AUTONOMOUS IRRIGATION UNDER NOISY SENSOR-ACTUATOR FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Short-horizon soil-moisture forecasts are potentially useful for irrigation control, but field records often combine heterogeneous signals whose timestamps, units, and physical correspondence are uncertain. We study conditional variational forecasting under this constraint using a structured conditioning vector with environmental–crop context (64 dimensions), control history (32), and plan/phenology information (8). A data audit of the supplied aligned table identifies 3,000 complete paired rows, a soil-moisture target range of 120–700 sensor units, extreme light-intensity observations, and weak cross-stream correlations. In the reported conditioning ablation, context-only conditioning gives the lowest one-step error, whereas naive 104-dimensional fusion performs substantially worse. A reported offset-and-denoising rerun reduces full-fusion MAE by 33.7% relative to naive fusion, but does not surpass context-only conditioning. We formalize why this negative result is informative: auxiliary variables cannot increase population Bayes risk under unrestricted prediction, so empirical degradation indicates missynchronization, finite-sample error, noise, optimization, regularization, or model mismatch. We therefore position the contribution as a synchronization-aware evaluation protocol rather than a novel CVAE. The study also specifies calibration, robustness, and irrigation-utility evaluations required before claims of uncertainty-aware autonomous control can be made.

1 INTRODUCTION

Precision irrigation uses forecasts of root-zone water status to support pump and valve scheduling while respecting agronomic and hardware constraints. Classical scheduling is grounded in evapotranspiration and crop coefficients (Allen et al., 1998); predictive control instead treats estimation, forecasting, and constrained actuation as a repeated decision problem (Camacho and Bordons, 2007). In either formulation, forecast usefulness depends not only on average error but also on temporal validity, uncertainty calibration, threshold-event detection, and robustness to imperfect actuation.

Learning from field data is difficult because soil moisture is partially observed, nonlinear, regime dependent, and delayed relative to irrigation. Additional sensing can help only when the auxiliary stream refers to a compatible place, time, scale, and physical process. Concatenating weakly paired channels can therefore increase estimation error even when the channels are individually informative. This issue is central here: the available tomato and drip-irrigation records do not expose a certified shared timestamp key, so row-wise pairing is an experimental reconstruction rather than verified synchronization.

We use a conditional variational autoencoder (CVAE) (Kingma and Welling, 2014; Rezende et al., 2014; Sohn et al., 2015) as a probabilistic forecasting instrument, not as a methodological novelty claim. The scientific question is narrower and more defensible: *when does structured multimodal conditioning help or harm one-step soil-moisture forecasting under weak synchronization?*

Contributions.

1. We define one nonredundant $64 + 32 + 8$ conditioning representation separating environmental–crop context, control history, and plan/phenology information.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2. We audit the supplied aligned table and quantify missing worksheet rows, target scale, sensor artifacts, pump-state imbalance, crop-stage imbalance, and lag-correlation weakness.
3. We compare context-only, control-only, plan-only, and full-fusion CVAEs, and summarize the reported benefit of offset correction and denoising without overstating causal alignment.
4. We establish a population-risk interpretation of fusion degradation and an identifiability limitation for offset alignment without shared clocks.
5. We provide a reviewer-driven evaluation protocol covering deterministic and probabilistic baselines, calibration, multiple seeds, robustness, multi-horizon forecasting, and irrigation-relevant utility.

2 Related Work and Positioning

Irrigation forecasting and control. FAO-56 water-balance abstractions remain useful because they connect evapotranspiration, crop coefficients, and crop stage to interpretable irrigation demand (Allen et al., 1998). Model predictive control complements this view by optimizing actions over a receding horizon under state and actuator constraints (Camacho and Bordons, 2007). Our study does not claim deployed control performance; it evaluates a forecaster that could supply predictive state information to a separately validated controller.

Conditional generative forecasting. VAEs optimize a variational lower bound using amortized inference (Kingma and Welling, 2014; Rezende et al., 2014). CVAEs condition the latent prior and decoder on observed covariates, allowing multiple predictive samples for the same context (Sohn et al., 2015). A latent variable, however, does not by itself guarantee calibrated uncertainty. Coverage, sharpness, proper scoring rules, and subgroup behavior must be measured explicitly (Gneiting and Raftery, 2007).

