Uncertainty-Aware Neural Runoff Forecasting for Robust Low-Flow Decisions

Celia Izaguirre*[1][0009-0009-1664-9713], Sandra Ranilla-Cortina[2][0000-0002-5326-5657], Inés Couso[1][0000-0002-1675-6203], and Luciano Sánchez[2][0000-0002-2446-1915]

[1] Statistics Department, Universidad de Oviedo, Spain
{izaguirrecelia,couso}@uniovi.es
[2] Computer Science Department, Universidad de Oviedo, Spain
{luciano,ranillasandra}@uniovi.es

Abstract. Thermal power stations depend on cooling water, but river abstraction is regulated and increasingly exposed to low-flow episodes. When discharge falls below a regulatory minimum, operators must rely on stored water; unnecessary pumping is costly, whereas missed pumping may cause production stoppage. The goal is therefore not average runoff accuracy, but reliable threshold decisions near the abstraction limit.
To target this decision directly, we couple a variational recurrent Seq2Seq runoff forecaster with a scenario-wise robust decision layer. The encoder defines a Gaussian posterior over latent basin state; recurrent decoders predict reservoir volume and discharge, and Monte-Carlo sampling yields predictive distributions under baseline, drought, flash-event and no-rain rainfall regimes. Candidate pumping thresholds are scored by asymmetric operational cost and selected by a decision criterion over scenarios.
We compare single-scenario tuning with maximin, maximax and Savage min–max regret rules on three synthetic basins and on the Ebro-at-Tudela record. The Savage criterion yields the lowest maximum regret on every non-trivial dataset. On the largest synthetic basin, robust selection reduces missed-pumping events from 668 under the baseline rule to 3; on the real basin, Savage reduces cost by about 7–10% and reduces missed events from 751 to 688. The results show that neural runoff models are more useful for low-flow management when evaluated as inputs to robust decisions, not only by forecasting error.

Keywords: Runoff forecasting · Decision under uncertainty · Variational autoencoder · Sequence-to-sequence · Min–max regret · Climate uncertainty.

1 Introduction

Thermal power stations require reliable access to cooling water, but river abstraction is regulated and may be restricted under low-flow conditions to preserve

* Corresponding author.

higher-priority uses such as human consumption and irrigation. Climate change increases the frequency and severity of drought and low-flow episodes [7], so plants that depend on river water must operate with a buffer: when discharge is sufficiently high, water can be pumped into on-site storage and later used if abstraction is no longer authorised. This creates an asymmetric preventive decision problem. Each day, the operator must decide whether to store water before discharge crosses a low-flow threshold during the following ten days. A false positive causes unnecessary pumping; a false negative may force production to stop. The relevant prediction target is therefore concentrated near the abstraction limit. Large errors at high flows may not affect the decision, whereas small errors near the limit can dominate operational cost. The task is closer to cost-sensitive classification around a threshold than to unconstrained runoff regression [12].

Neural rainfall–runoff models, especially LSTM-based architectures, provide strong multi-step discharge forecasts [6]. However, a point forecast does not define a pumping policy. The operator must still choose how conservative the alarm threshold should be. In conventional deployment, this offset is tuned post-hoc, often against a single reference future such as the historical baseline or a no-rainfall scenario. Such tuning hides the operator's risk attitude inside one parameter and may over-specialise the rule to that scenario. We address this deployment problem by coupling an uncertainty-aware Seq2Seq runoff forecaster with a robust decision layer. The encoder defines a variational latent basin state; recurrent decoders predict reservoir volume and discharge; and latent sampling yields predictive distributions under alternative rainfall scenarios. For each scenario, candidate pumping thresholds are evaluated through their asymmetric operational cost.

The deployed threshold is then selected by an explicit decision criterion over scenarios, rather than by single-scenario tuning. We compare the naive baseline with Wald's maximin criterion, an optimistic maximax rule and Savage min–max regret [11]. These criteria encode different risk attitudes: minimising worst-case cost, favouring the best-case scenario, or limiting the largest opportunity loss relative to the scenario-wise optimum.

