Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Measuring and Controlling Internal Activity in Variational Neural Units

Yves Ruffenach

yves@ruffenach.net

Conservatoire National des Arts et Métiers https://orcid.org/0009-0009-4737-0555

Method Article

Keywords: neural networks, machine learning, deep learning, probabilistic modeling, uncertainty quantification, variational inference, forecasting, model interpretability

Posted Date: May 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9585590/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Measuring and Controlling Internal Activity in Variational Neural Units

Yves Ruffenach
Conservatoire National des Arts et Métiers
Strasbourg, France

Abstract

Neural networks usually compute with deterministic internal activations. Probabilistic uncertainty is often modeled at the level of weights, global latent variables, or predictive outputs. This motivates a complementary design question: can probabilistic inference be placed inside the units that carry the forward pass? We study whether a neuron can be written as a small variational inference module with measurable internal probabilistic activity.

We introduce a local variational neuron. Each unit has a scalar latent variable, a standard normal prior, an amortized Gaussian posterior, a reparameterized sample, and a local Kullback–Leibler (KL) regularization term. Its internal activity can be tracked through local KL, squared posterior mean, out-of-range fraction, and optional autoregressive (AR) persistence.

We evaluate the unit on long-horizon forecasting tasks. DET denotes the matched deterministic baseline and provides the point-accuracy reference. EVE denotes the local variational unit and adds internal measurements induced by the local posterior: KL activity, squared posterior mean, out-of-range fraction, and optional AR persistence. These measurements are active, respond to simple controls, and relate to prediction error across runs. The contribution is a variational computational unit for measuring and controlling local probabilistic activity inside neural networks.

1 Introduction

Neural networks usually compute with deterministic internal activations. A unit receives an input, applies a learned transformation, and passes a value to the next layer. This design is effective for prediction. It also motivates a unit-level question: can the unit itself expose local uncertainty, activity, and stability?

Probabilistic neural models usually place uncertainty at the level of weights, global latent variables, or predictive outputs. Bayesian neural networks place it on weights [3]. Variational autoencoders (VAEs) use model-level latent variables and optimize a global evidence lower bound (ELBO) [13, 17]. Calibration and out-of-distribution methods focus on predictive uncertainty [6, 7]. Sequential latent-variable models add stochastic dynamics at the sequence or state level [5, 14]. These approaches motivate the design studied here: placing probabilistic inference inside the computational unit.

We define a local variational neuron. Each unit has a scalar latent variable z, a prior $p(z)$, an amortized posterior $q_\phi(z \mid h)$, a reparameterized sample, and a local KL term. The unit emits an activation from this local latent representation. In the small-variance limit, the unit reduces to a deterministic scalar mapping. If the posterior mean is parameterized as a standard affine map followed by a nonlinearity, this recovers the usual deterministic neuron.

Terminology. We use EVE for the proposed local variational unit and DET for the matched deterministic baseline. KL denotes Kullback–Leibler divergence. AR denotes autoregressive temporal dependence. Internal probabilistic activity refers to quantities induced by the local posterior and prior: posterior mean, posterior variance, local KL, sampled latent state, and derived activity statistics. Latent activity is measured mainly through the squared posterior mean $\mu_j(h)^2$ and, when useful, through local KL. An activity band is an interval $[\ell_j, u_j]$ for the unit-level proxy $m_j = \mathbb{E}_h[\mu_j(h)^2]$. The out-of-range fraction is the fraction of units with $m_j < \ell_j$ or $m_j > u_j$. The labels `projOFF`, `homeo`, and `projON` denote the three experimental settings: diagnostic-only measurement, soft band regularization, and hard post-update projection.

The goal is measurement and control. A local variational neuron exposes local KL activity, squared posterior mean, out-of-range fraction, and optional AR persistence. These quantities show whether latent activity is present, below the activity band, inside the activity band, above the activity band, changing, or persistent over time. The focus is local probabilistic measurement and control at the level of the computational unit.

This measurement is useful for sequential models. In long-horizon forecasting, internal activity can evolve over many steps. Prediction error gives one view of the model. Unit-level measurements give another. They can support analysis, configuration search, early selection of promising configurations, and stability monitoring.