Multimodal learning under weak alignment. Multimodal systems can benefit from complementary channels, but early concatenation is sensitive to asynchronous sampling, missing intervals, and representation mismatch. Learned alignment and cross-modal attention have been developed for unaligned sequences (Tsai et al., 2019). We use simpler offset diagnostics because the available table is already reconstructed; the limitations of this choice are made explicit.

Position of this work. The strongest evidence is a structured negative result: naive multimodal fusion is unreliable when auxiliary temporal semantics are weak. The contribution is therefore an evaluation and reporting framework for synchronization-aware irrigation forecasting, not a new variational architecture or evidence of autonomous field deployment.

3 Data, Task, and Evidence Status

3.1 Supplied aligned table

The supplied spreadsheet contains 10,519 worksheet rows, of which 3,000 rows contain complete aligned measurements and 6,919 are empty outside the populated table. The complete records contain drip-side soil, temperature, humidity, light intensity, field-operation status, and pump state, together with tomato-side soil moisture, evapotranspiration, days after planting, and crop stage. The target ranges from 120 to 700 in undocumented sensor units; because a physical calibration is unavailable, we do not relabel these values as volumetric water content. Light intensity reaches 25,219, supporting explicit artifact and saturation diagnostics.

Pump state takes values 0, 1, and 2 with counts 550, 508, and 1,942, respectively. Crop-stage counts are also imbalanced: 167 initial, 129 development, 2,501 mid-stage, and 203 late-stage records. These imbalances motivate stratified reporting rather than a single aggregate score.

3.2 Forecasting task and weak pairing

For a conditioning vector h_t constructed from information indexed no later than t, the target is

$$y_{t+1} = \mathrm{SoilMoisture}_{t+1}, \qquad p_\theta(y_{t+1} \mid h_t). \tag{1}$$

