# Online Technical Appendix

Computational Pipeline, Validation Record, and Replication Map

Mátyás Farkas

July 2026

**Abstract**

This appendix documents the computation behind the paper. It answers four questions: what is computed, how it is computed, what is checked, and what the checks do not prove. The pipeline has three separate objects: the direct nonlinear transition computed by stochastic extended path (SEP), the ROM1-residual surrogate that approximates that transition, and the profiled inversion objective used for empirical criterion comparisons in the application. The appendix is intentionally more comprehensive than the paper: it records successful diagnostics, failed diagnostics, non-promoted pilots, artifact locations, and replication commands.

## Contents

# 1 Scope

This appendix is a technical companion to the paper. It is not a second empirical paper. The purpose is to make the computation easy to audit. Each section states the object being computed, the procedure used to compute it, and the evidence produced by the corresponding diagnostic.

The core computational claim is narrow:

$$\text{ROM1} + \widehat{\Delta} \quad \text{can approximate} \quad \text{SEP}$$

on explicitly documented state–shock–parameter support, where $\widehat{\Delta}$ is a learned residual correction. The empirical comparisons in the paper use this approximation inside a profiled inversion objective. They are therefore not marginal-likelihood comparisons. Determinant terms alone would not close the gap: a Laplace approximation would also require the shock prior evaluated at the profiled shock path and the Hessian of the shock-level criterion, while an exact change-of-variables likelihood would require a separately derived Jacobian. The validation exercises below are designed around that distinction.

**Lower-bound terminology.** This appendix uses *ELB* for the economic effective lower bound. It uses *OBC* only for the occasionally-binding-constraint solver implementation, and *actual-floor path* only for the Galí validation design in which the simulated policy rate binds at the maintained floor.

**How to read this appendix.** The main paper should be read as the economic argument. This appendix should be read as the computation record. It is written to make it possible to answer three practical questions without reading the code first.

1. If a number appears in the paper, which artifact produced it?
2. If a validation passed, what exactly did it validate?
3. If a diagnostic failed or was not promoted, why was it excluded?

**Maintained components.** The maintained replication package contains four layers.

1. A direct nonlinear solver layer based on stochastic extended path [Fair and Taylor, 1983, Adjemian and Juillard, 2025].
2. A ROM1-residual surrogate layer trained on $g_\theta^{\text{SEP}} - g_\theta^{\text{ROM1}}$, not on the full transition.

3. An online inference layer using shock inversion, optional regime switching, and NUTS-HMC [Hoffman and Gelman, 2014].

4. A target-guided support-refinement layer that turns HMC-visited out-of-support gate points into additional SEP teacher labels before a promoted empirical run.

**Main limitation.** The full 18-parameter Smets–Wouters–HLT application is not validated by a full direct-SEP HMC run. The medium-scale bridge has two maintained parts. The first is a finite-support, one-period known-feature comparison over the investment block. It is strong evidence that the residual target is learned on the relevant nonlinear channel. The second is a short dynamic inversion bridge. The fixed one-step NN residual fails the harder eight-period diagnostic, but a path-calibrated residual trained on dynamic ROM1 states and recovered shocks passes on held-out anchors. This is evidence for the dynamic residual target on the reduced HLT bridge; it is not a complete certificate for the full 18-parameter multi-period posterior.

# 2 Claim Map

This section links the paper's claims to the evidence required to support them. The table is deliberately conservative. It separates statements that are direct measurements from statements that depend on the surrogate, the gate, or the posterior sampler. Table 1 reports the resulting claim map.

**Claims deliberately not made.** The paper does not claim that the surrogate objective is an exact marginal likelihood. It does not claim a Bayes factor against the linear model. It does not claim that the reported warm-started mode is globally dominant. It does not claim that the Galí two-parameter validation proves the full 18-parameter HLT posterior. It does not claim that the SEP sensitivity check is a global solver-invariance proof or that SEP convergence alone proves global stability outside the audited corridor. It does not claim that the NN residual correction alone raises HLT's published likelihood; the matched filter-free audit is mixed and out of support. It does not use the pruned-ROM2 pilot as evidence, because that pilot failed its sanity check.

Table 1: Claim Map

| Claim | Evidence | Scope |
|---|---|---|
| Investment is the largest matched SEP–ROM1 discrepancy among the seven observables under raw and SEP-standardized units | Parameter-matched normalization audit: 66.2–78.4 percent across those two metrics; investment is second among all 51 outputs, at 5.8–26.3 percent across reported normalizations | Descriptive location of a transition gap, not a causal equation attribution or normalization-invariant theorem |
| The residual surrogate improves on parameter-matched ungated ROM1 | Twelve held-out 18-parameter vectors and 2,208 transitions; raw pooled RMSE falls from 2.1879 to 1.2373 and all seven observables improve | Finite normal-state transition paths; no gate, inversion filter, posterior, or global certificate |
| Supervised SEP labels select one locally audited numerical branch | One direct SEP teacher label per accepted input; failed cells mapped; 40-vector parameter ablation converges at shock scales 0.1, 0.25, and 0.5 | Finite numerical branch consistency, not a mathematical proof of local or global equilibrium uniqueness |
| The 18-parameter HMC target can be diagnosed within its initialized basin | Four chains; rank, trace, autocorrelation, interval, correlation, energy, E-BFMI, ESS, MCSE, divergence, and tree-depth diagnostics | Profiled statistical target; does not test global economic stability or rule out other posterior modes |
| The dynamic HLT bridge requires path-aware residuals | The fixed one-step NN fails the eight-period diagnostic; a path-calibrated residual trained on five dynamic anchors passes on five held-out anchors, reducing held-out centered surface RMSE from 1.437 under ROM1 to 0.703 | Reduced HLT bridge, not full 18-parameter direct-SEP HMC |
| The filter-free HLT gate-shock audit does not support an NN-only likelihood claim | Matched 100-draw HMC audit: ROM1+NN raises observation fit on its own draws by 30.1 nats but lowers the sampled joint log posterior by 33.4 nats; same-draw likelihood gains are positive on ROM1+NN draws and negative on ROM1-only draws | Full 18-parameter real-data audit; evidence of local usefulness and support limits, not a robust likelihood improvement |
| The ELB/OBC pipeline passes the reported small-model diagnostics | Galí actual-floor ELB validation package; two-parameter matched direct/surrogate HMC over $(\sigma_z, \sigma_a)$ passes | Small-model ELB validation, not full HLT posterior validation |
| The warm-started extended-sample surrogate criterion exceeds the linear criterion | 2-by-2 criterion accounting: $-1817.59$, $-1852.63$, $-1830.44$, $-1781.68$ | Mixed exact Kalman and profiled-inversion criterion values; accounting exercise only, not a Bayes factor |
| The warm-started surrogate mode has a high profiled objective | Four warm-started surrogate chains have posterior-mean objectives from $-1787.38$ to $-1780.80$; cold starts remain in a low-objective basin | Conditional on mode-search evidence currently available |
| The gate needs caution in calm periods | Quiet-sample OOS failures under the original gate; q95 padded diagnostic improves fixed-posterior ratio to 1.04 and re-estimated full-gate ratio to 1.09 | Gate-design diagnostic; not a fully optimized forecasting rule |
| SEP solver settings do not overturn the 265-period sensitivity probe | Full 265-period SEP sensitivity run over five posterior draws and six $(\texttt{accept\_tol}, K)$ cells | Finite-domain corridor-stability diagnostic, not a theorem of global asymptotic stability or uniqueness |

*Notes*: The table separates direct SEP measurements, finite-support surrogate validations, and empirical profiled-criterion results. Exact Kalman likelihood values and profiled surrogate objective values are recorded as criterion values under the maintained normalization; the table does not report marginal likelihoods or Bayes factors. Artifact paths are listed in Section 15.3.

# 3 Notation and Objects

Let $s_t \in \mathbb{R}^{n_s}$ denote the model state, $\varepsilon_t \in \mathbb{R}^{n_\varepsilon}$ the structural shocks, $y_t \in \mathbb{R}^{n_y}$ the observables, and $\theta \in \Theta \subset \mathbb{R}^p$ the estimated parameter vector. The nonlinear one-step transition is

$$(x_t^{\mathrm{SEP}}, s_t^{\mathrm{SEP}}) = g_\theta^{\mathrm{SEP}}(s_{t-1}, \varepsilon_t),$$

where $x_t^{\mathrm{SEP}}$ stacks the quantities retained for the surrogate target. The first-order perturbation approximation is

$$(x_t^{\mathrm{ROM1}}, s_t^{\mathrm{ROM1}}) = g_\theta^{\mathrm{ROM1}}(s_{t-1}, \varepsilon_t).$$

The surrogate does not approximate $g_\theta^{\mathrm{SEP}}$ directly. It approximates the residual

$$\Delta_\theta(s_{t-1}, \varepsilon_t) = x_t^{\mathrm{SEP}} - x_t^{\mathrm{ROM1}}. \tag{1}$$

The online predictor is therefore

$$\widehat{x}_t = x_t^{\mathrm{ROM1}} + \widehat{\Delta}_w(s_{t-1}, \varepsilon_t, \theta), \tag{2}$$

with neural-network weights $w$.

The observation equation used for the inversion objective is

$$y_t^{\mathrm{obs}} = H x_t + u_t, \qquad u_t \sim \mathcal{N}(0, \Sigma_y). \tag{3}$$

In the HLT bridge and Galí ELB/OBC checks, $\Sigma_y$ is fixed by the validation design. In the empirical HLT estimation, $\Sigma_y$ is a calibrated measurement-error scale, with robustness checks tied to surrogate validation RMSE.

# 4 End-to-End Data Flow

**Algorithm 1**: Pipeline

1. Choose a model, parameter block, parameter support, observables, and shock/state design.
2. For each design point, compute a direct SEP transition and the corresponding ROM1 transition at the same $(s_{t-1}, \varepsilon_t, \theta)$.
3. Save a training payload with inputs $X = [s_{t-1}; \varepsilon_t; \theta]$, direct outputs $Y$, ROM1 outputs $Y^{\mathrm{ROM1}}$, solver residual diagnostics, observable indices, parameter names, and git

provenance.

4. Train a neural network on $Y - Y^{\mathrm{ROM1}}$. In the HLT bridge the reported target is observation-only; in the full empirical estimator the bundle can also carry state corrections needed for propagation.

5. Validate the surrogate out of sample by comparing $Y^{\mathrm{ROM1}} + \widehat{\Delta}$ to $Y$, and by comparing criterion surfaces against direct SEP on finite support.

6. Run a preliminary empirical chain or gated shock audit. Score the visited gate-period inputs against the surrogate normalization statistics. If the sampler leaves support, label those visited points with one-step SEP, append them to the residual dataset, and retrain.

7. In empirical estimation, recover shocks period by period using the ROM1 inversion filter, evaluate the surrogate correction on gate periods, use exact ROM1/Kalman predictive log-likelihood contributions on calm periods, use profiled observation-fit contributions on gate periods, and sample $\theta$ with NUTS-HMC.

The output-file convention uses raw payloads serialized as Julia `.jls` files under `.local_artifacts/`. Human-readable summaries are stored as `SUMMARY.md` or `SUPPORT_REPORT.md`. Tables promoted to the paper are stored as `.tex` fragments under either the paper directory or the relevant output directory. The paper should not report a numerical result unless a matching raw payload and human-readable summary exist.

# 5 Target-Guided Support Refinement

The surrogate is trained on a finite set of SEP solves. A later HMC run may visit a state, shock, and parameter combination that was not well covered by that set. The support-refinement step handles this directly. It uses the preliminary HMC path to decide where more SEP labels are needed, then retrains the surrogate on the enlarged dataset.

Let

$$z_j(x) = \left| \frac{x_j - \mu_j^X}{\sigma_j^X} \right|$$

be the feature-by-feature distance from the training distribution stored in the surrogate bundle. For each gate-active period $t$, I form

$$x_t = [s_{t-1}; \varepsilon_t; \theta]$$

from the preliminary chain or gated shock audit. I record the largest distance $\max_j z_j(x_t)$, the 95th percentile of the distances, and the feature that is farthest from the training set. Gate periods with large distances are sent back to SEP. The one-step teacher label solves

$$g_\theta^{\mathrm{SEP}}(s_{t-1}, \varepsilon_t)$$

from the visited state and shock. ROM1 is computed at the same input. The new training target is still the residual in equation (1). Only the region of support changes.