We evaluate the full forecast–decision pipeline on three synthetic basins of increasing routing complexity and on the real Ebro-at-Tudela record. The synthetic basins identify when scenario-aware deployment matters; the Ebro case tests whether the same mechanism persists on observed data. Across the non-trivial datasets, Savage min–max regret sharply reduces regret relative to the single-scenario rule and yields fewer missed-pumping stoppage events.

The paper makes four contributions. First, it formulates thermal-plant low-flow management as an asymmetric preventive pumping problem. Second, it introduces a robust deployment layer over neural runoff forecasts. Third, it adapts a recurrent Seq2Seq forecaster with a variational latent basin state for uncertainty-aware low-flow prediction. Fourth, it evaluates the resulting forecast–decision pipeline on controlled synthetic basins and on the real Ebro record, showing when robust threshold selection improves over single-scenario deployment.

2 Related Work

Thermal power generation is exposed to hydrological variability through cooling-water demand, abstraction constraints and competition with other water uses [9]. At the operational scale considered here, this exposure becomes a threshold problem: abstraction is possible while discharge remains above the regulatory limit, but stored water is needed once low-flow conditions make abstraction unavailable. Forecasting therefore matters most near the abstraction threshold, where errors determine whether preventive pumping is triggered in time. Neural rainfall–runoff models [13], and, in particular, LSTM-based models perform strongly in both single-basin and large-sample settings [6]. Their hidden states can encode storage-like hydrological processes [8], and encoder–decoder architectures extend them to multi-step forecasting [4].

Probabilistic hydrological forecasting moves from trajectories toward decision support by representing uncertainty. Aleatoric and epistemic uncertainty provide a useful distinction [2], and rainfall–runoff models can include probabilistic outputs, Monte-Carlo dropout and related mechanisms [5]. Generative approaches, including conditional variational autoencoders, extend this idea to multi-step streamflow forecasting [3]. These methods provide distributions of plausible discharges, but the pumping decision still requires mapping that distribution to action under asymmetric costs and assumed future rainfall regimes.

Forecast-value research makes the decision dependence explicit. Cost–loss models quantify the economic value of ensemble streamflow predictions [12], with extensions to risk-averse users and broader utility formulations [7]. Recent work shows that forecast skill and forecast value can diverge, especially in drought-sensitive contexts [10]. Scenario-based and robust decision methods address how decisions should perform across plausible futures. In water-resource planning, this has led to robustness-oriented approaches for climate uncertainty [1], including minimax-regret analyses of flood-risk and infrastructure strategies [11].

In our setting, this motivates a scenario-conditioned cost surface over candidate pumping thresholds. The surface links the probabilistic runoff model, the rainfall scenarios and the asymmetric operational loss. Robust criteria such as maximin, maximax and Savage min–max regret then select the deployed threshold from this surface, moving the evaluation from forecast trajectories to operational decisions.

These strands motivate our setting but leave a gap at the operational interface: forecast-value studies usually evaluate given probabilistic forecasts, whereas robust water-resource planning often targets longer-term policy or infrastructure choices. Preventive cooling-water pumping is instead a short-horizon regulatory-threshold problem with asymmetric loss, for which we couple uncertainty-aware neural runoff forecasts with explicit robust threshold selection.

3 Method

Following the decision-centred view above, the forecast is used through the pumping rule it induces. Let $Q(t)$ be discharge at the control point and $Q_{\min}$

the minimum discharge for authorised abstraction. Given a T-day forecast $\widehat{Q}(t+1), \ldots, \widehat{Q}(t+T)$, preventive pumping is triggered when the predicted discharge falls below a shifted threshold:

$$\widehat{a}_\tau(\delta) = \mathbb{1}\Big[\widehat{Q}(t+\tau) \leq Q_{\min} + \delta\Big]. \tag{1}$$

The offset δ is the deployment variable. Larger values pump earlier and reduce missed low-flow events; smaller values avoid unnecessary pumping but increase exposure to false negatives.

The operational loss is asymmetric. False positives are unnecessary pumping actions; false negatives are missed preventive actions that may force production to stop. Over an evaluation window,

$$\mathcal{C} = c_{\mathrm{FP}}\mathrm{FP} + c_{\mathrm{FN}}\mathrm{FN},$$

with $c_{\mathrm{FN}} \gg c_{\mathrm{FP}}$; in the experiments $c_{\mathrm{FN}}/c_{\mathrm{FP}} = 100$. Since future rainfall is unknown, each candidate threshold is evaluated under a scenario library $\mathcal{S}$. For each scenario, the runoff model generates forecasts, the thresholded actions are scored, and the decision layer selects δ from the resulting scenario-wise costs.