Main claim and scope. EVE turns a scalar computational unit into a local variational module whose posterior activity can be measured, controlled, and related to model behavior. The paper studies this claim in the long-horizon forecasting protocol in Section 4. The scope is local probabilistic activity inside the unit, its measurement, and its response to simple controls.

Contributions.

1. We define a local variational neuron with a scalar latent variable, prior, amortized posterior, reparameterized sample, and local KL term.
2. We define unit-level measurements of probabilistic activity: local KL, squared posterior mean, out-of-range fraction, and optional AR persistence.
3. We compare diagnostic-only measurement, soft band regularization, and hard projection as simple control settings.
4. We evaluate these quantities on long-horizon forecasting and use DET as the point-accuracy reference.

2 Related Work

Bayesian neural networks and uncertainty quantification. Bayesian neural networks represent uncertainty through distributions over weights [3]. Uncertainty quantification also studies calibration, confidence, and out-of-distribution behavior at the predictive output [6, 7]. Our setting is complementary. The weights may remain deterministic, while each unit carries a local latent variable and exposes local probabilistic measurements.

Variational autoencoders and hierarchical latent models. VAEs combine a prior, an amortized posterior, and an ELBO [13, 17]. Standard and hierarchical VAEs place latent variables at the model, representation, or layer level. Our construction changes the granularity. Each computational unit receives its own scalar latent variable:

$$q_\phi(z \mid h) = \prod_{i=1}^{N} q_{\phi_i}(z_i \mid h), \qquad p(z) = \prod_{i=1}^{N} p_i(z_i). \tag{1}$$

The KL term decomposes across units. This makes unit-level activity directly measurable.

Stochastic neurons and local objectives. Stochastic computation in neural networks includes wake–sleep, score-function estimators, straight-through estimators, and continuous relaxations [2, 8, 12, 15, 19]. Local and layerwise objectives can also make credit assignment more local [1, 9, 11, 16]. The local variational neuron combines these ideas at unit level: each unit has an explicit prior, posterior, reparameterized sample, and local KL measurement.

Sequential dynamics and forecasting architectures. Sequential latent-variable models add temporal stochastic structure at the model or state level [5, 14]. Long short-term memory networks (LSTMs) and gated recurrent units (GRUs) control information flow over time [4, 10]. Attention and transformer architectures use content-dependent routing between tokens or time steps [18]. These mechanisms operate at the architectural level. EVE operates at the unit level. It adds a posterior distribution, a local KL term, and internal measurements to the unit that carries the forward pass.

Position of this work. The contribution is a change of level. The prior, posterior, reparameterization, KL term, and internal measurements are placed inside a single computational unit. Related areas provide the ingredients. This work combines them as a neuron-level variational unit with measurable activity and optional local temporal persistence.

Table 1: Positioning of EVE relative to common uncertainty mechanisms. EVE places variational structure inside the computational unit rather than only on weights, outputs, or global latent variables.

Approach	**Uncertainty location**	**Computed unit**	**Training signal**	**Temporal dynamics**
Bayesian neural network	Weights	Scalar activation	Variational objective or likelihood-based training	Optional or model-level
VAE	Global latent variable	Scalar decoder units	Global ELBO	Possible sequential prior
Stochastic layer	Activations or injected noise	Scalar activation plus noise	Task loss or approximate inference	Usually implicit
Sequential latent model	State or sequence latent variable	Scalar network units	ELBO or likelihood	Model-level dynamics
EVE	Unit-level latent variable	Local posterior with sampled latent state	Task loss plus local KL	Optional unit-level AR prior

3 Method

3.1 From a deterministic unit to a local variational unit

A standard scalar neuron computes

$$a = f(w^\top h + b), \tag{2}$$

where h is the input, w and b are parameters, and f is a nonlinearity.

