NEURAL BELIEF-STATE FILTERING WITH CROSS-VARIABLE COUPLING FOR IRREGULAR TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

While many sequence models assume evenly spaced observations, in domains such as intensive care different measurements and interventions occur at irregular, asynchronous times, and the underlying state is only partially observed. We introduce Neural Belief-State Filtering (NBSF), a continuous-discrete filter that maintains a probabilistic belief over observed and latent variables, propagates it across irregular gaps with a structured exponential-decay backbone and a gap-dependent nonlinear residual, and assimilates each observation through a learned, innovation-driven update. The distinguishing feature of NBSF is that an observation of one variable can directly revise the belief about other, currently unobserved variables at event time rather than only influencing them later through temporal propagation and it does so without maintaining a dense filtering covariance. We study three assimilation variants: a local-coordinate control that updates only the observed variable, a learned dense cross-variable gain, and a nonlinear cross-variable update. Across four clinical benchmarks and a controlled sensing task, event-time cross-variable assimilation is most useful when direct target observations are scarce, and its advantage grows under target suppression; the dense-gain variant provides the most consistent benefit, while the nonlinear update shows dataset-dependent gains. Parameter-matched ablations show these gains are not explained by model size. Our central finding is that event-time cross-variable information flow is the more consistent mechanism under target scarcity, whereas additional nonlinear update complexity is not consistently beneficial. NBSF uses only $\sim$140K–170K parameters, provides an explicit probabilistic predictive belief whose calibration is evaluated in-distribution, and achieves competitive performance across established clinical irregular-time-series benchmarks.

1 INTRODUCTION

Clinical time-series data are inherently irregular, sparse, and asynchronous across variables. Measurements are often noisy and collected at variable frequencies, resulting in partially observed trajectories that require models to reason under both temporal dynamics and uncertainty. Importantly, this irregularity is not always random; observation patterns themselves reflect clinical decision-making and can be informative about patient state (Lipton et al., 2016; Che et al., 2018).

Existing approaches address these challenges in different ways. Recurrent models such as GRU-D (Che et al., 2018) incorporate missingness heuristics and time gaps but maintain deterministic hidden states. Attention-based methods such as mTAND (Shukla & Marlin, 2021) capture long-range dependencies but rely on time encoding or resampling to handle irregularity. Continuous-time models, including Neural ODEs (Chen et al., 2018) and Neural CDEs (Kidger et al., 2020), provide a principled treatment of irregular sampling but treat observations as part of a continuous trajectory and do not explicitly separate state evolution from observation updates. More broadly, many approaches handle missing data implicitly or through preprocessing rather than modeling uncertainty directly.

The clinical setting naturally motivates a probabilistic filtering perspective. A model should maintain a belief over patient state, propagate uncertainty between observations, and update this belief when new measurements arrive. Classical filtering methods such as the Kalman filter (Kalman, 1960; Simon, 2006) formalize this structure, but these methods rely on linear or analytically specified dynamics and do not readily accommodate high-dimensional, heterogeneous clinical data. Learnable

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.

nonlinear extensions such as deep Kalman filters (Krishnan et al., 2015) provide more flexible latent dynamics, but are typically formulated around discrete-time latent transitions rather than explicit continuous-discrete filtering under asynchronous observation events.

We introduce Neural Belief-State Filtering (NBSF), a neural continuous-discrete filtering framework that combines probabilistic filtering structure with learned nonlinear dynamics. The model maintains a distribution over latent state and evolves it through two operations: (i) time-aware propagation between observations, and (ii) explicit updates at observation times. A central component of the model is a learned cross-variable update that maps the innovation from a single observation to a dense correction across latent and observed dimensions, enabling direct cross-variable information flow and uncertainty adjustment in unobserved variables (e.g., Figure 1). Rather than relying on a predefined covariance structure for observation coupling, NBSF learns observation-conditioned cross-variable corrections directly from data.

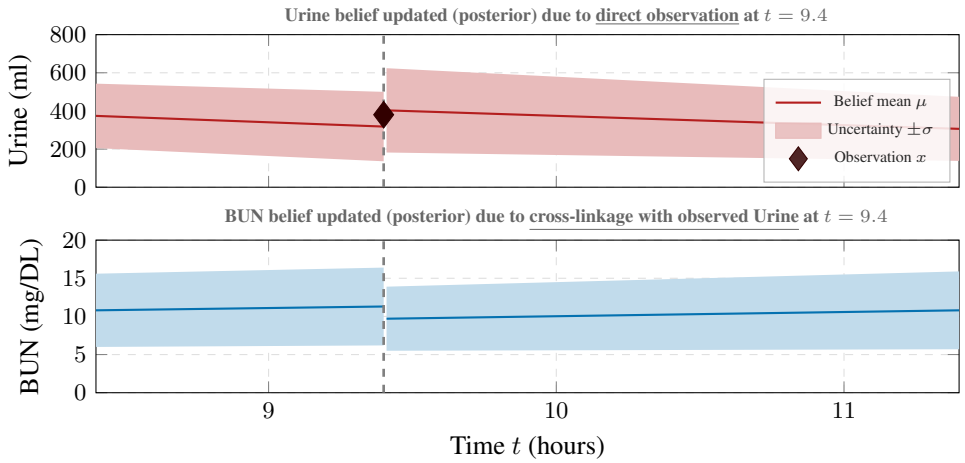


Figure 1: Proposed NBSF propagates beliefs about variables (μ and σ) between irregular observation times. Observations (e.g., urine output) directly update the observed variable and can simultaneously modify the beliefs of related unobserved variables through the learned cross-variable update (e.g., blood urea nitrogen (BUN)). Actual test results from patient ID 132936 in the PhysioNet 2012 dataset are shown.

Our contributions are:

- We introduce NBSF, a continuous-discrete probabilistic framework with learned propagation and innovation-based updates for irregular time series.
- We develop a cross-variable update that allows an observation to directly modify multiple belief dimensions without maintaining a dense filtering covariance.
- We combine structured nonlinear mean dynamics with an exact Ornstein–Uhlenbeck variance transition for irregular time gaps.
- Across clinical and controlled sensing experiments, we show that access to cross-variable information is most useful when direct target information is sparse or unavailable, while the benefit of nonlinear update complexity is dataset dependent.

2 Problem Formulation and Related Work

2.1 Problem Formulation

We represent an irregular multivariate time series as asynchronous events

$$\mathcal{D} = \{(t_i, k_i, x_i)\}_{i=1}^{N},$$

where t_i is the observation time, k_i identifies the observed variable, and x_i is its value. Missingness is represented by the absence of an event.

Rather than maintaining a deterministic hidden state, we maintain a probabilistic belief $q_\theta(z(t) \mid \mathcal{F}_t)$. Between events the belief is propagated through gap-aware continuous-discrete dynamics; when an observation arrives, its innovation updates the belief. Downstream predictions are obtained from the resulting posterior state. Section 3 specifies these operators.

2.2 Existing Approaches as Special Cases

Many existing approaches can be interpreted as restricted instances of this formulation, differing in how they represent state, handle time, and incorporate observations, as summarized in Table 1.

Deterministic sequence models: Recurrent models such as GRU-D (Che et al., 2018) combine propagation and update into a single deterministic transition, without maintaining an explicit probabilistic belief over state. Attention-based models such as mTAND (Shukla & Marlin, 2021) model sequences via attention but similarly do not maintain an explicit belief distribution over the evolving state. APN (Liu et al., 2026b) uses adaptive patch boundaries and time-aware averaging for irregular sequences. Patching-based methods such as t-PatchGNN (Zhang et al., 2024) transform irregular sequences into locally structured temporal segments, enabling modeling without explicit interpolation.

Continuous-time models: Neural ODEs (Chen et al., 2018) and ODE-RNN (Rubanova et al., 2019) model continuous dynamics, while Neural CDEs (Kidger et al., 2020) extend this idea to controlled differential equations. These approaches naturally handle irregular sampling, although their treatment of observations differs from the explicit continuous-discrete filtering formulation considered here. In particular, state evolution and uncertainty-aware observation assimilation are not generally represented as separate probabilistic filtering operations.

Graph and interaction-based models: Graph-based approaches such as Raindrop (Zhang et al., 2022) model inter-variable dependencies through learned graph structures. GARLIC (Li et al., 2026) further improves irregular clinical modeling by learning relational structure across variables. MTFP-DG (Liu et al., 2026a) combines temporal-frequency patching with dynamic graph modeling for asynchronous medical data. While expressive, these approaches primarily focus on deterministic predictive representations rather than maintaining an explicit probabilistic belief over the evolving state.

Foundation models for healthcare: Recent work such as MIRA (Li et al., 2025) introduces medical foundation models that use continuous-time rotary positional encoding to natively handle irregular intervals and heterogeneous sampling. These models provide expressive sequence representations but do not explicitly maintain a probabilistic filtering belief that is propagated and updated at individual observation events.

Probabilistic state-space models: Classical filters such as the Kalman filter and its continuous-discrete and nonlinear variants (Kalman, 1960; Simon, 2006; Julier & Uhlmann, 1997) maintain an explicit belief but assume linear or analytically specified models; Gaussian processes (Futoma et al., 2017) add uncertainty at growing cost; and Deep Kalman Filters and deep state-space models (Krishnan et al., 2015; 2017) learn nonlinear latent dynamics through amortized inference. Related probabilistic models for irregular observations include Irregular-Time Bayesian Networks (Ramati & Shahar, 2010) and continuous-time Bayesian networks (Liu et al., 2019). Related models such as GRU-ODE-Bayes (De Brouwer et al., 2019) combine continuous-time dynamics with Bayesian observation updates for sporadically observed time series. These methods establish important probabilistic treatments of irregular observations, but differ from NBSF in how cross-variable observation information is incorporated into the belief state. In particular, rather than a Bayesian update tied to per-coordinate dynamics, NBSF applies a learned operator in which the innovation from a single observed variable directly corrects the belief of other observed and latent coordinates, while keeping a factorized marginal belief and never forming a dense filtering covariance. More recent continuous-discrete models are closer in spirit: Continuous Recurrent Units derive a recurrent architecture from the continuous-discrete Kalman filter (Schirmer et al., 2022), Neural Continuous-Discrete State Space Models perform Bayesian inference for irregularly sampled series (Ansari et al., 2023), and ACSSM applies amortized control of a continuous state-space model with data assimilation (Park et al., 2025). These methods realize probabilistic assimilation through different Bayesian or latent-inference

Table 1: Comparison of representative approaches for irregular time-series modeling. The methods differ in their treatment of continuous time, observation events, uncertainty, and cross-variable information.

Model	State time	Obs time	Propagation	Latent state	Uncertainty	Update	Cross-var
Kalman Filter (Kalman, 1960)	Discrete	Discrete	Linear	**Explicit**	**Probabilistic**	Bayesian	Covariance
Hybrid KF (Simon, 2006)	**Continuous**	**Irregular**	Linear	**Explicit**	**Probabilistic**	Bayesian	Covariance
EKF / UKF (Simon, 2006; Julier & Uhlmann, 1997)	Discrete	Discrete	**Nonlinear**	**Explicit**	**Probabilistic**	Bayesian	Covariance
Deep Kalman Filter (Krishnan et al., 2015)	Discrete	Discrete	**Nonlinear**	Yes	**Probabilistic**	Implicit	Latent
Gaussian Process (Futoma et al., 2017)	**Continuous**	**Irregular**	Non-param.	No	**Probabilistic**	Implicit	Kernel
Deep SSM (Krishnan et al., 2017)	Discrete	Discrete	**Nonlinear**	**Explicit**	**Probabilistic**	Implicit	Latent
Neural ODE (Chen et al., 2018)	**Continuous**	**Irregular**	**Nonlinear**	**Explicit**	None	None	None
GRU-D (Che et al., 2018)	Discrete	**Irregular**	**Nonlinear**	Implicit	None	Implicit	Indirect
Neural CDE (Kidger et al., 2020)	**Continuous**	**Irregular**	**Nonlinear**	**Explicit**	None	Implicit	Implicit
ODE-RNN (Rubanova et al., 2019)	**Continuous**	**Irregular**	**Nonlinear**	**Explicit**	None	RNN	Implicit
GRU-ODE-Bayes (De Brouwer et al., 2019)	**Continuous**	**Irregular**	**Nonlinear**	**Explicit**	**Probabilistic**	Bayesian	Implicit
CRU (Schirmer et al., 2022)	**Continuous**	**Irregular**	Linear	**Explicit**	**Probabilistic**	Bayesian	Latent
NCDSSM (Ansari et al., 2023)	**Continuous**	**Irregular**	**Nonlinear**	**Explicit**	**Probabilistic**	Bayesian	Latent
ACSSM (Park et al., 2025)	**Continuous**	**Irregular**	**Nonlinear**	**Explicit**	**Probabilistic**	Bayesian	Latent
mTAND (Shukla & Marlin, 2021)	Discrete	**Irregular**	**Nonlinear**	Implicit	None	Attention	Attention
Raindrop (Zhang et al., 2022)	Discrete	**Irregular**	**Nonlinear**	Implicit	None	Implicit	Graph
GARLIC (Li et al., 2026)	Discrete	**Irregular**	**Nonlinear**	Implicit	None	Implicit	Graph
NBSF (ours)	**Continuous**	**Irregular**	**Nonlinear**	**Explicit**	**Probabilistic**	**Learned innovation**	**Learned**

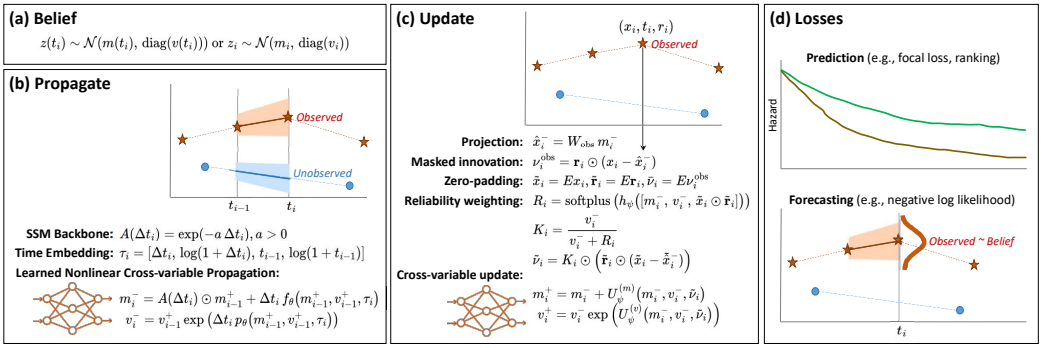


Figure 2: NBSF visual explanation: (a) Belief is maintained as mean $m(t)$ and variance $v(t)$ for observed and unobserved states. (b) It is propagated between observation instances using a structured state-space backbone and learnable nonlinear cross-variable residual function, with an exact Ornstein–Uhlenbeck variance transition for uncertainty propagation. (c) When a variable is observed at instance t_i, a learnable nonlinear cross-variable function updates the posterior belief using the observation innovation. (d) A multi-task head is used for forecasting and prediction. Some details are omitted for brevity.

constructions. NBSF instead maintains a factorized marginal belief and applies an explicit learned event-time operator that transfers the innovation from one observed variable to other observed and latent coordinates, without maintaining a full filtering covariance. Our controlled variants further isolate when this cross-coordinate transfer is responsible for the gains.