This step is deterministic once the preliminary chain and gate are fixed. A promoted run must report four diagnostics:

1. the number of visited gate periods selected for new SEP labels;
2. the direct SEP success count and residual distribution at those points;
3. the held-out RMSE of $Y^{\mathrm{ROM1}} + \widehat{\Delta}$ on the augmented dataset, relative to ROM1;
4. the post-retraining HMC comparison against the linear+gate/no-NN baseline.

The last item is essential. Better support does not by itself prove that the NN correction improves the empirical criterion. If the retrained gate+NN run still loses to the linear+gate baseline, the problem is not only extrapolation. It lies in the objective, the gate, or the residual target.

The current implementation is:

```
scripts/hlt_endogenous_support_dataset.jl
scripts/run_hlt_endogenous_support_retrain.sh
scripts/hlt_posterior_support_refinement_validate.jl
```

The first script builds the active labels and the combined dataset. The second script runs the full loop: active labeling, residual-surrogate retraining, short HMC check, and the matched gated shock–theta audit. The third script validates the local object learned by the refinement loop: direct SEP one-step labels on target-guided support points, compared against ROM1 and ROM1 plus scaled NN residual corrections.

# 6 Direct SEP Data Generation

The direct benchmark solves the nonlinear model with stochastic extended path. For each design point, the solver receives the current parameter vector, a finite shock path, and the previous state. It returns a nonlinear transition and a residual diagnostic. A sample is

accepted for training or validation only if the relevant quantities are finite and the final SEP residual is below the stage-specific acceptance tolerance.

## 6.1 Corridor-Stability Diagnostic

The SEP benchmark is used as a finite-domain stability diagnostic, not as a formal global-stability theorem. The distinction matters. In nonlinear dynamic models, local stability can coexist with limited basins of attraction or "corridors" in which the economy returns to a stable path, while larger perturbations can move the system to another regime or no admissible path [Semmler and Sieveking, 1993, Schleer and Semmler, 2015]. A local perturbation solution or a residual-trained neural solution can therefore look accurate near an ergodic cloud and still miss another solution branch.

The diagnostic fixes an audited domain

$$\mathcal{D} = \{(s_{t-1}, \varepsilon_t, \theta) : s_{t-1} \in \mathcal{S},\ \varepsilon_t \in \mathcal{E},\ \theta \in \Theta_{\text{audit}}\}.$$

For each design point in $\mathcal{D}$, the direct SEP solve is treated as a teacher label only if it passes the following checks:

1. *Finite path and residual*: all states, controls, observables, and likelihood inputs are finite, and the final SEP residual is below the stage-specific tolerance.
2. *Initialization robustness*: cold starts and available warm starts lead to the same early path within the reported tolerance.
3. *Horizon and terminal-condition robustness*: changing the SEP horizon or terminal rule does not materially change the first-period transition used for the likelihood.
4. *Finite-perturbation return*: shocks and states near the accepted point remain inside the same corridor, or failures are explicitly mapped.
5. *Regime coverage*: occasionally binding constraints, gates, and other discrete regimes are recorded so that a surrogate is not trained on an unlabeled mixture of regimes.

The output is a support map, not an imputed surface. Cells that fail the direct SEP solve are saved as failures and either excluded from the promoted support or used to define a separate stress region. This is the sense in which SEP adds discipline relative to an unsupervised or purely residual-based neural solver: the nonlinear path solver supplies a falsifiable label and a failure map before the neural network is trained. The network can still be wrong between audited points, so the paper additionally reports held-out prediction errors,

criterion-surface comparisons, and matched-HMC diagnostics. The corridor certificate says that the promoted domain has passed these numerical stress tests; it does not say that the model has a unique globally attracting solution for every state and shock.

## 6.2 Input Design

The HLT training and bridge scripts use three types of parameter designs.

1. *Prior/Sobol designs* for broad surrogate coverage.
2. *Deterministic grids* for support mapping and criterion-surface validation.
3. *Trimmed supported grids* after direct SEP reveals infeasible high-stress corners.

For the medium-scale HLT bridge, the maintained parameter block is

$$(\rho_b, \rho_{qs}, \sigma_b, \sigma_{qs}) \equiv (\texttt{crhob}, \texttt{crhoqs}, \texttt{z\_eb}, \texttt{z\_eqs}).$$

The supported bridge trims the risk-premium shock scale to

$$z_{eb} \in [1.20, 1.85],$$

because the support map showed systematic direct-SEP failure at the high-stress $z_{eb} = 2.50$ edge. This is not imputation. The benchmark support is redefined to the region where direct nonlinear solutions are available.

## 6.3 ROM1 and SEP at Common Inputs

For every accepted sample $j$, the dataset stores

$$X_j = [s_{j,-}; \varepsilon_j; \theta_j], \qquad Y_j = g^{\mathrm{SEP}}_{\theta_j}(s_{j,-}, \varepsilon_j), \qquad Y_j^{\mathrm{ROM1}} = g^{\mathrm{ROM1}}_{\theta_j}(s_{j,-}, \varepsilon_j).$$

The residual target is then $Y_j - Y_j^{\mathrm{ROM1}}$. This common-input design is the reason the validation RMSE is meaningful: the network is judged against the direct nonlinear correction at the same state, shock, and parameter vector.

## 6.4 Warm Starts and Failure Handling

Across deterministic grids, the solver uses the previous converged solution as an initial guess for the next point. The warm start is accepted only when the previous solve converged. If

the previous solve failed or exceeded tolerance, the next solve falls back to a cold start from the perturbation solution. This prevents failed trajectories from propagating across the grid.

Nonfinite samples are dropped before training. Failed direct-SEP cells are recorded in support reports rather than silently replaced. A grid is promoted to validation only if its support is explicit: either every cell solves or the accepted support is trimmed and re-run.

# 7 ROM1-Residual Surrogate

## 7.1 Residual Target

The trained object is a map

$$\widehat{\Delta}_w : \mathbb{R}^{n_s+n_\varepsilon+p} \to \mathbb{R}^{d_o},$$

where $d_o$ is either the number of observables or the number of retained state/observable outputs. The HLT bridge uses observation-only output,

$$\widehat{\Delta}_{w,o} \approx H\{g_\theta^{\mathrm{SEP}}(s,\varepsilon) - g_\theta^{\mathrm{ROM1}}(s,\varepsilon)\},$$

because the posterior-grid comparison is a known-feature, one-period observable comparison. The empirical HLT estimation bundle also contains the state-propagation machinery needed by the inversion filter.

Learning the residual rather than the full transition has two practical consequences. First, the target has a smaller scale because ROM1 captures the local dynamics near steady state. Second, a zero residual is an interpretable ablation: it gives the same gate and inversion architecture with no neural nonlinear correction.

## 7.2 Training Objective

Let $D_{\mathrm{train}}$ be the training set. Inputs and targets are standardized dimension by dimension:

$$\widetilde{X}_{j,k} = \frac{X_{j,k} - \bar{X}_k}{s_{X,k}}, \qquad \widetilde{\Delta}_{j,\ell} = \frac{\Delta_{j,\ell} - \bar{\Delta}_\ell}{s_{\Delta,\ell}}.$$

The default objective is mean squared error,

$$\min_w \frac{1}{|D_{\mathrm{train}}|} \sum_{j \in D_{\mathrm{train}}} \left\| \widehat{\Delta}_w(\widetilde{X}_j) - \widetilde{\Delta}_j \right\|^2. \tag{4}$$

Training uses AdamW with gradient clipping, cosine learning-rate decay with warmup, and SiLU activation. The maintained MLP has two hidden layers in the HLT bridge runs; the utility layer also supports a residual network architecture for wider training designs.

### 7.3 Validation Metrics

For each output coordinate $k$, the validation RMSE is

$$\mathrm{RMSE}_k^{\mathrm{sur}} = \left[ \frac{1}{|D_{\mathrm{val}}|} \sum_{j \in D_{\mathrm{val}}} \left( Y_{j,k} - Y_{j,k}^{\mathrm{ROM1}} - \widehat{\Delta}_{w,k}(X_j) \right)^2 \right]^{1/2}.$$

The improvement over ROM1 is

$$1 - \frac{\mathrm{RMSE}_k^{\mathrm{sur}}}{\mathrm{RMSE}_k^{\mathrm{ROM1}}}.$$

The paper reports the aggregate RMSE and surface-RMSE diagnostics only after the validation point is held out from training.

## 8 Regime-Switching Inversion Objective

The empirical HLT target uses a profiled inversion objective. The state dimension and shock sequence are too large to sample jointly with the 18-dimensional parameter vector in the current implementation. The inversion filter reduces the online sampler to $\theta$ by recovering shocks period by period.

### 8.1 Shock Recovery

Given $s_{t-1}$, $\theta$, and observed $y_t^{\mathrm{obs}}$, the ROM1 inversion step solves

$$\widehat{\varepsilon}_t(\theta) = \arg\min_{\varepsilon} \left\| \frac{y_t^{\mathrm{obs}} - H g_\theta^{\mathrm{ROM1}}(s_{t-1}, \varepsilon)}{\sigma_y} \right\|^2 + \left\| \frac{\varepsilon}{\sigma_\varepsilon(\theta)} \right\|^2. \tag{5}$$

The same recovered shock can then be evaluated under ROM1, direct SEP in validation exercises, or ROM1 plus the surrogate residual. This common-shock design is essential: it isolates the transition approximation from shock recovery. The shock penalty in (5) regularizes the recovery problem and fixes the common shock path; the profiled observation objective below does not subsequently add the shock-density term as a joint likelihood contribution.

## 8.2 Profiled Observation-Fit Objective

On a surrogate-evaluated period, the prediction is

$$\widehat{y}_t^{\mathrm{sur}}(\theta) = H\left[g_\theta^{\mathrm{ROM1}}(s_{t-1}, \widehat{\varepsilon}_t) + \widehat{\Delta}_w(s_{t-1}, \widehat{\varepsilon}_t, \theta)\right].$$

The period contribution is the Gaussian observation objective

$$\ell_t^{\mathrm{sur}}(\theta) = -\frac{1}{2}\left[(y_t^{\mathrm{obs}} - \widehat{y}_t^{\mathrm{sur}})^\top \Sigma_y^{-1}(y_t^{\mathrm{obs}} - \widehat{y}_t^{\mathrm{sur}}) + \log|2\pi\Sigma_y|\right]. \tag{6}$$

Because $\widehat{\varepsilon}_t(\theta)$ is profiled out, this is not an exact joint likelihood or marginal likelihood. The missing pieces are not only a determinant: a Laplace-style profiled likelihood would also require the shock log density evaluated at $\widehat{\varepsilon}_t(\theta)$ and the Hessian determinant of the shock-level criterion, while an exact change-of-variables calculation would require a separately derived Jacobian for the observation to shock mapping. The paper therefore refers to the reported values as profiled observation-fit objectives.

## 8.3 Gate

The regime switch is deterministic conditional on a gate-calibration payload. Let $m_t \in \{0, 1\}$ denote the hard gate, with $m_t = 1$ meaning that the surrogate correction is active. The gate is built from shock and forecast-error statistics, padded before and after detected stress periods, and filtered to avoid isolated one-period switches. With a hard gate,

$$\ell_T(\theta) = \sum_{t:m_t=1} \ell_t^{\mathrm{sur}}(\theta) + \sum_{t:m_t=0} \ell_t^K(\theta),$$

where $\ell_t^K$ is the Kalman contribution under ROM1. With a soft gate, the implementation uses precomputed probabilities $p_t \in [10^{-4}, 1 - 10^{-4}]$ and combines period contributions by the maintained mixing routine. The q95 gate diagnostic in the paper recalibrates this object to avoid classifying all quiet-sample periods as stressed. Because this object combines exact Kalman predictive contributions on calm periods with profiled observation-fit contributions on stress periods, it is a hybrid criterion unless the full joint density and state-update recursion are normalized under one likelihood architecture.

## 8.4 Exact Calculation Reported as a Profiled Objective

For a fixed parameter vector $\theta$, the empirical objective is computed by the following deterministic sequence.

