecological, energy, and equity outcomes. Captured runoff is not automatically counted as benefit if receiving-water impact worsens. Recharge volume is reported with recovery and quality assumptions. Avoided overflow is accompanied by downstream stage or pollutant consequences. Nature-based outcomes include inspection and condition indicators because short monitoring records may not reveal long-term ecological response. This balanced scorecard prevents optimization from shifting an unmeasured burden outside the model boundary.

TABLE XXVII. INTEGRATED GREEN-GREY DECISION CHECKS

Decision	Potential benefit	Mandatory safeguard
Detention pre-release	Create storm storage	Downstream stage/quality
Wetland retention	Treatment + habitat	Emergency freeboard
Recharge routing	Aquifer support	Source quality + mounding
Tank turnover	Lower water age	Supply reserve
Pump rescheduling	Energy/carbon reduction	Pressure + reliability
Maintenance trigger	Restore capacity	Field confirmation

APPENDIX J. PERFORMANCE MEASUREMENT AND REPORTING TEMPLATE

Evaluation begins with a pre-registered baseline, decision horizon, comparison strategy, inclusion rule, metric definition, and statistical method. Rule-based control is replayed under exactly the same realized forcing and initial state as the proposed strategy. Events are paired so weather and demand variability do not masquerade as control benefit. Failed simulations, fallback events, and rejected recommendations remain in the analysis with stated treatment. Metric scripts and scenario seeds are versioned before final results are calculated.

Forecast performance is reported by variable, lead time, operating regime, and event type. Suitable measures include bias, MAE or RMSE, NSE where meaningful, peak error, timing error, interval coverage, and sharpness. Aggregate skill is accompanied by critical-tail performance because operational value may depend on rare extremes. A forecast can be statistically strong yet unsafe if it misses the few events that activate a hard constraint. Conversely, a modest point forecast may support good decisions when uncertainty is well calibrated.

Operational outcomes include flood or spill volume, duration, maximum level, unmet demand, minimum pressure, storage reserve, energy, pump starts, leakage volume, detection delay, localization error, water age, and quality exceedance. Each metric defines spatial boundary, time aggregation, denominator, and whether it is observed, assimilated, simulated, or counterfactual. Results report absolute values and paired differences where disclosure permits. Percent improvement is never shown without the baseline and uncertainty.

Reliability is evaluated across ordinary, disturbed, and extreme regimes. The report lists hard-constraint violations, solver failures, fallback entries, stale-data periods, missed acknowledgements, and time outside the validated envelope. A strategy that achieves attractive median savings by failing in the tail is not acceptable. Conditional value-at-risk and worst-case summaries complement mean results. The system receives credit for safe fallback when it correctly refuses a recommendation that cannot be verified.

Human performance is part of the system result. Measures include time to understand, approval and rejection frequency, override reason, alarm burden, response time, recovery completion, and operator-rated clarity. These measures are interpreted carefully: high acceptance is not automatically good, and frequent rejection may reveal missing constraints or poor interface design. The report distinguishes recommendations that were never actionable from those declined for operational context not represented in the twin.

Equity reporting disaggregates minimum service, pressure deficits, interruption, flood burden, water quality, and recovery time by service zone and critical-customer class. Where demographic data are lawfully and ethically available, impact distribution can be examined without exposing household identity. Emergency priorities and minimum domestic-service rules are documented in advance. Aggregate optimization benefit does not justify a repeated burden on peripheral or historically underserved areas.

Economic reporting separates measured, modelled, and contingent value. Metered energy and verified avoided production may be directly estimated; avoided leakage combines reconciliation and repair evidence; avoided flood

loss requires probability and depth-damage analysis; deferred capital depends on sustained service and updated demand. Lifecycle cost includes instrumentation, communications, integration, compute, licensing, model stewardship, cybersecurity, operator time, maintenance, and independent assurance. Sensitivity ranges are reported instead of a single universal payback claim.