We replace this scalar activation with a small variational inference module. Each unit has a scalar latent variable $z \in \mathbb{R}$. The posterior parameters are predicted from the input:

$$q_\phi(z \mid h) = \mathcal{N}(\mu_\phi(h), \sigma_\phi(h)^2). \tag{3}$$

The prior is

$$p(z) = \mathcal{N}(0, 1). \tag{4}$$

The latent variable is sampled with the reparameterization trick:

$$z = \mu_\phi(h) + \sigma_\phi(h)\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 1). \tag{5}$$

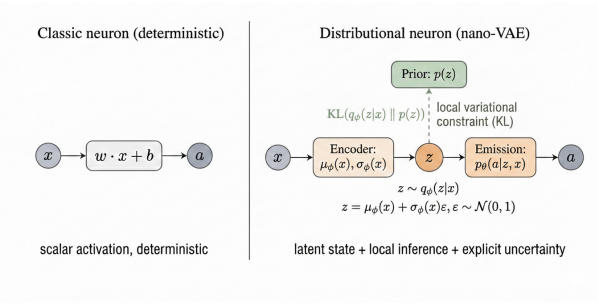
The unit emits an activation a. In general, this can be written as an emission model $p_\theta(a \mid z, h)$. In our experiments, the likelihood density itself is not used as the activation. The propagated message is either the sampled latent value z or the posterior mean $\mu_\phi(h)$. Gradients flow through the reparameterized latent variable. The local KL term regularizes the posterior toward the prior.

In the small-variance limit,

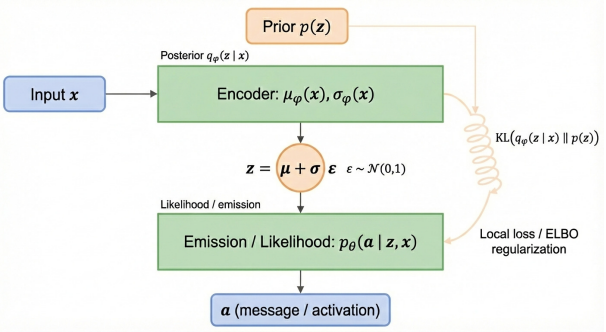
$$z = \mu_\phi(h) + \sigma_\phi(h)\varepsilon \to \mu_\phi(h). \tag{6}$$

The unit then propagates a single scalar value. If $\mu_\phi(h)$ is parameterized as a standard affine transformation followed by a nonlinearity, this recovers the usual deterministic neuron.

Figure 1 summarizes the construction.



(a) Deterministic and local variational units.



(b) Local variational mechanism.

Figure 1: Two views of the local variational neuron. Panel (a) shows how EVE extends a deterministic neuron with a local latent distribution. Panel (b) shows posterior inference, reparameterized sampling, emission, and KL regularization.

3.2 Training objective

For one unit, the variational term is

$$\mathrm{KL}(q_\phi(z \mid h) \,\|\, p(z)). \tag{7}$$

For a layer with N local variational units, we use the mean KL:

$$\mathrm{KL}_{\mathrm{mean}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{KL}_i. \tag{8}$$

This keeps the scale of the regularization stable as width changes.

The practical training loss is

$$L = L_{\mathrm{task}}(\hat{y}, y) + \beta \, \mathrm{KL}_{\mathrm{mean}} + \lambda_{\mathrm{band}} L_{\mathrm{band}} + \alpha_{\mathrm{AR}} L_{\mathrm{AR}}. \tag{9}$$

Here L_{task} is the prediction loss. The coefficient β controls the KL term. The band term controls the activity range of the latent means. The AR term controls temporal persistence.

We use Equation 9 as a supervised variational objective. If the task loss is read as a negative log-likelihood, the objective has the usual reconstruction-plus-KL form. In the forecasting experiments, the task term is prediction error and the variational term is the mean local KL over units.

3.3 Scalar latent dimension

We use one scalar latent variable per unit. This is the smallest stochastic representation. It isolates the effect of local variational inference without adding a larger hidden state.

The unit-level factorization applies to the local variational regularizer. It keeps the KL decomposable and measurable at unit level. Each unit receives a learned dense input h. Therefore, $\mu_j(h)$ and $\sigma_j(h)$ can depend on entangled upstream representations.

This scalar setting gives a simple diagnostic unit. Richer variants include correlated priors, group-wise priors, low-rank covariance structures, and $k > 1$ local latent vectors.