**Algorithm 2**: Empirical objective evaluation

1. Load the transformed data, steady state, measurement map, shock standard deviations, and gate-calibration payload.
2. Initialize the state using the same convention as the corresponding chain payload.
3. For period $t$, recover $\widehat{\varepsilon}_t(\theta)$ from the ROM1 inversion problem in (5).
4. If the gate is calm, add the Kalman/ROM1 period contribution $\ell_t^K(\theta)$.
5. If the gate is stressed, form the ROM1 prediction plus either a zero correction (linear+gate ablation) or the neural residual correction, and add the Gaussian observation contribution in (6).
6. Propagate the state with the same transition used for the period prediction.
7. Sum period contributions and add the parameter prior and transform Jacobian when evaluating the HMC target. This target omits a shock-prior contribution and the determinant/Laplace terms that would be needed for a full likelihood.

The word "profiled" has a precise meaning here. The shocks $\widehat{\varepsilon}_t(\theta)$ are chosen to fit the observation at each period rather than integrated out. The objective therefore compares models under a common shock-recovery rule. It is useful for studying whether the nonlinear transition improves fit under that rule. It is not a likelihood in the sense required for a Bayes factor unless the omitted shock-prior contribution, Jacobian or Hessian determinant, and a unified treatment of calm and stress periods are derived and included.

# 9 Approximation Error Logic

The paper uses one formal calculation only in a narrow way: it bounds the effect of replacing a direct SEP transition with a surrogate transition in an exact Gaussian observation likelihood. This is not a proof that the empirical profiled hybrid objective is an exact posterior. It is a diagnostic for the approximation component alone.

Let $s_t^F$ be the SEP state and $\widehat{s}_t$ the surrogate state at the same initial condition, shocks, and parameter vector. With

$$y_t = H s_t^F + \eta_t, \qquad \eta_t \sim N(0, \Sigma_y),$$

define the observable approximation error

$$e_t = H(s_t^F - \widehat{s}_t), \qquad r_t = y_t - \frac{H(s_t^F + \widehat{s}_t)}{2}.$$

The exact Gaussian likelihood difference satisfies

$$\ell_t^F(\theta) - \ell_t^S(\theta) = r_t^\top \Sigma_y^{-1} e_t.$$

This identity implies a deterministic pathwise bound,

$$|\ell_t^F - \ell_t^S| \le \lambda_{\min}(\Sigma_y)^{-1} \|r_t\| \, \|e_t\|.$$

Under the orthogonality condition

$$E[\eta_t^\top \Sigma_y^{-1} e_t] = 0,$$

the expected bias is second order:

$$|E[\ell_t^F - \ell_t^S]| \le \frac{1}{2\lambda_{\min}(\Sigma_y)} E\|e_t\|^2.$$

Thus the calculation controls the error from the learned transition on covered support. It does not control errors from shock profiling, gate design, omitted determinant terms, omitted shock-density terms, or mode search. In the 18-parameter application, the non-binding validation RRMSE of 0.0012 makes this approximation component small under the exact Gaussian benchmark, but the empirical 35.92-nat comparison remains a profiled-criterion comparison.

# 10 Matched No-NN Ablations

The maintained empirical script supports a no-NN ablation flag. This keeps the inversion filter, gate, state propagation convention, priors, and sampler structure, but sets $\widehat{\Delta}_w = 0$. Conceptually, this linear-plus-gate branch retains HLT's regime-switching linear approximation and the maintained inversion architecture while removing the additional learned residual. The promoted four-chain current-ROM result is reported in Section 13.2. This section first records two earlier bounded checks that informed that production run.

After rebuilding the HLT validation payload with unit shock scaling, I run a small dynamic real-data check with the same payload, gate, priors, seeds, and NUTS settings under

the full gate+NN objective and the linear+gate objective. Each run uses 100 adaptation draws and 100 post-warmup draws. Both runs complete with zero post-warmup divergences:

| | Mean objective | Mean acceptance |
|---|---|---|
| Full gate+NN | −1382.805 | 0.840 |
| Linear+gate, no NN | −1382.877 | 0.812. |

I also evaluate both objectives at the two target means. At the full gate+NN target mean, the full objective is −1382.805307 and the linear+gate objective is −1382.799759. At the linear+gate target mean, the full objective is −1382.882119 and the linear+gate objective is −1382.876532. The fixed-theta differences are about 0.006 nats in absolute value.

This preliminary check confirmed the gate, inversion filter, observation scaling, state propagation, and HMC wiring as an end-to-end dynamic empirical pipeline. It does not show a meaningful incremental real-data gain from the NN residual correction over the same gate and inversion filter. The relevant artifact is `.local_artifacts/hlt_18param_validation_unitshock_mvp_20260620/`.

I then run the stricter filter-free gate-shock audit from `scripts/hlt_filter_free_shock_hmc_audit.jl`. This audit samples $\theta$ and the structural shocks in gate-active periods; calm periods keep the ROM1 shock path. The full gate+NN and linear+gate branches use the same payload, theta source, gate calibration, random seed, 100 adaptation draws, and 100 post-warmup draws. Both branches complete with zero divergences. Averaging over saved draws gives

| | Full gate+NN | Linear+gate, no NN |
|---|---|---|
| Observation log likelihood | 810.734 | 780.604 |
| Shock prior | −2510.759 | −2449.850 |
| Parameter prior plus Jacobian | 4.077 | 6.690 |
| Joint log posterior | −1695.948 | −1662.556. |

Thus the NN branch improves observation fit on its own sampled draws by 30.1 nats, but it uses larger shocks and has a 33.4-nat lower joint log posterior. The same-draw check is also local: on ROM1+NN draws, applying the NN correction raises the observation likelihood by 5.73 nats on average; on ROM1-only draws, it lowers it by 14.45 nats on average. Feature-support diagnostics remain a binding caveat: all 244 ROM1+NN target-mean periods are out of the audited support by the stored feature-$z$ diagnostic. The relevant artifact is `.local_artifacts/hlt_18param_unitshock_3day_coverage_20260622_143524/publication_empirical_blockers_20260709_filterfree_paper/hlt_filterfree_paper_summary.`

`md`.

# 11 Prior, Parameter Transform, and HMC

Each estimated parameter has finite support $[a_i, b_i]$. NUTS operates on an unconstrained vector $z \in \mathbb{R}^p$, with

$$\theta_i(z_i) = a_i + (b_i - a_i)\Lambda(z_i), \qquad \Lambda(z) = \frac{1}{1 + \exp(-z)}.$$

The log posterior in unconstrained coordinates is

$$\log \pi(z \mid y) = \ell_T(\theta(z)) + \log p_\Theta(\theta(z)) + \sum_{i=1}^{p} \log\{(b_i - a_i)\Lambda(z_i)[1 - \Lambda(z_i)]\}. \tag{7}$$

Gradients are computed by finite differences in the maintained HLT estimation script. The sampler is AdvancedHMC.jl's multinomial NUTS implementation. The reported diagnostics are divergent transitions, numerical-error proposals, acceptance rate, split $\widehat{R}$, and rank-normalized effective sample size [Vehtari et al., 2021].

**Protocol 1** (HMC comparability)**.** *When comparing direct and surrogate objectives in validation, the following quantities are held fixed: data, priors, bounded-to-unbounded transform, initial values, seeds, warmup length, post-warmup draws, finite-difference step, target acceptance rate, and tree-depth limit. A validation failure is therefore interpreted as an objective difference or numerical-stability difference, not as a sampler-design difference.*

# 12 Validation Package

The validation package has two complementary roles. The Galí ELB/OBC exercise validates pipeline layers in a small model where direct nonlinear evaluation is feasible. The reduced HLT bridge validates the residual target in the medium-scale model used for the main application.

## 12.1 Galí ELB/OBC Package

The Galí package uses the current repository's Galí OBC implementation and three observables: output, inflation, and the policy rate. It imposes an adverse $\varepsilon_z$ shock block that makes the ELB bind for multiple periods, with output and inflation falling during the binding window. The package contains five checks.

1. A stress-path check verifying that the actual-floor ELB path is economically meaningful.
2. A known-shock residual-grid check over $\sigma_z$.
3. A one-parameter inversion-grid check using common ROM1-recovered shocks.
4. A one-parameter matched-HMC check using direct and surrogate objectives.
5. A two-parameter matched-HMC diagnostic over $(\sigma_z, \sigma_a)$.

The extended two-parameter run is the most extensive Galí validation currently reported. It uses a 24-period actual-floor path, four chains per objective, and 4,000 post-warmup draws per objective. Direct and surrogate 90 percent intervals overlap for both parameters, both true values are covered, and posterior mean differences are below one combined-MCSE unit in absolute value. Probes involving $\sigma_\nu$ are retained as identification diagnostics: direct and surrogate objectives agree, but the parameter is pushed to the local grid boundary. Table 2 summarizes the reported checks and their interpretation.

The important point is not that every attempted Galí exercise succeeded. They did not. The $\sigma_\nu$ exercises showed a local identification problem. The full three-parameter direct-SEP HMC version was too expensive to promote. The current evidence is therefore stated as an ELB/OBC pipeline audit: it validates the residual target, ROM1 inversion layer, and matched-HMC comparison in a small model over the two shock scales that the design identifies.

## 12.2 Parameter-Matched HLT Residual

The main paper's correction benchmark begins with the 18-parameter SEP checkpoint used for the decomposition. The checkpoint stores 120 parameter vectors and 184 transitions per vector. Its original ROM1 output was evaluated at the baseline calibration, so I do not use that stored output for the matched test. Instead, `scripts/hlt_decomposition_normalization_audit.jl` re-solves ROM1 at every parameter vector. Three vectors have no stable first-order solution and remain recorded as failures, leaving 117 vectors and 21,528 common-support transitions.

I train an observable residual on 105 complete parameter vectors and reserve twelve complete vectors for validation. This group split prevents transitions from the same parameter vector from appearing on both sides. The validation set contains 2,208 transitions. Table 3 reports the result.

This is the paper's direct correction benchmark. It shows that the learned residual improves on plain parameter-matched ROM1 over a finite normal-state corridor. It is not a gate comparison, an inversion-filter comparison, or a direct nonlinear posterior.

Table 2: Galí ELB/OBC Validation Package

| Check | Status | Interpretation |
|---|---|---|
| Stress path | Pass | The actual-floor ELB path binds for six of 24 periods; output and inflation fall in the forced window. |
| Known-shock residual grid over $\sigma_z$ | Pass | Direct ELB/OBC and surrogate predictions are finite, and the holdout posterior comparison passes. |
| ROM1-inversion grid over $\sigma_z$ | Pass | The common inversion layer is finite for direct and surrogate objectives. |
| One-parameter matched HMC | Pass | Direct and surrogate HMC intervals overlap; mean difference is 0.2067 combined MCSE units. |
| Two-parameter grid over $(\sigma_z, \sigma_a)$ | Pass | Direct and surrogate surfaces are finite and cover the two targeted parameters. |
| Two-parameter matched HMC | Pass | Four chains per objective; 4,000 post-warmup draws per objective; interval overlap and true-value coverage for both parameters. |
| $\sigma_\nu$ probes | Expected failure | Direct and surrogate agree, but the short design does not identify the parameter away from the boundary. |
| Three-parameter direct-SEP check | Pass as a finite-objective check | Finite objective and finite-gradient check; not used as a posterior recovery result. |

*Notes*: "Pass" means the check satisfies the package-specific finite-objective, interval-overlap, or diagnostic criterion described in the surrounding text; it is not a claim about the full HLT posterior. "Actual-floor" denotes the simulated ELB-binding validation path. "OBC" denotes the occasionally-binding-constraint implementation. Matched HMC uses common data, priors, seeds, and chain design across direct and surrogate objectives. MCSE is Monte Carlo standard error.

Table 3: Held-Out Accuracy of the Parameter-Matched HLT Residual

| Observable | ROM1 RMSE | Corrected RMSE | Reduction |
|---|---|---|---|
| Output growth | 0.7727 | 0.3950 | 48.9% |
| Consumption growth | 2.3068 | 1.1050 | 52.1% |
| Investment growth | 5.2089 | 3.0128 | 42.2% |
| Hours | 0.3530 | 0.3226 | 8.6% |
| Inflation | 0.2829 | 0.2127 | 24.8% |
| Wage growth | 0.3973 | 0.1565 | 60.6% |
| Policy rate | 0.3051 | 0.2987 | 2.1% |
| Raw pooled RMSE | 2.1879 | 1.2373 | 43.4% |

*Notes*: Direct SEP, ROM1, and ROM1 plus the residual share the state, shock, and 18-parameter vector. ROM1 is ungated and is re-solved at each vector. The twelve validation vectors are absent from training. Raw pooled RMSE is coordinate dependent; the observable rows show that all seven series improve.