3.1 Variational runoff forecaster

The forecasting component is a recurrent sequence-to-sequence model that produces scenario-conditioned discharge distributions. Its input is the past H days of covariates $\boldsymbol{x}_\tau = (P_\tau^\Sigma, V_\tau^\Sigma, Q_\tau)$: aggregated rainfall, aggregate reservoir storage and control-point discharge, with Σ denoting the corresponding station or reservoir aggregation and τ the input-history day. The key modelling choice is to treat the basin state inferred from this history as uncertain. Aggregated observations do not uniquely determine soil moisture, routing delays, reservoir conditions or upstream contributions, and this ambiguity is most consequential near the abstraction threshold. We therefore replace the deterministic encoder state by a Gaussian latent state:

$$q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{\mu}_\phi(\boldsymbol{x}), \operatorname{diag} \boldsymbol{\sigma}_\phi^2(\boldsymbol{x})\big). \tag{2}$$

An encoder LSTM reads the history, and two linear heads produce $\boldsymbol{\mu}_\phi$ and $\log \boldsymbol{\sigma}_\phi^2$. Training uses the standard reparameterisation $\boldsymbol{z} = \boldsymbol{\mu}_\phi + \boldsymbol{\sigma}_\phi \odot \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Samples from q_ϕ initialise the two decoders in Figure 1.

The reservoir decoder predicts the future aggregate storage trajectory. The discharge decoder is conditioned on future rainfall and on the predicted reservoir trajectory. Instead of regressing discharge directly, it outputs a compliance score $\widehat{p}_\tau \in [0, 1]$, which is mapped to discharge units by a monotone PCHIP calibrator. This gives a Wiener-like structure: a dynamic recurrent block followed by a static monotone output map. The calibrator preserves the ordering of discharge predictions and avoids letting high-flow squared errors dominate the low-flow decision region.

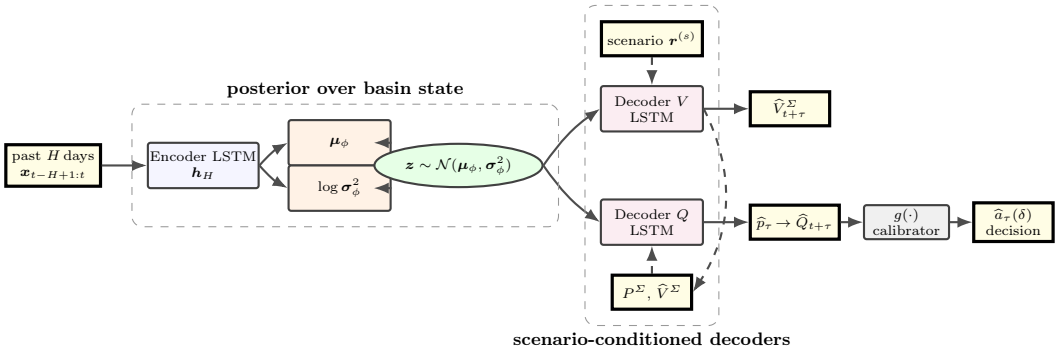


Fig. 1. Variational Seq2Seq architecture. The encoder maps the history $\boldsymbol{x}_{t-H+1:t}$ to a Gaussian posterior over the latent basin state $\boldsymbol{z}$. Samples initialise the reservoir and discharge decoders. The reservoir decoder predicts $\widehat{V}^{\Sigma}_{t+\tau}$ under scenario rainfall $\boldsymbol{r}^{(s)}$; the discharge decoder outputs a compliance score $\widehat{p}_\tau$, mapped to discharge by a monotone calibrator $g(\cdot)$. Monte-Carlo samples over latent states and rainfall trajectories induce the cost surface $L(\delta, s)$.