3.4 Internal measurements

The local posterior gives direct internal measurements. Table 2 lists the quantities used in the experiments.

Table 2: Internal measurements used in the experiments.

Measurement	**Definition**	**Interpretation**
Mean KL	$\frac{1}{N}\sum_i \mathrm{KL}_i$	Average local posterior activity.
Effective KL	$\max(\mathrm{KL} - \tau, 0)$	KL above a small activity threshold.
Squared posterior mean	$\frac{1}{N}\sum_i \mathbb{E}[\mu_i(h)^2]$	Proxy for latent activity when variance is controlled.
Out-of-range fraction	Fraction of units with activity below or above the prescribed band.	Detects weak or high local latent activity.
Projection correction	Size of the correction applied by hard projection.	Measures how often the hard control acts.
Sampling deviation	$\mathbb{E}[\lvert z - \mu \rvert]$	Checks whether sampling behaves as expected.
AR contribution	Relative contribution of the AR loss or measured local temporal dependence.	Measures local temporal persistence.

3.5 Activity band

For unit j, we define the activity proxy

$$m_j = \mathbb{E}_h \left[\mu_j(h)^2 \right]. \tag{10}$$

The low-rate, high-rate, and total out-of-range fraction are

$$\text{Low} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}[m_j < \ell_j], \qquad \text{High} = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}[m_j > u_j], \qquad \text{Out} = \text{Low} + \text{High}. \tag{11}$$

For a Gaussian posterior and a standard normal prior,

$$\text{KL}\left(\mathcal{N}(\mu, \sigma^2) \,\|\, \mathcal{N}(0,1) \right) = \frac{1}{2} \left(\mu^2 + \sigma^2 - \log \sigma^2 - 1 \right). \tag{12}$$

When σ^2 changes slowly or is constrained, μ^2 is a useful proxy for latent activity. We use this relation as a diagnostic proxy, not as a general theoretical guarantee.

We prescribe an interval $[\ell_j, u_j]$ and penalize values outside it:

$$L_{\text{band}} = \frac{1}{N} \sum_{j=1}^{N} \left[\max(0, \ell_j - m_j)^2 + \max(0, m_j - u_j)^2 \right]. \tag{13}$$

This penalty makes weak and high latent activity measurable. It also provides a direct control signal.

3.6 Control mechanisms

We compare three experimental settings. Their labels are run-log names, not new general-purpose algorithms. `projOFF` is the passive diagnostic setting: posterior sampling and KL regularization are active, but band control and hard projection are disabled. `homeo` adds soft activity-band regularization. It uses the band penalty to steer the squared posterior mean toward the prescribed activity range, without hard retraction. `projON` adds hard post-update projection. It uses the same band penalty and then rescales units whose activity leaves the prescribed range.

This separation is important. `projOFF` tests whether local probabilistic measurements exist without activity control. `homeo` tests soft control. `projON` tests hard correction.

For the soft activity-band setting, mini-batch activity is estimated by

$$\widehat{m}_j = \frac{1}{|B|} \sum_{h \in B} \mu_j(h)^2. \tag{14}$$

The band penalty is

$$L_{\text{band}} = \frac{1}{N} \sum_{j=1}^{N} \left[\max(0, \ell_j - \widehat{m}_j)^2 + \max(0, \widehat{m}_j - u_j)^2 \right]. \tag{15}$$

In `projOFF`, L_{band} is not used. The model logs $\widehat{m}_j$, low-rate, high-rate, and out-of-range fraction as diagnostics only. In `homeo`, these quantities enter the loss through L_{band}. This makes `homeo` a soft control setting rather than a passive diagnostic setting.

For hard projection, used only in `projON`, if m_j leaves its interval, we rescale the parameters that produce $\mu_j(h)$ by a factor s_j:

$$\mu_j(h) \leftarrow s_j \mu_j(h), \qquad m_j \leftarrow s_j^2 m_j. \tag{16}$$

The scaling factor is

$$s_j = \begin{cases} \sqrt{\ell_j / m_j} & \text{if } m_j < \ell_j, \\ \sqrt{u_j / m_j} & \text{if } m_j > u_j, \\ 1 & \text{otherwise.} \end{cases} \tag{17}$$

The denominator uses $\max(m_j, \varepsilon_{\text{num}})$ for numerical stability.