## 12.3 Fixed-Anchor Reduced HLT Bridge

The reduced HLT bridge targets the investment and risk-premium block

$$(\texttt{crhob}, \texttt{crhoqs}, \texttt{z\_eb}, \texttt{z\_eqs}),$$

with observables

$$\texttt{dy}, \quad \texttt{dc}, \quad \texttt{dinve}, \quad \texttt{labobs}, \quad \texttt{pinfobs}, \quad \texttt{dwobs}, \quad \texttt{robs}.$$

The maintained 10-by-10-by-10-by-10 supported grid solves all 10,000 direct SEP cells. The support report records SEP residual min/median/max equal to

$$2.700 \times 10^{-13}, \qquad 1.571 \times 10^{-9}, \qquad 9.996 \times 10^{-6}.$$

The stored bridge metadata records `rom_mode=baseline`. Direct SEP varies with the four grid parameters, while its ROM1 anchor remains at the baseline calibration. The posterior-grid comparison selects a held-out truth point, constructs a Gaussian observation scale, and summarizes the direct and surrogate criterion surfaces by discrete weights. It is therefore a test of a composite parameter-to-SEP emulator around a fixed linear anchor, not a parameter-matched SEP-minus-ROM1 residual.

The maintained result is:

$$\mathrm{RMSE}(\mathrm{ROM1}, \mathrm{SEP}) = 0.143574, \qquad \mathrm{RMSE}(\mathrm{sur}, \mathrm{SEP}) = 0.000749395,$$

and

$$\mathrm{RMSE}_{\log p}(\mathrm{ROM1}, \mathrm{SEP}) = 88.0719, \qquad \mathrm{RMSE}_{\log p}(\mathrm{sur}, \mathrm{SEP}) = 0.0534526.$$

Surrogate 90 percent intervals overlap direct SEP intervals for all four bridge parameters. This validates the configured fixed-anchor composite map on finite HLT support. It does not establish improvement over ROM1 re-solved at each grid point and does not validate full multi-period inversion-filter HMC for the 18-parameter empirical target. Tables 4 and 5 report the posterior-grid marginal comparison and the observation-level prediction errors.

Table 4: Reduced HLT Bridge: Posterior Marginals on the 10,000-Cell Grid

| Parameter | Truth | Direct mean | Surrogate mean | Common 90% interval |
|---|---|---|---|---|
| `crhob` | 0.583333 | 0.578232 | 0.578138 | [0.483333, 0.683333] |
| `crhoqs` | 0.711111 | 0.717703 | 0.717704 | [0.627778, 0.822222] |
| `z_eb` | 1.850000 | 1.687480 | 1.687460 | [1.416670, 1.850000] |
| `z_eqs` | 0.594444 | 0.595980 | 0.595913 | [0.411111, 0.777778] |

*Notes*: The grid contains $10^4$ supported cells over (`crhob`, `crhoqs`, `z_eb`, `z_eqs`). The truth point is held out from training. Direct and surrogate posterior summaries use the same discrete grid weights under the validation observation scale; the reported interval is the common 5th–95th percentile support implied by those weights.

Table 5: Reduced HLT Bridge: Prediction RMSE by Observable

| Observable | Fixed ROM1 RMSE | Surrogate RMSE | Improvement |
|---|---|---|---|
| `dy` | 0.193920 | 0.000923 | 99.52% |
| `dc` | 0.200742 | 0.001032 | 99.49% |
| `dinve` | 0.208037 | 0.001220 | 99.41% |
| `labobs` | 0.132777 | 0.000663 | 99.50% |
| `pinfobs` | 0.039289 | 0.000131 | 99.67% |
| `dwobs` | 0.031020 | 0.000109 | 99.65% |
| `robs` | 0.054567 | 0.000236 | 99.57% |

*Notes*: RMSE is computed against direct SEP observation targets on the HLT bridge grid, using the validation observable scale. Observables are output growth (`dy`), consumption growth (`dc`), investment growth (`dinve`), labor (`labobs`), inflation (`pinfobs`), wage growth (`dwobs`), and the policy rate (`robs`). The ROM1 column uses the fixed baseline anchor recorded in the bridge metadata. Improvement is $1 - \mathrm{RMSE}^{\mathrm{sur}}/\mathrm{RMSE}^{\mathrm{ROM1}}$.

## 12.4 How to Read the HLT Bridge

The reduced bridge shows that a supervised network can reproduce the direct SEP surface over 10,000 supported cells. Because its linear anchor is fixed, the fall in prediction RMSE

from 0.143574 to 0.000749395 should be read as fixed-anchor emulation accuracy. It is not the estimate of correction accuracy used in the main paper. Table 3 supplies that test with ROM1 re-solved at each parameter vector.

The bridge is not a full empirical posterior validation. It is one period, the features are known, the observation scale is fixed by the validation design, and the posterior is a discrete grid surface. The direct inference problem in the paper is harder: it is multi-period, uses recovered shocks, contains a gate, and samples an 18-dimensional parameter vector. The correct interpretation is narrower: the grid validates a four-parameter composite map, while the parameter-matched experiment validates the residual itself. Neither is a full inversion-filter HMC bridge.

A corrected dynamic bridge now exists as an explicit diagnostic. A four-period generator-path check passes on three nearby anchors. A harder eight-period, ten-anchor run shows why the dynamic correction must be path-aware: the fixed one-step NN residual keeps all direct-SEP objectives finite but fails the surface-shape criterion, with centered RMSE 1.872 for the surrogate and 1.395 for ROM1. The path-calibrated version fits a ridge residual on five dynamic anchors and evaluates five held-out anchors. It has held-out residual RMSE 0.0118, local MAP agreement with direct SEP, interval overlap for all four bridge parameters, and held-out centered surface RMSE 0.703 versus 1.437 for ROM1. This validates the reduced dynamic bridge, while leaving the full 18-parameter direct-SEP HMC comparison as a scaling target.

# 13 Empirical Pipeline Record

This section records the empirical runs used to interpret the 1959–2025 criterion comparison. The purpose is not to add more headline results. The purpose is to make the chain of evidence auditable.

## 13.1 Earlier Extended-Sample Objective Decomposition

This diagnostic predates the promoted matched current-ROM ablation. It is retained to document the earlier target comparison, not to identify the NN residual. The extended-sample comparison uses $T = 265$ quarters and seven observables. Let $Q_{\text{lin}}$ denote the exact ROM1/Kalman log likelihood evaluated at a fixed parameter vector, and let $Q_{\text{surr}}$ denote the regime-switching profiled observation-fit objective evaluated under the surrogate inversion architecture. These are criterion values, not two entries from a single marginal likelihood.

The maintained 2-by-2 criterion table is

| | $\theta_{\text{linear}}$ | $\theta_{\text{surrogate}}$ |
|---|---|---|
| Linear Kalman | $-1817.59$ | $-1852.63$ |
| Regime-switching surrogate | $-1830.44$ | $-1781.68$. |

The headline criterion gap is

$$-1781.68 - (-1817.59) = 35.92$$

nats. A symmetric decomposition assigns 29.05 nats, or 80.9 percent, to model change and 6.86 nats, or 19.1 percent, to parameter relocation. This is an accounting decomposition of the reported criteria only. It addresses one specific concern: the full criterion gap is not only the result of moving parameters to a different target mean. It does not answer the larger question of global posterior mass or Bayes factors, because the surrogate criterion omits the shock-prior, determinant/Jacobian, and Laplace terms discussed in Section 8.4, and because the surrogate result is still conditional on the reported warm-started basin. It also does not identify the NN residual as the source of the gain; the corrected linear+gate ablation in Section 10 finds no meaningful incremental real-data gain from the NN residual alone.

## 13.2 Four-Chain Current-ROM Headline Check

The promoted current-ROM check uses the full 1959Q1–2025Q1 sample ($T = 265$), all 18 HLT parameters, and four matched seeds per branch. Each chain has 500 warmup and 1,000 retained transitions, target acceptance 0.90, and maximum tree depth 8. ROM1 is re-solved at every proposal. The only branch difference is whether the target-guided NN residual enters at scale 0.25. The residual bundle metadata records `rom_mode=baseline`, while the HMC evaluator re-solves ROM1 at the current proposal. The branch comparison thus isolates the maintained correction inside the current-ROM evaluator, but it is not the ideal counterfactual in which the residual was trained against ROM1 at each training parameter vector.

Table 6 reports the full corrected-target marginals. These are summaries of the maintained profiled target, not credible intervals from a normalized likelihood posterior. They nevertheless show what the four-chain HMC run adds beyond a point estimate: joint uncertainty, simulation error, and the parameters that are hardest to explore.

For ROM1 plus NN, maximum $\widehat{R}$ is 1.0093 and minimum ESS is 653. For linear plus gate, the corresponding values are 1.0026 and 1,741. No parameter in either branch exceeds

Table 6: Corrected 18-parameter profiled-target marginals

| Parameter | Mean | 90% interval | $\widehat{R}$ | Bulk ESS | MCSE/SD |
|---|---|---|---|---|---|
| $\rho_a$ | 0.973 | [0.944, 0.989] | 1.0011 | 1906 | 0.020 |
| $\rho_b$ | 0.922 | [0.747, 0.988] | 1.0016 | 1583 | 0.023 |
| $\rho_g$ | 0.983 | [0.969, 0.990] | 1.0010 | 1819 | 0.019 |
| $\rho_{qs}$ | 0.930 | [0.821, 0.987] | 1.0016 | 1714 | 0.020 |
| $\rho_\pi$ | 0.936 | [0.749, 0.989] | 1.0064 | 865 | 0.033 |
| $\rho_w$ | 0.892 | [0.627, 0.986] | 1.0027 | 1365 | 0.029 |
| $\rho_{ms}$ | 0.466 | [0.033, 0.950] | 1.0020 | 978 | 0.030 |
| $\sigma_a$ | 4.417 | [3.531, 4.958] | 0.9999 | 2224 | 0.017 |
| $\sigma_b$ | 1.896 | [0.680, 4.167] | 1.0028 | 1791 | 0.026 |
| $\sigma_g$ | 4.185 | [3.137, 4.941] | 1.0001 | 1902 | 0.020 |
| $\sigma_{qs}$ | 3.851 | [2.472, 4.899] | 1.0030 | 1962 | 0.020 |
| $\sigma_\pi$ | 0.518 | [0.083, 1.769] | 1.0093 | 1067 | 0.030 |
| $\sigma_w$ | 1.736 | [0.399, 4.190] | 1.0040 | 1253 | 0.030 |
| $\sigma_{ms}$ | 2.889 | [1.521, 4.650] | 1.0003 | 2384 | 0.022 |
| $\xi_p$ | 0.839 | [0.618, 0.946] | 1.0090 | 957 | 0.031 |
| $\iota_p$ | 0.475 | [0.035, 0.963] | 1.0045 | 653 | 0.037 |
| $\varepsilon_p$ | 68.476 | [28.895, 109.171] | 1.0022 | 4707 | 0.015 |
| $\xi_w$ | 0.862 | [0.688, 0.945] | 1.0035 | 1564 | 0.025 |

*Notes*: Four chains, 500 warmup and 1,000 retained draws per chain. The interval is the pooled 5th–95th percentile range. MCSE/SD is the Monte Carlo standard error divided by the pooled marginal standard deviation. The target uses the parameter units maintained by the current-ROM HMC artifact.

$\widehat{R} = 1.01$, and neither branch records a current-ROM domain rejection. The geometry is nevertheless much harder with the residual correction: 3,163 of 4,000 retained transitions reach depth 8, and seed 20260722 records two numerical-error/divergent transitions. Linear plus gate has no retained numerical errors, no divergences, and no depth-8 transition.

At the four chain-specific target means, the mean profiled observation-fit criterion is $-393.049$ for ROM1 plus NN and $-392.946$ for linear plus gate. The matched NN-minus-linear differences are negative for every seed and average $-0.102$ nat. Thus the check establishes convergence within the maintained initialized basin and confirms that the current-ROM computation is executable at headline scale. Since the gate, inversion filter, observation scale, and current-ROM solution are identical across branches, the ablation assigns no incremental empirical fit to the maintained NN correction. It instead points to the shared HLT-style switching/inversion architecture, or its interaction with the profiled criterion, as the operative part of the maintained empirical target. That interpretation is implementation-specific: the maintained soft gate and profiled observation-fit calculation are not asserted to reproduce every element of HLT's published estimator. The check does not establish global mode uniqueness or equivalence with a direct-SEP posterior. The artifact and recovery record is stored under `.local_artifacts/hlt_18param_unitshock_3day_coverage_20260622_143524/hlt_18param_headline_T265_4chain_500x1000_20260710/`.