2.3 Limitations of Existing Approaches

These approaches differ mainly in how an observation updates the state. In local-update formulations an observation corrects only its own coordinate, with cross-variable information entering indirectly through later propagation; classical multivariate filters transfer it through covariance structure; and neural graph and interaction-based models capture it within deterministic hidden representations rather than an explicit probabilistic assimilation step. NBSF instead learns both propagation and observation-update operators, so each observation induces a structured cross-variable correction over observed and latent dimensions while maintaining an explicit probabilistic belief.

3 Proposed Approach

We instantiate the general filtering formulation using a Gaussian belief-state parameterization with learned propagation and update operators. This corresponds to a continuous-discrete filtering set-

ting: the latent belief evolves between observation times and is updated discretely when irregular observations arrive. See Figure 2 for a visual description.

3.1 Belief-State Representation

We represent the observed and latent variables of a patient as a factorized Gaussian belief state:

$$z(t) \sim \mathcal{N}(m(t),\, \mathrm{diag}(v(t)))\,, \tag{1}$$

where $m(t), v(t) \in \mathbb{R}^S$ and $S = D + L$, with D observed variables and L latent dimensions. For numerical stability, variance is parameterized in log-space:

$$\ell(t) = \log v(t), \qquad \ell(t) \leftarrow \mathrm{clamp}(\ell(t), c_{\min}, c_{\max}), \qquad v(t) = \exp(\ell(t)). \tag{2}$$

The factorized representation requires $O(S)$ storage for marginal uncertainty and avoids maintaining an $S \times S$ filtering covariance; cross-variable information enters through the learned update rather than a dense covariance matrix. We use $-$ and $+$ to denote the prior and posterior beliefs at an observation event, respectively.

3.2 Nonlinear Belief Propagation

The belief evolves over an inter-observation interval $\Delta t_i = t_i - t_{i-1}$ using a hybrid propagation model. For the mean, we combine a structured state-space decay

$$A(\Delta t_i) = \exp(-a\,\Delta t_i), \qquad a > 0, \tag{3}$$

with a neural residual function f_θ that models nonlinear cross-variable dynamics:

$$m_i^- = A(\Delta t_i) \odot m_{i-1}^+ + \Delta t_i\, f_\theta\big(m_{i-1}^+, \exp(\ell_{i-1}^+), \boldsymbol{\tau}_i\big)\,. \tag{4}$$

For the variance, rather than adding a learned increment proportional to Δt_i, we use the exact transition of a coordinate-wise Ornstein–Uhlenbeck process. Let $q > 0$ denote the process-noise rate. The prior variance is

$$v_i^- = A(\Delta t_i)^2 \odot v_{i-1}^+ + \frac{q}{2a} \odot \big(1 - A(\Delta t_i)^2\big)\,, \tag{5}$$

where all operations are element-wise. We then set

$$\ell_i^- = \mathrm{clamp}\big(\log(v_i^- + \epsilon), c_{\min}, c_{\max}\big)\,. \tag{6}$$

Equation 5 is the exact variance transition for the coordinate-wise OU process when a and q are constant over the inter-observation interval. Unlike an additive variance increment, it approaches the stationary variance $q/(2a)$ as the observation gap increases, preventing uncertainty from growing without bound solely because Δt_i is large.

We use the time encoding $\boldsymbol{\tau}_i = [\Delta t_i, \log(1 + \Delta t_i), t_{i-1}, \log(1 + t_{i-1})]$. Further properties of the OU transition and hybrid propagation are given in Appendices D.2 and D.3.

3.3 Relational Neural Update

At observation time t_i, we receive a partially observed vector $x_i \in \mathbb{R}^D$ with mask $\mathbf{r}_i \in \{0, 1\}^D$. The mask indicates which variables are observed at time t_i.

Let $W_{\mathrm{obs}} \in \mathbb{R}^{D \times S}$ denote the projection from the full belief state to the observed-variable space. We first project the prior belief mean into observation space:

$$\hat{x}_i^- = W_{\mathrm{obs}}\, m_i^-, \qquad \hat{x}_i^- \in \mathbb{R}^D. \tag{7}$$

The observation-space masked innovation is

$$\nu_i^{\mathrm{obs}} = \mathbf{r}_i \odot \mathrm{clip}\big(x_i - \hat{x}_i^-, -10, 10\big)\,, \qquad \nu_i^{\mathrm{obs}} \in \mathbb{R}^D. \tag{8}$$

Observation-space quantities are embedded into the full belief space. Let $E \in \mathbb{R}^{S \times D}$ denote the zero-padding embedding operator. We define

$$\tilde{x}_i = E x_i, \qquad \tilde{\mathbf{r}}_i = E \mathbf{r}_i, \qquad \bar{\nu}_i = E \nu_i^{\mathrm{obs}}, \tag{9}$$

where $\tilde{x}_i, \tilde{\mathbf{r}}_i, \bar{\nu}_i \in \mathbb{R}^S$.

Reliability-weighted innovation: We learn a gain-like reliability term in the full belief space:

$$R_i = \mathrm{softplus}\big(h_\psi\big([m_i^-,\, \exp(\ell_i^-),\, \tilde{x}_i \odot \tilde{\mathbf{r}}_i]\big)\big) + \epsilon, \qquad R_i \in \mathbb{R}^S, \tag{10}$$

$$K_i = \mathrm{clamp}\left(\frac{\exp(\ell_i^-)}{\exp(\ell_i^-) + R_i}, 0,\, K_{\max}\right), \qquad K_i \in \mathbb{R}^S. \tag{11}$$

The gain therefore controls the magnitude of information passed from the observation innovation according to the current predictive uncertainty and learned reliability term. The gain-weighted innovation is

$$\tilde{\nu}_i = K_i \odot \bar{\nu}_i. \tag{12}$$

Cross-variable update: The posterior correction is produced by a neural update. We write the mean and log-variance components of the update as U_ψ^m and U_ψ^ℓ, respectively. To ensure that a zero innovation cannot induce an arbitrary shift in the belief mean, we use an innovation-anchored mean correction:

$$\Delta m_i = U_\psi^m\big(m_i^-, \ell_i^-, \tilde{\nu}_i\big) - U_\psi^m\big(m_i^-, \ell_i^-, 0\big), \tag{13}$$

$$\Delta \ell_i = U_\psi^\ell\big(m_i^-, \ell_i^-, \tilde{\nu}_i\big), \tag{14}$$

$$m_i^+ = m_i^- + \Delta m_i, \tag{15}$$

$$\ell_i^+ = \mathrm{clamp}\big(\ell_i^- + \mathrm{clip}(\Delta \ell_i, -2, 0.2), c_{\min}, c_{\max}\big). \tag{16}$$

The anchoring gives $\Delta m_i = 0$ exactly when the innovation is zero. Unlike a local coordinate update, the neural mapping can use an innovation from one observed variable to modify multiple observed and latent belief dimensions, enabling instantaneous cross-variable information flow without maintaining a dense covariance matrix in the filtering state. Only observed variables contribute to the innovation term.

Update variants. We study three instantiations of the observation update, holding propagation and all other components fixed. The *local-coordinate* update is a diagonal, coordinate-wise correction in which an observed variable updates only its own belief dimension, with no direct cross-variable flow at observation time; it serves as our control. Cross-variable coupling still occurs during propagation through f_θ; the local-coordinate control removes such coupling only at observation time, isolating the effect of *event-time* cross-variable assimilation. The *cross-variable* update is the nonlinear mapping U_ψ defined above, which lets one observation correct multiple observed and latent dimensions. The *dense-gain* update replaces U_ψ with a single learned dense linear gain applied to the embedded innovation, providing cross-variable flow through an explicit dense gain rather than a nonlinear map. Both the cross-variable and dense-gain updates share observation information across variables; the local-coordinate update does not do so during assimilation.

The learned gain K_i and correction U_ψ make this assimilation filtering-inspired rather than an exact Bayesian posterior update; the variance propagation in Eq. (5) by contrast is the exact coordinate-wise OU transition. For a joint Gaussian belief with $\Sigma_{ij} \neq 0$, observing coordinate j shifts the posterior mean of an unobserved coordinate i by $\Sigma_{ij}(\Sigma_{jj} + \sigma_j^2)^{-1}(y_j - \mu_j^-)$, a correction that a coordinate-local update cannot make. NBSF learns an analogue of this cross-coordinate correction directly, without maintaining the full covariance.

3.4 Target-Dropout Training

The standard model assimilates every available observation. We additionally train a target-dropout variant that withholds a subset of available target observations from the assimilation mask while retaining them for the forecasting objective. This encourages the filter to use temporal and cross-variable information when direct target evidence is unavailable; full settings are reported in Appendix B.13.

3.5 Training and Loss

The model is trained end-to-end using a joint objective:

$$\mathcal{L} = \mathcal{L}_{\text{prediction}} + \lambda\, \mathcal{L}_{\text{forecasting}}, \tag{17}$$

Table 2: Characteristics of datasets used in the experiments.

Dataset	Variables	Patients/Series	Task
PhysioNet 2012 (Silva et al., 2012)	36	11,988	Stay >3 days; mortality
PhysioNet 2019 (Reyna et al., 2019)	34	39,104	Sepsis prediction
MIMIC-III (Johnson et al., 2016)	21	49,162	In-hospital mortality
eICU (Pollard et al., 2019)	27	195,495	In-hospital mortality
PAMAP2 (Reiss & Stricker, 2012)	31	1,208	Controlled sensing evaluation

where $\mathcal{L}_{\text{prediction}}$ is the downstream prediction loss (binary cross-entropy for the mortality and sepsis classification tasks), and $\mathcal{L}_{\text{forecasting}}$ is a forecasting loss, such as the negative log likelihood of an observation under the predictive belief distribution. The hyperparameter $\lambda \geq 0$ controls the relative contribution of the forecasting objective.

Downstream predictions use the final posterior belief $h_\omega([m_N^+; \ell_N^+])$.

3.6 Frozen-Marginal Joint Uncertainty

The filtering state uses a diagonal covariance for efficiency. For joint prediction we optionally augment it with a frozen-marginal correlation factor $\Sigma_i = D_i^{1/2} R_i D_i^{1/2}$ (with $D_i = \text{diag}(v_i)$, $\text{diag}(R_i) = 1$, and low-rank R_i), so that $\text{diag}(\Sigma_i) = v_i$ exactly and joint dependence can change without altering marginal predictions (Appendix D.5).

3.7 Theoretical Insights

The exponential backbone yields the first-order dynamics $\dot{m} = -a \odot m + f_\theta(m, v, t)$, while the exact OU variance transition approaches the finite stationary variance $q/(2a)$. Under bounded residual dynamics and bounded inter-observation gaps, the propagated mean remains uniformly bounded. Formal derivations and proofs are provided in Appendices D.2 and D.3.

4 Experiments and Results

4.1 Datasets, Tasks, Metrics, and Settings

We evaluate NBSF on irregularly sampled clinical ICU time-series datasets summarized in Table 2. PhysioNet 2012 (P12), PhysioNet 2019 (P19), MIMIC-III, and eICU are used for clinical evaluation, while PAMAP2 is used as a controlled non-clinical sensing experiment. For P12 and P19, we use five stratified folds. For MIMIC-III and eICU, encounters from the same patient are kept within the same split to avoid subject-level leakage. PAMAP2 is evaluated using leave-one-subject-out validation.

For prediction, we report AUROC and AUPRC; belief quality is evaluated using CRPS and NLL, with Brier score and ECE for classification calibration. Preprocessing and optimization details are provided in Appendices A.2 and A.3.

4.2 Prediction using Complete Data

Table 3 compares NBSF with representative irregular time-series baselines under the exact benchmark protocol of Li et al. (2026) (identical task definitions, preprocessing, and data splits); P12 LOS uses the task-matched Raindrop benchmark (Zhang et al., 2022), and baseline values are as reported for these benchmarks (AUROC/AUPRC, percentage points). Under this matched comparison, NBSF (dense-gain) attains the best reported AUROC and AUPRC on three of the four tasks (P12 LOS, P19, and MIMIC-III) and is second only on P12 mortality, at 134–170K parameters and under 1.5 GB peak memory, while scaling to the 195K-patient eICU cohort (Appendix A.5). Because baseline numbers are taken as reported rather than re-executed on our folds, we base our mechanism-level conclusions on the controlled within-model experiments that follow.

No single NBSF update wins everywhere: the dense-gain variant is strongest on the three tasks above, while the nonlinear cross-variable update is the stronger variant on P12 mortality and patient-grouped eICU (Appendix B.3). Both are cross-variable variants; the local-coordinate control that removes

Table 3: **Prediction results.** Five-fold AUROC/AUPRC (%, mean±SD). Bold and underline denote the best and second-best reported values per task.

	P12 LOS		P12 Mortality		P19		MIMIC-III	
Method	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
GRU-D (Che et al., 2018)	67.2±3.6	95.9±2.1	84.45±1.90	50.74±4.96	88.73±0.86	47.81±3.37	86.36±1.68	56.14±1.87
mTAND (Shukla & Marlin, 2021)	65.3±1.7	96.5±1.2	84.30±1.69	50.05±3.96	81.73±1.53	37.27±3.93	88.00±0.31	57.73±2.21
Raindrop (Zhang et al., 2022)	72.1±1.3	97.0±0.4	83.03±1.37	45.91±3.55	87.41±1.13	46.33±3.09	87.18±1.51	57.06±6.01
GARLIC (Li et al., 2026)	—	—	**86.40**±0.86	**56.89**±1.75	90.96±0.84	55.29±2.45	<u>90.09</u>±0.45	64.85±1.68
NBSF (dense-gain)	**81.53**±2.98	**97.89**±0.59	83.12±1.09	49.06±2.12	**93.00**±0.54	**62.39**±3.04	**90.21**±0.53	**66.59**±2.31
NBSF (cross-var.)	<u>79.36</u>±1.25	<u>97.75</u>±0.18	<u>84.67</u>±0.90	<u>51.60</u>±2.83	<u>92.11</u>±0.77	<u>61.17</u>±2.35	89.95±0.53	<u>65.82</u>±2.40

direct cross-variable flow at observation time is evaluated below, where its advantage is clearest under limited direct observation. These differences motivate direct evaluation of belief quality and controlled tests of when cross-variable information helps.