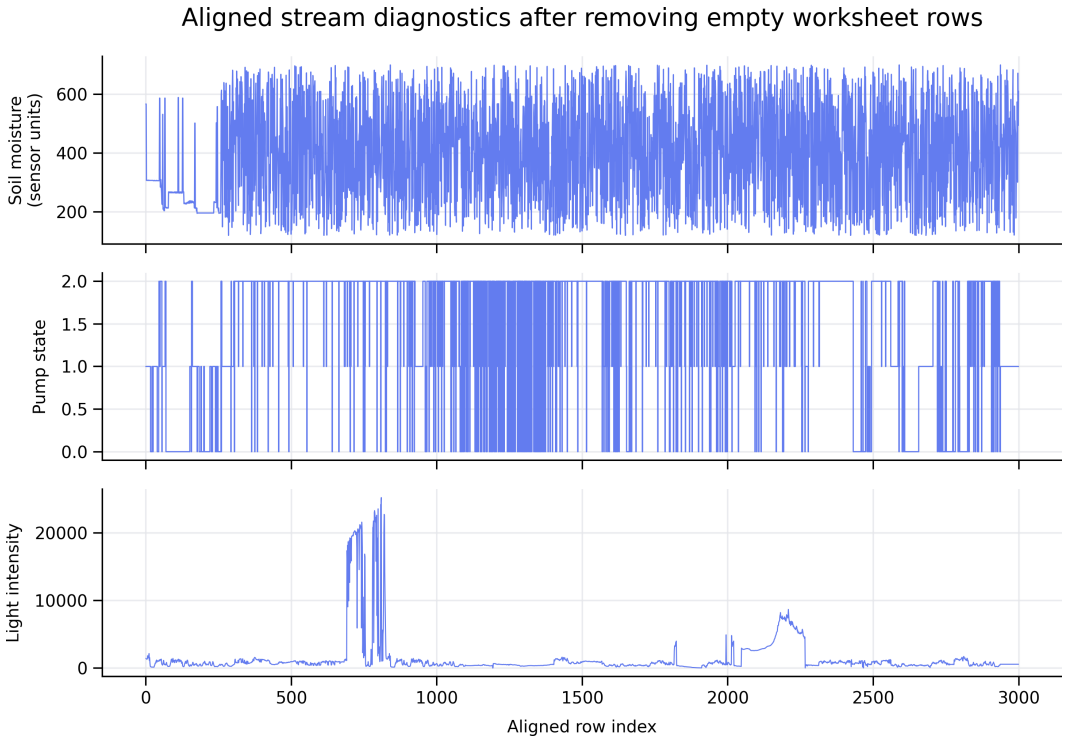

Figure 1: Diagnostics from the 3,000 complete rows of the supplied aligned table. The target exhibits abrupt regime changes, pump state is imbalanced, and light intensity contains extreme excursions. Values are plotted in the units supplied by the data source.

The normalized index increments uniformly within the populated table, but it does not establish a physical sampling interval. We therefore report a one-step horizon and do not translate it into minutes without source documentation.

Same-row Pearson correlations are weak: tomato soil moisture versus drip soil is -0.072, tomato ET versus drip temperature is 0.127, and tomato soil moisture versus pump state is 0.125. Searching lags from -60 to 60 rows yields pair-dependent maxima rather than a common clock correction (Figure 2). This supports the term *weak pairing* rather than synchronized sensor–actuator fusion.

4 Structured Conditional Variational Model

4.1 Conditioning representation

We define the conditioning vector once as

$$h_t = [c_t; \psi_t^{\mathrm{ctrl}}; \pi_t] \in \mathbb{R}^{104}, \quad c_t \in \mathbb{R}^{64}, \quad \psi_t^{\mathrm{ctrl}} \in \mathbb{R}^{32}, \quad \pi_t \in \mathbb{R}^{8}. \tag{2}$$

The context block contains four lags of 14 environmental and crop variables (56 features) plus eight prespecified first-difference and rolling-summary features. The control block contains eight lags of pump and field-operation channels (16 features) plus 16 switching, rolling, interaction, and sensor-response summaries. The plan block contains crop coefficient, days after planting, four stage indicators, lagged crop coefficient, and a planting-time difference. A machine-readable feature manifest should accompany the final code release.

4.2 Conditional prior, posterior, and decoder

The model uses

$$q_\phi(z_t \mid y_{t+1}, h_t), \qquad p_\theta(z_t \mid h_t), \qquad p_\theta(y_{t+1} \mid z_t, h_t), \tag{3}$$

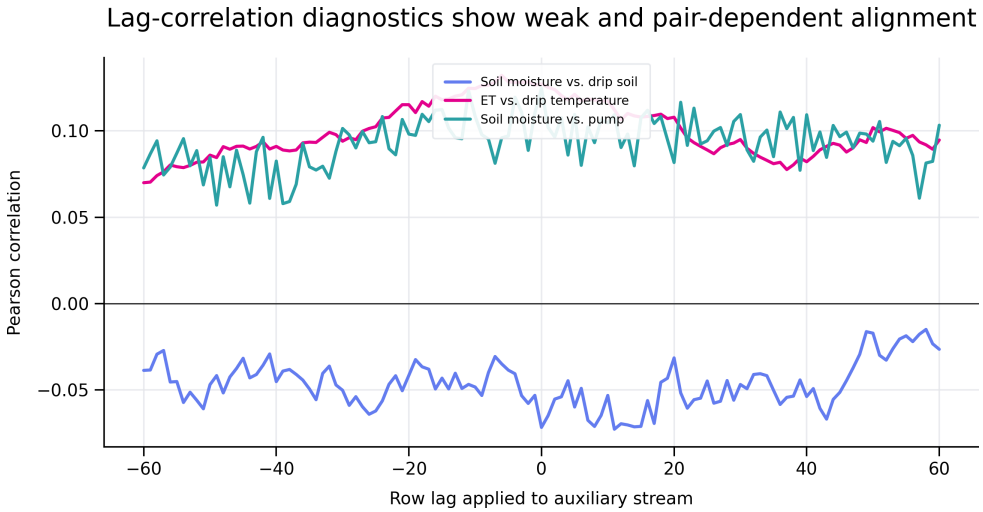

Figure 2: Cross-stream lag correlations over $[-60, 60]$ rows. The strongest absolute correlations are small and occur at different lags, so a single offset is not physically identified by these diagnostics.

