
Learning hydrological error states for process-model correction

Hans Korving
Deltares
Delft, The Netherlands
hans.korving@deltares.nl

Abstract

End-to-end sequence models can jointly learn temporal aggregation, state representations and prediction, but the resulting hydrological state is often difficult to inspect. We revisit the Warren River residual-correction case of Hammad et al. (2026) and replace their prescribed dry/wet residual states with a task-informed representation learned using a vector-quantized variational autoencoder (VQ-VAE). The VQ-VAE yields both a continuous latent state and a set of recurrent discrete prototypes, and a generalized additive model (GAM) is used to predict the GR4J residual correction. The resulting GR4J–VQ-VAE–GAM model performs in the same range as the published state-dependent dual-LSTM correction. Adding the continuous latent representation to explicit predictors improves the GAM readout, with the clearest benefit at high flows. Discrete prototype identity contributes little additional predictive skill, although the prototypes themselves show distinct hydroclimatic profiles, residual distributions and temporal occupancy. We use this case to examine whether hydrological error state can be represented explicitly, inspected, and used for state-dependent correction.

1 Introduction

Process-based hydrological models give an explicit account of rainfall–runoff behaviour, yet their residuals may vary with antecedent conditions, flow regime and meteorological state. Low-dimensional learned representations have been used to identify interpretable structure in hydrology and other physical-science applications [Bassi et al., 2024, Piras et al., 2025]. By contrast, flexible models fitted directly to residuals can be difficult to inspect.

Hammad et al. [2026] make hydrological state relevant to residual correction by separating negligible-rainfall and wet conditions and fitting separate LSTM corrections to GR4J [Perrin et al., 2003]. They report substantial improvements, particularly for poorer calibration and at high flows. Their dry/wet state distinction, however, is specified before the residual models are fitted. Residual-learning approaches typically predict process-model error directly or condition correction on prescribed states; here we instead learn the error-state representation itself.

The central question is whether hydrological error state can be learned while retaining competitive residual-correction skill. We consider two questions: whether a learned residual-state representation can support a comparatively simple readout with performance close to the published dual-LSTM correction, and whether the resulting latent space exhibits hydrologically expected structure.

We retain GR4J and learn a VQ-VAE representation [van den Oord et al., 2017] of conditions associated with its residuals. Continuous pre-quantization latents are used for prediction with a GAM (Wood, 2025), while the discrete codebook is retained for state analysis. This workflow represents

process-model discrepancy as an explicit error-state representation, separating continuous predictive structure from recurrent quantized states that can be inspected hydrologically.

2 Learning residual state

Let $\hat{Q}_t^{\text{GR4J}}$ denote GR4J discharge and

$$r_t = Q_t - \hat{Q}_t^{\text{GR4J}} \quad (1)$$

the model residual. The VQ-VAE representation is learned with task supervision from the residual-correction objective; the resulting latent states should therefore be interpreted as task-informed error states rather than unsupervised hydrological regimes.

The encoder maps standardized hydroclimatic inputs to a continuous state,

$$z_t = E_\theta(x_t), \quad (2)$$

where $z_t \in \mathbb{R}^d$. Vector quantization assigns each observation to its nearest codebook vector,

$$k(t) = \arg \min_k \|z_t - c_k\|^2. \quad (3)$$

The decoder reconstructs the original feature space from the quantized representation,

$$\hat{x}_t = D_\phi(c_{k(t)}). \quad (4)$$

The selected architecture uses a five-dimensional latent space, six codebook vectors, and fully connected encoder and decoder networks with one hidden Tanh layer of width 10. A separate regression head, with the same hidden-layer structure, predicts the standardized GR4J residual directly from the continuous latent state. The model is trained using a weighted combination of reconstruction, supervised prediction, codebook, and commitment losses:

$$\begin{aligned} \mathcal{L} = & \lambda_{\text{rec}} \text{MSE}(x, \hat{x}) + \lambda_{\text{sup}} \text{MSE}(r, \tilde{r}) \\ & + \lambda_{\text{vq}} [\text{MSE}(\text{sg}[z], c_{k(t)}) + \beta \text{MSE}(z, \text{sg}[c_{k(t)}])], \end{aligned} \quad (5)$$

where $\tilde{r}$ is the prediction from the supervised regression head, sg denotes stop-gradient, and β controls the commitment loss. A straight-through estimator propagates reconstruction gradients through the quantization operation.