The release summary contains the study boundary, model and data versions, validation period, principal assumptions, accepted and rejected events, performance with confidence intervals, known limitations, unresolved hazards, operator feedback, benefit classification, and next evidence gate. A failure appendix lists near-misses and adverse results. This reporting format encourages learning and prevents selective presentation of only favorable scenarios. The synthetic benchmark in this paper would therefore be reported as development evidence suitable for a shadow pilot, not as proof of field savings.

TABLE XXVIII. MINIMUM EVIDENCE DASHBOARD

Dimension	Core indicator	Required qualifier
Forecast	Skill + interval coverage	Lead time/event class
Physics	Mass/quality balance	Tolerance + failures
Control	Paired benefit + tail loss	Baseline + uncertainty
Safety	Hard violations + fallback	Operating envelope
Human	Decision time + overrides	Reason classification
Equity	Zone-level service burden	Critical classes
Economics	Measured/modelled value	Lifecycle cost + range
Governance	Version + approval status	Evidence expiry

APPENDIX K. FIELD-PILOT MONITORING AND SAMPLING PLAN

A field pilot should be designed prospectively around the claims it is intended to test. The utility, independent evaluator, and research team agree on primary outcomes, observation period, control authority, comparison strategy, exclusion rules, and stopping conditions before advisory operation begins. A full wet and dry season is preferable where monsoon variability is material. The plan identifies which measurements are operational records, which require temporary reference instruments, and which outcomes can only be estimated by calibrated counterfactual analysis.

Baseline monitoring begins before the twin influences decisions. It records existing rules, operator interventions, maintenance activity, pump availability, source limitations, telemetry quality, rainfall coverage, service complaints, and abnormal events. This period establishes natural variability and reveals whether the proposed metrics are observable. If mass balance, rainfall, or asset-status records are too incomplete to support causal comparison, instrumentation and data governance are corrected before benefit claims are attempted.

Reference instruments are deployed selectively at high-information locations. Temporary calibrated pressure loggers verify critical and peripheral zones; portable flow measurement checks trunk and district meters; independent level observations verify tank and basin scaling; additional rain gauges assess spatial bias; and pump power measurements verify energy baselines. Installation, zero checks, calibration certificates, clock alignment, and removal dates are documented. Reference devices are not assumed perfect; their uncertainty enters the evaluation.

Wet-weather sampling follows event triggers based on forecast, antecedent condition, basin storage, and staff safety. Measurements may include rainfall depth and intensity, inflow and outflow, peak stage, overflow occurrence, water-quality screening, and post-event storage recovery. The sampling plan anticipates missed events and access restrictions. Automatic samplers or remote observations are preferred where manual collection would expose staff to unsafe conditions. Field notes preserve obstructions, bypasses, debris, and unusual boundary stages.

Supply monitoring covers representative weekday, weekend, seasonal, restoration, and emergency regimes. The plan records production, transfers, district inflow, tank level, pump status and energy, critical pressure, source quality, and planned outages. Controlled tests are scheduled only with operational approval and customer safeguards. A recommendation that changes pumping or storage is evaluated against the same service and safety requirements as the existing rule, not only against energy or average pressure.

Operator evaluation uses structured but low-burden records. For each actionable recommendation, the interface captures review time, decision, reason, missing context, perceived clarity, and whether the action was operationally possible. Short interviews and after-action reviews explore trust, workload, coordination, and training needs without treating acceptance as the desired answer. Observed workarounds are mapped into requirements or constraints. Participation and privacy arrangements are approved institutionally before data collection.

Independent review is planned at defined milestones. The evaluator receives frozen model versions, event lists, metric scripts, accepted exclusions, and access to raw or appropriately protected evidence. Interim review can stop the pilot for safety or data-quality reasons without biasing the final analysis. Disagreements are logged before results are unblinded. The utility retains authority over operations, while the evaluator retains authority over whether evidence supports the claimed performance.