Figures 1–5 show the underlying diagnostics. The trace and rank plots check whether chains occupy the same within-basin distribution. The autocorrelation plot shows the persistence that drives effective sample size. The interval and correlation plots summarize the joint target. The energy panel reports HMC-specific exploration and makes the depth-cap saturation visible. These figures diagnose the sampler conditional on the selected transition. They are not tests of equilibrium uniqueness.

## 13.3 Mode Sensitivity

The extended-sample chain artifacts show two distinct surrogate regions. Four linear chains initialized from the calibrated interior cluster near $-1818$. Three cold-started surrogate chains remain near $-14535$, with small shock volatilities and lower investment persistence. Four warm-started surrogate chains initialized from the linear posterior mean reach a much higher objective, between $-1787.38$ and $-1780.80$ at chain means. The main paper therefore describes the nonlinear criterion comparison as conditional on the warm-started high-objective mode.

The compact chain-level record is:

1. Kalman, four interior chains: mean objectives from $-1817.97$ to $-1817.48$, zero diver-

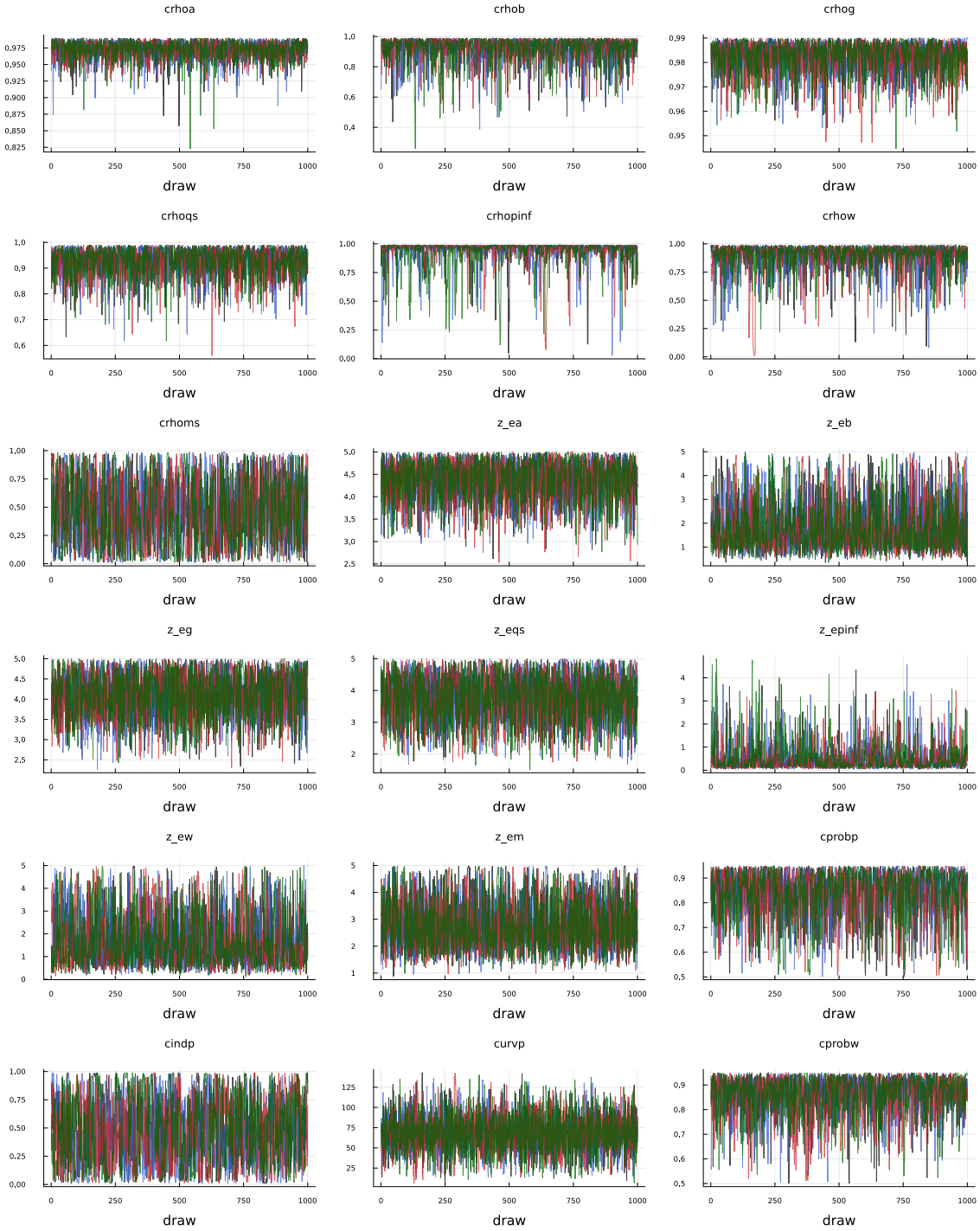



Figure 1: Corrected-target traces. Each color is one matched seed; each panel is one of the 18 sampled parameters.

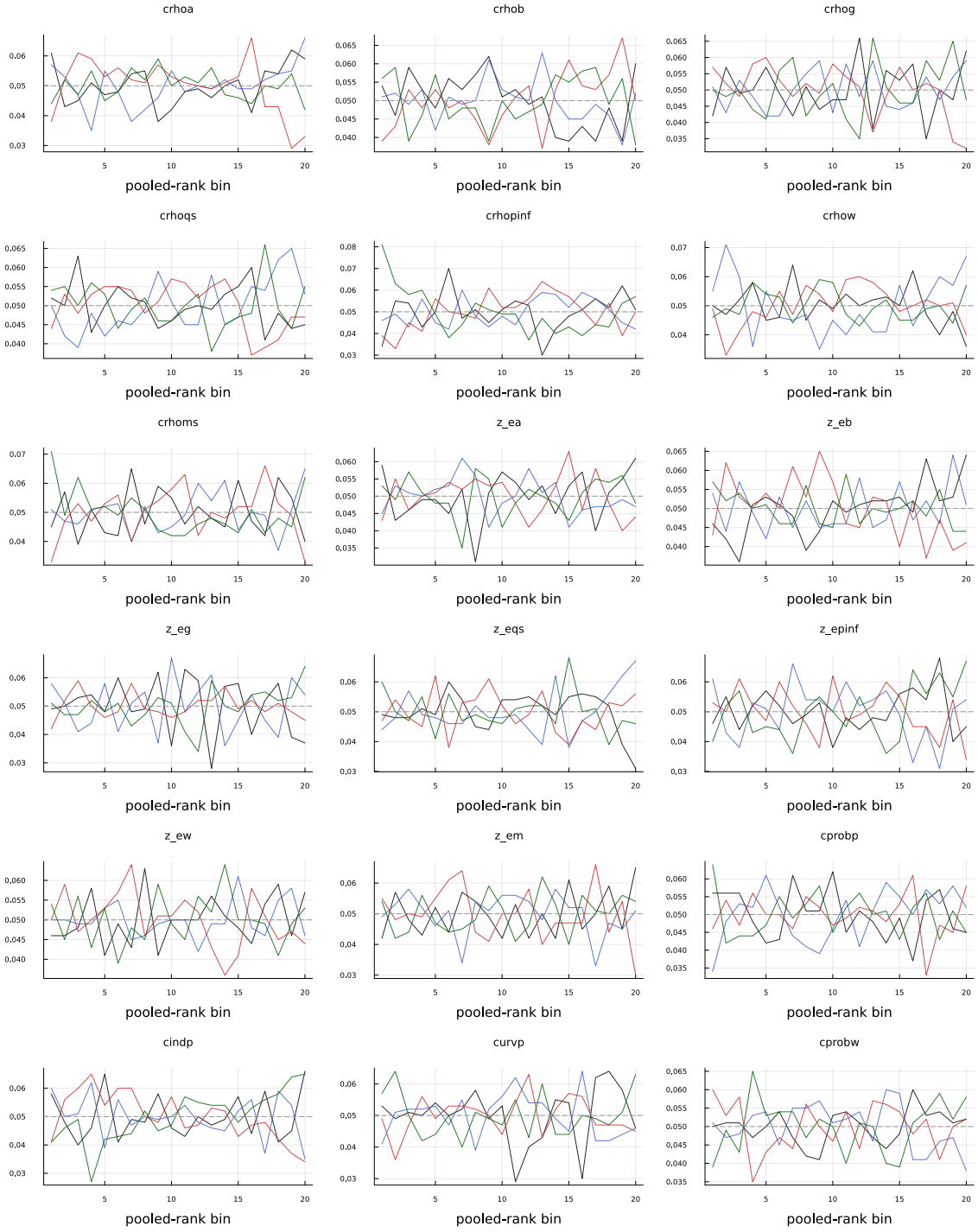



Figure 2: Corrected-target rank plots. Flat, overlapping chain ranks are the visual counterpart to the reported split-rank convergence statistics.

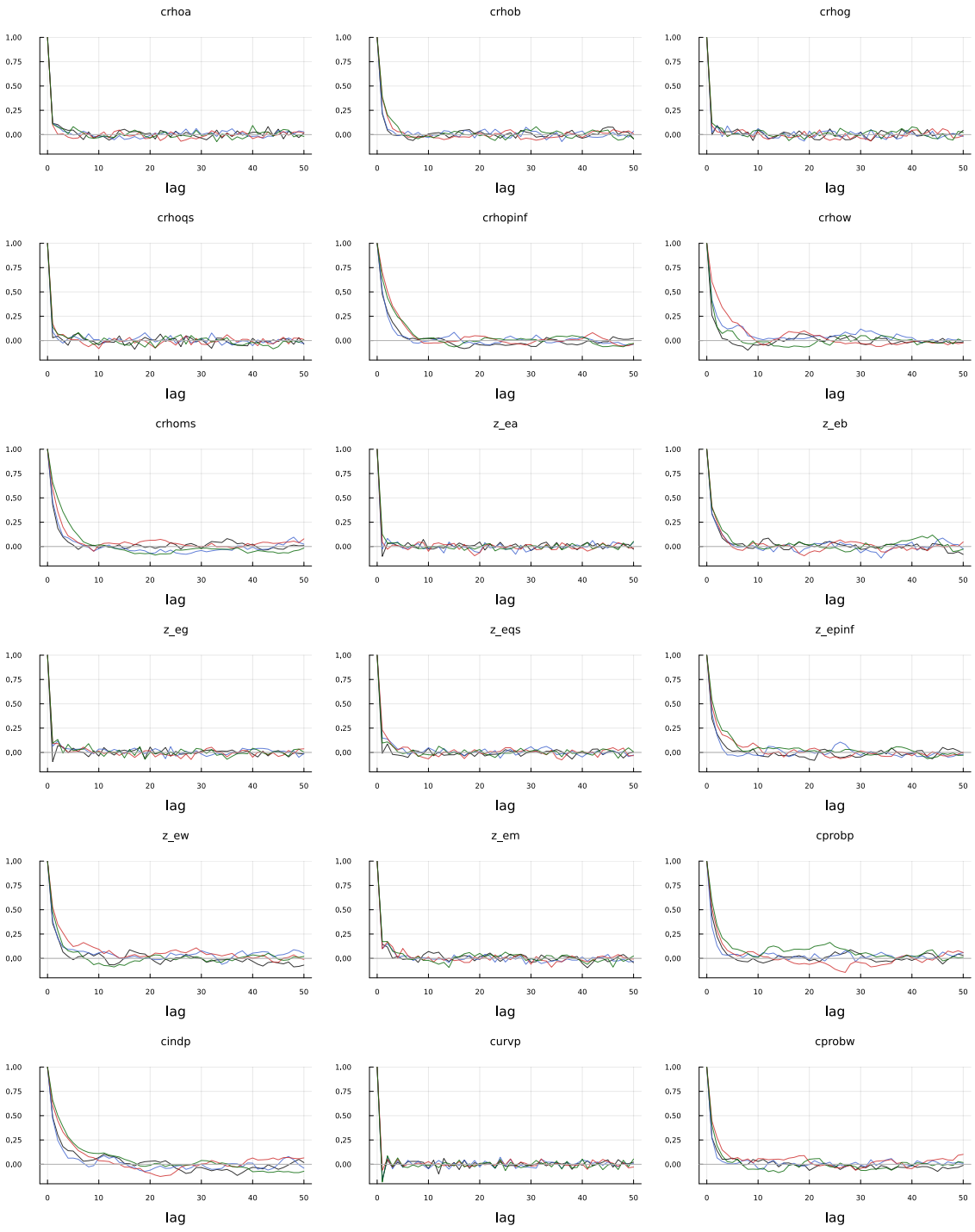
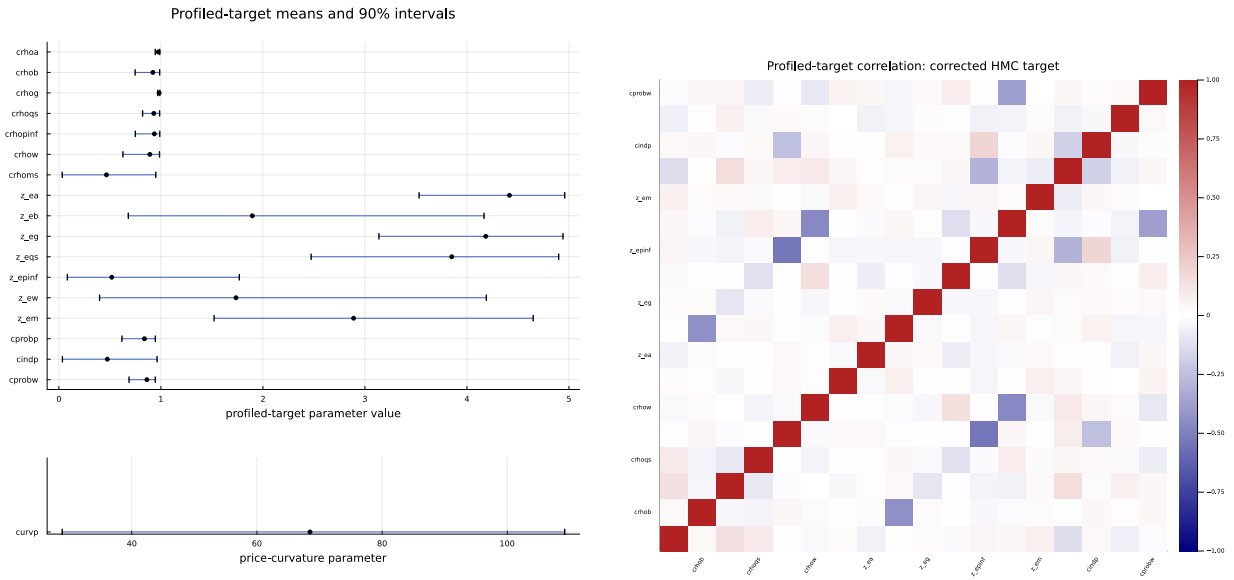