Architecture and training hyperparameters were selected using forest-based Bayesian optimization with up to 100 evaluations. The selection criterion combines residual-prediction performance, reconstruction quality, latent-space separation, codebook utilization, quantization error, and neighbourhood preservation. Additional penalties discourage sparsely occupied, infrequently recurring, and strongly imbalanced prototypes. Candidate models were trained for 1,000 epochs and the selected configuration was refitted for 3,000 epochs in the cross-validation experiments.

The continuous state z_t supports prediction, while the discrete prototypes $k(t)$ enable state analysis.

For prototype k , temporal occupancy is summarized by

$$p_k(\tau) = \frac{1}{N_\tau} \sum_{t \in \tau} \mathbf{1}[k(t) = k]. \quad (6)$$

We examine these occupancy ratios together with hydroclimatic feature profiles of assigned observations and the conditional residual distribution,

$$p(r | k). \quad (7)$$

This allows a prototype to be described by the conditions it represents, when those conditions occur, and whether GR4J shows a systematic error within them.

2.1 Residual readout

Residual corrections are predicted using a generalized additive model (GAM). For the continuous pre-quantization representation, the readout takes the form

$$\hat{r}_t = \beta_0 + \sum_{j=1}^d f_j(z_{tj}), \quad (8)$$

where f_j are smooth functions of the latent coordinates. The corrected discharge becomes

$$\hat{Q}_t = \hat{Q}_t^{\text{GR4J}} + \hat{r}_t. \quad (9)$$

The use of a GAM is deliberate. Nonlinear transformations have already occurred in the encoder, so the downstream model tests how much additional complexity is required once residual state has been organized.

We compare five feature sets: explicit hydroclimatic inputs; these inputs augmented with discrete prototype identity; the inputs augmented with the continuous latent state z_t ; the inputs augmented with both z_t and prototype identity; and the continuous latent state alone.

Each feature set is evaluated with and without three short-term residual-history inputs (r_{t-1} , r_{t-2} , r_{t-3}). When included, these lags enter the VQ-VAE encoder and are also available as explicit GAM predictors in configurations using the original inputs. The latent-only GAM uses residual-history information exclusively through the learned representation. Table 1 reports results with residual history, while Table A1 presents the complete ablation.

These comparisons distinguish the predictive contribution of the continuous representation from that of discrete prototype identity and assess how much residual-history information is retained in the learned state.

3 Warren River case

We focus on the uncalibrated Warren configuration reported by Hammad et al. [2026], for which process-model error is pronounced. Their published overall RMSE falls from approximately 0.22 mm d^{-1} for GR4J to 0.02 mm d^{-1} after their state-dependent dual-LSTM correction. Above the 95th and 99th flow percentiles, the corresponding corrected RMSE values are approximately 0.08 and 0.11 mm d^{-1} . We use these published values as a reference, not as a controlled architecture benchmark, because the dual-LSTM models are not rebuilt and the reported values are available only at their published precision. Our own results are means across 15 blocked cross-validation models; Fig. A3 shows the split structure.

3.1 Learned representation supports a simple residual readout

The GR4J-VQ-VAE-GAM decomposition performs in the same range as the published state-dependent LSTM correction for the uncalibrated Warren case (Table 1).

Adding the continuous latent state to the explicit inputs reduces overall RMSE by 10.9%, with larger improvements of 19.0% and 12.5% above the 95th and 99th flow percentiles, respectively. At high flows, the latent-only readout performs slightly better, reducing RMSE by 24.2% and 13.2% relative to the explicit-input model. This suggests that the encoder may retain a compact combination of forcing and antecedent information relevant under extreme-flow conditions.

Discrete prototype identity adds little predictive skill: including it changes overall RMSE by less than 1%, whether or not the continuous latent representation is included.

Short-term residual history remains important. These terms are retained in the closest comparison because residual-history information is also available to the models in Hammad et al. [2026]. Removing the residual lags substantially reduces absolute skill. Without these terms, however, the latent-only readout performs best across all three flow regimes, reducing RMSE relative to explicit inputs by 21.6% overall, 40.4% above the 95th percentile, and 31.6% above the 99th percentile (Table A1). Several latent coordinates and the first two residual lags also show broadly consistent GAM smooths across models (Fig. A4).