Pilot stopping conditions include a hard-constraint violation linked to the recommendation, repeated unexplained divergence, unbounded command behavior, loss of independent interlock, cybersecurity compromise, inability to reconstruct an action, or sustained data quality below the acceptance gate. Stopping does not imply permanent failure; it triggers containment, investigation, corrective action, regression testing, and a documented decision about resumption. This procedure makes safety and transparency stronger than pressure to complete a demonstration.

At completion, the pilot report separates technical feasibility, operational usefulness, verified benefit, and readiness for greater authority. A system may forecast accurately yet fail workflow integration, or improve operations while remaining unsuitable for automation. The next phase is approved only when the evidence addresses the stated claims and residual risks. All other conclusions are labelled hypotheses for further monitoring rather than converted into broad deployment promises.

TABLE XXIX. PROSPECTIVE PILOT MEASUREMENT SCHEDULE

Evidence	Frequency	Purpose
SCADA and quality flags	Native interval	State and action replay
Reference pressure/flow	Campaign + spot checks	Bias and uncertainty
Rainfall and basin level	Continuous/event	Storm balance

Evidence	Frequency	Purpose
Pump energy	Interval meter	Verified energy baseline
Maintenance and topology	At every change	Model validity
Operator decision log	Every recommendation	Human-system outcome
Independent review	Quarterly + incident	Evidence assurance

APPENDIX L. STATISTICAL ANALYSIS AND CLAIM DISCIPLINE

The statistical analysis plan is finalized before the final holdout or field comparison is examined. It names one or more primary outcomes and limits secondary exploratory claims. Paired comparisons are preferred because each strategy faces the same rainfall, demand, initial state, and asset availability. The analysis unit is defined explicitly as event, day, district, or control interval. Repeated observations from one event are not treated as independent samples merely because they occur at many timesteps.

Effect estimates include an absolute difference, a relative difference where the denominator is stable, and an uncertainty interval. Bootstrap resampling is performed at the event or scenario level to preserve within-event dependence. When sample size is limited, exact or permutation procedures may supplement asymptotic tests. Statistical significance is not treated as engineering importance. A small energy saving can be irrelevant if it increases pressure risk, while a rare avoided violation can justify modest average cost.

Multiplicity is controlled by pre-specifying the primary family of outcomes and treating remaining metrics as supportive or exploratory. The report does not search across many objective weights, scenario subsets, or event definitions and present only the most favorable result. Sensitivity analyses repeat the comparison under reasonable inclusion, missing-data, and counterfactual assumptions. Any post hoc subgroup is labelled clearly and accompanied by the number of events and a plausible mechanism.

Missingness is characterized by cause, duration, affected variable, and relationship to event severity. Short random gaps may be imputed for state estimation, but the performance analysis preserves an indicator and tests alternative treatments. Long or safety-critical gaps can make an event non-evaluable for a specific metric without removing it from the operational-failure record. Events are never excluded solely because the proposed method performed poorly or entered fallback.

Forecast and decision evaluations are separated. Forecast skill is calculated from predictions issued before the observation became available, with no retrospective state correction. Decision value is evaluated from the action and information available at the decision time. A good realized outcome after an unexpectedly mild event does not prove the recommendation was good; the action is judged against the contemporaneous scenario distribution and constraints. Counterfactual replay is labelled model-based evidence.

Robustness claims require a defined perturbation range. Sensor outage, bias, latency, demand shift, and asset failure are sampled jointly according to the scenario design, and performance is stratified by severity. Extrapolation beyond the tested range is prohibited. The analysis reports where benefit becomes indistinguishable from baseline, where fallback frequency rises, and where assumptions fail. This

boundary is as important as the best median result because it defines the validated operating envelope.

Economic analysis propagates technical uncertainty into monetary ranges. Energy price, water production cost, repair response, damage function, event frequency, discount rate, and lifecycle cost are varied transparently. Benefits that share the same physical mechanism are checked for double counting. Avoided production and avoided purchase, for example, may overlap. Deferred capital is reported separately from operating savings and requires an approved planning linkage rather than a simple modelled capacity margin.

