Al-Hikmah Journal of Pure and Applied Sciences
©2026
Volume: (9) Issue (2): 173-189

On Adaptive Bayesian Updating for Misspecified Likelihood Models

***[1]Afolabi, B. T, [2]Garba, M. K. and [1]Ishaq, A. A.**

[1]Department of Physical Science, Al-Hikmah University, Ilorin, Nigeria
[2]Department of Statistics, University of Ilorin, Ilorin, Nigeria

ARTICLE INFO

Article history

Received: 01/09/2026
Revised: 07/10/2026
Accepted: 15/10/2026
Published: 06/11/2026

Doi:https://doi.org/10.5281/zenodo.24219228

Keywords:

Bayesian updating;
model misspecification;
generalized posteriors;
power tempering;
Robust divergence;
Posterior calibration

Corresponding Author

Email:
bilikisolaafolabi@gmail.com

Phone:+2348186220007

ABSTRACT

Bayesian updating is commonly based on the assumption that the working likelihood adequately represents the true data-generating process. In practice, this assumption may be violated, resulting in poorly calibrated posterior distributions even when point estimates appear reasonable. This study compares standard Bayesian updating, fixed power tempering, a data-driven adaptive learning-rate posterior, and a density-power-divergence (DPD) generalized posterior with fixed tuning parameter β = 0.5. A Monte Carlo simulation was conducted for one specific setting: a Gaussian location model with known variance, sample size n = 60, and symmetric scale contamination, with contamination proportions from 0 to 0.30. Bias, mean squared error (MSE), and empirical coverage of 95% credible intervals were evaluated over 1000 replications for each scenario. Under standard Bayesian updating, MSE increased from 0.017 to 0.338 because the sampling variability of the posterior mean grew with contamination, whereas coverage fell from 0.951 to 0.327 because the posterior standard deviation (0.129) did not respond to the contamination. Adaptive tempering restored coverage to between 0.948 and 0.996 by widening the credible intervals up to about 4.5-fold, but it left MSE essentially unchanged relative to standard Bayes. The DPD posterior reduced MSE to 0.032 at 30% contamination, and maintained high, generally conservative coverage (0.996 to 0.999). Sensitivity analyses over fixed $\eta \in \{0.25, 0.50, 0.75\}$ and $\beta \in \{0.25, 0.50, 0.75\}$ did not change this ranking. The findings apply to the Gaussian scale-contamination setting examined here and should not be extrapolated to other forms of misspecification without further study.

Citation: Afolabi, B.T., Garba, M.K. and Ishaq, A.A. (2026). On Adaptive Bayesian Updating for Misspecified Likelihood Models. *AJPAS.* 9(2): 173-189

1.0 Introduction

Bayes' theorem specifies a unique, coherent mechanism for updating a prior distribution into a posterior once data are observed, provided the likelihood function correctly describes the process that generated the data. This assumption is convenient for theory but rarely holds outside of controlled settings. Bissiri et al. [1] argued that modern applied problems make it increasingly difficult for a modeller to specify a likelihood that captures the true data-generating mechanism, and proposed replacing the log-likelihood in Bayes' rule with a general loss function, recovering standard Bayesian updating only as a special case. When the working model family does not contain the true distribution, the posterior does not target the truth; instead, under mild regularity conditions it concentrates on the pseudo-true parameter that minimises the Kullback-Leibler divergence between the model and the data-generating process [2, 3].

This shift in the target of inference has consequences that extend beyond a simple relocation of the posterior mode. Grünwald and van Ommen [4] demonstrated using ordinary Bayesian ridge regression and model averaging, that a misspecified Bayesian model can not only mislocate its estimate but can fail to concentrate at all, a phenomenon they term hyper-compression, which produces posteriors that are simultaneously overconfident and inconsistent. Owhadi et al. [5] reached a related conclusion from a different direction, showing that posterior inferences can be extremely brittle to small perturbations of the assumed model under a mismatch between the finite information available and the continuum of possible data-generating processes. Holmes and Walker [6] and Wu and Martin [7] noted that credible intervals built from a misspecified posterior routinely fail to attain their nominal frequentist coverage, because the curvature of the log-likelihood no longer matches the sampling variability of the score under the true distribution.