The comparison with Hammad et al. therefore concerns downstream model complexity, not the removal of residual memory.

3.2 The latent space recovers expected hydrological structure

Quantization partitions the continuous latent space into six recurrent prototypes (Fig. A2), which exhibit distinct hydroclimatic profiles and GR4J residual distributions (Fig. 1). Their codebook

Table 1: Predictive performance in the uncalibrated Warren River case, overall and at high flows. All GAM configurations use residual-history information ($r_{t-1}, r_{t-2}, r_{t-3}$): directly as predictors in the explicit-input models, or indirectly through the learned latent representation in the latent-only model. Published Hammad et al. [2026] results are included for reference.

Regime	Model / representation	RMSE ↓	MAE ↓	NSE ↑	KGE ↑	R^2 ↑
Overall	Hammad: GR4J, uncalibrated	0.22	0.16	-0.13	-0.06	0.23
	Hammad: state-dependent dual LSTM	0.02	0.01	0.96	0.92	0.94
	GAM: explicit inputs	0.0294	0.0125	0.9793	0.9153	0.9855
	GAM: explicit inputs + latent	0.0262	0.0117	0.9833	0.9348	0.9870
	GAM: latent only	0.0419	0.0197	0.9571	0.9023	0.9634
$q > 95\%$	Hammad: GR4J, uncalibrated	0.59	0.56	-8.20	0.22	0.21
	Hammad: state-dependent dual LSTM	0.08	0.05	0.91	0.95	0.91
	GAM: explicit inputs	0.1031	0.0779	0.8105	0.8835	0.8737
	GAM: explicit inputs + latent	0.0835	0.0606	0.8741	0.9237	0.9020
	GAM: latent only	0.0781	0.0564	0.8917	0.9322	0.8976
$q > 99\%$	Hammad: GR4J, uncalibrated	0.74	0.71	-0.66	-0.06	0.16
	Hammad: state-dependent dual LSTM	0.11	0.08	0.47	0.75	0.58
	GAM: explicit inputs	0.1246	0.0981	0.1254	0.6136	0.5340
	GAM: explicit inputs + latent	0.1090	0.0837	0.3274	0.6878	0.5593
	GAM: latent only	0.1082	0.0831	0.3425	0.7210	0.5718

Note: Hammad et al. values are read from their published uncalibrated Warren River results and are reported at two-decimal precision; our results are means across 15 blocked cross-validation models. The comparison is therefore indicative rather than a controlled architecture benchmark. Results without residual-history inputs are reported in Table A1.

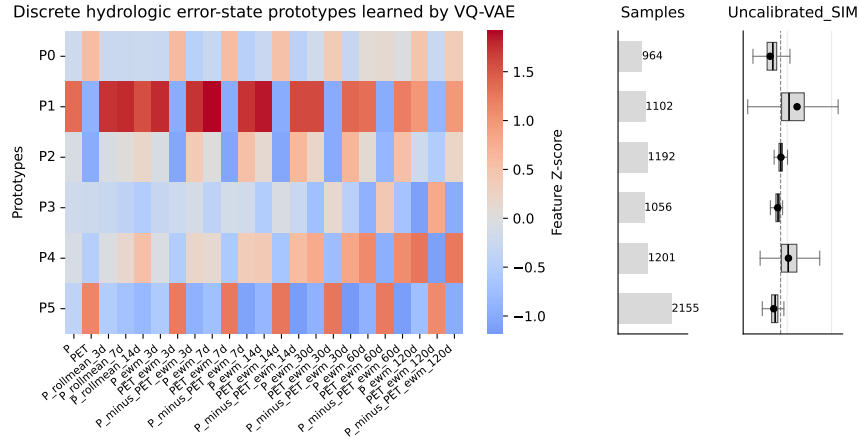


Figure 1: Learned discrete error-state prototypes: standardized hydroclimatic feature profiles (left), assigned observations (centre), and GR4J residual distributions (right).

vectors provide complementary decoded profiles (Fig. A6). Prototype occupancy ranges from 964 to 2155 observations, with median residuals between -0.16 and 0.02 mm d^{-1} .

The codebook also has a structured geometry (Fig. 2a): its first two principal components explain 75.6% and 17.9% of the codebook variation. Recurrent one-step transitions occur between a restricted subset of states, although geometric proximity does not necessarily imply temporal accessibility. Persistence varies strongly through the year (Fig. 2b), with different prototypes persistent in distinct seasonal windows; cells with fewer than 20 outgoing transitions are treated as low-support estimates. The 2009 example in Fig. A1 further illustrates prototype persistence and switching alongside discharge and the GR4J residual.