4.3 Belief Quality under Limited Direct Observation

We next evaluate whether cross-variable information improves the maintained belief when direct measurements of a target variable are limited. We consider two settings: the belief before a variable is first observed and the belief immediately before an observation is assimilated.

On P12, the cross-variable model improves before-first CRPS from 0.5956 for the local-coordinate update to 0.5707, and pre-assimilation CRPS from 0.3753 to 0.3510. The parameter-matched wide-local model obtains 0.5905 and 0.3729, respectively, indicating that increased capacity alone does not reproduce the same improvement. On P19, cross-variable assimilation improves before-first CRPS relative to both local and wide-local controls, although pre-assimilation performance is mixed. Dense-gain gives the lowest CRPS in all four core dataset–endpoint comparisons, suggesting that access to cross-variable information is more important than nonlinear update complexity itself. Detailed five-fold results are reported in Appendix B.7.

The cross-variable before-first benefit is also reproduced on patient-grouped MIMIC-III, improving CRPS relative to the local-coordinate model in all five folds with patient-level bootstrap intervals excluding zero (Appendix B.11).

4.4 Cross-variable information under reduced target availability

To test this hypothesis directly, we progressively suppress direct target observations on patient-grouped MIMIC-III and P19 while keeping the trained models fixed. Figure 3 shows the AUROC differences relative to local-coordinate on both datasets; the full three-arm AUROC and AUPRC results are reported in Appendix B.12.

On MIMIC-III, the nonlinear cross-variable arm becomes favorable relative to local-coordinate in AUROC as direct target information is removed, with the bootstrap interval excluding zero from 65% suppression onward. Dense-gain shows a more consistent effect: its AUROC advantage over local-coordinate excludes zero from 30% suppression onward. On P19, dense-gain outperforms both local-coordinate and nonlinear cross-variable updates in AUROC at every suppression level, whereas the nonlinear cross-variable advantage over local-coordinate is not reliable. AUPRC provides complementary evidence in the same direction (Appendix B.12). These results indicate that event-time cross-variable assimilation, rather than nonlinear update complexity, is the more consistent mechanism. Conversely, progressively permuting non-target variables across patients erodes and then reverses both cross-variable arms' advantage on all three datasets while the local variant degrades least (Appendix B.20), bracketing the mechanism together with the suppression sweep. Target-dropout further improves all six P12/P19 belief-quality comparisons and improves patient-grouped eICU AUROC/AUPRC from $0.8844/0.5791$ to $0.8893/0.5903$ (Appendix B.13).

Longer P19 histories improve performance, and the nonlinear cross-variable arm degrades more slowly than local-coordinate under strong value-dependent missingness; full truncated-window, MNAR, MCAR, sensor-dropout, and subsystem-holdout results, PAMAP2 before-first CRPS, and zero-shot cross-hospital transfer appear in Appendices B.8, B.18, B.14, B.15, and B.9.

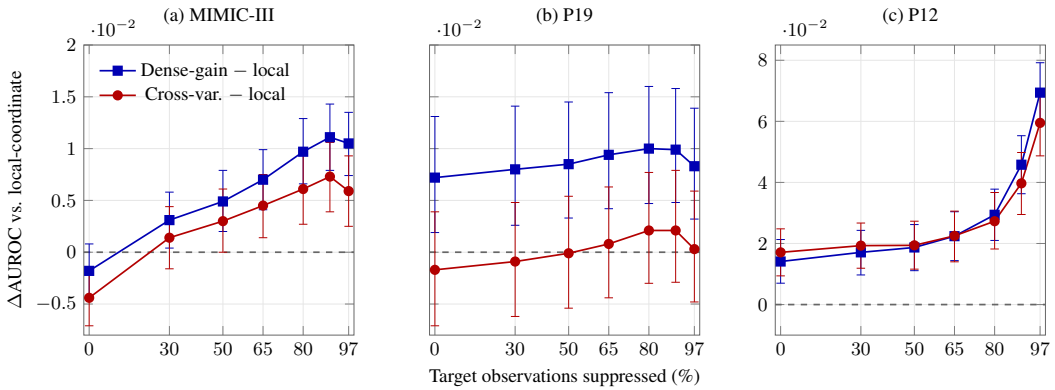


Figure 3: Target-observation suppression on patient-grouped MIMIC-III, P19, and P12. Curves show AUROC relative to the local-coordinate control; positive values favor the corresponding cross-variable assimilation variant over the local-coordinate control. Error bars are patient-level bootstrap 95% CIs. Panel (c) uses a wider y-range, as P12 effects are several times larger.

4.5 Cross-Variable Predictive Dependencies

The learned update matrices are asymmetric and capture event-conditioned predictive rather than causal relationships; the learned matrices characterize predictive information transfer for the filtering objective and are not causal or physiological estimates. Several recovered dependencies match established clinical relationships; full MIMIC-III and P12 update matrices and correlation references are provided in Appendix B.22.

4.6 Additional Results

Exact-OU corrects the long-gap variance inflation of the additive formulation (pooled five-fold $\Delta\text{CRPS} = -0.0672$, 95% CI $[-0.0873, -0.0499]$; Appendix B.16). The frozen-marginal rank-8 model improves joint NLL by 13.8% on P12 and 11.2% on P19 without changing marginal predictions (Appendix B.17). Parameter-matched, update-mechanism, propagation, and latent-dimension controls show that the observed effects are not explained simply by additional model capacity (Appendices B.4, B.5, B.6, and B.7).

5 Conclusion, Broader Impact, Limitations, Future Work

We introduced Neural Belief-State Filtering (NBSF), a continuous-discrete framework for irregular multivariate time series that combines continuous-discrete, gap-aware propagation with event-time cross-variable assimilation, and achieves competitive predictive performance across four clinical benchmark tasks under a matched protocol. Across belief-quality and target-suppression experiments, event-time cross-variable assimilation is more consistently useful than nonlinear update complexity: dense-gain is often strongest, while the nonlinear cross-variable update shows dataset-dependent benefits, with the clearest gains when direct target information is sparse or unavailable. NBSF uses a factorized Gaussian state and fixed-step mean propagation, and its learned cross-variable dependencies are predictive rather than causal.

AI Use Statement

Generative AI tools were used to assist with manuscript writing, language editing, formatting, and research-workflow support. The authors retained responsibility for the model design, experimental decisions, verification of data processing and results, interpretation of the analyses, and the final content of the manuscript. All AI-assisted material was reviewed and verified by the authors before inclusion.

Ethics Statement

This work uses only previously collected, de-identified clinical datasets that are publicly available to credentialed researchers (PhysioNet 2012 and 2019, MIMIC-III, and eICU), together with the PAMAP2 activity-recognition dataset. All datasets were accessed under their respective data use agreements; no new patient data were collected and no interventions were performed. The datasets are de-identified by their providers, and we did not attempt re-identification. NBSF is a research prototype for modeling irregular clinical time series; it is not validated for clinical use, and its predictions should not guide patient care without prospective evaluation and regulatory approval. We report mixed and negative results alongside positive ones to avoid overstating clinical applicability, and the learned cross-variable dependencies are predictive rather than causal or physiological.

Reproducibility Statement

An anonymized implementation of NBSF, including training and evaluation scripts for all reported experiments, is available at `https://anonymous.4open.science/r/NBSF-B820/`. Model, optimization, and data-processing details are provided in Appendices A.3 and A.2.

REFERENCES

Abdul Fatir Ansari, Alvin Heng, Andre Lim, and Harold Soh. Neural continuous-discrete state space models for irregularly-sampled time series. In *International Conference on Machine Learning (ICML)*, 2023.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018. doi: 10.1038/s41598-018-24271-9.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, volume 29, pp. 3504–3512, 2016.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Joseph Futoma, Sanjay Hariharan, and Katherine Heller. Learning to detect sepsis with a multitask Gaussian process RNN classifier. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1174–1182. PMLR, 2017. URL https://proceedings.mlr.press/v70/futoma17a.html.

Tian Guo, Tao Lin, and Nino Antulov-Fantulin. Exploring interpretable lstm neural networks over multi-variable data. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2494–2504, 2019.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016. doi: 10.1038/sdata.2016.35.

Simon J. Julier and Jeffrey K. Uhlmann. A new extension of the Kalman filter to nonlinear systems. In *Signal Processing, Sensor Fusion, and Target Recognition VI*, volume 3068, pp. 182–193. SPIE, 1997. doi: 10.1117/12.280797.

Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960. doi: 10.1115/1.3662552.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Rahul G. Krishnan, Uri Shalit, and David Sontag. Deep Kalman filters. *arXiv preprint arXiv:1511.05121*, 2015. URL https://arxiv.org/abs/1511.05121.

Rahul G. Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. doi: 10.1609/aaai.v31i1.10779.

Hao Li, Bowen Deng, Chang Xu, Zhiyuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, and Jiang Bian. MIRA: Medical time series foundation model for real-world health data. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3296.

Yanke Li, Ruirui Wang, Manuel Günther, and Diego Paez-Granados. GARLIC: Graph attention-based relational learning of multivariate time series in intensive care. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4ZAwmIaA9y.

Zachary C. Lipton, David C. Kale, and Randall Wetzel. Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, pp. 253–270. PMLR, 2016. URL https://proceedings.mlr.press/v56/Lipton16.html.

Manxia Liu, Fabio Stella, Arjen Hommersom, Peter J. F. Lucas, Lonneke Boer, and Erik Bischoff. A comparison between discrete and continuous time bayesian networks in learning from clinical time series data with irregularity. *Artificial Intelligence in Medicine*, 95:104–117, 2019. doi: 10.1016/j.artmed.2018.10.002.

Xueping Liu, Tianyi Gong, Youru Li, Na Li, and Silu Ding. Medical irregular multivariate time series forecasting based on multi-scale temporal-frequency domain patch fusion and dynamic graph. *Frontiers in Physiology*, 17:1767037, 2026a. doi: 10.3389/fphys.2026.1767037.

Xvyuan Liu, Xiangfei Qiu, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Jilin Hu, and Bin Yang. Rethinking irregular time series forecasting: A simple yet effective baseline. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23873–23881, 2026b. doi: 10.1609/aaai.v40i28.39563.

Byoungwoo Park, Hyungi Lee, and Juho Lee. Amortized control of continuous state space feynman-kac model for irregular time series. In *International Conference on Learning Representations (ICLR)*, 2025.

Tom Pollard, Alistair Johnson, Jesse Raffa, Leo Anthony Celi, Omar Badawi, and Roger Mark. eICU Collaborative Research Database. *PhysioNet*, April 2019. doi: 10.13026/C2WM1R. URL https://doi.org/10.13026/C2WM1R. Version 2.0.

Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, and Garrison W. Cottrell. A dual-stage attention-based recurrent neural network for time series prediction. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 2627–2633, 2017. doi: 10.24963/ijcai.2017/366.

Michael Ramati and Yuval Shahar. Irregular-time bayesian networks. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pp. 484–491, 2010.

Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *Proceedings of the 16th IEEE International Symposium on Wearable Computers*, pp. 108–109. IEEE, 2012. doi: 10.1109/ISWC.2012.13.

Matthew Reyna, Chris Josef, Russell Jeter, Supreeth Shashikumar, Benjamin Moody, M. Brandon Westover, Ashish Sharma, Shamim Nemati, and Gari D. Clifford. Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019. *PhysioNet*, August 2019. doi: 10.13026/v64v-d857. Version 1.0.0.

Yulia Rubanova, Ricky T. Q. Chen, and David K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Mona Schirmer, Mazin Eltayeb, Stefan Lessmann, and Maja Rudolph. Modeling irregular time series with continuous recurrent units. In *International Conference on Machine Learning (ICML)*, 2022.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=4c0J6lwQ4_.

Ikaro Silva, George Moody, Daniel J. Scott, Leo A. Celi, and Roger G. Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *Computing in Cardiology*, volume 39, pp. 245–248, 2012.

Dan Simon. *Optimal State Estimation: Kalman, H-Infinity, and Nonlinear Approaches*. Wiley, 2006.

Qingxiong Tan, Mang Ye, Baoyao Yang, Siqi Liu, Andy Jinhua Ma, Terry Cheuk-Fung Yip, Grace Lai-Hung Wong, and PongChi Yuen. DATA-GRU: Dual-attention time-aware gated recurrent unit for irregular multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 930–937, 2020. doi: 10.1609/aaai.v34i01.5440.

Sindhu Tipirneni and Chandan K. Reddy. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery from Data*, 16(6):1–17, 2022. doi: 10.1145/3516367.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 753–763, 2020. doi: 10.1145/3394486.3403118.

Jiawen Zhang, Shun Zheng, Wei Cao, Jiang Bian, and Jia Li. Warpformer: A multi-scale modeling approach for irregular clinical time series. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3273–3285, 2023. doi: 10.1145/3580305.3599543.

Weijia Zhang, Chenlong Yin, Hao Liu, Xiaofang Zhou, and Hui Xiong. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60179–60196. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhang24bw.html.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Kwm8I7dU-l5.

Liangwei Nathan Zheng, Zhengyang Li, Chang George Dong, Wei Emma Zhang, Lin Yue, Miao Xu, Olaf Maennel, and Weitong Chen. Irregularity-informed time series analysis: Adaptive modelling of spatial and temporal dynamics. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 3405–3414, 2024. doi: 10.1145/3627673.3679716.

A Implementation Details

A.1 Unified Variable Space

Table 4 lists the canonical 36-variable clinical space used for PhysioNet 2012. Other clinical datasets use dataset-specific subsets of the same general variable families.

Table 4: **Canonical clinical variable groups for PhysioNet 2012.**

Category	Count	Variables
Vitals	8	HR, SBP, DBP, MAP, RR, SpO_2, Temp, FiO_2
Metabolic	8	Glucose, Lactate, BUN, Creatinine, Potassium, Sodium, Chloride, Bicarbonate
Hematology	4	Hemoglobin, Hematocrit, Platelets, WBC
Blood Gas	3	pH, $PaCO_2$, PaO_2
Liver/Cardiac	6	Bilirubin, Albumin, AST, ALT, TroponinI, TroponinT
Other	5	Calcium, Magnesium, Phosphate, INR, PTT
Neuro/Vent	2	GCS, MechVent

A.2 Preprocessing and Evaluation Protocol

For every fold, preprocessing statistics are fit only on the training split. Continuous variables are z-scored using the training-set mean and standard deviation. Variables with skewness greater than 2 are log-transformed before standardization. Missing values are represented by observation masks; NBSF does not require forward filling, mean imputation, or linear interpolation.