A family of remedies has grown out of this diagnosis, generally organised around the idea of tempering or re-weighting the likelihood contribution to the posterior. Raising the likelihood to a fractional power, often called the learning rate or tempering parameter, discounts the influence of the data relative to the prior and has been shown by Grünwald and van Ommen [4], to repair both hyper-compression and posterior inconsistency in linear models when the exponent is chosen sufficiently small. Because the correct exponent depends on unknown features of the true data-generating process, several data-driven selection procedures have since been proposed, including the SafeBayes sequential risk minimisation of Grünwald and van Ommen [4], and Heide et al. [8], the information-matching approach of Holmes and Walker [6], the loss-likelihood bootstrap of Lyddon et al. [3], and the generalized posterior calibration algorithm of Syring and Martin [9]. Because the learning rate (or, more generally, the scaling constant of a generalized posterior) directly controls posterior spread, the calibration of generalized posteriors is an important issue whenever posterior uncertainty is interpreted [7, 9]. A separate strand of work departs from uniform tempering altogether and instead re-weights individual observations according to a robust divergence, as in the general divergence framework of Jewson et al. [10], and the optimisation-centric generalized variational view of Knoblauch et al. [11], building on the classical density power divergence estimator of Basu et al. [12]. The Bayesian counterpart of that estimator, the R(α)-posterior of Ghosh and Basu [22], is a direct antecedent of the DPD generalized posterior studied here.

Although adaptive and robust generalized Bayesian methods are now well established as theoretical proposals, comparative evidence on what each method actually repairs is fragmented across separate papers with different data-generating processes, different metrics, and different notions of what "robustness" means. Wu and Martin [7] compared several learning-rate selection procedures directly and found that no single method uniformly resolves severe misspecification, which suggests that the choice of adaptive rule is not interchangeable and depends on the inferential quantity a practitioner actually cares about. What remains underexplored in an accessible, side-by-side form is the distinction between two qualitatively different repair mechanisms: uniform tempering of the whole likelihood, which rescales the overall contribution of the data without discriminating among observations, and divergence-based re-weighting, which can discount individual discrepant observations. Practitioners choosing between these families need direct evidence on whether either approach delivers point-estimation accuracy, calibration, or both, and at what efficiency cost. Therefore, this study addresses

part of that gap using a controlled simulation of one fully specified setting, a Gaussian location model with symmetric scale contamination, in which the ground truth and the contamination mechanism are both known, so that bias, mean squared error, and coverage can be measured exactly rather than inferred indirectly.

2.0 Materials and Methods

This section describes the materials and methods used in this study and is divided into sub-sections.

2.1 Key Theories and Concepts

The theoretical starting point for this study is the general Bayesian updating framework of Bissiri et al. [1], which derives, from a decision-theoretic coherency argument, a unique rule for updating a prior belief distribution using any loss function that connects data to a parameter of interest. When the loss function is chosen to be the negative log-likelihood, ordinary Bayes' rule is recovered exactly, so the framework nests standard Bayesian inference as a special case rather than replacing it. The resulting object is often called a general Bayesian posterior or a Gibbs posterior, because its exponential form resembles a Gibbs measure with the loss function playing the role of an energy. Zellner [13] is an earlier information-based contribution: he showed that Bayes' theorem can be obtained as the solution of an optimal information-processing problem, a variational characterisation that anticipates loss-based updating.

A second theoretical strand concerns what happens when a genuine likelihood is used but the likelihood itself is wrong. Kleijn and van der Vaart [2] extended the Bernstein-von Mises theorem to the misspecified case and showed that, under regularity conditions, the posterior still converges to a Gaussian shape but centred at the Kullback-Leibler minimiser rather than the true parameter, with a covariance matrix that generally differs from the inverse Fisher information evaluated at that point. This mismatch between the posterior's spread and the true sampling variability of the corresponding point estimator is the formal reason that credible intervals lose their frequentist coverage under misspecification, a point developed further by Syring and Martin [9] in their construction of calibrated general posterior credible regions.

The idea of tempering the likelihood, that is, raising it to a power strictly between zero and one before combining it with the prior, has a long history summarised by Grünwald and van Ommen [4], who trace it to Barron and Cover [14]. Walker and Hjort [15] demonstrated that consistency of a fractional posterior can be established under only local prior conditions, in contrast with the global entropy conditions ordinarily required for consistency at a learning rate of one. Grünwald and van Ommen [4] reinterpreted the tempered posterior as an exact Bayesian posterior for a hypothetical, non-misspecified model obtained by raising the true density to the same power, which shows that tempering can, for a sufficiently small exponent, eliminate both hyper-compression and inconsistency. Because the exponent depends on unobservable features of the true process, they propose SafeBayes, a sequential procedure that selects the learning rate by minimising an accumulated out-of-sample predictive risk in the spirit of Dawid's prequential principle [16].