3.7 Autoregressive prior

For sequential data, we test a local AR prior. For unit j at time t,

$$\begin{aligned} p(z_{j,t} \mid z_{j,t-1}) &= \mathcal{N}\left(\phi_j z_{j,t-1}, \sigma_{\text{AR}}^2\right), \\ \phi_j &= \exp(-\Delta t / \tau_j). \end{aligned} \tag{18}$$

A larger ϕ_j means stronger local temporal persistence. In implementation, we use an equivalent causal penalty on posterior means:

$$L_{\text{AR}} = \frac{1}{N(T-k)} \sum_{j=1}^{N} \sum_{t=k+1}^{T} \left(\mu_{j,t} - \phi_j \mu_{j,t-k}\right)^2. \tag{19}$$

3.8 Readout and optimization

A layer contains N local variational units. The predictor aggregates their posterior means or sampled values. We use a normalized readout:

$$\hat{y} = \frac{w^\top \mu(h)}{\sqrt{N}} + b_0. \tag{20}$$

The factor $1/\sqrt{N}$ keeps the readout scale stable as width changes.

Checkpoint selection uses validation data only. For controlled EVE variants, the validation score is

$$S_{\text{val}} = \text{MSE}_{\text{val}} + \lambda_{\text{band}} L_{\text{band,val}} + \alpha_{\text{AR}} L_{\text{AR,val}}. \tag{21}$$

This score combines prediction quality and internal activity control. For `projOFF`, the band term is not optimized and the logged band statistics remain diagnostics.

All models use Adam or AdamW, mini-batches, finite-value checks, and gradient clipping. We use one Monte Carlo sample per unit unless stated otherwise. Results are reported over five random seeds.

4 Experiments

The experiments test whether EVE provides internal measurements that are active, controllable, and related to prediction error.

Table 3: Reproducibility checklist. Values are fixed before test evaluation and reported in the supplemental notebooks or appendix.

Item	**Reported value**
Optimizer	Adam or AdamW, with exact learning rate and weight decay.
Training budget	Number of epochs or iterations, batch size, and early stopping patience.
Latent dimension	$k = 1$ for all local variational units.
Posterior parameterization	Functions used to produce $\mu_\phi(h)$ and $\sigma_\phi(h)$, including the positivity transform for $\sigma_\phi(h)$.
KL coefficient	Exact value or schedule for β.
Band control	Values of ℓ_j, u_j, and λ_{band}.
Projection	Whether hard projection is enabled and how often it is applied.
AR term	Values of ϕ_j, stride k, σ_{AR}, and α_{AR} when AR is enabled.
Monte Carlo samples	Number of latent samples per unit during training and evaluation.
Seeds	Random seeds used for all reported mean and standard deviation values.
Checkpoint selection	Exact validation score used for model selection.

We use long-horizon multivariate forecasting with input length 336 and horizon 96 on Electricity (ECL), TrafficL, Weather, and Exchange. Results are reported over five seeds. Checkpoints are selected on validation data and evaluated once on test data.

DET is the point-accuracy reference. EVE variants add posterior parameters, local KL, squared posterior mean, out-of-range fraction, and optional AR persistence. The three settings are passive diagnostics (`projOFF`), soft band regularization (`homeo`), and hard projection (`projON`).

External metrics are mean squared error (MSE) and mean absolute error (MAE). Internal metrics are mean KL, squared posterior mean, out-of-range fraction, band penalty, and AR contribution.