with a 16-dimensional latent variable and a Gaussian decoder. The training criterion is the negative β-ELBO,

$$\mathcal{L}_\beta(\theta, \phi) = -\,\mathbb{E}_{q_\phi}[\log p_\theta(y_{t+1} \mid z_t, h_t)] \tag{4}$$

$$+\,\beta\, D_{\mathrm{KL}}(q_\phi(z_t \mid y_{t+1}, h_t) \,\|\, p_\theta(z_t \mid h_t))\,. \tag{5}$$

The reported protocol uses Adam with learning rate 10^{-3}, batch size 64, 100 epochs, linear KL annealing over 30 epochs, and a chronological 70/15/15 split. Final reporting should use at least five fixed seeds and identical model-selection rules across conditioning regimes.

At inference, $z_t^{(s)} \sim p_\theta(z_t \mid h_t)$ and $y_{t+1}^{(s)} \sim p_\theta(y_{t+1} \mid z_t^{(s)}, h_t)$. Point forecasts use the Monte Carlo predictive mean. Predictive NLL must be distinguished from negative ELBO; the final results table should state the estimator and number of samples.

4.3 Offset Correction and Denoising

The reported preprocessing rerun uses a global offset of 21 rows, selected from cross-correlations among nominally related variables, followed by a three-point rolling median and exponentially weighted smoothing for continuous channels. Pump and field-operation variables remain discrete. Because the reported alignment score is only 0.0646 and different pairs favor different lags, this is a sensitivity analysis, not verified synchronization.

5 Theoretical Interpretation

Proposition 1 (Auxiliary information and Bayes risk)**.** *Let X denote trusted context, A an auxiliary modality, and Y the target. Under squared loss and unrestricted measurable predictors,*

$$\inf_f \mathbb{E}[(Y - f(X, A))^2] \le \inf_g \mathbb{E}[(Y - g(X))^2].$$

Proof. Every predictor $g(X)$ is contained in the larger class through $f(X, A) = g(X)$. Taking the infimum over a superset cannot increase risk. □

Thus, worse empirical full fusion does not prove that actuation history is intrinsically useless. It identifies a practical failure mode caused by finite samples, incorrect pairing, noise, optimization, regularization, feature scaling, or restricted model capacity.

Proposition 2 (Non-identifiability of a global row offset)**.** *Without shared timestamps or a generative relation between the acquisition clocks, a row offset is not identified solely by low cross-correlation among a finite set of paired channels.*

Proof. When cross-correlation surfaces are flat, weak, or pair dependent, multiple offsets produce statistically indistinguishable alignment scores. No unique physical time correspondence follows without additional clock or process assumptions. □

Proposition 3 (Monte Carlo predictive consistency)**.** *For a fixed trained model and integrable decoder statistic $g(z)$, if $z^{(1)}, \ldots, z^{(S)}$ are independent draws from $p_\theta(z \mid h)$, then $S^{-1} \sum_{s=1}^{S} g(z^{(s)})$ converges almost surely to $\mathbb{E}[g(z) \mid h]$ as $S \to \infty$.*

Proof. This is the strong law of large numbers. It establishes consistency of Monte Carlo integration for the fitted model, not calibration or correctness of that model. □

6 Experiments

We evaluate whether structured multimodal conditioning improves one-step soil-moisture forecasting when the environmental, agronomic, and actuation streams are only weakly synchronized. All experiments use the same chronological data partition and predict

$$y_{t+1} = \mathrm{SoilMoisture}_{t+1}$$

from information available at or before time index t. Because the physical sampling interval and target calibration are not independently verified, the forecast horizon is reported as one step and the prediction errors are expressed in the supplied sensor units.