PhysioNet 2012 (P12) and PhysioNet 2019 (P19) use five stratified train/validation/test folds. MIMIC-III and eICU use patient-grouped folds so that encounters belonging to the same patient cannot appear in both training and test data. We additionally evaluate eICU with hospital-grouped folds as a site-level robustness check. PAMAP2 is evaluated with eight-subject leave-one-subject-out (LOSO) splits for the controlled missing-information experiments. The fold-specific standardizer is reused for validation and test examples. Table 5 summarizes the resulting dataset sizes, temporal gaps, and empirical observation coverage.

Table 5: **Dataset regimes used in the final experiments.** $C(\rho)$ denotes the empirical fraction of variables with coverage at least ρ under the dataset-specific coverage calculation. PAMAP2 is naturally fully observed and is used as a controlled missing-information probe.

Dataset	N	D	Median gap (h)	Gap IQR (h)	$C(0.2)$	$C(0.4)$	Full coverage
P12	11,988	36	1.000	0.600	0.5914	0.6676	0.7173
P19	39,104	34	1.000	0.000	0.4651	0.5581	0.6464
MIMIC-III	49,162	21	1.000	0.250	0.8088	0.9209	0.9584
eICU	195,495	27	0.083	0.000	0.3580	0.4647	0.6107
PAMAP2	1,208	31	1.000	0.000	1.0000	1.0000	1.0000

A.3 Architecture and Hyperparameters

Table 6 summarizes the default NBSF architecture and training configuration.

The state dimension is $S = D + 16$. Typical model sizes are approximately 170K parameters for P12/P19, 140K for MIMIC-III, and 150K for eICU.

Additional hyperparameters. The total objective is the unweighted sum of the forecasting and label losses ($\mathcal{L} = \mathcal{L}_{\text{forecast}} + \mathcal{L}_{\text{label}}$, i.e. $\lambda = 1$ rather than a tuned coefficient), and the label loss is standard binary cross-entropy. The gain is clamped with $K_{\max} = 0.95$, and the observation-noise floor is $\epsilon = 0.01$, with a separate posterior-variance floor of 10^{-4} applied after each update. The observation projection W_{obs} is learned (a linear map from the state dimension to the observed dimension), not a fixed selection matrix. The belief is initialized at $m_0 = \mathbf{0}$ and $v_0 = 1$, both

Table 6: **Default NBSF architecture and training hyperparameters.**

Component	Value
Observed variables	Dataset-specific D
Latent dimensions	16
Belief dimension	$S = D + 16$
Hidden width	128
Mean propagation	SSM decay + neural residual
Variance propagation	Exact coordinate-wise OU transition
Observation update	Neural cross-variable update
Prediction head input	$[m_N^+; \ell_N^+]$
Optimizer	Adam
Learning rate	2×10^{-4}
Batch size	512
Maximum epochs	200
Early stopping patience	25 epochs
Log-variance clamp	$[-4, 4]$; tuned runs use $[-3, 3]$

learnable so that training can move the initial belief. The residual networks f_θ and h_ψ and the cross-variable update network U_ψ are each two-hidden-layer MLPs of width 128 with ReLU activations. Target-dropout samples a per-instance drop fraction $\text{drop_frac} \sim \text{Unif}(0.1, 0.5)$.

Wide-local. The wide-local control keeps the coordinate-local observation update (no cross-variable assimilation) but widens the observation-update hidden dimension from 128 to a dataset-specific value chosen to match the cross-variable arm's parameter count at that dataset's state dimension: 367 on P12 (169,918 parameters versus 169,991 for cross-variable) and 370 on P19 (165,351 versus 165,449). The propagation network is unchanged; only the observation-update MLP is widened.

A.4 Training Stability

An earlier P19 configuration used a learning rate of 5×10^{-4} and a log-variance clamp of $[-6, 6]$. Although it reached a low validation NLL, the run later became numerically unstable. The final training pipeline uses a tighter log-variance clamp ($[-4, 4]$ for standard runs and $[-3, 3]$ for tuned runs), NaN guards, gradient clipping, and best/last-safe checkpoints. These changes were used for the final experiments reported in the paper.

A.5 Computational Efficiency

Table 7 reports an implementation-level timing comparison between sequential and batched NBSF code on a single NVIDIA RTX PRO 6000 Blackwell GPU. These measurements compare two implementations of the same model and should not be interpreted as a cross-model hardware benchmark.

Table 7: **Implementation timing for sequential and batched NBSF.**

Implementation	Time/epoch	Total (150 ep.)	GPU memory
Sequential	$\sim$26 min	$\sim$65 h	< 1 GB
Batched	$\sim$25 sec	$\sim$1 h	< 1 GB
Observed speedup		$\sim 60\times$	

Table 8 reports NBSF uses lower peak memory than every baseline in this benchmark and fewer parameters than the graph-attention (GARLIC) and transformer (Warpformer) models, while remaining the only model evaluated at eICU scale (195K patients); Warpformer exceeds memory on MIMIC-III.

Table 8: **Model size, memory, and scaling.** NBSF measured in this work; baseline parameter counts and peak GPU memory are from the efficiency benchmark of Li et al. (2026) (their Table 3 / Appendix A.2.2) under matched batch size and hyperparameters. "OOM": out of memory; "—": not run in that benchmark.

	NBSF	GARLIC	Warpformer	mTAND
Parameters	134–170K	∼500K	∼1M	∼200K
Peak GPU memory	<1.5 GB	3.3 GB	8.6 GB	2.9 GB
MIMIC-III (49K patients)	✓	✓	OOM	✓
eICU (195K patients)	✓	—	—	—

B Additional Experimental Results

B.1 Complete Benchmark Comparison

Table 9 gives the complete benchmark comparison for P12 mortality, P19 sepsis, and MIMIC-III mortality. Baseline results are reproduced from the controlled benchmark of Li et al. (2026); NBSF is evaluated using the same dataset construction, preprocessing, train/validation/test splits, and task definitions. The comparison therefore follows the same task definitions and evaluation protocol used in the main paper. For MIMIC-III, the final NBSF value is the patient-grouped result; the earlier non-grouped value is not used.

The benchmark results show strong downstream performance but do not motivate a claim of uniform superiority. NBSF is strongest on P19 in this comparison and has the highest MIMIC-III AUPRC, while GARLIC has the highest P12 mortality metrics and MIMIC-III AUROC. The main mechanistic claims therefore rely on the belief-quality and controlled information-removal experiments below rather than on classification ranking alone.

B.2 P12 Length-of-Stay >3 Days

The P12 LOS>3d task is retained as a task-matched comparison with the Raindrop benchmark. Table 10 reports the three NBSF update variants from the final five-fold evaluation. The cross-variable arm is the default NBSF architecture used in the main comparison; the dense-gain arm is numerically strongest on this task.

The cross-variable arm has the smallest fold-to-fold standard deviation, but we do not interpret this post-hoc stability difference as a primary advantage.

B.3 Final eICU Evaluation

All final eICU experiments use grouped splits. Table 11 shows the patient-grouped comparison of the three update mechanisms.

Unlike MIMIC-III, eICU favors the cross-variable arm on AUROC, AUPRC, and Brier score. Dense-gain has the lowest ECE. This dataset-specific difference is reported without averaging the two cohorts into a single directional claim.

A hospital-grouped evaluation of the cross-variable model gives AUROC 0.8804 ± 0.0036, AUPRC 0.5565 ± 0.0222, Brier 0.0552 ± 0.0009, and ECE 0.0086 ± 0.0028. Thus the patient-grouped result is not dependent on a single grouping strategy. Target-dropout training further improves the patient-grouped cross-variable model to AUROC 0.8893 ± 0.0033, AUPRC 0.5903 ± 0.0075, Brier 0.0533 ± 0.0006, and ECE 0.0055 ± 0.0015.

B.4 Update-Mechanism Ablation

We compare local-coordinate, dense-gain, and cross-variable observation updates across five matched splits on P12 and P19. Table 12 reports the corresponding classification and calibration results.

Table 9: **Classification results on P12 mortality, P19 sepsis, and MIMIC-III** (AUROC/AUPRC in %, mean ± standard deviation). Baseline results follow the controlled benchmark of Li et al. (2026); NBSF uses the same evaluation protocol. Best results are shown in **bold** and second-best results are underlined. "—" denotes not reported.

	P12 Mortality		P19 Sepsis		MIMIC-III Mortality	
Method	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
RNN-based						
RNN-Δ_t	82.46±1.57	47.29±4.33	89.21±0.63	51.20±2.50	82.98±2.27	50.31±4.22
RNN-Mean	82.38±2.11	47.80±3.98	88.08±0.88	45.80±2.10	84.18±1.70	50.06±5.17
RNN-Decay	84.26±1.10	49.81±3.31	89.15±1.15	49.73±3.46	87.63±0.79	56.29±3.29
GRU-D (Che et al., 2018)	84.45±1.90	50.74±4.96	88.73±0.86	47.81±3.37	86.36±1.68	56.14±1.87
ODE-based						
ODE-RNN (Rubanova et al., 2019)	83.02±1.72	48.64±2.68	89.97±0.86	54.29±2.76	88.07±0.77	61.03±2.00
L-ODE-RNN (Chen et al., 2018)	80.59±2.06	41.40±4.15	89.63±1.31	51.02±3.77	86.98±0.58	56.88±2.20
L-ODE-ODE (Rubanova et al., 2019)	83.93±1.14	49.55±1.43	90.01±2.04	53.54±6.12	87.36±0.94	59.09±1.93
Attention / Graph / Transformer						
MTGNN (Wu et al., 2020)	80.95±1.71	43.61±4.93	86.94±1.65	40.97±5.26	86.24±0.67	51.74±0.92
mTAND (Shukla & Marlin, 2021)	84.30±1.69	50.05±3.96	81.73±1.53	37.27±3.93	88.00±0.31	57.73±2.21
Raindrop (Zhang et al., 2022)	83.03±1.37	45.91±3.55	87.41±1.13	46.33±3.09	87.18±1.51	57.06±6.01
Warpformer (Zhang et al., 2023)	84.58±0.48	50.62±1.25	89.95±0.57	54.10±2.38	89.17±0.57	61.52±1.93
MTSFormer (Zheng et al., 2024)	83.65±1.70	50.31±4.30	87.88±0.61	48.80±2.97	88.14±0.50	61.09±0.95
Interpretable / Attention-based						
IMV-LSTM (Guo et al., 2019)	84.02±1.30	49.08±3.28	84.80±2.76	42.87±8.27	86.85±1.33	54.96±3.61
DA-RNN (Qin et al., 2017)	79.84±1.36	42.70±2.09	74.36±3.50	20.89±2.73	82.56±1.25	46.33±4.08
RETAIN (Choi et al., 2016)	83.08±1.11	49.27±3.01	78.09±2.22	26.04±1.35	82.40±0.94	46.55±1.96
GARLIC (Li et al., 2026)	**86.40±0.86**	**56.89±1.75**	90.96±0.84	55.29±2.45	**90.09±0.45**	64.85±1.68
NBSF (cross-variable)	84.67±0.90	51.60±2.83	**92.11±0.77**	**61.17±2.35**	89.95±0.53	**65.82±2.40**

Table 10: **P12 LOS>3d, NBSF update variants** (five folds).

Update	**AUROC**	**AUPRC**
Local-coordinate	0.8101 ± 0.0206	0.9771 ± 0.0061
Dense-gain	$\mathbf{0.8153 \pm 0.0298}$	$\mathbf{0.9789 \pm 0.0059}$
Cross-variable	0.7936 ± 0.0125	0.9775 ± 0.0018

Table 11: **eICU patient-grouped classification** (five folds). Lower is better for Brier and ECE.

Arm	**AUROC**	**AUPRC**	**Brier ↓**	**ECE ↓**
Local-coordinate	0.8829 ± 0.0033	0.5733 ± 0.0074	0.0546 ± 0.0006	0.0047 ± 0.0024
Dense-gain	0.8788 ± 0.0055	0.5661 ± 0.0121	0.0550 ± 0.0009	$\mathbf{0.0037 \pm 0.0011}$
Cross-variable	$\mathbf{0.8844 \pm 0.0033}$	$\mathbf{0.5791 \pm 0.0070}$	$\mathbf{0.0541 \pm 0.0005}$	0.0058 ± 0.0017

Table 12: **Update-mechanism ablation** over five matched splits (mean ± standard deviation). Higher AUROC/AUPRC and lower Brier/ECE are better.

Dataset	**Update**	**AUROC**	**AUPRC**	**Brier ↓**	**ECE ↓**
P12	Local-coordinate	82.96 ± 0.95	46.01 ± 1.90	0.1016 ± 0.0039	0.0457 ± 0.0095
	Cross-variable	$\mathbf{84.67 \pm 0.90}$	$\mathbf{51.60 \pm 2.83}$	$\mathbf{0.0940 \pm 0.0027}$	0.0297 ± 0.0045
	Dense-gain	83.12 ± 1.09	49.06 ± 2.12	0.0961 ± 0.0017	$\mathbf{0.0276 \pm 0.0032}$
P19	Local-coordinate	92.28 ± 1.63	63.54 ± 2.96	0.0299 ± 0.0014	0.0127 ± 0.0039
	Cross-variable	92.11 ± 0.77	$\mathbf{61.17 \pm 2.35}$	$\mathbf{0.0289 \pm 0.0015}$	$\mathbf{0.0116 \pm 0.0031}$
	Dense-gain	$\mathbf{93.00 \pm 0.54}$	62.39 ± 3.04	0.0285 ± 0.0022	0.0127 ± 0.0032

The cross-variable update gives the strongest P12 AUROC/AUPRC. On P19, dense-gain gives the highest AUROC and lowest Brier score, while local-coordinate gives the highest AUPRC and the nonlinear cross-variable update gives the lowest ECE. Thus, no update mechanism is uniformly best across datasets and metrics, motivating the belief-quality and controlled information-removal analyses below.

B.5 Propagation-Mechanism Ablation

We compare the hybrid propagation model with MLP-only and SSM-only alternatives under identical splits and training settings. The five-fold results are summarized in Table 13.

Table 13: **Propagation-mechanism ablation** over five matched splits (mean ± standard deviation).

Dataset	Propagation	AUROC	AUPRC
P12	Hybrid	$\mathbf{84.67 \pm 0.90}$	51.60 ± 2.83
	MLP	83.85 ± 1.39	49.74 ± 2.92
	SSM	84.65 ± 0.85	$\mathbf{51.79 \pm 2.45}$
P19	Hybrid	92.11 ± 0.77	61.17 ± 2.35
	MLP	92.24 ± 0.62	61.88 ± 1.64
	SSM	$\mathbf{93.12 \pm 0.58}$	$\mathbf{63.17 \pm 2.35}$

No propagation mechanism consistently dominates across both datasets and metrics. We retain the hybrid formulation because it combines structured continuous-time decay with nonlinear residual dynamics, rather than because it is uniformly superior in downstream classification.