Figure 3: Chain-specific autocorrelation functions for the corrected target.




(a) Means and 90 percent intervals.

(b) Pooled target correlations.

Figure 4: Marginal and joint geometry of the corrected profiled target.

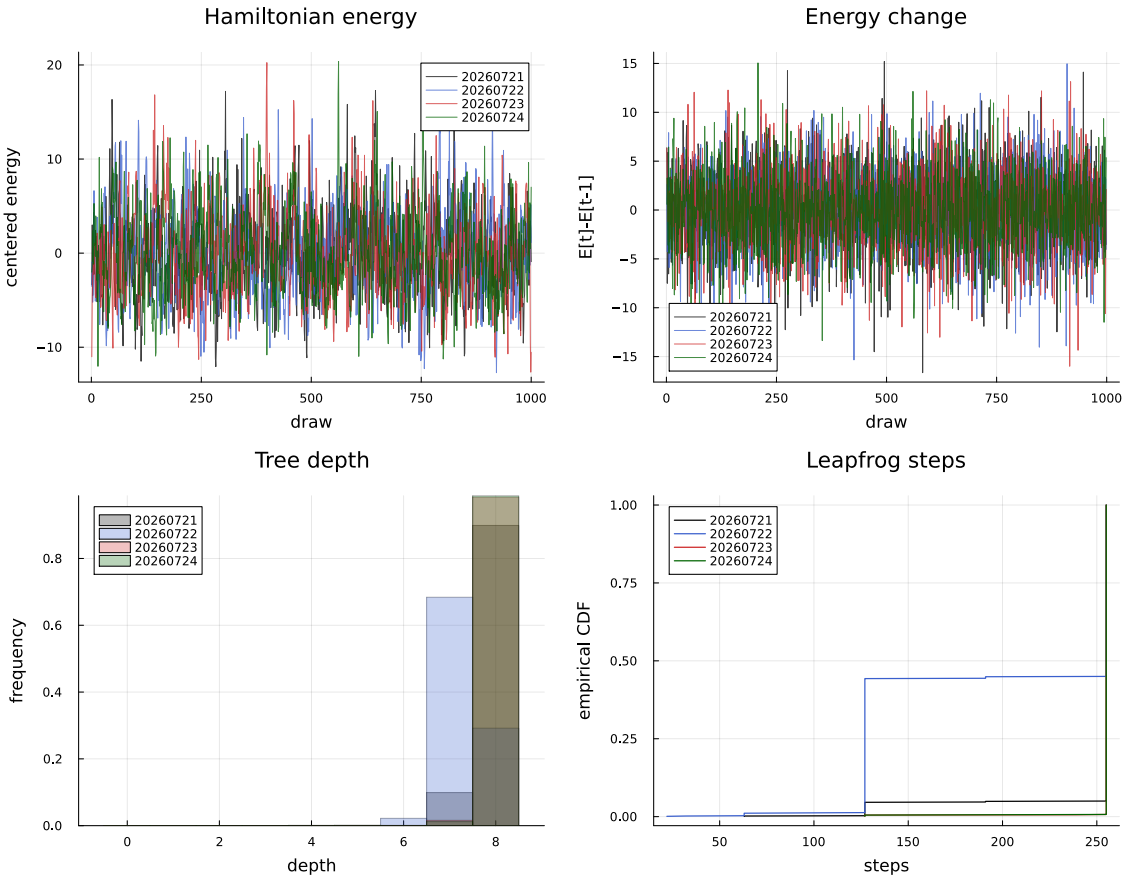



Figure 5: Corrected-target HMC energy, energy changes, tree depth, and leapfrog steps. The tree-depth panel shows the concentration at the maintained cap of 8.

gences, $\sigma_b = 0.166$–$0.179$, $\rho_b = 0.731$–$0.764$, and $\rho_w = 0.157$–$0.179$.

2. Surrogate cold starts, three interior chains: mean objective $-14534.52$, zero divergences, $\sigma_b = 0.094$–$0.098$, $\rho_b = 0.606$–$0.610$, and $\rho_w = 0.606$–$0.622$.

3. Surrogate warm starts, four linear-mean chains: mean objectives from $-1787.38$ to $-1780.80$, zero divergences, $\sigma_b = 0.091$–$0.092$, $\rho_b = 0.897$–$0.901$, and $\rho_w = 0.226$–$0.236$.

The cold-start result is a warning, not a secondary headline. It means the surrogate objective has difficult geometry. It also means that claims about global posterior dominance or posterior mass would require a broader mode-search exercise. The paper does not make that claim.

## 13.4 ARMA Markup Confound

The HLT baseline suppresses the ARMA(1,1) price- and wage-markup shock terms used in the original Smets–Wouters specification. This matters because the markup block is known to create flat posterior targets in the full model [Farkas and Tatár, 2020]. The local artifact comparing HLT with the restored markup MA terms shows an improvement of about 34 nats in the linear likelihood on the baseline sample. That number is close to the 35.92-nat extended-sample surrogate objective gap. The implication is limited: the reported parameter shifts are criterion-based evidence relative to the HLT benchmark, not a universal statement about every Smets–Wouters specification.

## 13.5 Quiet-Sample Forecasting and Gate Recalibration

The quiet-sample out-of-sample exercise estimates through 1994Q4 and evaluates 1995Q1–2007Q4. It is a gate-design diagnostic. Under the original gate, the switching surrogate classifies too many calm periods as stressed and performs poorly. Holding posterior draws fixed and recalibrating the gate to a q95 padded rule reduces the full-window RMSE ratio to 1.04, with a hard-gate share of 1.9 percent. Re-estimating the surrogate chain under the q95 padded gate and evaluating with the full gate gives:

$$\text{early quiet ratio} = 0.70, \qquad \text{full OOS ratio} = 1.09, \qquad \text{remaining ratio} = 1.19.$$

Thus the failure is localized but not eliminated. The gate can be made much less damaging in calm periods, and it improves the early quiet window, but it does not beat ROM1 over the full calm holdout.

The diagnostic rows are:

1. Base no padding: full ratio 1.180, early ratio 1.476, remaining ratio 1.062, hard-gate share 1.9 percent.
2. q95 padded, fixed posterior: full ratio 1.038, early ratio 1.072, remaining ratio 1.027, hard-gate share 1.9 percent.
3. q95 padded, re-estimated, full gate: full ratio 1.090, early ratio 0.700, remaining ratio 1.190, hard-gate share 1.9 percent.

## 13.6 SEP Sensitivity

The maintained SEP sensitivity table comes from `.local_artifacts/sep_sensitivity/full_265_20260521/`. It supersedes the earlier three-draw, 40-period pilot. The rerun covers the full 265-period extended payload for five posterior draws selected by Mahalanobis distance from the pooled surrogate chains. It evaluates

$$\texttt{accept\_tol} \in \{0.01, 0.1, 0.35\}, \qquad K \in \{3, 5\}.$$

Table 7 reports the reconciled provenance, convergence rate, Euler residual, and prediction-error distance from the $(\texttt{accept\_tol}, K) = (0.01, 5)$ reference cell.

Table 7: SEP Solver Sensitivity to Acceptance Tolerance and Quadrature Nodes

| Tolerance | $K$ | Convergence | Periods | Max Euler error | $\log_{10}$(error) | RMSE vs. ref. |
|---|---|---|---|---|---|---|
| 0.01 | 3 | 1.00 | 265 | $9.87 \times 10^{-8}$ | -7.01 | $1.08 \times 10^{-2}$ |
| 0.10 | 3 | 1.00 | 265 | $9.87 \times 10^{-8}$ | -7.01 | $1.08 \times 10^{-2}$ |
| 0.35 | 3 | 1.00 | 265 | $9.87 \times 10^{-8}$ | -7.01 | $1.08 \times 10^{-2}$ |
| 0.01 | 5 | 1.00 | 265 | $9.74 \times 10^{-8}$ | -7.01 | ref. |
| 0.10 | 5 | 1.00 | 265 | $9.74 \times 10^{-8}$ | -7.01 | 0.00 |
| 0.35 | 5 | 1.00 | 265 | $9.74 \times 10^{-8}$ | -7.01 | 0.00 |

*Notes*: Runs use five posterior draws stratified by Mahalanobis distance from the pooled warm-started surrogate HMC chains (seeds 42, 5, 6, 7). "Convergence" is the fraction of draws where SEP completed all 265 periods. "Max Euler error" is the maximum SEP residual ($\ell_\infty$ norm of the model's first-order-condition residuals) across all periods, averaged over draws. "RMSE vs. ref." is the root-mean-squared observation prediction error against the reference cell (acceptance tolerance = 0.01, $K = 5$), pooled over draws and 265 periods. The source artifact is `.local_artifacts/sep_sensitivity/full_265_20260521/`; this table supersedes the earlier three-draw, 40-period pilot.

All six cells converge for all selected periods. The $K = 5$ cells are identical to the reference at the reported precision. Moving from $K = 5$ to $K = 3$ changes the observation path, with aggregate RMSE $1.08 \times 10^{-2}$ against the reference and the largest per-variable RMSE in labor. The result is not that SEP is completely insensitive to every setting. The result is

that the reference quadrature choice is stable across tolerance values in this rerun, and the documented $K = 3$ versus $K = 5$ difference is now explicit.

The investment-shock benchmark adds checks that the earlier table did not contain. It holds the HLT calibration and one-standard-normal investment shock fixed, then varies the path horizon, terminal approximation, and linear algebra used inside the same nonlinear solver. Table 8 compares each design with the 24-quarter, first-order-terminal reference.