6.1 Conditioning Regimes

The primary experiment compares four semantically distinct conditioning configurations:

1. **Context only:**
$$h_t = c_t \in \mathbb{R}^{64},$$
containing environmental, crop-state, and local sensor-history features.
2. **Control history only:**
$$h_t = \psi_t^{\mathrm{ctrl}} \in \mathbb{R}^{32},$$
containing lagged pump, field-operation, switching, and actuation-summary features.
3. **Plan and phenology only:**
$$h_t = \pi_t \in \mathbb{R}^{8},$$
containing crop coefficient, planting progression, and crop-stage descriptors.
4. **Naive full fusion:**
$$h_t = [c_t; \psi_t^{\mathrm{ctrl}}; \pi_t] \in \mathbb{R}^{104},$$
formed by direct concatenation of all three blocks.

This ablation isolates the predictive contribution of proximal environmental context, recent control history, and slowly evolving crop-regime information. It also tests whether increasing the number of conditioning variables improves forecasting when cross-stream synchronization is uncertain.

6.2 Alignment and Denoising Rerun

To determine whether the poor performance of naive full fusion is caused by irrelevant modalities or by preprocessing limitations, we evaluate two additional full-fusion configurations. The first configuration applies a global row-offset correction estimated from nominally related cross-stream variables. The second applies the same offset correction followed by a rolling-median filter and exponentially weighted smoothing of continuous channels. Discrete pump and field-operation

Table 1: Reported one-step CVAE conditioning ablation. MAE and RMSE are expressed in the supplied soil-moisture sensor units. The archived likelihood quantity is labeled reported NLL because its estimator requires verification. Lower values are better.

Conditioning regime	Dim.	MAE	RMSE	Reported NLL
Context only	64	**139.23**	**165.18**	**6.5909**
Control history only	32	146.50	177.05	6.6845
Plan and phenology only	8	163.30	199.08	29.0736
Naive full fusion	104	243.24	296.93	84.8890

variables are retained without smoothing to preserve switching events. The resulting configurations are:

naive full fusion,

offset-corrected full fusion,

offset-corrected and denoised full fusion.

The selected offset and its weak correlation score are treated as an empirical sensitivity setting rather than verified physical synchronization. Accordingly, these experiments evaluate the robustness of multimodal fusion to preprocessing choices and do not establish a causal relationship between pump action and subsequent soil-moisture response.

6.3 Training and Evaluation

Each conditioning regime uses the same conditional-prior CVAE formulation, optimization budget, chronological split, and model-selection criterion. The model is trained on the first 70% of the sequence, validated on the next 15%, and evaluated on the final 15%. Detailed architectural settings, feature definitions, leakage controls, random-seed protocol, and preprocessing specifications are provided in Appendix **??**. Point-forecast performance is evaluated using mean absolute error and root mean squared error:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y}_i|\,, \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}. \tag{7}$$

The prior experimental records also contain a quantity labeled NLL. Because the archived runs do not unambiguously establish whether this quantity is predictive negative log-likelihood, negative ELBO, or an annealed β-ELBO objective, it is retained as *reported NLL* and interpreted cautiously. Appendix **??** specifies the predictive-likelihood estimator required for the final rerun. The broader benchmark suite, uncertainty-calibration metrics, irrigation-oriented utility measures, negative controls, and statistical-comparison procedures are also deferred to Appendix A.

7 Results

7.1 Conditioning Ablation

Table 1 summarizes the reported one-step forecasting results. Context-only conditioning provides the lowest MAE, RMSE, and reported NLL. Control-history conditioning ranks second, while plan-only conditioning provides weaker one-step predictive information. Naive full fusion performs substantially worse than every individual conditioning block. The observed ranking is

$$c_t \succ \psi_t^{\mathrm{ctrl}} \succ \pi_t \succ [c_t; \psi_t^{\mathrm{ctrl}}; \pi_t],$$

where $\succ$ denotes lower forecasting error. Relative to context-only conditioning, naive full fusion increases MAE by