Robust-divergence approaches constitute a related but mechanistically distinct branch of theory. Basu et al. [12] introduced the density power divergence as a family of discrepancy measures, indexed by a tuning parameter, that interpolates between the Kullback-Leibler divergence and increasingly outlier-resistant alternatives as the tuning parameter grows. Jewson et al. [10] formalised the use of such divergences inside the Bissiri et al. [1] framework, arguing that once a modeller accepts that every candidate likelihood is misspecified relative to the true process, inference should target the minimiser of a divergence chosen for its robustness properties in the tails rather than the Kullback-Leibler minimiser implied by ordinary Bayes. Knoblauch et al. [11] generalised this further within an optimisation-centric view of Bayes' rule, showing that classical Bayes, general Bayes, and variational inference all arise as solutions to a common regularised optimisation problem that differs only in the choice of loss, divergence, and constraint set.

2.2 Empirical and Methodological Findings from Prior Studies

Grünwald and van Ommen [4] provided empirical evidence, in Bayesian ridge regression and model-selection settings, that a SafeBayes-selected learning rate is competitive with standard Bayes when the model happens to be correct and substantially outperforms it when the model is misspecified, measured by predictive Kullback-Leibler loss. Heide et al. [8] extended the SafeBayes machinery to generalized linear regression and confirm the same qualitative pattern outside the Gaussian case. Holmes and Walker [6] proposed an alternative, non-sequential learning-rate selection rule based on matching the expected information gain of a single observation under the tempered and untempered updates, which is computationally far less demanding than SafeBayes.

Wu and Martin [7] conducted a direct, head-to-head comparison of several learning-rate selection methods, including SafeBayes, the Holmes and Walker [6] information-matching rule, the Lyddon et al. [3] loss-likelihood bootstrap, and their own generalized posterior calibration algorithm, across a range of misspecified scenarios. Their central empirical finding is that credible-region coverage is the metric most methods fail to protect: several rules perform well when misspecification is mild but none uniformly restores nominal coverage once misspecification becomes severe, and the generalized posterior calibration algorithm tends to be the most reliable specifically for coverage among the methods they test. Syring and Martin [9] report a closely related result, showing that their calibration algorithm can match nominal frequentist coverage even in settings where the naive general Bayes posterior badly under-covers.

On the divergence side, Jewson et al. [10] demonstrated, in a random-effects probit model with a misspecified random-effects distribution, that a divergence-based generalized posterior can remain close to the exact posterior obtained under correct specification, whereas the ordinary Bayesian posterior is visibly displaced. Matsubara et al. [17] extended robust generalized Bayesian inference to settings with intractable likelihoods, showing that divergence-based scores retain their outlier-resistance properties even when the normalising constant of the model cannot be evaluated directly. Huggins and Miller [18] studied a computationally distinct robustification strategy, BayesBag, which averages posteriors computed on bootstrap resamples of the data, and show that it achieves near-optimal robustness under log-loss but offers no comparable guarantee once a different, application-specific loss function is of interest, a caveat that echoes the target-dependence emphasised throughout this literature. Miller and Dunson [19] approached robustness from yet another angle, conditioning on the event that the model generates data distributionally close to what was observed rather than on the observed data exactly; they show that for relative-entropy coarsening this construction reduces, after an asymptotic approximation, to a power posterior, which links coarsening formally back to the tempering literature. Wenzel et al. [20] documented a related but distinct empirical puzzle in Bayesian neural networks, the cold posterior effect, in which tempering the posterior below its theoretically motivated temperature improves predictive performance, which they attribute partly to model misspecification and partly to inadequacies of standard priors in high-dimensional networks. Yonekura and Sugasawa [21] proposed a further adaptive tuning-parameter selection method for robust-divergence posteriors, reinforcing that the learning-rate or tuning-parameter selection problem recurs across both the tempering and divergence branches of this literature.