The model is trained with a β-weighted variational objective:

$$\mathcal{L} = \mathcal{L}_V + \lambda_Q \mathcal{L}_Q + \beta \,\mathrm{KL}\big(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})\big), \tag{3}$$

where $\mathcal{L}_V$ combines reservoir error with a smoothness penalty, $\mathcal{L}_Q$ is a class-weighted binary cross-entropy on the low-flow alarm label, and the KL term regularises the posterior.

At inference time, the model is run under each rainfall scenario s. For a future rainfall trajectory $\boldsymbol{r}^{(s)}$, the predictive score is approximated by Monte-Carlo sampling:

$$\widehat{p}^{(s)}_\tau(\boldsymbol{x}) \;\approx\; \frac{1}{K} \sum_{k=1}^{K} \widehat{p}_\tau\big(\boldsymbol{x};\, \boldsymbol{z}^{(k)}, \boldsymbol{r}^{(s)}\big), \quad \boldsymbol{z}^{(k)} \sim q_\phi(\boldsymbol{z} \mid \boldsymbol{x}). \tag{4}$$

Repeating this over rainfall scenarios and candidate thresholds yields the scenario-wise costs used by the robust decision criteria.

The predictive distribution is used through thresholded actions, not discharge error directly: for each δ, sampled forecasts become pumping alarms and are compared with whether discharge falls below $Q_{\min}$ within the horizon. With $c_{\mathrm{FP}} = 1$ and $c_{\mathrm{FN}} = 100$, $L(\delta, s)$ records the scenario-wise trade-off between unnecessary alarms and missed low-flow events.

4 Scenario-Based Deployment

The variational forecaster provides predictive distributions, but the deployed object is a pumping threshold. This threshold must account for future rainfall

Table 1. Rainfall scenario library $\mathcal{S}$. Multiplicative factors on wet-day frequency, mean wet-day intensity and Gamma shape parameter (lower $k \Rightarrow$ heavier tail). s_0 matches the training rainfall regime; s_4 is the no-rain anchor.

s	p_{wet}	μ_P	k	operational interpretation
s_0 baseline	$\times 1.00$	$\times 1.00$	$\times 1.00$	training-distribution match
s_1 mild drought	$\times 0.85$	$\times 0.95$	$\times 1.00$	moderate dry stress test
s_2 severe drought	$\times 0.60$	$\times 0.80$	$\times 1.00$	severe dry-tail stress test
s_3 flashy	$\times 0.70$	$\times 1.70$	$\times 0.60$	fewer, larger rainfall events
s_4 no rain	$\times 0.00$	–	–	short-horizon no-rain anchor

uncertainty, asymmetric pumping costs and the operator's risk attitude. Selecting it from a single reference future is fragile: a baseline continuation may underprepare the plant for drought, whereas a no-rain assumption may over-trigger pumping in ordinary conditions. We therefore evaluate each candidate threshold over a finite set of rainfall regimes and select the deployed rule from the resulting cost surface.

The scenario library $\mathcal{S}$ is a controlled stress-test basis for short-horizon low-flow operation, not a calibrated climate projection. In the synthetic basins, rainfall follows a two-state wet/dry process with seasonal modulation and Gamma-distributed wet-day intensities. The scenarios in Table 1 modify three statistics of this process: wet-day frequency p_{wet}, mean wet-day intensity μ_P, and Gamma shape k, which controls event concentration. In the real-basin evaluation, the same definitions are applied as perturbations of the observed future rainfall pattern. The five regimes cover the training-distribution baseline, two dry stress tests, a flash-event regime with fewer but heavier events, and a no-rain anchor. They are used here to make the robustness experiment controlled and reproducible; in deployment, they could be replaced by regional climate projections, downscaled products or meteorological ensembles.

For each rolling evaluation day and scenario, we generate M rainfall trajectories. In the synthetic basins, the scenario parameters modify the simulator rainfall process; in the real basin, they perturb the observed future rainfall sequence over the forecast horizon. For each trajectory, the forecaster is evaluated under K samples from the latent basin posterior. The sampled forecasts are then thresholded, converted into pumping alarms and scored by the asymmetric operational loss. In the experiments, each test day is evaluated over five scenarios, $M = 20$ rainfall trajectories, $K = 100$ latent states and 121 candidate offsets. The result is an empirical cost surface over thresholds, rainfall futures and latent basin states.