Negative and null results remain visible. The paper reports scenarios in which strategies are similar, recommendations are rejected, computation times out, benefit reverses, or operators provide essential context. Failure analysis distinguishes model, data, optimization, interface, asset, and institutional causes. This prevents the system from learning the wrong lesson and improves external validity. A credible research paper is strengthened, not weakened, by showing the boundary of useful performance.

Claim wording follows the evidence level. Synthetic results “indicate” performance within the benchmark; shadow operation can “demonstrate predictive and workflow feasibility”; advisory trials can “estimate realized benefit” under monitored decisions; and bounded closed-loop trials can support a site-specific operational claim. None of these automatically supports another city. Transfer claims require new topology verification, calibration, holdout testing, and governance approval. The conclusions in this paper intentionally remain at the benchmark-evidence level.

TABLE XXX. CLAIM-TO-EVIDENCE HIERARCHY

Claim level	Minimum design	Allowed conclusion
Concept	Architecture + equations	Internally coherent proposal
Synthetic benchmark	Paired scenarios + stress tests	Simulation-domain benefit
Shadow pilot	Prospective live prediction	Site feasibility
Advisory trial	Logged human decisions	Estimated operational value
Bounded automation	Independent safety case	Site-specific controlled use
Transfer	New-site validation	Conditional generalization

APPENDIX M. INDEPENDENT ENGINEERING REVIEW CHECKLIST

Independent review begins by confirming the decision boundary. Reviewers identify the physical systems, geographic extent, timescale, controllable assets, excluded processes, validated regimes, and parties affected. They verify that the title, abstract, conclusions, and benefit statements do not exceed this boundary. The synthetic nature of the present case must remain visible wherever numerical performance is interpreted. Reviewers also confirm that the proposed innovation is compared with credible rules and deterministic control rather than with an artificially weak baseline.

Hydraulic and hydrologic review checks topology, elevations, storage curves, pump and valve behavior, runoff transformation, infiltration, boundary conditions, water balance, and numerical convergence. Calibration parameters are examined for physical plausibility and identifiability. The review reproduces representative ordinary, disturbed, and extreme events from archived inputs. Any mismatch between the surrogate and the detailed process model is assessed at the

decision level, including whether it could alter constraint activation or action ranking.

Data and software review traces at least one result from source observation through quality control, state estimation, forecast, optimization, verification, operator display, and reported metric. Versions, units, timestamps, quality flags, random seeds, dependencies, and configuration are checked. The reviewer confirms that holdout events were not used for tuning and that failed runs are not silently omitted. Reproduction on a clean offline environment should recover the stated metric within an agreed numerical tolerance.

Control review challenges objectives and constraints. Weights are varied to expose trade-offs, while hard limits remain invariant. Reviewers test infeasibility, timeout, stale data, incorrect topology, command rejection, missing acknowledgement, and simultaneous manual action. They confirm that fallback is reachable, understandable, and periodically exercised. Candidate commands must remain bounded, expiring, reversible where claimed, and subject to independent physical interlocks. A safe refusal is recorded as correct behavior when evidence is insufficient.

Security, privacy, and equity review verifies least privilege, segmentation, signed configuration, audit logging, backup recovery, retention, and incident ownership. It examines whether household or critical-infrastructure data are minimized and appropriately protected. Service and flood outcomes are disaggregated by zone and critical customer class. Reviewers test whether aggregate optimization could repeatedly transfer pressure loss, overflow, poor quality, or delayed recovery to the same community and whether transparent constraints prevent that outcome.

Operational review is performed with control-room, field, maintenance, planning, information-technology, and public-health representatives. It confirms that asset changes reach the model, alarms have owners, recommended actions are physically possible, explanations support timely judgment, and recovery procedures match actual staffing. Unresolved assumptions receive named owners and deadlines. Approval is conditional on training, drills, maintenance budget, calibration capability, and governance continuing after the research team or vendor leaves.