2.3 The Gap this Study Fills

Two gaps motivate the present study. First, although Wu and Martin [7] compared several learning-rate selection rules against one another, and Jewson et al. [10], separately demonstrate the value of divergence-based updating, no single controlled comparison places a sequential, prequential adaptive tempering rule of the SafeBayes type side by side with a robust-divergence generalized posterior under an identical, fully known contamination mechanism, tracking bias, mean squared error, and coverage simultaneously. Because tempering and divergence-based reweighting operate through different mechanisms, uniform discounting of the whole sample in the first case and selective discounting of individual discrepant observations in the second, it is unclear from the existing, separately reported results whether the two families are substitutes or address different failure modes. Second, much of the applied literature discusses tempering and calibration as though repairing coverage were equivalent to repairing point-estimation accuracy; the extent to which these are, in fact,

separable objectives has not been demonstrated transparently in an accessible, low-dimensional setting where the exact posterior is available in closed form. This study addresses both gaps, within one specific setting, by running the four update rules on a common Huber-contamination Gaussian design and reporting bias, mean squared error, coverage, and interval width together, and by adding a modest sensitivity analysis over the tuning parameters. This makes it possible to see directly whether a given remedy repairs calibration, accuracy, both, or neither in that setting.

2.4 Research Design

This study adopts a controlled Monte Carlo simulation design rather than an observational or survey design, because the research questions concern the statistical operating characteristics of competing updating rules under a known, experimenter-controlled form of likelihood misspecification. A simulation design is the appropriate methodology here for the same reason it is standard throughout the methodological literature this study builds on, including Grünwald and van Ommen [4], Wu and Martin [7], and Jewson et al. [10]; only when the true data-generating process and the true parameter value are known exactly can bias, mean squared error, and empirical coverage be computed without approximation, which is a precondition for isolating the mechanism by which each updating rule succeeds or fails.

2.5 Data-Generating Process and Working Model

Data were generated from a Huber-type contamination mixture, a standard and well-understood misspecification device in the robustness literature. For a contamination proportion ε, each observation was drawn independently as $x_i \sim (1-\varepsilon)\cdot N(\mu_0, 1) + \varepsilon \cdot N(\mu_0, \tau^2)$, with $\mu_0 = 2$ and $\tau = 8$, so that a proportion ε of the sample is drawn from a distribution with the same mean but a substantially larger variance than the bulk of the data. Each observation therefore has variance given by Eq. (1), which equals 19.9 when $\varepsilon = 0.30$. The working model used for inference was the simple Gaussian location family $N(\mu, 1)$ with known unit variance, which is exactly correct only when $\varepsilon = 0$; for every $\varepsilon > 0$ the working likelihood is misspecified in its tail behaviour, even though the central tendency of the contaminating component is not itself displaced. This design isolates the effect of tail misspecification, of the kind emphasised by Jewson et al. [10], as the central concern motivating divergence-based robustness, from the confound of a shifted contamination mean, and it allows a closed-form conjugate posterior to be computed for every tempered updating rule, which removes Markov Chain Monte Carlo approximation error from the comparison entirely. All conclusions of this study are conditional on this single design: one Gaussian location model with known variance, one sample size ($n = 60$), and one symmetric scale-contamination mechanism.

$$\sigma_\varepsilon^2 = 1 - \varepsilon + \varepsilon\tau^2 \quad (1)$$

Table 1: Simulation parameters used in the Monte Carlo study

Parameter	Value / Specification
True location parameter (μ_0)	2.0
Working likelihood	$N(\mu, 1)$, variance assumed known
Contamination distribution	$N(\mu_0, 8^2)$, symmetric scale contamination (Huber-type mixture)
Contamination proportions (ε)	0.00, 0.05, 0.10, 0.20, 0.30
Prior	$\mu \sim N(0, 10^2)$ (variance 100)
Sample size per replication (n)	60
Monte Carlo replications per condition (R)	1000
Fixed tempering exponent	$\eta = 0.5$ (sensitivity: 0.25, 0.50, 0.75)
Adaptive learning-rate grid	$\eta \in \{0.05, 0.10, \ldots, 1.00\}$
Adaptive selection criterion	Cumulative prequential predictive log-score, Eq. (3); fixed observation order; ties → smallest η
Robust divergence tuning parameter	$\beta = 0.5$ (density power divergence; sensitivity: 0.25, 0.50, 0.75)
DPD posterior evaluation	4001-point grid on $[-6, 10]$; trapezoidal quadrature; 2.5% and 97.5% posterior quantiles
Random-number seed	20260929 (NumPy default_rng; one SeedSequence child per ε)
Software	Python 3.12.3; NumPy 2.4.4; SciPy 1.17.1
Nominal credible level	95%

A weakly informative conjugate prior, $\mu \sim N(0, 10^2)$, that is, a normal prior with mean 0 and variance 100,, was placed on the location parameter for every updating rule so that differences in performance could be attributed to the updating mechanism rather than to differences in prior information. Table 1 summarises the full parameter specification of the simulation.