B.6 Latent-Dimension Ablation

To test whether downstream performance is driven primarily by additional latent capacity, we remove the latent block ($L = 0$) and retrain the model across the same five P19 folds. The reduced model contains 129,961 parameters. Table 14 compares the reduced and default cross-variable models.

Table 14: **Latent-dimension ablation on P19** (mean ± standard deviation across five matched folds).

Model	AUROC	AUPRC	Brier ↓	ECE ↓
Cross-variable NBSF ($L = 16$)	0.9211 ± 0.0077	0.6117 ± 0.0235	0.0289 ± 0.0015	$\mathbf{0.0116 \pm 0.0031}$
$L = 0$	$\mathbf{0.9249 \pm 0.0074}$	$\mathbf{0.6199 \pm 0.0340}$	$\mathbf{0.0284 \pm 0.0019}$	0.0125 ± 0.0023

Removing the latent block does not reduce downstream classification performance on P19; AUROC increases slightly from 0.9211 to 0.9249, with similar behavior for AUPRC and Brier score. This result indicates that downstream classification performance is not explained simply by additional latent-state capacity. We therefore treat $L = 16$ as a modeling choice rather than as a necessary source of predictive performance.

B.7 Core Belief Quality

We evaluate belief quality before the first direct target observation (BF) and immediately before assimilation of an incoming observation (PA). Table 15 reports the five-fold results. CRPS is the primary metric because P19 NLL contains large fold-level outliers.

Dense-gain is numerically strongest on CRPS in all four dataset–endpoint cells. The parameter-matched wide-local comparison addresses a different question: cross-variable beats wide-local in three of four settings (P12 BF, P12 PA, and P19 BF) but is worse on P19 PA. This supports a capacity-control claim, not a claim that the nonlinear cross-variable arm is uniformly best.

B.8 Multi-Horizon Early Prediction

We evaluate the three update variants at increasing observation-window lengths (five folds, AUROC, mean±SD). This probes a different axis from the target-suppression experiment of Section 4: here

Table 15: **Core belief-quality results** (five-fold mean ± standard deviation). Lower is better.

Dataset	Model	BF CRPS	BF NLL	PA CRPS	PA NLL
P12	Local-coordinate	0.5956 ± 0.0099	1.9134 ± 0.4348	0.3753 ± 0.0103	1.2283 ± 0.1342
	Wide-local	0.5905 ± 0.0066	1.9103 ± 0.4988	0.3729 ± 0.0078	1.2165 ± 0.1376
	Dense-gain	$\mathbf{0.5698 \pm 0.0047}$	2.5007 ± 0.6101	$\mathbf{0.3497 \pm 0.0026}$	1.1693 ± 0.1350
	Cross-variable	0.5707 ± 0.0039	2.2788 ± 0.4679	0.3510 ± 0.0027	$\mathbf{1.1447 \pm 0.1257}$
P19	Local-coordinate	0.5575 ± 0.0586	223.82 ± 497.34	0.3430 ± 0.0060	21.45 ± 45.79
	Wide-local	0.5585 ± 0.0589	231.77 ± 515.12	0.3428 ± 0.0079	22.18 ± 47.43
	Dense-gain	$\mathbf{0.5501 \pm 0.0592}$	$\mathbf{135.64 \pm 300.18}$	$\mathbf{0.3348 \pm 0.0061}$	$\mathbf{13.31 \pm 27.64}$
	Cross-variable	0.5549 ± 0.0595	232.71 ± 517.21	0.3465 ± 0.0065	22.27 ± 47.63

the window is shortened but the target variable remains observed, whereas suppression keeps the full window and removes the target. The cross-variable and dense-gain arms improve over local-coordinate on P12 at every horizon, but on P19 and MIMIC-III the local-coordinate arm is best (or tied) at the early horizons and the cross-variable advantage does not appear. This is consistent with, and sharpens, our main finding: the benefit of cross-variable assimilation is specific to *target scarcity* (the target unobserved), not to reduced data in general (a shorter window with the target still observed).

Table 16: **P12 mortality, multi-horizon AUROC** (five folds).

Horizon	Local-coordinate	Dense-gain	Cross-variable
6h	0.6585±0.0193	0.6824±0.0168	**0.6870±0.0302**
12h	0.7365±0.0121	0.7476±0.0133	**0.7481±0.0197**
24h	0.7914±0.0115	0.8033±0.0096	**0.8052±0.0139**
48h/full	0.8296±0.0095	0.8438±0.0095	**0.8467±0.0090**

Table 17: **P19 sepsis, multi-horizon AUROC** (five folds).

Horizon	Local-coordinate	Dense-gain	Cross-variable
6h	**0.6315±0.0130**	0.6284±0.0119	0.6241±0.0074
12h	**0.7550±0.0078**	0.7468±0.0194	0.7417±0.0163
24h	**0.8523±0.0057**	0.8488±0.0065	0.8459±0.0061
48h	0.9145±0.0166	**0.9212±0.0078**	0.9139±0.0128
full	0.9228±0.0146	**0.9300±0.0048**	0.9211±0.0077

Table 18: **MIMIC-III mortality, multi-horizon AUROC** (five folds).

Horizon	Local-coordinate	Dense-gain	Cross-variable
6h	**0.7624±0.0062**	0.7465±0.0051	0.7361±0.0118
12h	**0.8176±0.0059**	0.8064±0.0074	0.7935±0.0122
24h	**0.8619±0.0050**	0.8544±0.0048	0.8524±0.0044
48h/full	**0.9040±0.0058**	0.9021±0.0047	0.8995±0.0047

B.9 Cross-Hospital Zero-Shot Transfer

We test whether the learned belief state transfers across hospitals without any target-domain training. A model trained on a source dataset is evaluated zero-shot on MIMIC-III under a leak-free, patient-grouped protocol; variables absent in the target are treated as permanently unobserved, so only the shared subset drives the belief. Despite large sensor mismatch (18 of 36 variables absent for P12→MIMIC-III), the transferred model still produces non-trivial predictions (Table 19). We report this as a capability of the belief-state formulation, which maintains a distribution over all modeled variables and can be queried even when many channels are never observed in the target domain.

Table 19: **Cross-hospital zero-shot transfer to MIMIC-III** (patient-grouped, leak-free).

Transfer	Shared vars	Missing sensors	AUROC	AUPRC
P12 $\rightarrow$ MIMIC-III	18/36	18	0.7795	0.3496
eICU $\rightarrow$ MIMIC-III	21/27	6	0.7813	0.4618

B.10 Persistence Baseline

We compare NBSF with a simple carry-forward persistence baseline on P19. For each target variable, the baseline uses its most recent available measurement as the prediction for the next observation. Evaluation is performed across the same five folds used for the NBSF belief-quality experiments. The five-fold pre-assimilation CRPS comparison is reported in Table 20.

Table 20: **P19 pre-assimilation belief quality compared with a carry-forward persistence baseline** (five-fold mean $\pm$ standard deviation). Lower CRPS is better.

Model	**Pre-assimilation CRPS $\downarrow$**
Persistence	0.3520 ± 0.0054
Local-coordinate NBSF	0.3430 ± 0.0060
Dense-gain NBSF	$\mathbf{0.3348 \pm 0.0061}$
Cross-variable NBSF	0.3465 ± 0.0065

All three learned NBSF update variants improve on the carry-forward persistence baseline. Dense-gain gives the lowest pre-assimilation CRPS, while both the local-coordinate and cross-variable models also outperform persistence. This provides a simple sanity check that the maintained belief contains predictive information beyond carrying forward the most recent measurement.

B.11 MIMIC-III Before-First Evaluation

Table 21 reports patient-paired bootstrap comparisons for all five grouped MIMIC-III folds. Negative values indicate lower CRPS for the first-named model.

Table 21: **MIMIC-III grouped before-first CRPS differences** with patient-level 95% bootstrap intervals.

Fold	**Cross–Local**	**Dense–Local**	**Dense–Cross**
1	-0.0032 $[-0.0036, -0.0027]$	-0.0036 $[-0.0040, -0.0031]$	-0.0004 $[-0.0008, +0.0000]$
2	-0.0028 $[-0.0033, -0.0024]$	-0.0040 $[-0.0044, -0.0035]$	-0.0011 $[-0.0015, -0.0007]$
3	-0.0034 $[-0.0039, -0.0029]$	-0.0035 $[-0.0040, -0.0030]$	-0.0001 $[-0.0005, +0.0003]$
4	-0.0025 $[-0.0029, -0.0020]$	-0.0001 $[-0.0006, +0.0004]$	$+0.0023$ $[+0.0018, +0.0028]$
5	-0.0012 $[-0.0017, -0.0008]$	$+0.0021$ $[+0.0016, +0.0027]$	$+0.0034$ $[+0.0029, +0.0038]$

Cross-variable improves over local-coordinate in all five folds with bootstrap intervals excluding zero. Dense-vs-local and dense-vs-cross change direction across folds and are therefore not treated as consistent effects.

B.12 Controlled Target-Observation Suppression

We progressively suppress direct target observations at test time while keeping the trained local-coordinate, dense-gain, and nonlinear cross-variable models fixed. The same suppression realizations are used across update mechanisms. Table 22 reports AUROC together with patient-level paired-bootstrap 95% confidence intervals for all three pairwise comparisons, while Table 23 reports the corresponding AUPRC values.

On MIMIC-III, dense-gain becomes significantly better than local-coordinate in AUROC from 30% suppression onward, while the nonlinear cross-variable arm becomes significantly better than

Table 22: **Controlled target-observation suppression on MIMIC-III and P19.** $\Delta_{\mathrm{C-L}}$, $\Delta_{\mathrm{D-L}}$, and $\Delta_{\mathrm{C-D}}$ denote Cross−Local, Dense−Local, and Cross−Dense AUROC differences, respectively. Brackets give patient-level paired-bootstrap 95% confidence intervals.

Dataset	Supp.	Local	Dense	Cross	$\Delta_{\mathrm{C-L}}$ [95% CI]	$\Delta_{\mathrm{D-L}}$ [95% CI]	$\Delta_{\mathrm{C-D}}$ [95% CI]
MIMIC-III	0%	0.9040	0.9021	0.8995	-0.0044 [-0.0071, -0.0017]	-0.0018 [-0.0042, $+0.0008$]	-0.0026 [-0.0051, -0.0001]
MIMIC-III	30%	0.8974	0.9005	0.8988	$+0.0014$ [-0.0016, $+0.0044$]	$+0.0031$ [$+0.0004$, $+0.0058$]	-0.0017 [-0.0042, $+0.0008$]
MIMIC-III	50%	0.8945	0.8994	0.8976	$+0.0030$ [-0.0000, $+0.0061$]	$+0.0049$ [$+0.0020$, $+0.0079$]	-0.0019 [-0.0043, $+0.0005$]
MIMIC-III	65%	0.8908	0.8978	0.8953	$+0.0045$ [$+0.0014$, $+0.0075$]	$+0.0070$ [$+0.0041$, $+0.0099$]	-0.0025 [-0.0050, -0.0000]
MIMIC-III	80%	0.8864	0.8961	0.8925	$+0.0061$ [$+0.0027$, $+0.0094$]	$+0.0097$ [$+0.0066$, $+0.0129$]	-0.0036 [-0.0061, -0.0011]
MIMIC-III	90%	0.8832	0.8944	0.8905	$+0.0073$ [$+0.0039$, $+0.0106$]	$+0.0111$ [$+0.0079$, $+0.0143$]	-0.0038 [-0.0065, -0.0012]
MIMIC-III	97%	0.8824	0.8929	0.8883	$+0.0059$ [$+0.0025$, $+0.0093$]	$+0.0105$ [$+0.0074$, $+0.0135$]	-0.0046 [-0.0072, -0.0020]
P19	0%	0.9228	0.9300	0.9211	-0.0017 [-0.0071, $+0.0039$]	$+0.0072$ [$+0.0019$, $+0.0131$]	-0.0089 [-0.0134, -0.0045]
P19	30%	0.9212	0.9292	0.9203	-0.0009 [-0.0062, $+0.0048$]	$+0.0080$ [$+0.0026$, $+0.0141$]	-0.0089 [-0.0131, -0.0048]
P19	50%	0.9204	0.9289	0.9204	-0.0001 [-0.0054, $+0.0054$]	$+0.0085$ [$+0.0033$, $+0.0145$]	-0.0085 [-0.0125, -0.0045]
P19	65%	0.9195	0.9289	0.9202	$+0.0008$ [-0.0044, $+0.0063$]	$+0.0094$ [$+0.0042$, $+0.0154$]	-0.0086 [-0.0126, -0.0047]
P19	80%	0.9182	0.9282	0.9203	$+0.0021$ [-0.0030, $+0.0077$]	$+0.0100$ [$+0.0047$, $+0.0160$]	-0.0079 [-0.0119, -0.0041]
P19	90%	0.9171	0.9270	0.9192	$+0.0021$ [-0.0029, $+0.0079$]	$+0.0099$ [$+0.0048$, $+0.0158$]	-0.0078 [-0.0117, -0.0038]
P19	97%	0.9176	0.9259	0.9179	$+0.0003$ [-0.0048, $+0.0059$]	$+0.0083$ [$+0.0032$, $+0.0139$]	-0.0080 [-0.0120, -0.0041]

Table 23: **AUPRC under controlled target-observation suppression.**

Dataset	Supp.	Local	Dense	Cross
MIMIC-III	0%	0.6670	0.6659	0.6582
	30%	0.6556	0.6629	0.6569
	50%	0.6498	0.6614	0.6550
	65%	0.6469	0.6589	0.6517
	80%	0.6413	0.6567	0.6474
	90%	0.6366	0.6550	0.6448
	97%	0.6334	0.6522	0.6418
P19	0%	0.6354	0.6239	0.6117
	30%	0.6316	0.6235	0.6108
	50%	0.6284	0.6240	0.6124
	65%	0.6246	0.6242	0.6118
	80%	0.6199	0.6240	0.6082
	90%	0.6164	0.6219	0.6038
	97%	0.6152	0.6170	0.6006

local-coordinate from 65% onward. Dense-gain remains numerically above the nonlinear arm at every suppression level, with the gap widening under severe suppression. On P19, dense-gain is significantly better than both local-coordinate and nonlinear cross-variable updates in AUROC at every suppression level, whereas the Cross–Local interval contains zero throughout.