4.1 Cost surface and robust threshold selection

The deployment variable is the threshold offset δ. Larger offsets trigger pumping earlier and reduce missed low-flow events; smaller offsets reduce unnecessary pumping but increase exposure to false negatives. For each scenario s, the Monte-Carlo procedure estimates the expected cost of deploying δ:

$$L(\delta, s) \;=\; \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\cdot|\boldsymbol{x})}\Big[\, c_{\text{FP}}\,\text{FP}(\delta, s, \boldsymbol{z}) + c_{\text{FN}}\,\text{FN}(\delta, s, \boldsymbol{z}) \,\Big], \tag{5}$$

where false positives are unnecessary pumping actions and false negatives are missed preventive pumping actions. In practice, the expectation is approximated over latent samples and rainfall perturbations. The surface $L(\delta, s)$ is the interface between the forecaster and the decision rule.

The single-scenario baseline selects the threshold that minimises cost under the reference rainfall regime:

$$\delta^*_{\text{naive}} = \arg\min_{\delta} L(\delta, s_0). \tag{6}$$

The robust criteria instead select δ after aggregating costs over $\mathcal{S}$. Maximin minimises the largest realised cost,

$$\delta^*_{\text{maximin}} = \arg\min_{\delta} \max_{s\in\mathcal{S}} L(\delta, s), \tag{7}$$

whereas maximax selects the threshold favoured by the best scenario,

$$\delta^*_{\text{maximax}} = \arg\min_{\delta} \min_{s\in\mathcal{S}} L(\delta, s). \tag{8}$$

Finally, Savage's criterion minimises the largest scenario-wise regret,

$$\delta^*_{\text{Savage}} = \arg\min_{\delta} \max_{s\in\mathcal{S}} \big[L(\delta, s) - \min_{\delta'} L(\delta', s)\big]. \tag{9}$$

This compares each threshold with the best threshold that could have been chosen under the same scenario, which is useful when some rainfall regimes are intrinsically more costly than others.

5 Experimental Results and Discussion

We evaluate the method as a deployment procedure for preventive pumping. On each test day, the model and the rainfall scenarios induce a cost surface over candidate thresholds; each criterion then selects the threshold to deploy. The reported quantities therefore describe the cost and failure modes of the resulting rules, not forecast error alone. The evaluation combines controlled synthetic basins with a real Ebro case study. The synthetic basins vary routing complexity under a known data-generating process, allowing us to identify when scenario-aware threshold selection matters. The Ebro record tests whether the same mechanism persists on observed data, where rainfall, reservoir operation and discharge response are only partially represented.

5.1 Datasets and protocol

Three synthetic basins with $N_1 \in \{8, 16, 64\}$ routing nodes are generated by a discrete-time stochastic simulator. Each topology is a random directed tree rooted at the control point, with branching factors 4.0, 2.5 and 1.5, respectively; thus $N_1 = 8$ is wide and shallow, whereas $N_1 = 64$ is narrower and deeper. Three

reservoirs are placed at random headwater nodes, with total capacity calibrated to 30 days of mean discharge. Rainfall follows a seasonal two-state wet/dry process with Gamma-distributed wet-day intensities, and routing uses first-order edge-lag dynamics. Each basin yields 5479 daily observations, corresponding to 15 years. The real case is the Ebro basin upstream of Tudela, Spain. The record contains 3321 daily observations from nine rainfall stations, three reservoirs and the discharge gauge, covering nine years.

For each dataset, the variational Seq2Seq model is trained on the first 80% of the record, with half of that slice used for early stopping. We train for up to 300 epochs with latent dimension $d_z = 16$, $\beta = 10^{-2}$, $K = 100$ latent samples and Adam, where d_z is the dimension of the latent state $\boldsymbol{z}$ in Equation (2). Evaluation is rolling: each test day issues an independent ten-day forecast. For each day and scenario, $M = 20$ rainfall trajectories are sampled. The cost surface is evaluated on 121 offsets $\delta \in [-Q_{\min}, 3Q_{\min}]$, and the four criteria (6)–(9) each select a deployed threshold. Costs use $c_{\mathrm{FP}} = 1$ and $c_{\mathrm{FN}} = 100$. Thus, each test day evaluates five scenarios, 20 rainfall trajectories per scenario, 100 latent states per trajectory and 121 thresholds.