2.6 Competing Updating Rules

Four principal updating rules were compared for every simulated data set, each computed on the identical sample so that comparisons are not affected by Monte Carlo noise arising from different random draws across methods; tuning-parameter variants used in the sensitivity analyses were computed on the same samples. The ordinary Bayesian posterior was computed using the full, untempered Gaussian likelihood. Because the prior $N(m_0, v_0)$ with $m_0 = 0$ and $v_0 = 100$ is conjugate, the posterior is available in closed form as $N(m_1, v_1)$, given by Eq. (2) with $\eta = 1$. The posterior standard deviation depends on n, v_0 and the assumed variance but not on the observed spread of the data. This rule serves as the baseline against which every adaptive or robust alternative is judged, consistent with its role in Bissiri et al. [1], as the special case recovered when the general loss function is the negative log-likelihood.

$$v_\eta = \left(\frac{1}{v_0} + \eta n\right)^{-1}, \qquad m_\eta = v_\eta\left(\frac{m_0}{v_0} + \eta \sum_{i=1}^{n} x_i\right) \tag{2}$$

A naive, non-adaptive benchmark was constructed by raising the likelihood to a fixed power $\eta = 0.5$ before combining it with the prior, following the tempering construction formalised by Grünwald and van Ommen [4]. Because the tempered Gaussian likelihood remains conjugate to a Gaussian prior, with the data precision term scaled by η, this posterior is also available in closed form, namely Eq. (2) with $\eta = 0.5$. The values $\eta = 0.25$ and $\eta = 0.75$ were added in the sensitivity analysis (Section 3.4). This rule quantifies what a fixed, non-data-driven discount buys relative to standard Bayes, without the added complexity of a selection procedure.

The third rule selected a learning rate independently for every simulated data set by minimising a cumulative prequential (sequential, out-of-sample) negative log predictive risk, in the spirit of the SafeBayes procedure of Grünwald and van Ommen [4] and Dawid's prequential principle [16]. Let $m_{i-1}(\eta)$ and $v_{i-1}(\eta)$ denote the mean and variance of the η-tempered posterior in Eq. (2) computed from the first $i - 1$ observations only (with m_0 and v_0 when $i = 1$). The η-posterior predictive density of the next observation is $N(m_{i-1}(\eta), 1 + v_{i-1}(\eta))$, so the criterion is

$$S_n(\eta) = \sum_{i=1}^{n}\left[\frac{1}{2}\log\{2\pi(1 + v_{i-1}(\eta))\} + \frac{(x_i - m_{i-1}(\eta))^2}{2\{1 + v_{i-1}(\eta)\}}\right] \tag{3}$$

and the selected learning rate is the minimiser over the grid H given in Eq. (4). The tempered posterior of Eq. (2) computed on the full sample with $\eta = \hat{\eta}$ was reported as the adaptive posterior for that replication.

$$\hat{\eta} = \arg\min_{\eta \in H} S_n(\eta), \qquad H = \{0.05, 0.10, \ldots, 1.00\} \tag{4}$$

Several details are stated explicitly. (i) Variant. Eq. (3) scores each observation by the full η-posterior predictive density. This is the standard prequential log-score of the η-posterior. It is not the I-log variant of Grünwald and van Ommen [4], which scores the plug-in density $N(x_i; m_{i-1}(\eta), 1)$ evaluated at the posterior mean, nor their R-log variant, which uses the posterior-expected log loss. The rule is therefore "SafeBayes-style" only in the sense that η is chosen by cumulative prequential log-risk; the I-log and R-log variants were also implemented and are reported in Section 3.4. (ii) Order. Observations entered the criterion in the order in which they were generated. Because the data are independent and identically distributed this order is arbitrary, and a run with an independent random permutation in every replication is reported in Section 3.4. (iii) Ties. Ties among grid values were resolved by taking the smallest tied η, the most conservative learning rate; no exact ties occurred in any replication. (iv) Grid boundary. The lower grid value 0.05 is a boundary of the chosen grid. When the selected value equals 0.05 the criterion may prefer an even smaller learning rate, so the

selected value should be read as "0.05 or smaller" rather than as an estimate of an interior optimum; a check with the grid extended to 0.01 is given in Section 3.4.