AUPRC provides complementary evidence. On MIMIC-III, both cross-variable mechanisms become increasingly favorable relative to local-coordinate as target observations are removed, with dense-gain remaining above the nonlinear cross-variable arm at every suppression level. On P19, dense-gain remains above the nonlinear cross-variable arm throughout and closes its initial AUPRC deficit relative to local-coordinate, becoming numerically higher from 80% suppression onward; the nonlinear cross-variable arm remains below local-coordinate at every suppression level. Together, these results support event-time cross-variable assimilation, realized most consistently here by dense-gain, rather than uniform superiority of the nonlinear cross-variable update.

B.13 Target-Dropout Training

Target-dropout withholds a subset of available target observations from the assimilation mask during training while retaining them for the forecasting objective. Table 24 reports the six P12/P19 five-fold comparisons. We report the mean CRPS difference and fold agreement; we do not treat the arithmetic mean of fold-specific confidence limits as a pooled confidence interval.

Table 24: **Target-dropout belief-quality results.** Positive ΔCRPS denotes lower CRPS for target-dropout than for the plain cross-variable model.

Dataset	Endpoint	Cross	Target-dropout	ΔCRPS / folds
P12	Before-first	0.5707	0.5644	$+0.0061$ / 5 of 5
P12	Pre-assimilation	0.3510	0.3421	$+0.0081$ / 5 of 5
P12	Subsystem holdout	0.5769	0.5502	$+0.0217$ / 5 of 5
P19	Before-first	0.5549	0.5519	$+0.0030$ / 5 of 5
P19	Pre-assimilation	0.3465	0.3426	$+0.0027$ / 5 of 5
P19	Subsystem holdout	0.6574	0.5874	$+0.0574$ / 5 of 5

For P12 subsystem holdout, four of five fold-specific bootstrap intervals exclude zero even though the raw CRPS direction is favorable in all five folds. For the other rows, the main reproducibility statement is the consistent raw fold direction. Because $n = 5$ folds, a two-sided Wilcoxon signed-rank test cannot attain a value below 0.0625 when all five differences have the same sign; we therefore avoid using that test as evidence of significance.

The eICU result provides an additional classification-level check: target-dropout improves the patient-grouped cross-variable model from AUROC 0.8844 to 0.8893 and from AUPRC 0.5791 to 0.5903.

B.14 Subsystem Holdout

Whole physiological groups are withheld to test reconstruction from the remaining variables. Table 25 gives the five-fold mean CRPS.

Cross-variable is best in 3 of 10 dataset–group cells and dense-gain is best in the remaining 7. The result supports the importance of access to other variables under subsystem removal, but not a claim that the nonlinear cross-variable update is uniformly superior to a dense linear gain.

B.15 Controlled PAMAP2 Probe

PAMAP2 is naturally densely observed in the processed data used here. We therefore use it as a controlled missing-information experiment rather than as a naturally irregular benchmark or a direct reproduction of previously reported PAMAP2 classification numbers. Eight-subject LOSO evaluation is used throughout this subsection. Table 26 reports overall predictive NLL under controlled observation removal, while Table 27 evaluates sensor-group holdout.

Before the first direct observation of a target, cross-variable assimilation reduces mean CRPS from 0.5477 to 0.5411 at missingness 0.1 and from 0.5382 to 0.5204 at missingness 0.4. The direction is

Table 25: **Subsystem-holdout CRPS.** Lower is better. Bold marks the best of the three NBSF update mechanisms in each row.

Dataset	Group	Local	Dense	Cross
P12	Renal	0.9679	0.5825	**0.5663**
P12	Hemodynamic	0.9392	**0.5629**	0.5913
P12	Respiratory	0.6888	0.5761	**0.5309**
P12	Metabolic	0.5622	0.5294	**0.5245**
P12	Liver	0.5799	**0.5439**	0.5470
P19	Renal	0.5243	**0.4933**	0.5117
P19	Hemodynamic	0.9072	**0.6367**	0.7310
P19	Respiratory	0.9539	**0.5083**	0.5558
P19	Metabolic	0.5311	**0.5152**	0.5241
P19	Liver	0.5135	**0.4991**	0.5095

Table 26: **PAMAP2 LOSO test NLL** averaged over eight held-out subjects. Lower is better.

Missingness	Local	Wide-local	Cross	Target-dropout
0.1	0.2196	0.1709	**−0.0720**	-0.0604
0.4	0.2361	0.1993	**0.0549**	0.0593

favorable for all eight held-out subjects at both missingness levels (two-sided Wilcoxon $p = 0.0078$ in each setting). Across the 16 subject–missingness configurations, wide-local is never closer to cross-variable than to local-coordinate, supporting its use as a parameter-capacity control.

Table 27: **PAMAP2 sensor-group holdout CRPS** averaged over eight LOSO subjects. "Favorable" counts subjects for which cross-variable has lower CRPS than local-coordinate.

Missingness	Group	Local	Cross	Favorable
0.1	Hand	0.5489	0.5476	3/8
0.1	Chest	0.5044	0.4979	5/8
0.1	Ankle	0.5260	0.5137	5/8
0.4	Hand	0.5468	0.5233	7/8
0.4	Chest	0.5007	0.4649	7/8
0.4	Ankle	0.5247	0.4898	7/8

The sensor-group result is mixed at mild missingness and becomes more consistent at missingness 0.4. We therefore use PAMAP2 as controlled cross-domain support for the information-scarcity result, not as evidence of a universal sensor-holdout advantage.

B.16 Exact Ornstein–Uhlenbeck Variance Propagation

Table 28 compares the original baseline, the additive variance formulation (S1), and the exact coordinate-wise OU variance transition on P12. Values are pooled across five folds.

Exact-OU primarily corrects the long-gap variance inflation of S1. Relative to S1, the pooled five-fold CRPS difference is -0.0672 with a patient-level 95% bootstrap interval of $[-0.0873, -0.0499]$; all five fold differences are negative (-0.1006, -0.0276, -0.0176, -0.1431, and -0.0195). Exact-OU is not uniformly best against the original baseline: the baseline has lower CRPS in the 12–24h and $>$ 24h bins and has coverage close to the nominal 90% level there. The longest-gap bin is also comparatively small. We therefore interpret exact-OU as a correction to S1's variance behavior rather than as a universal improvement over every propagation variant in every gap regime.

B.17 FROZEN-MARGINAL JOINT UNCERTAINTY

The base filter uses diagonal marginal uncertainty. A rank-8 dependence head is therefore evaluated as a post-filter joint model while keeping the marginal means and variances fixed. Table 29 summarizes the five-fold joint-NLL improvement.

For the structured-noise base, the five-fold improvements are $13.50, 13.28, 16.05, 13.02, 13.13\%$ on P12 and $11.26, 11.05, 11.18, 10.92, 11.60\%$ on P19. The multivariate energy score also favors rank-8 in all 10 fold–dataset combinations. Because the marginal means and variances are fixed by construction, these improvements reflect better modeling of joint predictive dependence rather than changes in the marginal predictions.

B.18 ADDITIONAL MISSING-DATA ROBUSTNESS PROBES

The robustness analyses below are evaluated across five matched P19 folds. They include truncated observation windows, random removal of complete sensors, random removal of individual observations (MCAR), and controlled value-dependent missingness (MNAR). All perturbations are applied at test time while keeping the trained checkpoints fixed.

B.18.1 TRUNCATED OBSERVATION WINDOWS

We evaluate how predictive performance changes when only an initial portion of the P19 trajectory is available. Each trained checkpoint is evaluated using observations up to a fixed time horizon, with no retraining. Table 30 reports the five-fold results.

Performance improves progressively as additional observation history becomes available. By 48 hours, the model reaches AUROC 0.9139 ± 0.0143, close to the full-trajectory AUROC of 0.9211 ± 0.0086, while the 96-hour result is essentially indistinguishable from the full trajectory.

B.18.2 VALUE-DEPENDENT MNAR STRESS TEST

We evaluate value-dependent missingness on P19 across five matched folds. The probability that an observation is removed depends on its standardized observed value. Both cross-variable and local-coordinate models are evaluated under the same missingness realizations while keeping their trained checkpoints fixed. Table 31 reports the corresponding five-fold discrimination results across increasing missingness strengths.

At native coverage and mild missingness, the two models perform similarly. As value-dependent missingness becomes stronger, the cross-variable model degrades more slowly. At 73.3% realized missingness, its AUROC/AUPRC is $0.8690/0.4489$, compared with $0.7937/0.2975$ for the local-coordinate model. This provides five-fold evidence that direct cross-variable information can improve robustness when informative observations become increasingly unavailable.

B.18.3 RANDOM MISSING SENSORS

We evaluate structural sensor missingness by removing selected variables for the entire test trajectory while keeping the trained model fixed. This experiment is repeated across five matched P19 splits. Table 32 reports performance across increasing sensor-dropout levels.

Performance decreases progressively as complete sensors are removed. Across the five matched splits, NBSF retains higher mean AUROC and AUPRC than Raindrop at every evaluated sensor-dropout level.

B.18.4 RANDOM MISSING OBSERVATIONS (MCAR)

We independently remove observed entries at increasing dropout rates while keeping the trained cross-variable model fixed. Table 33 summarizes the five-fold classification, forecasting, and predictive-uncertainty statistics.

As additional observations are removed uniformly at random, discrimination and forecasting accuracy degrade progressively rather than collapsing abruptly. Even at $\rho = 0.7$, the cross-variable model retains AUROC 0.8756. This experiment provides a controlled five-fold characterization of robustness

Table 28: **P12 belief quality by inter-observation gap.** Cov90 and Cov95 are empirical predictive coverages.

Gap	Model	CRPS	NLL	Cov90	Cov95	Pred. SD
0–1h	Baseline	0.3697	1.2514	0.9067	0.9426	0.664
	S1	**0.3658**	1.2451	0.9084	0.9435	0.657
	Exact-OU	0.3662	**1.2342**	0.9101	0.9448	0.662
1–3h	Baseline	0.3107	0.9672	0.9179	0.9519	0.584
	S1	0.3060	0.9600	0.9184	0.9518	0.573
	Exact-OU	**0.3059**	**0.9556**	0.9173	0.9511	0.571
3–6h	Baseline	0.4371	1.3363	0.8587	0.9037	0.699
	S1	0.4453	**1.2543**	0.9077	0.9426	0.859
	Exact-OU	**0.4326**	1.2894	0.8640	0.9102	0.685
6–12h	Baseline	0.5841	1.5848	0.8649	0.9033	0.992
	S1	0.6449	1.6462	0.9428	0.9629	1.567
	Exact-OU	**0.5744**	**1.5790**	0.8818	0.9174	1.005
12–24h	Baseline	**0.5759**	**1.5551**	0.8992	0.9330	1.093
	S1	0.6533	1.6777	0.9515	0.9692	1.658
	Exact-OU	0.5840	1.5820	0.9323	0.9515	1.242
$>$ 24h	Baseline	**0.5441**	**1.4587**	0.8989	0.9438	1.091
	S1	0.5845	1.5328	0.9607	0.9775	1.485
	Exact-OU	0.5845	1.5893	0.9607	0.9831	1.532

Table 29: **Joint NLL improvement of rank-8 dependence over diagonal covariance.** Positive values indicate lower joint NLL for the rank-8 model.

Base	Dataset	Joint NLL improvement	Marginal change
Structured-noise	P12	13.80%	exactly 0
Structured-noise	P19	11.20%	exactly 0

Table 30: **P19 prediction under truncated observation windows** (mean $\pm$ standard deviation across five folds; no retraining).

Observation window	AUROC	AUPRC
6 h	0.6241 ± 0.0083	0.0736 ± 0.0022
12 h	0.7417 ± 0.0182	0.1465 ± 0.0214
24 h	0.8459 ± 0.0068	0.2923 ± 0.0397
48 h	0.9139 ± 0.0143	0.6030 ± 0.0336
96 h	$\mathbf{0.9216 \pm 0.0090}$	$\mathbf{0.6136 \pm 0.0276}$
Full	0.9211 ± 0.0086	0.6117 ± 0.0263

Table 31: **Value-dependent MNAR robustness on P19.** Values are means across five matched folds.

		Cross-variable		Local-coordinate	
Strength	Missing (%)	AUROC	AUPRC	AUROC	AUPRC
0.0	0.0	0.9211	0.6117	0.9228	0.6354
0.1	8.2	0.9183	0.6053	0.9192	0.6247
0.3	24.4	0.9118	0.5879	0.9097	0.5963
0.5	40.7	0.9041	0.5580	0.8945	0.5491
0.7	57.0	0.8898	0.5146	0.8678	0.4650
0.9	73.3	**0.8690**	**0.4489**	0.7937	0.2975

Table 32: **Robustness to random missing sensors on P19** (AUROC/AUPRC in %, mean ± standard deviation across five matched splits).

	NBSF		Raindrop	
Sensor dropout	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
0%	92.1 ± 0.8	61.2 ± 2.4	87.0 ± 2.3	51.8 ± 5.5
10%	91.5 ± 0.7	59.5 ± 2.4	83.5 ± 3.2	47.4 ± 7.0
30%	89.6 ± 0.9	54.7 ± 2.1	83.1 ± 1.5	48.2 ± 4.7
50%	88.5 ± 1.0	50.8 ± 2.1	80.9 ± 2.4	45.2 ± 6.9

Table 33: **Random observation dropout (MCAR) on P19.** Values are means across five matched folds for the cross-variable model.

Dropout ρ	**AUROC**	**AUPRC**	**Forecast RMSE ↓**	**Mean pred. uncertainty**
0.0	0.9211	0.6117	0.6862	0.6237
0.1	0.9192	0.6058	0.6935	0.6199
0.3	0.9110	0.5831	0.7136	0.6119
0.5	0.9010	0.5393	0.7455	0.6037
0.7	0.8756	0.4549	0.7988	0.5953
0.9	0.8032	0.3243	0.9173	0.5878

to increasing MCAR observation loss. These robustness experiments show gradual degradation under additional missing information. They do not, by themselves, identify cross-variable assimilation as the cause of robustness; that question is addressed by the controlled arm comparisons and suppression experiments above. Mean predictive uncertainty does not increase under MCAR removal despite the increase in forecasting error, indicating incomplete uncertainty sensitivity to this perturbation. We therefore do not interpret the MCAR experiment as evidence of improved uncertainty calibration under missingness.

B.19 Classification Calibration on Grouped Cohorts

Because the final MIMIC-III and eICU analyses use grouped splits, we report classification calibration for the final grouped arms rather than the earlier non-grouped summary. Table 34 summarizes the final five-fold classification and calibration results for the patient-grouped MIMIC-III and eICU evaluations.

Table 34: **Grouped classification metrics on MIMIC-III and eICU** (mean ± standard deviation across five folds). Higher AUROC/AUPRC and lower Brier/ECE are better.