Evidence for the central claim. The reported runs support the three parts of the claim. First, the internal measurements are active: mean KL is nonzero across EVE runs, and selected checkpoints show mean KL values ranging from 44.3 to 297.8 across datasets and control settings. Second, the activity is controllable: the out-of-range fraction changes with the control setting. On Weather, it decreases from 0.772 ± 0.012 without projection to 0.573 ± 0.022 with soft band regularization. On Exchange, it decreases from 0.594 ± 0.056 to 0.403 ± 0.036. Third, the internal measurement relates to external behavior: across $N = 60$ runs, the out-of-range fraction correlates with test MSE after dataset-wise standardization, with $r = 0.567$ and $p = 2.3 \times 10^{-6}$. These results show that EVE exposes internal probabilistic activity that is measurable, responsive to control, and informative about prediction error under this protocol.

Reproducibility. We provide an anonymized supplemental archive containing the notebooks used to reproduce the reported experiments.

4.1 Deterministic accuracy reference

DET uses the same task, input length, horizon, readout, optimizer family, training budget, checkpoint-selection split, and random seeds as EVE. It removes only the local variational mechanism: no posterior, no latent sampling, no local KL term, no activity-band control, and no AR latent prior.

The comparison has a specific role. DET is the matched point-accuracy reference. It measures how well the architecture predicts without local stochastic units. EVE keeps point prediction, but adds a posterior, a prior, a reparameterized latent state, a local KL term, and optional internal controls.

The three EVE settings separate measurement from control. `projOFF` activates the posterior, sampling, and local KL, while leaving activity-band control disabled. It is the passive diagnostic

setting. `homeo` adds a soft activity-band penalty and tests whether internal activity can be steered without hard retraction. `projON` adds a hard post-update projection and tests explicit correction into the prescribed activity band.

Thus, the comparison is not framed as a pure point-accuracy replacement claim. It is a trade-off study between matched predictive accuracy and internal probabilistic observability.

4.2 Control variants

We compare three EVE variants: passive diagnostics, soft band regularization, and hard projection. The goal is to test how each control setting changes prediction error and internal activity.

No projection gives the best EVE MSE on ECL and TrafficL. Soft band regularization gives the best EVE MSE on Weather and Exchange. It also gives the lowest out-of-range fraction on Weather and Exchange. The internal measurements therefore respond to the control mechanism.

4.3 Relation between internal measurements and prediction error

Across $N = 60$ runs, after standardizing within each dataset, the out-of-range fraction correlates with test MSE:

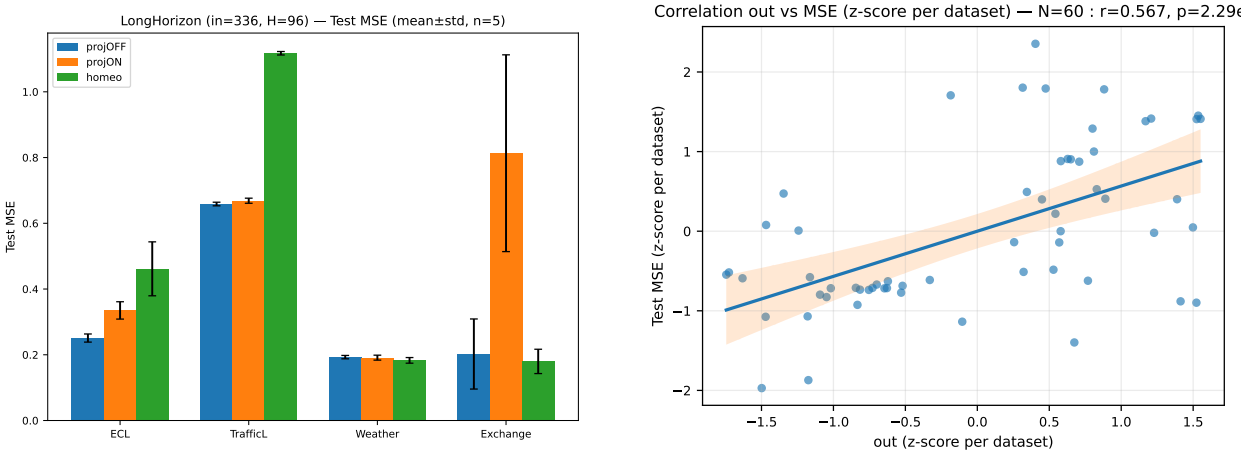
$$r = 0.567, \qquad p = 2.3 \times 10^{-6}.$$

Mean KL has a weaker relation:

$$r = 0.351, \qquad p = 6.0 \times 10^{-3}.$$

Among the tested internal measurements, the out-of-range fraction gives the clearest relation to test MSE.