The fourth rule replaces the tempered log-likelihood with a density power divergence (DPD) loss, following Basu et al. [12], its incorporation into general Bayesian updating by Jewson et al. [10], and the R(α)-posterior of Ghosh and Basu [22]. For a Gaussian working density φ_μ(x) = N(x; μ, 1) and tuning parameter β = 0.5, the per-observation DPD loss is

$$\ell_\beta(x;\mu) = -\frac{1}{\beta}\varphi_\mu(x)^\beta + \frac{1}{1+\beta}\int \varphi_\mu(u)^{1+\beta}\,du \tag{5}$$

For the N(μ, 1) family the integral term equals (2π)−β/2(1 + β)−1/2 and does not depend on μ, so it cancels on normalisation. Unlike the tempering rules, which raise the entire product likelihood to a common power and therefore re-weight the whole sample uniformly, the DPD loss re-weights each observation individually: the derivative of ℓ_β(x; μ) with respect to μ is −φ_μ(x)^β (x − μ), and the weight φ_μ(x)^β = (2π)−β/2 exp{−β(x − μ)²/2} decays with the distance between the observation and the candidate location, whereas the log-likelihood score (x − μ) grows linearly. Observations in the tails of the current candidate density, which are the observations most likely to be contaminants, are therefore attenuated automatically. The resulting generalized posterior, with unit loss weight and the same N(0, 10²) prior, is

$$\pi_\beta(\mu \mid x) \propto \exp\left\{\frac{1}{\beta}\sum_{i=1}^{n}\varphi_\mu(x_i)^\beta\right\}\pi_0(\mu) \tag{6}$$

This posterior is not conjugate. It was evaluated on a numerical grid of 4001 equally spaced points over μ ∈ [−6, 10] using scipy.integrate.trapezoid to normalise the density and to compute its mean and variance. The 95% credible interval for the DPD posterior was constructed directly from the numerical posterior as the interval between its 2.5th and 97.5th percentiles, obtained by cumulative trapezoidal integration (scipy.integrate.cumulative_trapezoid) followed by linear interpolation of the cumulative distribution function; the interval mean ± 1.96 SD is reported alongside for comparison (Section 3.5). The tuning parameter was fixed at β = 0.5, a moderate setting within the range explored by Jewson et al. [10] and Yonekura and Sugasawa [21]; no adaptive selection of β was performed. Because β was fixed, this rule is a fixed-tuning robust-divergence posterior and not an adaptive one. The values β = 0.25 and β = 0.75 were added in the sensitivity analysis (Section 3.4) so that the comparison with the adaptive tempering rule does not rest on a single arbitrary tuning choice. The posterior was not rescaled or calibrated, so its spread is whatever the DPD loss with unit weight implies.

2.7 Evaluation Metrics

For every combination of contamination level and updating rule, the following quantities were computed across the R = 1000 Monte Carlo replications, where m_r denotes the posterior mean in replication r: (i) bias, the average difference between the posterior mean and the true parameter μ_0; (ii) mean squared error (MSE), the average squared difference between the posterior mean and μ_0. By Eq. (7) MSE equals the squared bias plus the variance s_m² of the posterior mean across replications, so it measures the accuracy and sampling variability of the point estimator and is not a function of the posterior spread that a method reports; (iii) the empirical Monte Carlo standard deviation (Emp. SD) of the posterior mean, and the average posterior standard deviation (Post. SD), which measures how much uncertainty each rule assigns to its own estimate; (iv) empirical coverage of the nominal 95% credible interval, the proportion of replications in which the interval contained μ_0. For the conjugate rules the interval is mean ± 1.96 posterior SD, which is exact for a normal posterior; for the DPD posterior it is the interval between the 2.5th and 97.5th posterior quantiles; and (v) the average width of that interval, which quantifies the efficiency cost of any coverage gain. For the adaptive rule, the mean and standard deviation of the selected learning rate $\hat{\eta}$ across replications, and the percentage of replications at the grid boundary, were also recorded.

$$\mathrm{MSE} = \frac{1}{R}\sum_{r=1}^{R}(m_r - \mu_0)^2 = \mathrm{Bias}^2 + s_m^2 \quad (7)$$

Monte Carlo uncertainty. The Monte Carlo standard error (MCSE) of a coverage estimate p̂ is √{p̂(1 − p̂)/R}, which is about 0.007 near 0.95 with R = 1000 (about 0.013 with 300 replications, the number used in the first version of this study). The MCSE of the bias is s_m/√R and the MCSE of the MSE is the standard deviation of the squared errors divided by √R. MCSEs are reported for coverage in every table, and the error bars in Figs 1, 2, 4 and 5 are 95% Monte Carlo intervals (±1.96 MCSE). Differences of one or two percentage points in coverage are within Monte Carlo variation and are not interpreted.