5.2 Experimental results

Table 2 reports per-day worst-case cost and maximum regret. The $N_1 = 8$ basin acts as a negative control: all criteria coincide because the forecasts never create a scenario-dependent alarm close enough to the threshold. The robust layer therefore does not create artificial gains when the cost surface contains no real decision conflict. On the remaining datasets, the decision criterion matters. Savage min–max regret gives the lowest maximum regret on every non-trivial dataset: 7 on $N_1 = 16$, 9 on $N_1 = 64$ and 1 on Ebro, versus 299, 500 and 300 for the naive single-scenario rule. The naive threshold is therefore over-specialised to the reference rainfall regime: it can be near-optimal under baseline rainfall and still incur large opportunity loss when the regime changes.

Worst-case cost gives the complementary result. On the two non-trivial synthetic basins, Maximin and Savage tie on the maximum worst-case cost, since the driest regimes dominate the upper tail and both criteria choose similarly protective thresholds. On Ebro, all criteria share the same maximum because of one high-cost forecast day; the 95th percentile is more informative, with Maximin and Savage at 154 versus about 200 for naive and maximax. The decision layer cannot repair every forecasting failure, but it improves the upper tail when threshold choice can still change the outcome.

Table 3 shows the failure mode of single-scenario deployment. The naive rule is best on the baseline scenario by construction, but its non-baseline costs are 66%–162% higher than Savage on $N_1 = 16$ and 200%–670% higher on $N_1 = 64$. Thus, a threshold tuned to ordinary rainfall can become unsafe under dry or no-rain futures. Ebro preserves the same pattern at smaller scale: Savage is best in three of the four non-baseline scenarios and essentially tied with the naive rule on the baseline.

Table 2. Headline statistics across the rolling evaluation: per-day worst-case cost and max-regret summaries. Bold marks the per-dataset best; Savage is uniformly lowest on regret, while Maximin and Savage tie on worst-case statistics on every non-trivial dataset.

dataset	criterion	$\max \mathcal{C}^{\mathrm{wc}}$	$\mathrm{p}_{95}\mathcal{C}^{\mathrm{wc}}$	$\max \mathcal{R}$	$\mathrm{p}_{95}\mathcal{R}$	$\overline{\mathcal{R}}$
synth $N_1 = 8$	naive	10	0	0	0	0.00
	all	10	0	0	0	0.00
synth $N_1 = 16$	naive	403	101	299	100	7.09
	maximin	**400**	**7**	200	0	0.57
	maximax	403	101	398	100	9.34
	savage	**400**	**7**	**7**	**0**	**0.05**
synth $N_1 = 64$	naive	504	106	500	100	18.14
	maximin	**104**	**9**	**9**	**6**	**0.67**
	maximax	503	201	499	196	26.90
	savage	**104**	**9**	**9**	**6**	**0.67**
Ebro (real)	naive	703	200	300	0	3.85
	maximin	703	**154**	200	0	2.04
	maximax	703	201	300	0	3.77
	savage	703	**154**	**1**	**0**	**0.01**

Table 3. Cumulative scenario-wise cost over the rolling window. Bold marks the best criterion per cell; Savage is best on every non-baseline scenario in $N_1 \in \{16, 64\}$ and on three of four in Ebro.

dataset	criterion	baseline	mild dr.	severe dr.	flashy	no rain
synth $N_1 = 8$	all	424	439	471	410	506
synth $N_1 = 16$	naive	5 707	7 450	9 494	11 051	13 820
	maximin	6 029	5 871	5 806	6 074	5 647
	maximax	12 689	11 338	11 382	14 939	15 124
	savage	**5 755**	**5 596**	**5 336**	**5 797**	**5 278**
synth $N_1 = 64$	naive	**3 008**	16 633	22 182	9 009	30 922
	maximin	3 604	**3 709**	**3 864**	**3 576**	4 012
	maximax	28 147	29 395	30 755	20 567	39 277
	savage	3 604	**3 709**	**3 864**	**3 576**	**4 012**
Ebro	naive	**24 231**	21 638	14 794	30 826	11 910
	maximin	24 732	21 539	14 497	28 228	11 710
	maximax	26 230	22 138	15 096	31 026	**11 110**
	savage	24 234	**19 941**	**13 700**	**28 130**	11 114