Post-hoc one-versus-all SkopeRules analysis further reveals that the prototypes are distinguished by different combinations of meteorological forcing and hydrological memory (Fig. A5). For example, one prototype is characterized primarily by recent and accumulated precipitation (P1 in Fig. A5), whereas another is more strongly associated with antecedent observed discharge and longer-term memory variables (P4 in Fig. A5). Notably, both prototypes exhibit relatively wide GR4J residual distributions despite median residuals close to zero (Fig. 1). The one-versus-all rule models achieved F1-scores of approximately 0.9 or higher for most prototypes using a relatively small number of

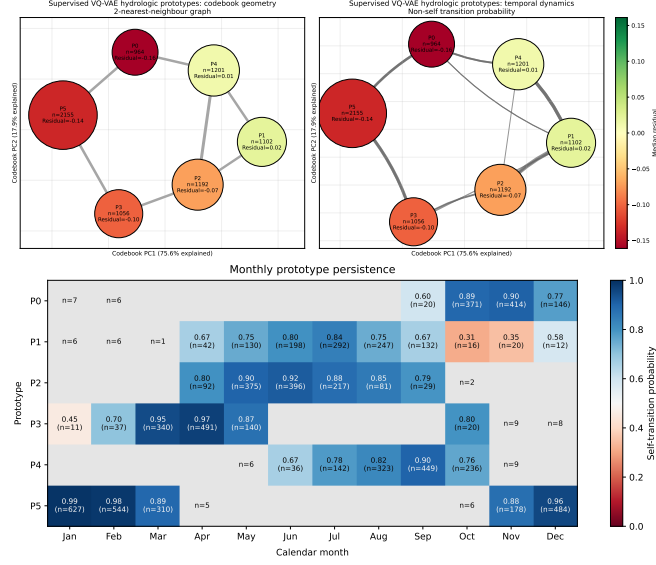


Figure 2: Geometry and temporal organization of the learned prototypes. (a) Codebook 2-nearest-neighbour geometry (left) and observed non-self transitions (right), shown in the first two codebook principal components; node size gives occupancy and colour median GR4J residual. (b) Monthly self-transition probability; grey cells have fewer than 20 outgoing transitions.

rules, indicating that the learned states can be approximated by compact, interpretable decision rules. These distinctions suggest that the learned error states capture different combinations of forcing and catchment-history information, rather than merely reproducing a prescribed dry/wet classification. Together with the prototype-specific GR4J residual distributions in Figure 1, the results provide a diagnostic view of the hydroclimatic conditions under which process-model errors differ. The associations should not, however, be interpreted as direct evidence of individual physical mechanisms.

4 Discussion

Hammad et al. showed that conditioning GR4J residual correction on hydrological state can improve prediction, using a dry/wet distinction specified before model fitting. Here, the state associated with the remaining process-model discrepancy is learned instead.

For Warren River, the continuous latent representation adds predictive information beyond the explicit predictors and supports a GAM readout in the same performance range as the published state-dependent LSTM correction, with relatively large improvements at high flows. Part of this skill depends on residual-history terms, which are also available in the information setting of Hammad et al. [2026]. The comparison therefore concerns downstream model complexity rather than removal of residual memory. The quantized prototypes differ in hydroclimatic profile and residual distribution and show structured geometry and seasonal persistence, but their identity adds little predictive skill once the continuous representation is available. This also creates a natural point for expert interaction: the discrete prototypes can be inspected, recognized or challenged as candidate error states, while the continuous latent state retains the richer representation used for prediction.

The learned states depend on the selected inputs, latent dimension, codebook size and VQ objective, and recognizable structure does not make them physical or causal. Because only one catchment is analysed, the experiment does not support a general claim that VQ-VAE is a preferred architecture.