2.8 Computational Implementation and Reproducibility

All simulations were implemented in Python 3.12.3 with NumPy 2.4.4 for random-number generation and vectorised conjugate updating, and SciPy 1.17.1 for numerical integration [23, 24]. Figures were produced with Matplotlib [25]. Random numbers were generated with numpy.random.default_rng (PCG64). The seed was 20260929; a numpy.random.SeedSequence with this entropy was spawned into five child sequences, one for each contamination level, so that every level has its own reproducible stream and all methods share the same simulated samples. The Python 3.12 environment, package versions, seed, and all code needed to reproduce every table and figure are provided as supplementary material (Supplementary_Code.zip; see Data and Code Availability). The conjugate rules require no iterative approximation. The DPD posterior was computed with scipy.integrate.trapezoid and scipy.integrate.cumulative_trapezoid. Numerical accuracy of the DPD posterior was assessed by (a) the maximum normalised posterior density at the two end points of the grid, which shows that no appreciable posterior mass is truncated, and (b) recomputing every replication for β = 0.5 on a finer and wider grid (12001 points on [−10, 14]) and comparing posterior means, standard deviations, quantiles and normalising constants with the main grid (Section 3.5).

3.0 Results

Table 2 reports bias, MSE, the empirical Monte Carlo SD of the posterior mean, average posterior SD, empirical 95% credible-interval coverage with its Monte Carlo standard error, and average interval width for all four updating rules at every contamination level. Figs 1 to 4 present the same information graphically with 95% Monte Carlo intervals, and the selected adaptive learning rate is shown in Fig. 3. Sensitivity analyses and numerical checks are reported in Sections 3.4 and 3.5.

3.1 Standard Bayes under Increasing Contamination

At ε = 0, where the working model is correctly specified, standard Bayes performed as theory predicts: bias was negligible (−0.004, MCSE 0.004), MSE was low (0.017), and coverage of the 95% credible interval was 0.951, indistinguishable from nominal given an MCSE of 0.007. As the contamination proportion increased, MSE rose sharply, reaching 0.128 at ε = 0.10 and 0.338 at ε = 0.30, an increase of about 20-fold relative to the correctly specified case (Fig. 1). This increase reflects greater sampling variability of the posterior mean: its empirical SD grew from 0.131 to 0.582, while the bias remained small (|bias| ≤ 0.030 in every condition, which is negligible next to the empirical SD). Contamination raises the variance of every observation from 1 to σ²_ε in Eq. (1), and the posterior mean is essentially the sample mean, whose variance is σ²_ε/n. Table 3 shows that the analytic values agree with the simulation within Monte Carlo error.

The failure of standard Bayes to respond to contamination is a property of its posterior spread. The average posterior SD stayed at 0.129 at every ε, a mechanical consequence of the conjugate updating equations, so the ratio of the average posterior SD to the empirical SD of the posterior mean fell from 0.98 at ε = 0 to 0.22 at ε = 0.30. The interval therefore became severely overconfident: coverage fell to 0.690 at ε = 0.05, 0.529 at ε = 0.10 and 0.327 at ε = 0.30 (Fig. 2), so under the most severe contamination the nominally 95% interval contained the true parameter about one third of the time. The non-response of the posterior SD explains the loss of coverage; it does not explain the increase in MSE.

Table 2: Simulation results across contamination levels (ε) and updating rules, 1000 Monte Carlo replications per cell.