Table 8: SEP Horizon, Terminal Rule, and Linear-Algebra Checks

| Design | Scaled RMSE | Investment RMSE | Max residual | Seconds |
|---|---|---|---|---|
| H=12 | 0.0769 | 0.0925 | 8.52e-08 | 19.0 |
| H=18 | 0.0451 | 0.0402 | 6.91e-08 | 29.1 |
| H=24 | 0.0000 | 0.0000 | 7.23e-08 | 22.2 |
| terminal order 0 | 0.0032 | 0.0006 | 3.24e-08 | 3.5 |
| QR linear solve | 0.0000 | 0.0000 | 9.96e-08 | 122.9 |

*Notes*: The model, calibration, initial state, shock, and twelve reported response quarters are fixed. Scaled RMSE normalizes each response by its root mean square under the 24-quarter reference. The QR row changes only the internal linear solve; it is not an independent nonlinear solution method. The benchmark uses the smooth HLT system without an active lower-bound equation. All maximum equilibrium residuals are below $10^{-7}$.

The horizon matters more than the terminal rule in this experiment. A 12-quarter horizon has scaled RMSE 0.0769 against the 24-quarter path; extending it to 18 quarters lowers the discrepancy to 0.0451. At horizon 24, replacing the first-order terminal rule with a steady-state terminal condition changes the scaled path by 0.0032. Solving the same Newton step by QR rather than normal equations changes it by $1.5 \times 10^{-5}$. These checks support the maintained terminal and algebra choices, but they also show why a short horizon should not be treated as numerically interchangeable with the reference.

## 13.7 Euler-Residual Diagnostic

The Euler-residual artifact evaluates residuals along realized trajectories at the extended-sample posterior mean. It is a useful numerical diagnostic but is not the main accuracy criterion for the empirical criterion comparison. In that artifact, the SEP trajectory has lower maximum residual than ROM1, while the ROM1+NN trajectory can have larger equation residuals, especially when the NN correction is forced everywhere. This does not overturn the paper's prediction RMSE or posterior-grid evidence because the NN is trained to correct observation-space transition errors on audited support, not to solve the full nonlinear equilibrium system when applied outside its gate.

### 13.8 Runs Not Promoted

Several artifacts are retained precisely because they prevent overstatement.

1. The HLT legacy three-parameter direct-SEP/MH runs are finite-objective checks. The reported production threshold is false, sample sizes are one draw in the maintained summaries, and interval comparisons are not meaningful.
2. The full Galí three-parameter direct-SEP HMC run is a scaling target. The $\sigma_\nu$ probes indicate local identification failure in the short ELB/OBC design, not a surrogate failure.
3. The original one-cell pruned-ROM2 pilot is superseded by the common-design HLT benchmark in the main paper. Its historical artifact remains unchanged. The production comparison includes three shock sizes and both pruned second and pruned third order under common inputs.
4. The q95 re-estimated quiet-sample diagnostic must be read with the correct gate-evaluation file. The full-gate evaluation gives the 1.09 ratio; a mismatched all-gate evaluation gives a much worse 3.05 ratio and is not the promoted diagnostic.

## 14 Acceptance Criteria

Table 9 states the criteria used to promote diagnostics from local artifacts into the paper record.

## 15 Output Files and Replication Entry Points

### 15.1 Main Scripts

The maintained scripts are organized by what they do. The exact file names are:

```
scripts/hlt_sep_surrogate_dataset_generate.jl
scripts/hlt_sep_surrogate_train.jl
scripts/hlt_parameter_matched_residual_dataset.jl
scripts/hlt_parameter_matched_residual_report.jl
scripts/hlt_investment_solution_benchmark.jl
scripts/hlt_investment_policy_counterfactual.jl
scripts/hlt_bridge_support_report.jl
scripts/hlt_bridge_posterior_grid_compare.jl
```

Table 9: Validation Criteria by Layer

| Layer | Criterion |
|---|---|
| SEP corridor support | Direct nonlinear cells are finite and either all solve on the reported support or failures are explicitly mapped and the support is trimmed; the result is a finite-domain corridor certificate, not a global stability proof. |
| Training payload | $X$, $Y$, $Y^{\mathrm{ROM1}}$, metadata, observable indices, parameter names, solver residuals, and git provenance are saved. |
| Surrogate accuracy | Held-out RMSE of $Y^{\mathrm{ROM1}} + \widehat{\Delta}$ is reported against direct SEP; ROM1 baseline RMSE is reported on the same split. |
| Posterior grid | Direct SEP and surrogate posterior intervals overlap for all reported bridge parameters; surrogate surface RMSE is below ROM1 surface RMSE. |
| Matched HMC | Direct and surrogate intervals overlap, posterior mean differences are within combined Monte Carlo error for the targeted parameters, and numerical-error proposals are zero or explicitly negligible. |
| Empirical use | Any paper table has a raw payload, human-readable summary, and provenance note. |

*Notes*: These are promotion criteria and audit rules, not empirical estimates. The Galí ELB/OBC package and the reduced HLT bridge meet the criteria stated for their respective layers; the full 18-parameter multi-period direct-SEP HMC bridge remains a scaling target rather than a passed criterion.

`scripts/hlt_bridge_robustness_sweep.jl`

`scripts/hlt_endogenous_support_dataset.jl`

`scripts/run_hlt_endogenous_support_retrain.sh`

`scripts/hlt_filter_free_shock_hmc_audit.jl`

`scripts/run_surrogate_hmc_advancedhmc.jl`

`scripts/gali_validation_package.jl`

`scripts/gali_obc_actual_floor_twoparam_inversion_grid_validation.jl`

Their roles are:

**Dataset generation**
: Generate direct SEP and ROM1 training payloads.

**Surrogate training**
: Train the ROM1-residual surrogate.

**Parameter-matched residual test**
: Re-solve ROM1 at each 18-parameter vector, hold out complete vectors, and report the correction relative to the resulting matched first-order path.

**Solution-order benchmark**
: Compare first, pruned second, and pruned third order with SEP under one HLT calibration, initial state, shock, and response window. The same script also checks the SEP horizon, terminal rule, and internal linear solve.

**Investment policy experiment**
: Compare baseline and stronger output responses after the same adverse investment shock under SEP and ROM1.

**Support report**
: Convert dataset metadata into finite-support reports.

**Grid comparison**
: Compare direct SEP, ROM1, and surrogate posterior-grid surfaces.

**Bridge sweep** Run the HLT bridge dataset, support report, training, and posterior-grid comparison in one command.

**Target-guided support**
: Convert HMC-visited gate-period inputs into additional one-step SEP labels and a combined retraining dataset.

**Target-guided retrain controller**

Run active labeling, retraining, a short HMC check, and the matched gated shock–theta audit.

**Target-guided validation**

Validate scaled NN residual corrections on SEP labels from posterior support points.

**Filter-free shock audit**

Sample shocks directly in selected gate periods, optionally together with parameters, to test whether the NN correction improves the objective once shocks are no longer profiled out everywhere.

**Empirical HMC**

Run empirical HLT surrogate NUTS-HMC with inversion filtering and gate support.

**Galí package audit**

Check the maintained Galí ELB/OBC validation output files.

**Galí HMC diagnostic**

Run the two-parameter actual-floor inversion/HMC diagnostic.

## 15.2 Output File Layout

For an HLT bridge run with identifier $r$, the expected layout is

```
.local_artifacts/hlt_reduced_bridge_validation/r/
  hlt_sep_surrogate_dataset.jls
  SUPPORT_REPORT.md
  hlt_sep_surrogate_trained_r.jls

.local_artifacts/hlt_reduced_bridge_validation/posterior_grid_compare_r/
  hlt_bridge_posterior_grid_comparison.jls
  SUMMARY.md
  comparison_table.tex
  ROBUSTNESS_SWEEP.md
```

For the Galí package, the consolidated audit is written under

```
.local_artifacts/gali_validation_package/
```

and individual layer outputs remain under their producing run directories.

## 15.3 Artifact Ledger

The following ledger records the artifacts that currently support the paper-facing numerical claims. Paths are relative to the repository root.

**HLT decomposition normalization audit.** `.local_artifacts/`

```
hlt_decomposition_normalization_audit/
  hlt_decomposition_normalization_audit.jls
  SUMMARY.md
  block_shares.csv
```

is produced by `scripts/hlt_decomposition_normalization_audit.jl`. It re-solves ROM1 at every stored parameter vector, reproduces the withdrawn baseline-ROM accounting as a reconstruction check, records the three unstable ROM1 parameter vectors, and reports every retained normalization on the same 21,528 observations.

**HLT parameter-matched residual validation.** `.local_artifacts/`

```
hlt_parameter_matched_residual/
production_20260802/
HLT_PARAMETER_MATCHED_RESIDUAL_VALIDATION.md
```

is produced by `scripts/hlt_parameter_matched_residual_dataset.jl`, `scripts/hlt_sep_surrogate_train.jl--split-by-theta`, and `scripts/hlt_parameter_matched_residual_report.jl`. ROM1 is re-solved at each parameter vector. The validation split holds out twelve complete vectors, or 2,208 transitions. Raw pooled observable RMSE falls from 2.1879 to 1.2373, and all seven observable errors fall.

**HLT solution-order and SEP-design benchmark.** `.local_artifacts/`

```
hlt_investment_solution_benchmark/
production_20260802/
HLT_INVESTMENT_SOLUTION_BENCHMARK.md
```

is produced by the solution-order benchmark script listed above. It compares ROM1, pruned ROM2, and pruned ROM3 with SEP at three investment-shock sizes. It also holds the experiment fixed while changing the SEP horizon, terminal rule, and internal linear solve.

**HLT investment policy experiment.** `.local_artifacts/`

```
hlt_investment_policy_counterfactual/
production_20260802/
HLT_INVESTMENT_POLICY_COUNTERFACTUAL.md
```

is produced by the investment policy script listed above. It holds the model, initial state, and adverse investment shock fixed and changes only the Taylor-rule response to the output gap. The artifact reports transition paths, not welfare or an optimal rule.

**HLT fixed-anchor bridge support.** `.local_artifacts/hlt_reduced_bridge_validation/investment4p_supported_grid10_full_sweep_20260527/SUPPORT_REPORT.md`

records the 10,000-cell support map, zero failures, and SEP residual min/median/max. The bridge varies the SEP side over four parameters while ROM1 remains at the baseline calibration. It therefore supports a finite-support composite-map benchmark, not a parameter-matched residual claim.

**HLT bridge posterior grid.** `.local_artifacts/hlt_reduced_bridge_validation/posterior_grid_compare_<run_id>/SUMMARY.md`

with run id `investment4p_supported_grid10_full_sweep_20260527`. It records the direct, fixed-anchor ROM1, and surrogate grid comparison. This supports the configured composite-map prediction-RMSE improvement from 0.143574 to 0.000749395 and the surface-RMSE improvement from 88.0719 to 0.0534526. It is not used as the parameter-matched correction benchmark.

**HLT high-Kimball neighborhood stress.** `.local_artifacts/kimball_curvature_sensitivity/hlt_reported_full_mapped_maxit400_20260606/KIMBALL_CURVATURE_SENSITIVITY_SUMMARY.md`

records a fixed-calibration decomposition at the Harding et al. [2022] high-curvature point $(\varphi_p, \xi_p, \varepsilon_p) = (1.20, 0.667, 77.3)$, with reported matching core and shock-process parameters mapped into the maintained code where the repo has a direct analogue. In that fixed raw accounting, the investment/capital block is largest. The result is a stress-point diagnostic and is not used as a normalization-invariant decomposition.

**HLT posterior-region local ablation.**
`.local_artifacts/counterfactual_decomposition/hlt_posterior_region_local10_40draws_scales012505_queued_20260609/COUNTERFACTUAL_ABLATION_SUMMARY.md` records the posterior-region local 10% ablation. The run uses 40 representative draws from the pooled 8,000-draw HLT surrogate

posterior and shock scales 0.1, 0.25, and 0.5. All SEP paths converge. Kimball curvature is the largest local log-MSE sensitivity in 5%, 0%, and 0% of posterior draws across the three scales, while investment-adjustment and capital-utilization curvature are largest in 95%, 100%, and 100%. This supports the statement that nonlinear Phillips-curve curvature is not the dominant local source of the SEP–ROM1 gap in the estimated posterior region.