Dataset	**Arm**	**AUROC**	**AUPRC**	**Brier ↓**	**ECE ↓**
MIMIC-III	Local-coordinate	$\mathbf{0.9040 \pm 0.0065}$	$\mathbf{0.6670 \pm 0.0259}$	0.0625 ± 0.0032	0.0163 ± 0.0068
	Dense-gain	0.9021 ± 0.0053	0.6659 ± 0.0231	$\mathbf{0.0623 \pm 0.0025}$	$\mathbf{0.0125 \pm 0.0029}$
	Cross-variable	0.8995 ± 0.0053	0.6582 ± 0.0240	0.0643 ± 0.0024	0.0205 ± 0.0042
eICU	Local-coordinate	0.8829 ± 0.0033	0.5733 ± 0.0074	0.0546 ± 0.0006	0.0047 ± 0.0024
	Dense-gain	0.8788 ± 0.0055	0.5661 ± 0.0121	0.0550 ± 0.0009	$\mathbf{0.0037 \pm 0.0011}$
	Cross-variable	0.8844 ± 0.0033	0.5791 ± 0.0070	0.0541 ± 0.0005	0.0058 ± 0.0017
	Target-dropout	$\mathbf{0.8893 \pm 0.0033}$	$\mathbf{0.5903 \pm 0.0075}$	$\mathbf{0.0533 \pm 0.0006}$	0.0055 ± 0.0015

The calibration results are mixed across variants. We therefore avoid a broad claim that cross-variable NBSF is uniformly better calibrated than the local or dense alternatives.

Figure 4 visualizes calibration directly: reliability curves for the cross-variable arm (a), per-mechanism ECE consistent with Table 34 (b), and empirical interval coverage on P12 and P19 (c).

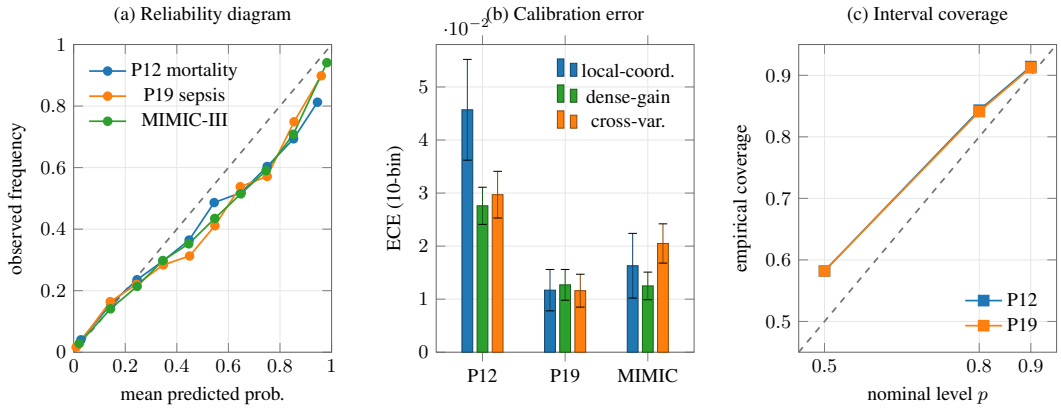


Figure 4: Calibration of NBSF on clean, in-distribution test data (pooled 5-fold). (a) Reliability curves for the cross-variable model track the diagonal, with mild overconfidence confined to high-risk bins on P12. (b) Expected calibration error is low across all three update mechanisms; differences fall within 5-fold standard deviation. (c) Empirical interval coverage meets or exceeds the nominal level, slightly conservative at p=0.5; MIMIC-III is absent from the coverage run. Error bars in (b) are 5-fold std.

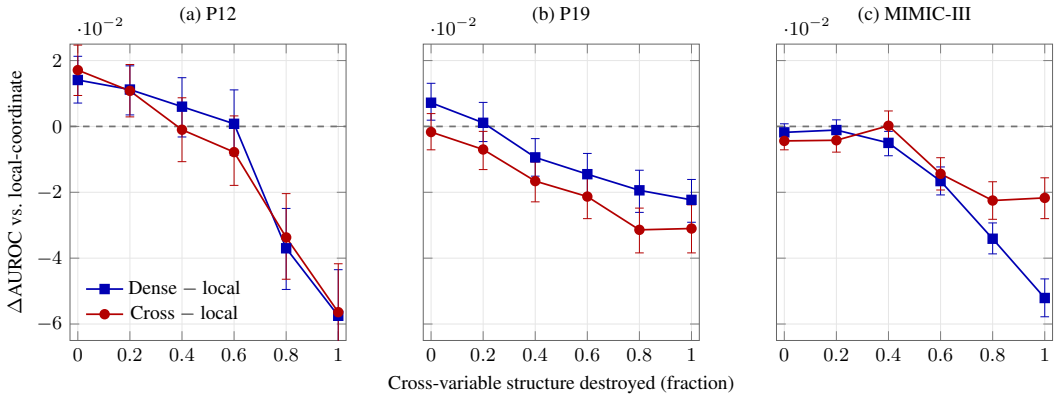


Figure 5: Cross-variable information destruction on P12, P19, and MIMIC-III. Non-target variables are permuted across patients at fraction f, destroying cross-variable structure while each trained model, the target variable, and the observation times are left unchanged. Curves show the AUROC of each cross-variable arm relative to the local-coordinate variant; error bars are patient-level bootstrap 95% CIs. On every dataset the advantage erodes and becomes negative as structure is destroyed—the mirror image of the target-suppression sweep (Figure 3), sharing the clean-data point at f=0. All arms degrade under permutation, but the local-coordinate variant degrades least (from f=0 to f=1: P12 local −0.198 vs. cross −0.271/dense −0.269; P19 −0.045 vs. −0.074/−0.074; MIMIC-III −0.190 vs. −0.207/−0.240), isolating the cross-variable pathway as the source of the effect.

B.20 Cross-Variable Information Destruction

As a controlled complement to the target-suppression sweep, we progressively destroy cross-variable structure at test time: for a fraction f of non-target variables (chosen at random per fold), we permute their values across patients, breaking inter-variable relationships while leaving each trained model, the target variable, and all observation times unchanged. Figure 5 plots the AUROC of each cross-variable arm relative to the local-coordinate variant as a function of f.

Every arm degrades under permutation, but the local-coordinate variant degrades least, so the effect cannot be attributed to additional capacity: with the model, its parameters, the target signal, and the observation times fixed, an arm loses performance from destroyed cross-variable structure only to the

extent it relied on that structure. This complements the parameter-matched controls and, together with the target-suppression analysis, brackets when cross-variable assimilation helps.

B.21 Comparison with STraTS and DATA-GRU

STraTS (Tipirneni & Reddy, 2022) and DATA-GRU (Tan et al., 2020) use different preprocessing pipelines and are therefore shown only as published operating points, not as controlled head-to-head rankings.

Table 35: **Comparison with STraTS and DATA-GRU under their published protocols** (AUROC/AUPRC in %). Absolute values are not directly comparable because the protocols differ.

		P12 Mortality		**MIMIC-III Mortality**	
Method	**Protocol**	AUROC	AUPRC	AUROC	AUPRC
DATA-GRU	Own protocol	85.2	—	85.2	—
STraTS	Own protocol	85.1	52.3	86.6	49.3
NBSF	Final protocol	84.67±0.90	51.60±2.83	89.95±0.53	65.82±2.40

The conceptual distinction is that NBSF maintains an explicit predictive belief state and performs event-wise innovation updates. We do not infer superiority from Table 35 because the protocols differ.

B.22 Cross-Variable Predictive Dependencies

The learned update can be summarized by the average normalized change in a column variable's belief when a row variable is observed. This quantity is asymmetric and describes how the trained model redistributes predictive information at an observation event. It is not a causal effect, nor does it identify a unique physiological graph.

For reference, we compare these update magnitudes with symmetric Pearson correlation after linear interpolation. Figures 6 and 7 show the correlation and learned update matrices for MIMIC-III, respectively, while Figures 8 and 9 show the corresponding comparison for P12. The interpolated correlation matrices are retrospective descriptive baselines because interpolation uses observations on both sides of a time point. The comparison is therefore intended to distinguish symmetric covariation from event-conditioned predictive belief updates, not to establish either representation as a ground-truth physiological or causal graph.

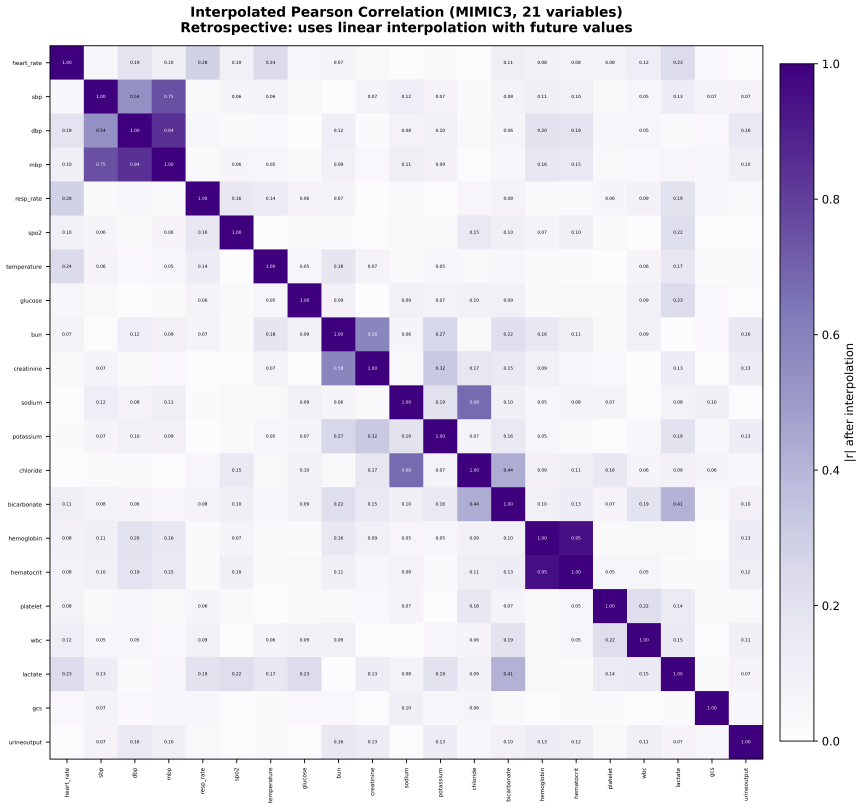


Figure 6: **MIMIC-III interpolated Pearson correlation.** Symmetric absolute correlation after linear interpolation.

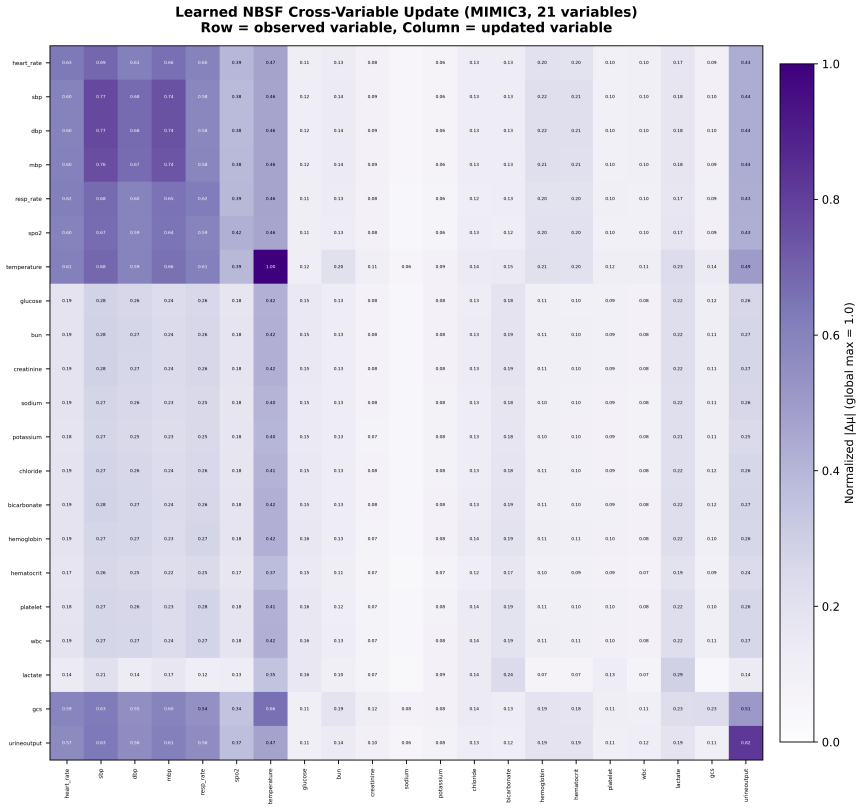


Figure 7: **MIMIC-III NBSF update magnitude.** Rows denote observed variables and columns denote belief dimensions updated at the same event. The matrix visualizes predictive update structure, not causal influence.

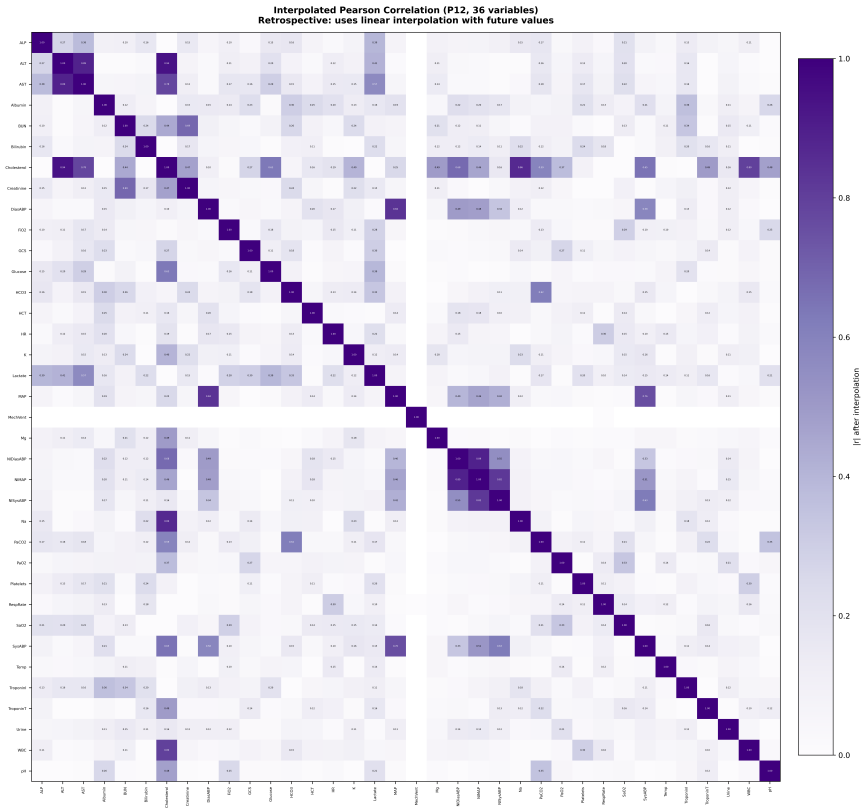


Figure 8: **P12 interpolated Pearson correlation.** Symmetric absolute correlation after linear interpolation.