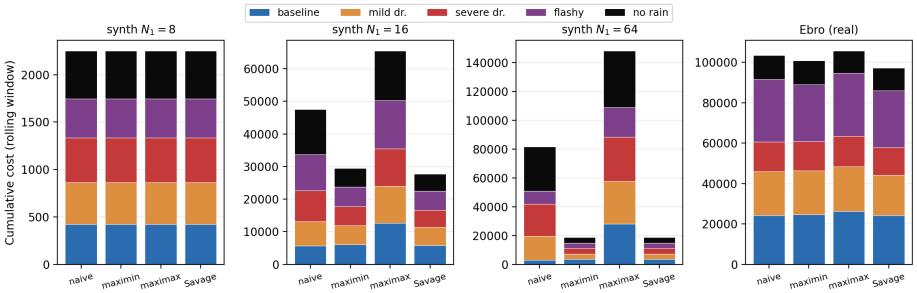


Fig. 2. Cumulative cost by scenario, dataset and criterion. The $N_1 = 8$ totals coincide, whereas $N_1 = 16$ and $N_1 = 64$ expose non-baseline cost under naive and maximax, with the same weaker pattern on Ebro.

Table 4. Cumulative missed-pumping stoppage events (false negatives) by dataset, criterion and scenario. The **total** column sums across scenarios.

dataset	criterion	baseline	mild dr.	severe dr.	flashy	no rain	**total**
synth $N_1 = 8$	all four	0	0	0	0	0	**0**
synth $N_1 = 16$	naive	40	57	77	94	120	**388**
	maximin	43	41	40	44	38	**206**
	maximax	110	96	96	133	133	**568**
	savage	**40**	**38**	**35**	**41**	**34**	**188**
synth $N_1 = 64$	naive	**1**	137	191	61	278	**668**
	maximin	**1**	**1**	**0**	**1**	**0**	**3**
	maximax	257	269	281	181	367	**1 355**
	savage	**1**	**1**	**0**	**1**	**0**	**3**
Ebro (real)	naive	195	165	87	264	40	**751**
	maximin	200	164	84	238	38	**724**
	maximax	215	170	90	266	**32**	**773**
	savage	**195**	**148**	**76**	**237**	**32**	**688**

The operationally critical error is the false negative: failing to trigger storage before a low-flow event. Table 4 shows that Savage reduces missed events from 388 to 188 on $N_1 = 16$, and that Maximin and Savage reduce them from 668 under the naive rule to only 3 on $N_1 = 64$. Thus, when the low-flow decision is difficult and scenario-dependent, robust deployment almost eliminates the catastrophic error type. On Ebro, Savage reduces missed events from 751 to 688 and is best or tied in the non-baseline scenarios, without sacrificing the baseline scenario.

5.3 Main findings

The experiments show that robust thresholds help only when low-flow decisions are scenario-dependent: the effect is absent on $N_1 = 8$, emerging on $N_1 = 16$ and decisive on $N_1 = 64$. Maximin targets worst-case cost, whereas Savage minimises maximum regret; naive and maximax reveal baseline overfitting and adverse-scenario exposure.

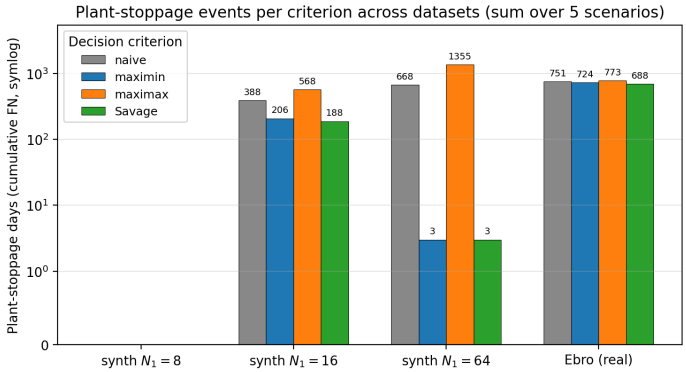


Fig. 3. Missed-pumping events by criterion, summed over scenarios on a symlog axis.