References

- A. Bassi, M. Höge, A. Mira, F. Fenicia, and C. Albert. Learning landscape features from streamflow with autoencoders. *Hydrology and Earth System Sciences*, 28:4971–4988, 2024. doi: 10.5194/hess-28-4971-2024.
- M. Hammad, R. Mehrotra, and A. Sharma. Can state-dependant residual modeling improve legacy hydrological model simulations? *Water Resources Research*, 62:e2026WR044285, 2026. doi: 10.1029/2026WR044285.
- C. Perrin, C. Michel, and V. Andréassian. Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, 279(1–4):275–289, 2003. doi: 10.1016/S0022-1694(03)00225-7.
- D. Piras, L. Herold, L. Lucie-Smith, and E. Komatsu. Λ CDM and early dark energy in latent space: A data-driven parametrization of the CMB temperature power spectrum. *Physical Review D*, 111: 083537, 2025. doi: 10.1103/PhysRevD.111.083537.
- A. van den Oord, O. Vinyals, and K. Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, pages 6306–6315, 2017.
- S. N. Wood. Generalized additive models. *Annual Review of Statistics and Its Application*, 12: 497–526, 2025. doi: 10.1146/annurev-statistics-112723-034249.

A Additional results

Table A1: Ablation of short-term residual-history information in the VQ-VAE-GAM pipeline. The same five feature sets are evaluated with and without three preceding observed GR4J residuals (r_{t-1} , r_{t-2} , r_{t-3}). When included, residual history enters the GAM directly through explicit inputs or indirectly through the learned latent representation. Removing residual history substantially reduces predictive skill. Bold values indicate the best result within each residual-history setting and flow regime.

Regime	Feature set	Lags	RMSE ↓	MAE ↓	NSE ↑	KGE ↑	R^2 ↑
Overall	Inputs	Yes	0.0294	0.0125	0.9793	0.9153	0.9855
	Inputs	No	0.1749	0.0883	0.2683	0.2268	0.7555
	Inputs + code	Yes	0.0295	0.0125	0.9791	0.9150	0.9855
	Inputs + code	No	0.1750	0.0883	0.2678	0.2255	0.7560
	Inputs + latent	Yes	0.0262	0.0117	0.9833	0.9348	0.9870
	Inputs + latent	No	0.1612	0.0801	0.3759	0.2931	0.7927
	Inputs + latent + code	Yes	0.0262	0.0117	0.9832	0.9345	0.9870
	Inputs + latent + code	No	0.1614	0.0802	0.3746	0.2919	0.7927
	Latent only	Yes	0.0419	0.0197	0.9571	0.9023	0.9634
	Latent only	No	0.1371	0.0764	0.5511	0.4725	0.7732
$q > 95\%$	Inputs	Yes	0.1031	0.0779	0.8105	0.8835	0.8737
	Inputs	No	0.3747	0.2819	-1.4976	0.1946	0.3243
	Inputs + code	Yes	0.1042	0.0786	0.8062	0.8808	0.8723
	Inputs + code	No	0.3761	0.2828	-1.5163	0.1903	0.3263
	Inputs + latent	Yes	0.0835	0.0606	0.8741	0.9237	0.9020
	Inputs + latent	No	0.3528	0.2529	-1.2595	0.2525	0.4771
	Inputs + latent + code	Yes	0.0836	0.0606	0.8737	0.9237	0.9018
	Inputs + latent + code	No	0.3534	0.2535	-1.2653	0.2512	0.4774
	Latent only	Yes	0.0781	0.0564	0.8917	0.9322	0.8976
	Latent only	No	0.2232	0.1691	0.1047	0.6040	0.5005
$q > 99\%$	Inputs	Yes	0.1246	0.0981	0.1254	0.6136	0.5340
	Inputs	No	0.4014	0.3469	-8.0729	-0.9588	0.1472
	Inputs + code	Yes	0.1262	0.0994	0.1015	0.6068	0.5305
	Inputs + code	No	0.4042	0.3490	-8.1951	-0.9655	0.1479
	Inputs + latent	Yes	0.1090	0.0837	0.3274	0.6878	0.5593
	Inputs + latent	No	0.4523	0.3841	-10.9431	-0.9852	0.1778
	Inputs + latent + code	Yes	0.1092	0.0838	0.3252	0.6869	0.5590
	Inputs + latent + code	No	0.4536	0.3855	-10.9980	-0.9902	0.1778
	Latent only	Yes	0.1082	0.0831	0.3425	0.7210	0.5718
	Latent only	No	0.2746	0.2424	-3.3243	-0.3080	0.1573