$$\frac{243.24 - 139.23}{139.23} \times 100 \approx 74.7\%,$$

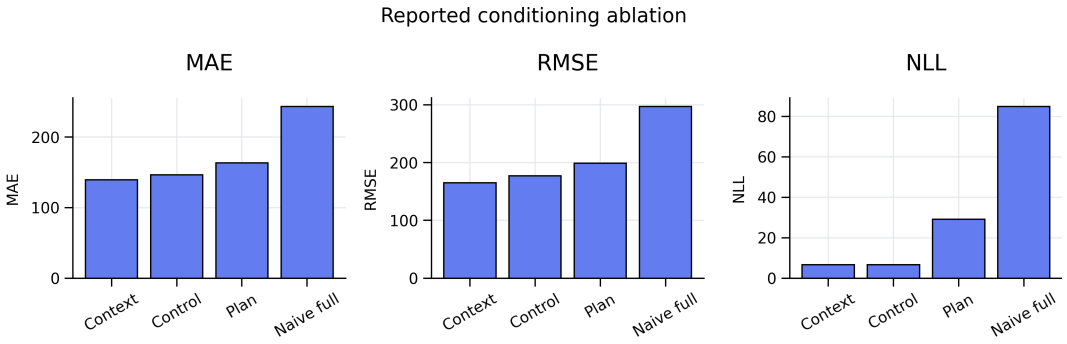

Figure 3: Reported conditioning-ablation results. Context-only conditioning provides the strongest one-step performance, whereas naive full fusion produces the largest errors. The figure is reconstructed from archived manuscript values and does not represent a new training run.

Table 2: Reported preprocessing sensitivity of the 104-dimensional full-fusion CVAE. The values are retained from archived experimental outputs and require regeneration from a unified run registry before final submission.

Full-fusion configuration	MAE	RMSE	Reported NLL
Naive fusion	243.24	296.93	84.89
Offset corrected	198.72	248.50	79.21
Offset corrected and denoised	**161.35**	**209.44**	**74.02**

and increases RMSE by

$$\frac{296.93 - 165.18}{165.18} \times 100 \approx 79.8\%.$$

These differences indicate that direct concatenation does not exploit the additional modalities effectively under the current weak-pairing assumption. The result should not be interpreted as evidence that control history is intrinsically irrelevant. Under unrestricted population prediction, adding valid information cannot increase Bayes risk. The empirical degradation is therefore more plausibly explained by temporal mismatch, noisy auxiliary channels, finite-sample estimation, model-capacity effects, optimization, or regularization.

7.2 Effect of Alignment and Denoising

Table 2 reports the sensitivity of the full-fusion model to offset correction and denoising. Offset correction decreases MAE from 243.24 to 198.72. Adding denoising further decreases MAE to 161.35. Relative to naive fusion, offset correction and denoising reduce MAE by

$$\frac{243.24 - 161.35}{243.24} \times 100 \approx 33.7\%,$$

and RMSE by

$$\frac{296.93 - 209.44}{296.93} \times 100 \approx 29.5\%.$$

These improvements show that full-fusion performance is highly sensitive to the quality of cross-stream reconstruction. Nevertheless, the improved full-fusion model does not outperform context-only conditioning. Its MAE remains 22.12 sensor units higher, and its RMSE remains 44.26 sensor units higher, than the corresponding context-only values. The appropriate interpretation is therefore partial recovery rather than fusion superiority. Offset correction and denoising reduce the damage caused by naive concatenation, but the current data do not establish that the auxiliary stream contains sufficiently synchronized information to improve one-step forecasting beyond the proximal context representation.