The final review opinion distinguishes conforming evidence, minor observations, major deficiencies, and stop conditions. It states the approved operating envelope, residual risks, compensating controls, evidence expiration, and next review trigger. Authors respond to each finding with a change, evidence, or documented disagreement. This structured challenge is not a ceremonial sign-off; it is the mechanism that converts an innovative prototype into a defensible civil-engineering decision system.

TABLE XXXI. INDEPENDENT REVIEW SIGN-OFF MATRIX

Review area	Evidence sample	Disposition
Scope and claims	Abstract, limits, conclusions	Accept/revise
Physics	Balance + extreme replay	Pass/fail
Data and code	Lineage + clean reproduction	Pass/fail
Control safety	Fault and fallback tests	Pass/fail
Security/privacy	Threat and access review	Accept/mitigate
Equity	Zone-level burden analysis	Accept/mitigate
Operations	Drill + ownership	Ready/not ready

REFERENCES

- [1] M. M. H. van der Laan et al., "From digital twin paradigm to digital water services," *Journal of Hydroinformatics*, vol. 25, no. 6, pp. 2444–2461, 2023.
- [2] P. Ghorbani Bam et al., "Digital Twin Applications in the Water Sector: A Review," *Water*, vol. 17, no. 20, Art. 2957, 2025, doi: 10.3390/w17202957.
- [3] Y. Yang et al., "Digital twinning of river basins towards full-scale, sustainable and equitable water management and disaster mitigation," *npj Natural Hazards*, vol. 1, Art. 43, 2024, doi: 10.1038/s44304-024-00047-2.
- [4] J. E. Pesantez, A. Alghamdi, S. Sabu, C. Mahinthakumar, and A. Berglund, "Using a digital twin to explore water infrastructure impacts during the COVID-19 pandemic," *Sustainable Cities and Society*, vol. 77, Art. 103520, 2022, doi: 10.1016/j.scs.2021.103520.
- [5] World Bank, "Introducing Digital Water," *Water Global Practice*, Washington, DC, 2024.
- [6] World Bank, *Utility of the Future 4.0: Taking Water and Sanitation Utilities Beyond the Next Level*. Washington, DC, 2024.
- [7] NITI Aayog, *Composite Water Management Index: A Tool for Water Management*. New Delhi, India, 2018.
- [8] NITI Aayog, *Reuse of Treated Wastewater in Urban/ Peri-Urban Agriculture in India*. New Delhi, India, 2023.
- [9] L. A. Rossman, *EPANET 2.2 User Manual*, EPA/600/R-20/133. U.S. Environmental Protection Agency, Cincinnati, OH, 2020.
- [10] L. A. Rossman and W. C. Huber, *Storm Water Management Model User's Manual Version 5.2*, EPA/600/R-22/030. U.S. Environmental Protection Agency, 2022.
- [11] P. Pandey et al., "Machine Learning-Based Digital Twin for Water Distribution Networks," *Engineering Proceedings*, vol. 69, no. 1, Art. 201, 2024.
- [12] N. S. V. Lund, A. K. V. Falk, M. Borup, H. Madsen, and P. S. Mikkelsen, "Model predictive control of urban drainage systems: A review and perspective towards smart real-time water management," *Critical Reviews in Environmental Science and Technology*, vol. 48, no. 3, pp. 279–339, 2018, doi: 10.1080/10643389.2018.1455484.
- [13] J. Oh and M. Bartos, "Model predictive control of stormwater basins coupled with real-time data assimilation enhances flood and pollution control under uncertainty," *Water Research*, vol. 235, Art. 119825, 2023, doi: 10.1016/j.watres.2023.119825.
- [14] A. Mullapudi, M. J. Lewis, C. L. Gruden, and B. Kerkez, "Deep reinforcement learning for the real time control of stormwater systems," *Advances in Water Resources*, vol. 140, Art. 103600, 2020, doi: 10.1016/j.advwatres.2020.103600.
- [15] L. Sun et al., "Integrated intra-storm predictive analysis and real-time control for urban stormwater storage to reduce flooding risk in cities," *Sustainable Cities and Society*, vol. 92, Art. 104506, 2023, doi: 10.1016/j.scs.2023.104506.
- [16] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019, doi: 10.1016/j.jcp.2018.10.045.
- [17] F. Kratzert et al., "Rainfall-runoff modelling using Long Short-Term Memory networks," *Hydrology and Earth System Sciences*, vol. 22, pp. 6005–6022, 2018, doi: 10.5194/hess-22-6005-2018.