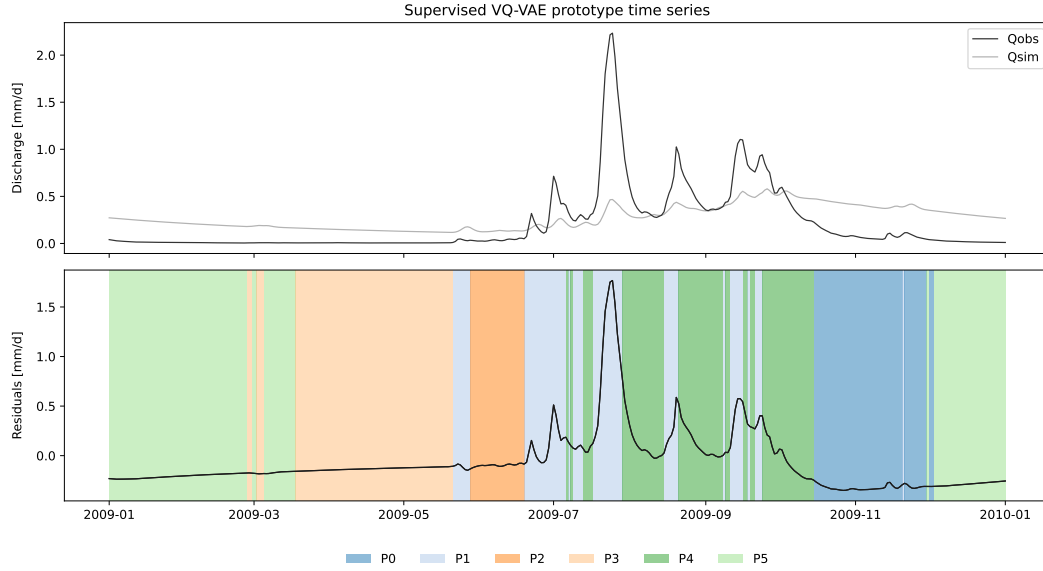


Figure A1: Example of prototype occupancy through time. Observed and uncalibrated GR4J discharge for 2009 (top), with the corresponding GR4J residual and VQ-VAE prototype assignments shown as coloured background intervals (bottom).

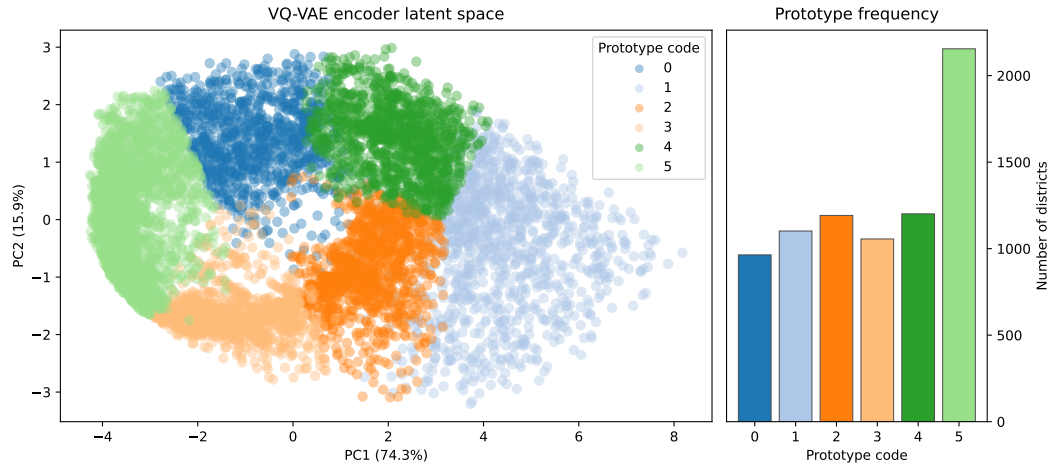


Figure A2: Continuous VQ-VAE latent space and discrete prototype assignments. The first two principal components of the five-dimensional encoder output explain 90.2% of its variance (left). Colours indicate nearest-codebook assignments; the right panel shows prototype frequencies.



Figure A3: Cross-validation scheme. Three repeats of blocked five-fold cross-validation are applied over the model-development period. Training and validation blocks are shown separately; the final test period is held out from model fitting and selection.