**HLT curvature-surface figures.**

`.local_artifacts/hlt_posterior_mean_curvature_surface/`
`sw07_hlt_baseline_surface_13x13_std1_shock01_1path_20260611/`
`sw07_hlt_baseline_surface_13x13_std2_bounded_shock01_1path_20260611/`
`hlt_highkimball_surface_13x13_std1_shock01_1path_20260611/`
`hlt_highkimball_surface_13x13_std2_shock01_1path_20260611/`
`CURVATURE_SURFACE_SUMMARY.md`
`docs/SurrogateNN_paper/figures/fig_hlt_curvature_surface_*_13x13_std*_shock01.png` records the figure-level local surfaces. The metric is the RMSE of one-step observable forecast errors, SEP minus ROM1, evaluated at the same SEP state, shocks, and parameter vector. Each figure uses common shock paths across cells, a 13-by-13 grid, and standard-deviation coordinates. For parameters estimated in the HLT chain, the scale is the posterior or prior scale reported in the configuration; for fixed calibration parameters, the scale is a documented local calibration proxy. The baseline $\pm 1$ s.d. ranges, in RMSE times 1,000, are 11.4 for the real-side surface and 1.4 for the Kimball surface. The baseline $\pm 2$ s.d. ranges are 23.5 and 2.6. At the high-Kimball stress center, the corresponding ranges are 37.0 and 54.7 for $\pm 1$ s.d., and 53.9 and 210.6 for $\pm 2$ s.d. All cells in the high-Kimball surfaces are accepted; one isolated baseline Kimball cell fails ROM1 construction and is filled only for plotting, while the raw CSV records the failure.

**Target-guided support refinement.**

Let `RUN_ROOT` denote `.local_artifacts/hlt_18param_unitshock_3day_coverage_20260622_143524`.

```
RUN_ROOT/
  publication_empirical_blockers_20260709_support_paper_.../
  ENDOGENOUS_SUPPORT_SUMMARY.md
  POSTERIOR_GUIDED_SUPPORT_REFINEMENT_VALIDATION.md
  hlt_endogenous_support_dataset.jls
```

```
hlt_sep_surrogate_dataset_with_endogenous_support.jls
hlt_sep_surrogate_trained_endogenous.jls

publication_empirical_blockers_20260709_filterfree_paper/
hlt_filterfree_paper_summary.md
hlt_filterfree_paper_fullnn.jls
hlt_filterfree_paper_lineargate.jls

endogenous_support_gate113_20260702/
ENDOGENOUS_SUPPORT_SUMMARY.md
hlt_endogenous_support_dataset.jls
hlt_sep_surrogate_dataset_with_endogenous_support.jls
hlt_sep_surrogate_trained_endogenous.jls
POSTERIOR_GUIDED_SUPPORT_REFINEMENT_VALIDATION.md

endogenous_support_drawcloud_zero_pass3_20260703/
POSTERIOR_GUIDED_SUPPORT_REFINEMENT_VALIDATION.md

posterior_guided_support_holdout_pass3full_20260703/
ENDOGENOUS_SUPPORT_SUMMARY.md
POSTERIOR_GUIDED_SUPPORT_REFINEMENT_VALIDATION.md
```

records the first and draw-cloud support-refinement passes from the gated shock–theta HMC audit. The first pass labels all 113 hard-gate periods with one-step SEP, repeats those labels in the retraining set, and retrains the observation-only ROM1-residual surrogate. The draw-cloud pass labels 250 draw-period inputs from the posterior region. In the promoted July 9 run, the old surrogate has residual RRMSE 0.988 on these one-step labels. The refined surrogate lowers residual RRMSE to 0.019 and reduces one-step RMSE by 98.1 percent relative to ROM1 at its best gate scale. This validates the local residual-learning technology on target-guided SEP labels. It does not by itself support a final empirical likelihood claim; the matched post-retraining filter-free audit is mixed and remains outside the audited feature support.

**Galí validation package.** `.local_artifacts/gali_validation_package/gali_validation_package_20260521/VALIDATION_PACKAGE_REPORT.md`

records the passing small-model ELB/OBC pipeline audit.

**Objective decomposition.**

`.local_artifacts/ll_decomposition/LL_DECOMPOSITION_SUMMARY.md` records the 35.92-nat gap and the 80.9/19.1 symmetric decomposition.

**Linear+gate ablation.**

`.local_artifacts/ll_decomposition/LINEAR_GATE_ABLATION_SUMMARY.md` records the fixed-parameter no-NN ablation. It is internal evidence on the gate and residual correction, not a posterior table.

**Mode sensitivity.**

`.local_artifacts/mode_sensitivity/MODE_SENSITIVITY_SUMMARY.md` records the chain-level mode table and the warm-start caveat.

**Gate recalibration.**

`.local_artifacts/oos_gate_recalibration/OOS_GATE_RECALIBRATION_SUMMARY.md` records the fixed-posterior quiet-sample gate-policy diagnostic.

**Quiet OOS re-estimation.** `.local_artifacts/oos_forecast/OOS_SUMMARY_quiet_1994_q95_reestimated_fullgate.md`

records the q95 padded re-estimation diagnostic: 0.70 early-window ratio, 1.09 full-window ratio, and 1.19 remaining-window ratio.

**SEP sensitivity.** `.local_artifacts/sep_sensitivity/full_265_20260521/SEP_SENSITIVITY_SUMMARY.md`

records the full 265-period SEP sensitivity rerun.

**Euler residuals.**

`.local_artifacts/euler_errors/EULER_ERRORS_SUMMARY.md` records the extended-sample Euler-residual diagnostic. It is not a likelihood-accuracy headline.

**Pruned ROM2 pilot.**

`.local_artifacts/pruned_rom2_benchmark/PRUNED_ROM2_BENCHMARK_SUMMARY.md` records a failed pilot and is explicitly not promoted.

The ledger should be updated whenever a number moves from an artifact into the paper. If a future result has no raw payload, no human-readable summary, or no clear scope statement, it should stay out of the submission draft.

## 15.4 Replication Commands

The parameter-matched residual benchmark is built, trained, and reported with:

```
julia --project=. scripts/hlt_parameter_matched_residual_dataset.jl \
  --out=<matched_dataset.jls>
julia --project=. scripts/hlt_sep_surrogate_train.jl \
  <matched_dataset.jls> \
  --epochs=400 --hidden=256 --hidden2=128 --batch=512 \
  --seed=20260802 --rom-residual=1 --obs-only \
  --split-by-theta --arch=mlp --out=<matched_surrogate.jls>
julia --project=. scripts/hlt_parameter_matched_residual_report.jl \
  --bundle=<matched_surrogate.jls> --report-dir=<report_dir>
```

The common HLT solution-order and policy exercises are:

```
julia --project=. scripts/hlt_investment_solution_benchmark.jl \
  --periods=12 --shock-sizes=0.5,1.0,2.0 \
  --reference-horizon=24 --include-third=true
julia --project=. scripts/hlt_investment_policy_counterfactual.jl \
  --periods=16 --shock-size=-1.0 --sep-horizon=24 \
  --strong-cry=0.25
```

The dense HLT bridge is expensive but has a single orchestration entry point:

```
julia --project=. scripts/hlt_bridge_robustness_sweep.jl \
  --stage=full \
  --run-id=<run_id> \
  --param-set=investment_4p_supported \
  --grid=10 \
  --epochs=400 \
  --hidden=256 \
  --hidden2=128 \
  --obs-sigma-scale=1.0 \
  --dgp-noise-scale=0.25
```

The Galí validation package audit is:

```
julia --project=. scripts/gali_validation_package.jl \
  --run-id=<run_id> \
  --strict=true
```

The empirical HLT surrogate chain is run through:

```
julia --project=. scripts/run_surrogate_hmc_advancedhmc.jl \
  --surrogate=<surrogate_bundle.jls> \
  --data=<data_payload.jls> \
  --gate-calibration=<gate_payload.jls> \
  --gate-mode=soft \
  --out=<chain_payload.jls>
```

The linear+gate ablation uses the same command with `-disable-nn-correction=true`.

# 16 Interpretation Rules

The pipeline deliberately separates approximation accuracy from economic interpretation.

1. A low residual RMSE means the network tracks direct SEP on the audited support. It does not by itself prove correctness of the empirical target away from that support.
2. A passing Galí two-parameter HMC diagnostic validates the ROM1-residual, inversion, and matched-HMC layers in an ELB/OBC small model. It does not establish that the 18-parameter HLT posterior has been globally explored.
3. The parameter-held-out HLT experiment validates the residual target against ROM1 resolved at each vector. The separate 10,000-cell bridge validates a fixed-anchor composite map. Neither supplies a full multi-period direct-SEP HMC benchmark.
4. Exact Kalman likelihood entries and profiled inversion-objective entries are reported as criterion values under the maintained normalization. They should not be described as one marginal likelihood or as a Bayes factor.
5. The 35.92-nat extended-sample criterion gap is an earlier conditional comparison of unlike statistical objects. It is not evidence that the NN residual alone improves HLT's published likelihood.
6. The promoted matched current-ROM ablation isolates the maintained NN correction under a common profiled criterion. Its negative NN-minus-linear differences assign no incremental observation fit to that correction in the maintained basin. Because the residual bundle records baseline-ROM training while the evaluator uses current ROM1, this is not a clean test of a parameter-matched residual-training design.
7. Warm-started HMC diagnostics certify sampling within the reported basin. They do not rule out other posterior basins unless a separate mode search establishes their mass.

# 17 Reproducibility Checklist

Before a result is promoted to the paper, the replication package should contain the following.

1. The exact command or orchestration script.
2. The git commit recorded in the payload.
3. A raw `.jls` payload containing draws, grids, objectives, or predictions.
4. A human-readable `SUMMARY.md` or support report.
5. A LaTeX table fragment if the result appears in a table.
6. A statement of support: full grid support, trimmed support, or mapped failure region.
7. A statement of scope: exact Kalman likelihood, profiled inversion objective, one-period known-feature comparison, or full HMC comparison.

# 18 Conclusion

The maintained pipeline is designed to be easy to audit. The surrogate is trained on a measured residual against direct nonlinear solutions. Validation compares ROM1, direct SEP, and the residual surrogate on common inputs. The HLT application samples a profiled inversion target, and the limits of that target are stated explicitly. The promoted current-ROM ablation establishes within-basin convergence but assigns no incremental observation fit to the NN correction used there. Its baseline-ROM training metadata is an explicit limit on interpretation. Numerical results promoted to the paper require raw output files. This structure is the basis for the paper's methodological claim. The remaining gap for a stronger submission is not hidden implementation detail but scale: the full multi-period direct-SEP HMC benchmark for the 18-parameter HLT model remains beyond the currently reported validation package, as does a fully normalized posterior comparison across competing basins.

# References

Stéphane Adjemian and Michel Juillard. Stochastic extended path. Working paper, Dynare Team, 2025. URL https://stephane-adjemian.fr/papers/sep-2025.pdf. Accessed: 2026-02-27.

Ray C. Fair and John B. Taylor. Solution and maximum likelihood estimation of dynamic nonlinear rational expectations models. *Econometrica*, 51(4):1169–1185, 1983. doi: 10.2307/1912057.

Mátyás Farkas and Bálint Tatár. Bayesian estimation of DSGE models with Hamiltonian Monte Carlo. IMFS Working Paper Series 144, Goethe University Frankfurt, Institute for Monetary and Financial Stability (IMFS), 2020. URL https://www.econstor.eu/bitstream/10419/223402/1/1728957516.pdf. Adapts HMC to DSGE estimation via Stan; documents a flat Smets–Wouters posterior target associated with markup ARMA dynamics.

Martín Harding, Jesper Lindé, and Mathias Trabandt. Resolving the missing deflation puzzle. *Journal of Monetary Economics*, 126:15–34, 2022. doi: 10.1016/j.jmoneco.2021.09.003.

Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

Frauke Schleer and Willi Semmler. Financial sector and output dynamics in the euro area: Non-linearities reconsidered. *Journal of Macroeconomics*, 46:235–263, 2015. doi: 10.1016/j.jmacro.2015.09.001.

Willi Semmler and Malte Sieveking. Nonlinear liquidity-growth dynamics with corridor-stability. *Journal of Economic Behavior & Organization*, 22(2):189–208, 1993. doi: 10.1016/0167-2681(93)90016-S.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with Discussion). *Bayesian Analysis*, 16(2):667–718, 2021. doi: 10.1214/20-BA1221.