This correlation is observational. It does not establish a causal relation between out-of-range fraction and prediction error. It supports the use of the out-of-range fraction as a diagnostic signal under this protocol.



(a) Test MSE across control variants.

(b) Out-of-range fraction versus test MSE.

Figure 2: External accuracy and internal measurement. Panel (a) reports test MSE for EVE variants. Panel (b) shows the relation between out-of-range fraction and test MSE after dataset-wise standardization.

4.4 Early validation signals

On ECL with soft band regularization, we run a staged hyperparameter search. It starts with 24 configurations for 3 epochs and keeps the most promising configurations for later stages. Early out-of-range activity correlates with the best validation MSE:

$$r = 0.582, \qquad p = 2.8 \times 10^{-3}.$$

The AR contribution anti-correlates with best validation MSE:

$$r = -0.606, \qquad p = 1.7 \times 10^{-3}.$$

These quantities can be used as candidate early validation signals for ranking configurations under this protocol.

4.5 Autoregressive ablation

We compare soft band regularization with and without the AR term. The AR term creates a measurable temporal signal. The AR contribution becomes nonzero when the term is enabled. On Exchange, AR-on gives lower average MSE. On ECL, AR-off gives lower average MSE. This shows that temporal persistence is measurable and can be studied as a separate control signal.

4.6 Weather training dynamics

We also inspect training dynamics on Weather. This view shows how prediction error and internal activity evolve during training.

Figure 3 supports reporting both external metrics and internal measurements.

5 Discussion

EVE exposes probabilistic measurements at the level of the computational unit. In the reported runs, local KL is active, squared posterior mean is measurable, and the out-of-range fraction changes across control settings.

DET gives the point-accuracy reference. EVE adds observable and controllable internal probabilistic structure. This trade-off is useful for monitoring, debugging, configuration search, and stability analysis.

The relation between out-of-range fraction and forecasting error is observational. It supports diagnostic use under this protocol, not a causal claim.

6 Limitations

This study isolates a local variational unit in a focused setting. Each unit uses one scalar latent variable, and the variational regularizer is factorized across units. This makes KL decomposable and measurable, but it does not model cross-unit posterior covariance. Extensions include correlated priors, group-wise priors, low-rank covariance, and $k > 1$ local latent vectors.

The experiments focus on long-horizon forecasting. This stresses drift, saturation, and temporal persistence. Broader tests should cover more domains, stronger sequence models, compute-matched hybrids, and larger benchmarks. The controls are simple by design: passive diagnostics, soft band regularization, hard projection, and an AR term.