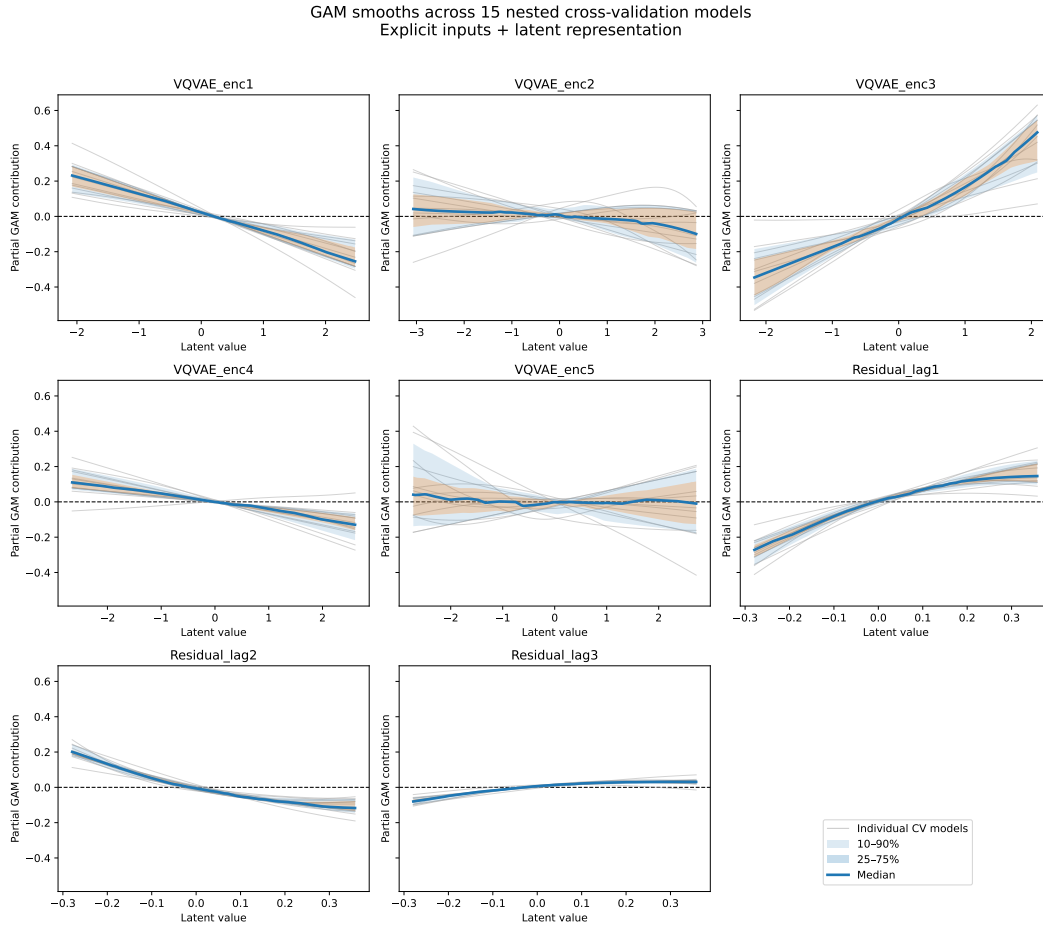


Figure A4: GAM partial smooths across the 15 cross-validation models for the explicit-input plus continuous-latent readout. Thin lines show individual fitted models, the solid line shows the median smooth, and shaded intervals show the 25–75% and 10–90% ranges across models. Panels show the five continuous VQ-VAE coordinates and the three short-term residual-history terms.