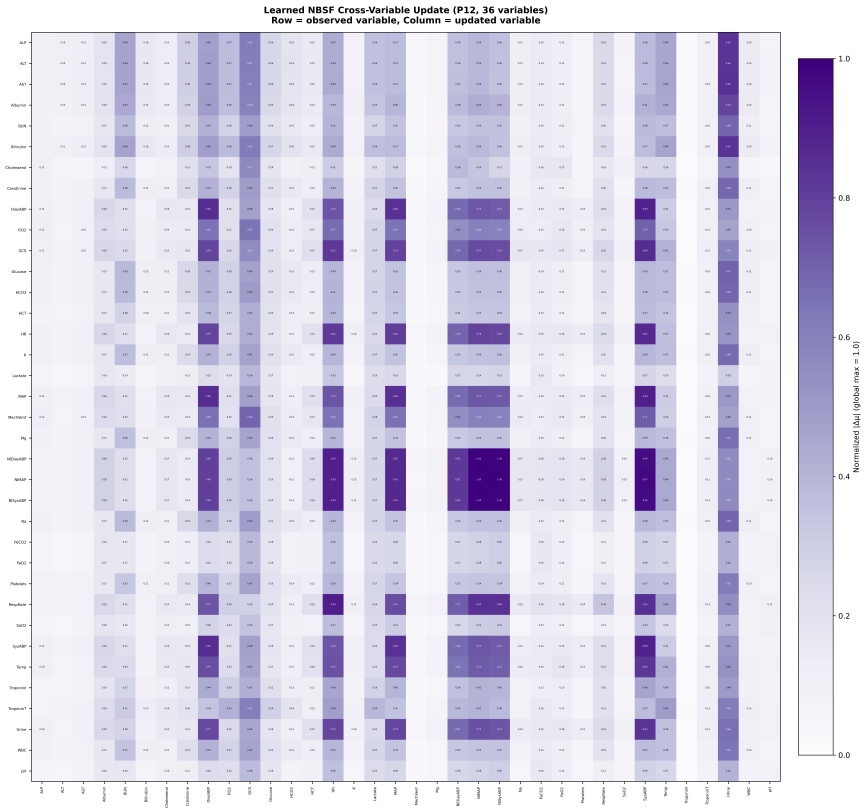


Figure 9: **P12 NBSF update magnitude.** Rows denote observed variables and columns denote belief dimensions updated at the same event. The asymmetry reflects learned predictive update behavior rather than identified physiology.

C FORWARD PASS AND COMPLEXITY

Algorithm 1 gives the final cross-variable NBSF forward pass. The two changes relative to the earlier implementation are the exact OU variance transition and an innovation-anchored mean update.

Algorithm 1 NBSF forward pass with exact-OU variance propagation and cross-variable update.

Require: Irregular trajectory $\{(t_i, \boldsymbol{x}_i, \boldsymbol{r}_i)\}_{i=1}^{N}$, $\boldsymbol{x}_i \in \mathbb{R}^D$, $\boldsymbol{r}_i \in \{0,1\}^D$
Require: $S = D + d_{\text{lat}}$; initial belief $(\boldsymbol{m}_0, \boldsymbol{v}_0)$; parameters $(\boldsymbol{a}, \boldsymbol{q}, \theta, \psi, \omega)$ with $\boldsymbol{a}, \boldsymbol{q} > 0$
$\boldsymbol{m} \leftarrow \boldsymbol{m}_0$; $\boldsymbol{v} \leftarrow \boldsymbol{v}_0$; $t_{\text{prev}} \leftarrow 0$; $\text{loss}_{\text{NLL}} \leftarrow 0$; $M \leftarrow 0$
for $i = 1, \dots, N$ **do**
  $\Delta t \leftarrow \max(t_i - t_{\text{prev}}, 0)$
  $\phi_i \leftarrow [\Delta t, \log(1 + \Delta t), t_{\text{prev}}, \log(1 + t_{\text{prev}})]$
  $A_i \leftarrow \exp(-\boldsymbol{a}\Delta t)$
  $\boldsymbol{m}^- \leftarrow A_i \odot \boldsymbol{m} + \Delta t\, f_\theta(\boldsymbol{m}, \boldsymbol{v}, \phi_i)$
  $\boldsymbol{v}^- \leftarrow A_i^2 \odot \boldsymbol{v} + \frac{\boldsymbol{q}}{2\boldsymbol{a}} \odot (1 - A_i^2)$
  $\log \boldsymbol{v}^- \leftarrow \text{clamp}(\log(\boldsymbol{v}^- + \epsilon), c_{\min}, c_{\max})$
  $\widehat{\boldsymbol{x}}_i^- \leftarrow W_{\text{obs}} \boldsymbol{m}^-$
  $\boldsymbol{\nu}_i^{\text{obs}} \leftarrow \boldsymbol{r}_i \odot \text{clip}(\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i^-, -10, 10)$
  $\widetilde{\boldsymbol{x}}_i \leftarrow [\boldsymbol{x}_i; \mathbf{0}_{S-D}]$; $\widetilde{\boldsymbol{r}}_i \leftarrow [\boldsymbol{r}_i; \mathbf{0}_{S-D}]$
  $\bar{\boldsymbol{\nu}}_i \leftarrow [\boldsymbol{\nu}_i^{\text{obs}}; \mathbf{0}_{S-D}]$
  $R_i \leftarrow \text{softplus}(h_\psi([\boldsymbol{m}^-; \boldsymbol{v}^-; \widetilde{\boldsymbol{x}}_i \odot \widetilde{\boldsymbol{r}}_i])) + \epsilon$
  $K_i \leftarrow \text{clamp}(\boldsymbol{v}^- \oslash (\boldsymbol{v}^- + R_i), 0, K_{\max})$
  $\widetilde{\boldsymbol{\nu}}_i \leftarrow K_i \odot \bar{\boldsymbol{\nu}}_i$
  $\Delta \boldsymbol{m}_i \leftarrow U_\psi^m(\boldsymbol{m}^-, \log \boldsymbol{v}^-, \widetilde{\boldsymbol{\nu}}_i) - U_\psi^m(\boldsymbol{m}^-, \log \boldsymbol{v}^-, \mathbf{0})$
  $\Delta \log \boldsymbol{v}_i \leftarrow U_\psi^\ell(\boldsymbol{m}^-, \log \boldsymbol{v}^-, \widetilde{\boldsymbol{\nu}}_i)$
  $\boldsymbol{m}^+ \leftarrow \boldsymbol{m}^- + \Delta \boldsymbol{m}_i$
  $\log \boldsymbol{v}^+ \leftarrow \text{clamp}(\log \boldsymbol{v}^- + \text{clip}(\Delta \log \boldsymbol{v}_i, -2, 0.2), c_{\min}, c_{\max})$
  $\boldsymbol{v} \leftarrow \exp(\log \boldsymbol{v}^+)$; $\boldsymbol{m} \leftarrow \boldsymbol{m}^+$; $t_{\text{prev}} \leftarrow t_i$
end for
return $\sigma(\text{MLP}_\omega([\boldsymbol{m}; \log \boldsymbol{v}]))$

Complexity. Each event requires an $O(S)$ exact-OU variance transition, a neural mean propagation step, an observation projection, and one neural update. For hidden width H, the dominant per-event cost is $O(SH + H^2 + DS)$, giving linear scaling in the number of observation events N. This statement concerns asymptotic event scaling and is not intended as a universal runtime comparison with every attention implementation.

D MATHEMATICAL DETAILS

Let $q_t(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{m}_t, \text{diag}(\boldsymbol{v}_t))$ denote the base factorized Gaussian belief with $\boldsymbol{m}_t \in \mathbb{R}^S$ and $\boldsymbol{v}_t \in \mathbb{R}^S_{>0}$.

D.1 LOCAL GAUSSIAN UPDATE AS A RESTRICTED CASE

Consider a scalar observation (t_i, k_i, x_i) with factorized Gaussian prior and observation model

$$q^-(\boldsymbol{z}) = \prod_{j=1}^{S} \mathcal{N}(z_j; m_j^-, v_j^-), \qquad p(x_i \mid \boldsymbol{z}) = \mathcal{N}(x_i; z_{k_i}, R_i).$$

For the observed coordinate, the scalar Gaussian update is

$$K_i = \frac{v_{k_i}^-}{v_{k_i}^- + R_i}, \qquad m_{k_i}^+ = m_{k_i}^- + K_i(x_i - m_{k_i}^-), \qquad v_{k_i}^+ = (1 - K_i) v_{k_i}^-.$$

All $j \neq k_i$ remain unchanged. A local-coordinate NBSF update is obtained by restricting the learned correction to this observed coordinate. The full NBSF removes that locality restriction: an innovation

from an observed variable can produce corrections in other observed and latent coordinates. The learned neural update is therefore filtering-inspired; it is not claimed to equal an analytic Bayesian posterior in the unrestricted nonlinear case.

D.2 Exact OU Variance Transition

For one coordinate, consider the linear stochastic differential equation

$$dz_t = -az_t\,dt + \sqrt{q}\,dW_t, \qquad a > 0,\ q > 0.$$

If the variance at the beginning of an interval is v_t, then for an interval of length Δt,

$$A(\Delta t) = e^{-a\Delta t}, \qquad v_{t+\Delta t} = A(\Delta t)^2 v_t + \frac{q}{2a}\left(1 - A(\Delta t)^2\right).$$

Applied coordinate-wise, this gives the variance propagation used by NBSF:

$$\boldsymbol{v}_{t+\Delta t} = A(\Delta t)^2 \odot \boldsymbol{v}_t + \frac{\boldsymbol{q}}{2\boldsymbol{a}} \odot \left(1 - A(\Delta t)^2\right).$$

The formula is exact for the coordinate-wise linear OU variance dynamics when $\boldsymbol{a}$ and $\boldsymbol{q}$ are constant over the interval. It is not an exact solution of the full nonlinear NBSF mean dynamics.

The transition has the semigroup property. For intervals Δt_1 and Δt_2, substituting the first transition into the second gives

$$\mathcal{T}_{\Delta t_2}(\mathcal{T}_{\Delta t_1}(v)) = e^{-2a(\Delta t_1+\Delta t_2)}v + \frac{q}{2a}\left(1 - e^{-2a(\Delta t_1+\Delta t_2)}\right) = \mathcal{T}_{\Delta t_1+\Delta t_2}(v).$$

As $\Delta t \to \infty$, the propagated variance approaches the finite stationary value $q/(2a)$.

D.3 Hybrid Mean Propagation and Stability

The mean propagator is

$$\boldsymbol{m}_{t+\Delta t} = A(\Delta t) \odot \boldsymbol{m}_t + \Delta t\, f_\theta(\boldsymbol{m}_t, \boldsymbol{v}_t, \phi(\Delta t, t)), \qquad A(\Delta t) = \exp(-\boldsymbol{a}\Delta t), \quad \boldsymbol{a} > 0.$$

For small Δt,

$$A(\Delta t) = 1 - \boldsymbol{a}\Delta t + O(\Delta t^2),$$

so the corresponding first-order continuous limit is

$$\dot{\boldsymbol{m}}(t) = -\boldsymbol{a} \odot \boldsymbol{m}(t) + f_\theta(\boldsymbol{m}(t), \boldsymbol{v}(t), t).$$

The negative sign follows directly from the decaying exponential transition.

Assumption 1 (Bounded residual and bounded inter-observation gaps)**.** *There exist $B_f < \infty$ and $\Delta t_{\max} < \infty$ such that*

$$\|f_\theta(\boldsymbol{m}, \boldsymbol{v}, \phi)\|_\infty \le B_f, \qquad 0 < \Delta t \le \Delta t_{\max}.$$

Let $a_{\min} = \min_j a_j > 0$.

Proposition 1 (Uniform bound for the propagated mean)**.** *Under the assumption above,*

$$\|\boldsymbol{m}_t\|_\infty \le \max\left\{\|\boldsymbol{m}_0\|_\infty, \frac{B_f \Delta t_{\max}}{1 - \exp(-a_{\min}\Delta t_{\max})}\right\}.$$

Proof. For one step,

$$\|\boldsymbol{m}_{t+\Delta t}\|_\infty \le e^{-a_{\min}\Delta t}\|\boldsymbol{m}_t\|_\infty + \Delta t B_f.$$

Define

$$R = \max\left\{\|\boldsymbol{m}_0\|_\infty, \sup_{0<\Delta t\le\Delta t_{\max}} \frac{B_f \Delta t}{1 - e^{-a_{\min}\Delta t}}\right\}.$$

The ratio $\Delta t/(1 - e^{-a_{\min}\Delta t})$ is increasing for $\Delta t > 0$, so the supremum is attained at $\Delta t_{\max}$. If $\|\boldsymbol{m}_t\|_\infty \le R$, then

$$\|\boldsymbol{m}_{t+\Delta t}\|_\infty \le e^{-a_{\min}\Delta t}R + \Delta t B_f \le R.$$

Induction gives the stated bound. □

D.4 Zero-Innovation Anchoring

A generic neural update can produce a nonzero correction even when the innovation is zero. We remove this degree of freedom from the mean update by defining

$$\Delta \boldsymbol{m}(\widetilde{\nu}) = U_\psi^m(\boldsymbol{m}^-, \ell^-, \widetilde{\nu}) - U_\psi^m(\boldsymbol{m}^-, \ell^-, 0).$$

Hence

$$\Delta \boldsymbol{m}(0) = 0$$

exactly. This property makes the mean correction explicitly innovation-driven without constraining the nonlinear response for nonzero innovations.

D.5 Frozen-Marginal Low-Rank Dependence Model

Let $D = \mathrm{diag}(\boldsymbol{v})$ denote the base marginal variance matrix and let $V \in \mathbb{R}^{S \times r}$ be a learned low-rank factor. Define

$$C = I + VV^\top, \qquad S_C = \mathrm{diag}(C), \qquad R = S_C^{-1/2} C S_C^{-1/2}.$$

Because $C \succeq I$, C is positive definite and R is a valid correlation matrix with unit diagonal. The joint covariance is

$$\Sigma = D^{1/2} R D^{1/2}.$$

Its marginal variances are unchanged:

$$\mathrm{diag}(\Sigma) = \boldsymbol{v}.$$

Thus the dependence head can alter joint likelihood and multivariate scoring without changing the base marginal means or variances. This is the basis for the exact marginal-identity checks reported in Section B.17.