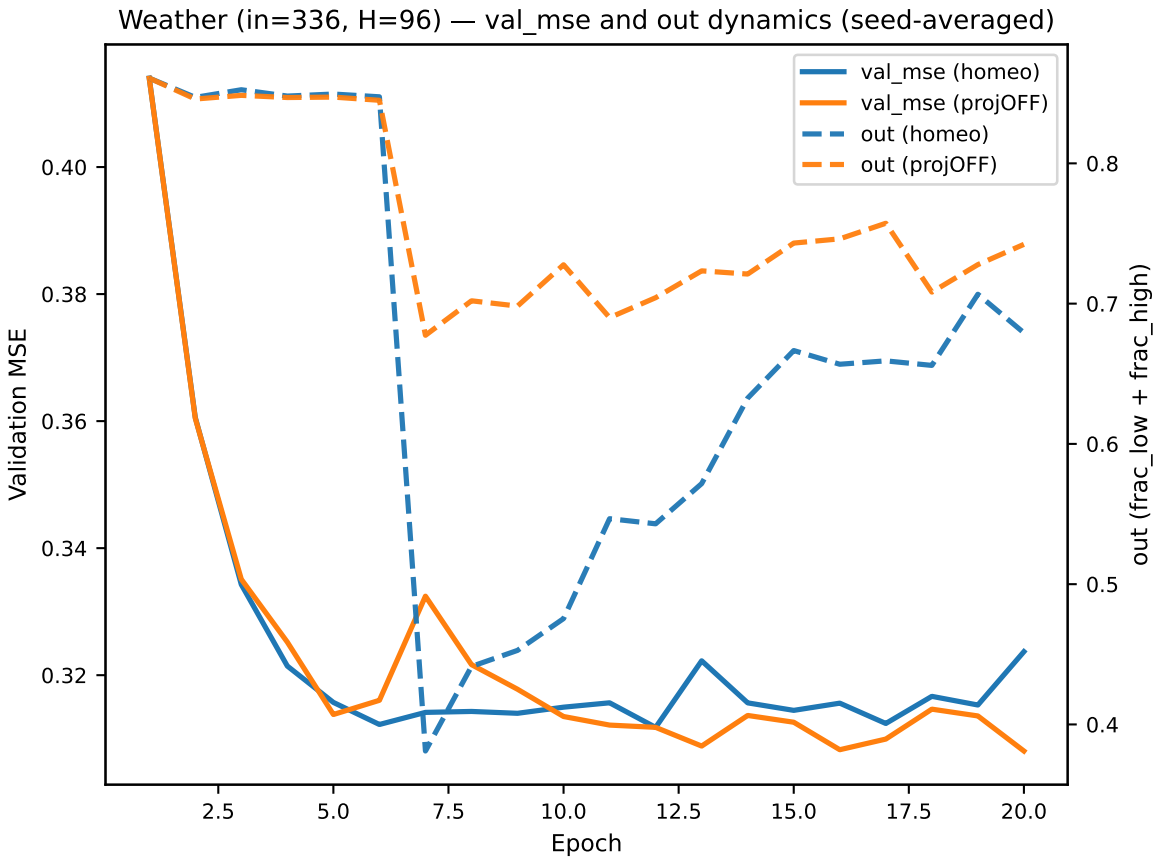


Figure 3: Weather validation dynamics. The plot compares validation MSE and out-of-range fraction for soft band regularization and no projection. Internal activity can change independently from prediction error.

EVE is more costly than DET because it computes posterior parameters, latent samples, local KL, and control terms. In the supplied notebooks, selected EVE variants are about 3–5× slower. Compute-matched scaling is left to future work.

7 Conclusion

We introduced a local variational neuron: a neural unit with a scalar latent variable, a prior, an amortized posterior, a reparameterized sample, and a local KL term. This moves variational inference into the computational unit itself. The unit exposes measurable internal activity through local KL, squared posterior mean, out-of-range fraction, and AR persistence.

These measurements respond to simple controls, and the out-of-range fraction relates to prediction error across runs. A matched deterministic model remains the point-accuracy reference. The main lesson is that unit-level variational inference creates measurable quantities for local probabilistic computation. This supports diagnosis, stability monitoring, configuration search, and controlled training strategies in future neural architectures.

References

[1] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

[2] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation, 2013. arXiv:1308.3432.

[3] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[4] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014.

[5] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[6] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[7] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[8] Geoffrey E. Hinton, Peter Dayan, Brendan J. Frey, and Radford M. Neal. The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158–1161, 1995. doi: 10.1126/science.7761831.

[9] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006. doi: 10.1162/neco.2006.18.7.1527.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9 (8):1735–1780, 1997.

[11] Max Jaderberg, Wojciech M. Czarnecki, Simon Osindero, Oriol Vinyals, Alex Graves, Koray Kavukcuoglu, and David Silver. Decoupled neural interfaces using synthetic gradients. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[12] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-softmax. In *International Conference on Learning Representations (ICLR)*, 2017.

[13] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[14] Rahul G. Krishnan, Uri Shalit, and David Sontag. Deep kalman filters, 2015. arXiv:1511.05121.

[15] Andriy Mnih and Danilo J. Rezende. Variational inference for monte carlo objectives. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

[16] Arild Nøkland and Lars Hiller Eidnes. Training neural networks with local error signals. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[17] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models, 2014. arXiv:1401.4082.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[19] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992. doi: 10.1007/BF00992696.