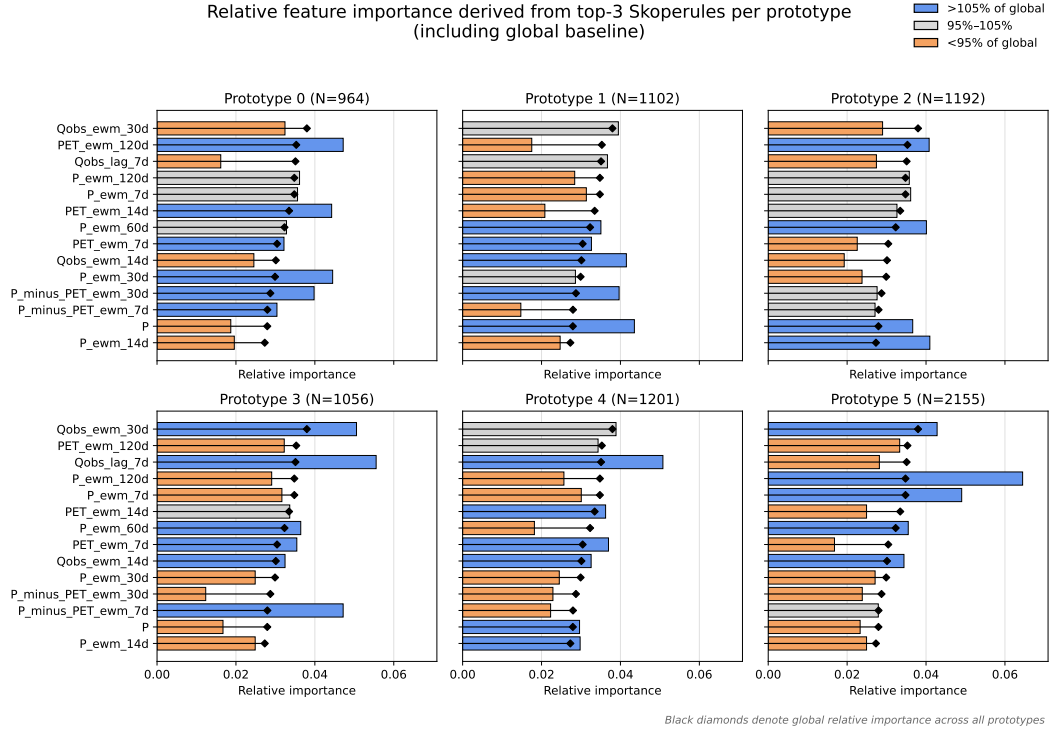


Figure A5: Prototype-specific feature importance derived from post-hoc one-versus-all SkopeRules analysis. Bars show the relative importance of hydroclimatic predictors for distinguishing each prototype from the remaining states. Colours indicate whether prototype-specific importance exceeds (blue), approximates (grey), or falls below (orange) 95–105% of the global reference importance (black diamonds). The results highlight differences in the combinations of meteorological forcing and antecedent hydrological conditions associated with individual error states.

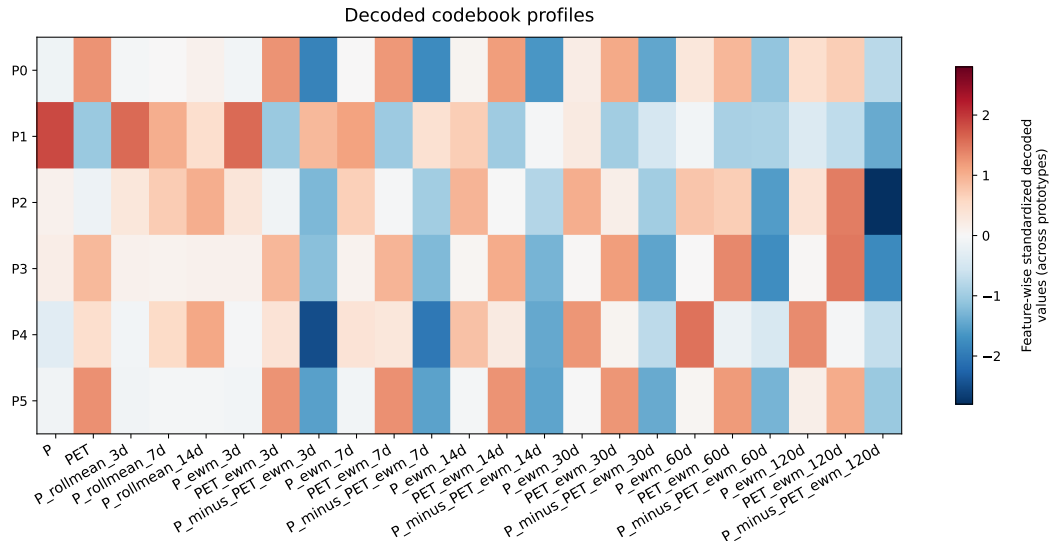


Figure A6: Decoded hydroclimatic profiles of the six learned prototypes. Each codebook vector is mapped back to the original input space through the trained VQ-VAE decoder. Feature values are standardized across prototypes to highlight their relative differences. The resulting profiles provide an interpretation of the information retained by the discrete representation, complementary to the post-hoc rule-based characterization.