The smaller Ebro gain is consistent with partial observability from limited rainfall coverage and unmodelled regulation, abstractions and reservoir management. Robust criteria require actionable scenario information; further validation should cover more catchments, weather/climate ensembles, cost ratios, deterministic baselines and multiple seeds.

6 Conclusion

In low-flow operation, forecasts matter through induced actions: a threshold tuned to one rainfall future may be unsafe when conditions change. We therefore select thresholds over a scenario-conditioned cost surface that combines a variational Seq2Seq runoff model, rainfall scenarios and robust decision criteria. The latent state represents basin-response uncertainty, scenarios encode rainfall alternatives and the criterion encodes asymmetric costs.

Experiments show benefits when decisions become scenario-dependent: $N_1 = 8$ yields the same safe decisions, but in the largest basin missed pumping falls from 668 under the naive rule to 3 under Maximin and Savage; on Ebro, Savage reduces missed events from 751 to 688 and improves non-baseline scenarios.

Thus, low-flow protection should replace point-forecast calibration with threshold selection against operating rainfall conditions and explicit risk attitude. Future work will use basin-specific weather or climate ensembles, make the decoder uncertainty-aware end to end and incorporate GIS structure with graph neural networks.

Acknowledgements

This work has been partially supported by the Ministry of Science, Innovation and Universities ("Ministerio de Ciencia, Innovación y Universidades") from

Spain/FEDER under grants PID2023-146257OB-I00 and PCI2025-163245 "GeoAI-based Augmentation of Multi-source Urban GIS" and by Principado de Asturias, grant IDE/2024/000734.

References

1. Herman, J.D., Quinn, J.D., Steinschneider, S., Giuliani, M., Fletcher, S.: Climate adaptation as a control problem: review and perspectives on dynamic water resources planning under uncertainty. Water Resources Research **56**(2), e2019WR025502 (2020)
2. Hüllermeier, E., Waegeman, W.: Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Machine Learning **110**(3), 457–506 (2021)
3. Jahangir, M.S., Quilty, J.: Generative deep learning for probabilistic streamflow forecasting: conditional variational auto-encoder. Journal of Hydrology **629**, 130498 (2024)
4. Jahangir, M.S., You, J., Quilty, J.: A quantile-based encoder–decoder framework for multi-step ahead runoff forecasting. Journal of Hydrology **619**, 129269 (2023)
5. Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., Nearing, G.: Uncertainty estimation with deep learning for rainfall–runoff modeling. Hydrology and Earth System Sciences **26**, 1673–1693 (2022)
6. Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrology and Earth System Sciences **23**(12), 5089–5110 (2019)
7. Laugesen, R., Thyer, M., McInerney, D., Kavetski, D.: Flexible forecast value metric suitable for a wide range of decisions: application using probabilistic subseasonal streamflow forecasts. Hydrology and Earth System Sciences **27**(4), 873–893 (2023)
8. Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., de Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., Dadson, S.J.: Hydrological concept formation inside long short-term memory (LSTM) networks. Hydrology and Earth System Sciences **26**(12), 3079–3101 (2022)
9. Lohrmann, A., Farfan, J., Caldera, U., Lohrmann, C., Breyer, C.: Global scenarios for significant water use reduction in thermal power plants based on cooling water demand estimation using satellite imagery. Nature Energy **4**, 1040–1048 (2019)
10. Modi, P.A., Carbone, J.C., Jennings, K.S., Kamen, H., Kasprzyk, J.R., Szafranski, B., Wobus, C.W., Livneh, B.: Understanding the relationship between streamflow forecast skill and value across the western US. Hydrology and Earth System Sciences **29**, 5593–5623 (2025)
11. van der Pol, T.D., Gabbert, S., Weikard, H.P., van Ierland, E.C., Hendrix, E.M.T.: A minimax regret analysis of flood risk management strategies under climate change uncertainty and emerging information. Environmental and Resource Economics **68**(4), 1087–1109 (2017)
12. Roulin, E.: Skill and relative economic value of medium-range hydrological ensemble predictions. Hydrology and Earth System Sciences **11**(2), 725–737 (2007)
13. Shen, C.: A transdisciplinary review of deep learning research and its relevance for water resources scientists. Water Resources Research **54**(11), 8558–8593 (2018)