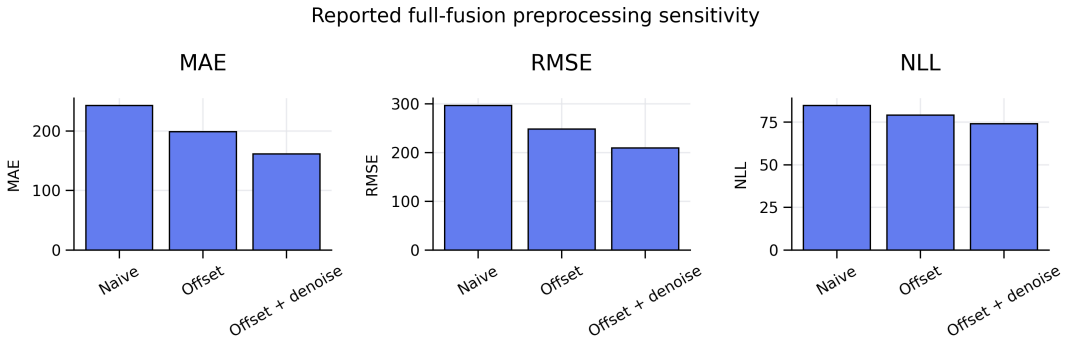

Figure 4: Reported sensitivity of full fusion to offset correction and denoising. Preprocessing substantially improves the fused model but does not surpass context-only conditioning. The results do not establish causal sensor–actuator synchronization.

7.3 Summary of Findings

The reported experiments support three principal findings. First, proximal environmental and crop-state context provides the strongest signal for one-step soil-moisture prediction. This is consistent with the short forecast horizon, for which the current hydrological state is more informative than slowly evolving phenology or incompletely aligned actuation history. Second, control history contains useful predictive information but does not outperform environmental context in isolation. Its value is likely mediated by soil state, weather, irrigation magnitude, infiltration delay, and hydraulic execution, none of which can be fully reconstructed from pump state alone. Third, naive multimodal fusion is not robust to weak temporal pairing. Offset correction and denoising recover a substantial fraction of the lost accuracy, indicating that the principal difficulty is not modality count alone but the temporal and statistical coherence of the auxiliary channels. Because the available results are single-point estimates from archived runs, the present analysis does not claim statistical significance, calibrated uncertainty, or closed-loop irrigation benefit. Those questions require the multi-seed, calibration, robustness, and control-oriented evaluations specified in Appendix A.

8 Discussion

The experiments support three conclusions. First, proximal context dominates this one-step task. Second, control history is not useless, but its value is obscured when the actuation stream is weakly paired with the target. Third, denoising and an assumed offset recover substantial full-fusion accuracy without establishing that the streams are physically synchronized.

The forecast traces in the prior draft showed regression toward the mean. This behavior should be tested directly through errors by target quantile, extreme-event recall, and calibrated predictive intervals. If the model continues to miss dry and wet extremes, average MAE improvements are insufficient for autonomous control.

The current study is best understood as pre-deployment forecasting analysis. A usable autonomous system would additionally require timestamp-verified actuation logs, flow and pressure feedback, validated moisture thresholds, multi-step forecasts, actuator constraints, and closed-loop evaluation against agronomic and safety objectives.

9 Limitations, Broader Impact, and Reproducibility

The main limitation is the absence of a certified shared clock and documented physical relationship between the two streams. The target unit and one-step duration are also undocumented. Results come from one reconstructed table and do not establish transfer to another crop, farm, soil, climate, or sensor suite. Alignment and denoising may remove genuine events as well as noise. Finally, a CVAE predictive distribution is not automatically calibrated.

Learning-enabled irrigation could reduce water and energy use, but poorly calibrated forecasts may trigger crop stress, saturation, pump cycling, or equipment damage. Deployment should retain rule-based safety interlocks, manual override, pressure and flow protection, and conservative fallback schedules.

The reproducibility release should contain immutable split indices, feature manifests, preprocessing and alignment code, fixed seeds, environment files, parameter counts, runtime and hardware, per-seed predictions, and one script that regenerates every table and figure.

10 Conclusion

This study reframes multimodal irrigation forecasting as a synchronization-aware learning problem. The strongest reported model uses environmental–crop context alone; naive full fusion fails, while offset correction and denoising partially recover performance. The result is not evidence against multimodality. It is evidence that auxiliary information must have credible temporal semantics, adequate data quality, and fair evaluation before it can improve probabilistic forecasting or support autonomous irrigation.

References

Richard G. Allen, Luis S. Pereira, Dirk Raes, and Martin Smith. *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Drainage Paper 56. Food and Agriculture Organization of the United Nations, 1998.