- [18] P. Bhasme, P. K. Choudhury, and A. K. Singh, "Physics-informed machine learning for hydrological processes," *Journal of Hydrology*, vol. 603, Art. 126767, 2021, doi: 10.1016/j.jhydrol.2021.126767.
- [19] G. Feng et al., "Developing a physics-informed deep learning model to simulate runoff response to climate and cryosphere change," *Water Resources Research*, vol. 59, 2023, doi: 10.1029/2022WR034118.
- [20] J. E. Nash and J. V. Sutcliffe, "River flow forecasting through conceptual models part I—A discussion of principles," *Journal of Hydrology*, vol. 10, no. 3, pp. 282–290, 1970, doi: 10.1016/0022-1694(70)90255-6.
- [21] D. P. Solomatine and A. Ostfeld, "Data-driven modelling: Some past experiences and new approaches," *Journal of Hydroinformatics*, vol. 10, no. 1, pp. 3–22, 2008.
- [22] L. García et al., "Modeling and real-time control of urban drainage systems: A review," *Advances in Water Resources*, vol. 85, pp. 120–132, 2015, doi: 10.1016/j.advwatres.2015.08.007.
- [23] S. P. Rimer et al., "pystorms: A simulation sandbox for the development and evaluation of stormwater control algorithms," *Environmental Modelling & Software*, vol. 162, Art. 105635, 2023, doi: 10.1016/j.envsoft.2023.105635.
- [24] J. M. Sadler et al., "Leveraging open source software and parallel computing for model predictive control of urban drainage systems using EPA-SWMM5," *Environmental Modelling & Software*, vol. 120, Art. 104484, 2019, doi: 10.1016/j.envsoft.2019.07.009.
- [25] B. P. Wong and B. Kerkez, "Real-time control of urban headwater catchments through linear feedback," *Water Resources Research*, vol. 54, pp. 7309–7330, 2018, doi: 10.1029/2018WR022657.
- [26] Intergovernmental Panel on Climate Change, *Climate Change 2022: Impacts, Adaptation and Vulnerability*. Cambridge, U.K.: Cambridge Univ. Press, 2022.
- [27] World Health Organization, *Guidelines for Drinking-water Quality*, 4th ed., incorporating the first and second addenda. Geneva, Switzerland, 2022.
- [28] World Bank, *Water Infrastructure Resilience*. Washington, DC: World Bank, 2020.
- [29] H. R. Maier et al., "Evolutionary algorithms and other metaheuristics in water resources: Current status, research challenges and future directions," *Environmental Modelling & Software*, vol. 62, pp. 271–299, 2014, doi: 10.1016/j.envsoft.2014.09.013.
- [30] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002, doi: 10.1109/4235.996017.
- [31] A. Saltelli et al., *Global Sensitivity Analysis: The Primer*. Chichester, U.K.: Wiley, 2008.
- [32] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960, doi: 10.1115/1.3662552.
- [33] G. D. McCarthy et al., "Digital water: Industry leaders chart the transformation journey," *Water Research Foundation*, Denver, CO, 2019.
- [34] Government of India, Ministry of Jal Shakti, "Strengthening Groundwater Management for India," Press Information Bureau, Jan. 2026.
- [35] Bhabha Atomic Research Centre, "Empowering local decisions with isotope hydrology," *BARC Newsletter*, 2025.