Eduardo F. Camacho and Carlos Bordons. *Model Predictive Control*. Springer, 2 edition, 2007.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4): 1748–1764, 2021.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, 2014.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, 2015.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

A Extended Experimental and Evaluation Protocol

A.1 Extended Conditioning and Negative-Control Experiments

In addition to the four principal conditioning regimes reported in the main paper, the final rerun should evaluate:

- context plus control history;
- context plus plan/phenology;

- control history plus plan/phenology;
- shuffled-control fusion;
- deliberately time-shifted control fusion;
- matched-noise control replacement;
- modality dropout;
- separate modality encoders with late fusion.

The shuffled, shifted, and matched-noise configurations are negative controls. These experiments test whether apparent improvements arise from meaningful temporal information or merely from higher input dimensionality and model capacity.

A.2 Forecasting Baselines

The final comparison should include:

- persistence and seasonal persistence;
- linear regression and ridge regression;
- random forest and gradient boosting;
- ARIMA or SARIMAX;
- LSTM and GRU;
- heteroscedastic Gaussian regression;
- quantile regression;
- deep ensembles;
- Temporal Fusion Transformer or another modern temporal model.

All methods must use identical chronological partitions, target definitions, causal histories, and test timestamps.

A.3 Uncertainty Calibration

Probabilistic evaluation should include:

- predictive negative log-likelihood;
- continuous ranked probability score;
- empirical coverage at 50%, 80%, 90%, and 95%;
- mean prediction-interval width;
- interval score;
- probability integral transform histograms;
- calibration error;
- sharpness-versus-coverage comparisons.

Predictive NLL must be computed from the normalized predictive distribution. Negative ELBO and annealed β-ELBO values must be reported separately.

A.4 Irrigation-Oriented Utility

Once physically calibrated moisture thresholds are available, the evaluation should report:

- dry-threshold precision, recall, and lead time;
- wet-threshold false-alarm rate;
- moisture-band violation duration;

- irrigation-trigger disagreement;
- pump-switching count;
- water-use and energy proxies;
- operational-cost CVaR.

These thresholds must be derived from agronomic calibration or domain specifications rather than arbitrary empirical quantiles.

A.5 Statistical Comparison

Each model should be trained using at least five fixed random seeds. Results should be reported as mean, standard deviation, and 95% confidence intervals. Pairwise comparisons should use paired bootstrap differences on common test timestamps, with Holm correction for multiple comparisons. Performance should also be reported by:

- crop stage;
- pump state;
- target quantile;
- dry, nominal, and wet operating regimes.

These stratified analyses are necessary to determine whether low aggregate error is achieved by regression toward the mean while missing operationally important extremes.

B Result Provenance and Integrity Controls

An earlier draft contained conflicting context-only RMSE and likelihood values across the main text and appendix. The duplicate values have been removed from the main manuscript. Before final submission, every result should be regenerated from a single machine-readable run registry containing:

- run identifier and random seed;
- source-data hash and split hash;
- preprocessing and feature-manifest version;
- temporal offset and denoising configuration;
- target scaling and measurement unit;
- model architecture and parameter count;
- checkpoint-selection rule;
- predictive-sample count;
- likelihood estimator;
- per-timestamp predictions and predictive samples.

No statistical-significance, uncertainty-calibration, or autonomous-control claim should be made from the current single-point archived estimates.

C Reproducibility Checklist for the Final Rerun

1. Verify common location, experimental period, timestamp fields, sampling frequencies, units, and missing intervals.
2. Publish the exact 64, 32, and 8 feature lists with lag and rolling-window definitions.
3. Regenerate all results with a chronological split and at least five seeds.
4. Compare deterministic, probabilistic, and modern temporal baselines under identical histories.

5. Separate predictive NLL, negative ELBO, and importance-weighted likelihood estimates.
6. Report calibration, stage-wise results, pump-state results, target-quantile results, and extreme-event performance.
7. Add timestamp jitter, sensor dropout, block missingness, spike, drift, and offset-misspecification stress tests.
8. Evaluate 1, 2, 4, 8, and 16 steps only after the physical sampling interval is verified.
9. Do not claim autonomous control benefit without a validated process model or closed-loop field trial.

D AI ASSISTANCE STATEMENT

AI-assisted tools were used for language editing.