ε	Method	Bias	MSE	Emp. SD	Post. SD	Coverage95 (MCSE)	CI Width
0.00	Standard Bayes (η=1)	−0.004	0.017	0.131	0.129	0.951 (0.007)	0.506
	Fixed tempering (η=0.5)	−0.005	0.017	0.131	0.183	0.995 (0.002)	0.716
	Adaptive tempering	−0.005	0.017	0.131	0.165	0.963 (0.006)	0.649
	Robust divergence (DPD)	−0.004	0.020	0.143	0.222	0.998 (0.001)	0.873
0.05	Standard Bayes (η=1)	−0.001	0.065	0.255	0.129	0.690 (0.015)	0.506
	Fixed tempering (η=0.5)	−0.002	0.065	0.255	0.183	0.837 (0.012)	0.716
	Adaptive tempering	−0.006	0.064	0.254	0.494	0.996 (0.002)	1.935
	Robust divergence (DPD)	−0.006	0.022	0.147	0.228	0.998 (0.001)	0.897
0.10	Standard Bayes (η=1)	0.004	0.128	0.357	0.129	0.529 (0.016)	0.506
	Fixed tempering (η=0.5)	0.004	0.128	0.357	0.183	0.692 (0.015)	0.716
	Adaptive tempering	−0.002	0.127	0.356	0.561	0.996 (0.002)	2.200
	Robust divergence (DPD)	0.001	0.022	0.149	0.235	0.999 (0.001)	0.923
0.20	Standard Bayes (η=1)	−0.030	0.213	0.461	0.129	0.399 (0.015)	0.506
	Fixed tempering (η=0.5)	−0.030	0.213	0.461	0.183	0.542 (0.016)	0.716
	Adaptive tempering	−0.036	0.212	0.459	0.576	0.988 (0.003)	2.258
	Robust divergence (DPD)	−0.007	0.028	0.168	0.249	0.996 (0.002)	0.979
0.30	Standard Bayes (η=1)	0.001	0.338	0.582	0.129	0.327 (0.015)	0.506
	Fixed tempering (η=0.5)	0.001	0.338	0.582	0.183	0.452 (0.016)	0.716
	Adaptive tempering	−0.005	0.336	0.580	0.576	0.948 (0.007)	2.259
	Robust divergence (DPD)	−0.005	0.032	0.178	0.267	0.999 (0.001)	1.047

Emp. SD is the Monte Carlo standard deviation of the posterior mean; Post. SD is the average posterior standard deviation; Coverage95 is the empirical coverage of the 95% credible interval with its Monte Carlo standard error in parentheses. DPD intervals are 2.5% and 97.5% posterior quantiles (β = 0.5).

Table 3: Analytic benchmark for standard Bayes and calibrating learning rate

ε	σ^2_ε	η□	MSE analytic	MSE simulated (MCSE)	Coverage analytic	Coverage simulated (MCSE)	Coverage at η□ (MCSE)
0.00	1.00	1.000	0.017	0.017 (0.001)	0.950	0.951 (0.007)	0.951 (0.007)
0.05	4.15	0.241	0.069	0.065 (0.003)	0.664	0.690 (0.015)	0.958 (0.006)
0.10	7.30	0.137	0.122	0.128 (0.006)	0.532	0.529 (0.016)	0.938 (0.008)
0.20	13.60	0.074	0.227	0.213 (0.009)	0.405	0.399 (0.015)	0.963 (0.006)
0.30	19.90	0.050	0.332	0.338 (0.015)	0.340	0.327 (0.015)	0.947 (0.007)

Analytic MSE and coverage use the exact normal distribution of the posterior mean (including prior shrinkage); η□ = $1/\sigma^2_\varepsilon$ is the learning rate at which the tempered posterior variance equals the sampling variance of the sample mean. The last column is the coverage of the tempering rule that uses η = η□ (an oracle that requires knowledge of the contamination).

3.2 Adaptive Tempering: Restoration of Coverage and Its Cost

The adaptive learning rate restored coverage close to or above the nominal 95% level at every contamination level examined: 0.963 at ε = 0, 0.996 at ε = 0.05, 0.996 at ε = 0.10, 0.988 at ε = 0.20, and 0.948 at ε = 0.30 (Fig. 2). With MCSEs between 0.002 and 0.007, the values at ε = 0.05 to 0.20 lie clearly above 0.95, so the rule restored coverage but tended to overcover rather than to calibrate exactly. This is explained by the selected learning rate (Fig. 3): the mean $\hat{\eta}$ fell from 0.765 at ε = 0 to 0.168 at ε = 0.05 and 0.071 at ε = 0.10, and reached the lower grid boundary of 0.05 at ε = 0.20 (99.8% of replications) and ε = 0.30 (100%). The tempered posterior is calibrated for the sample mean when η equals the value in Eq. (8), η□ = 0.241, 0.137, 0.074 and 0.050 at ε = 0.05, 0.10, 0.20 and 0.30, and the oracle rule that uses η□ attained coverage between 0.938 and 0.963 (Table 3). The data-driven $\hat{\eta}$ was smaller than η□ for ε = 0.05 to 0.20, which widened the intervals beyond what calibration requires (average posterior SD 1.94, 1.58 and 1.25 times the empirical SD).

$$\eta^* = \frac{1}{\sigma_\varepsilon^2} \qquad (8)$$

The value 0.05 is a boundary of the search grid and should be interpreted cautiously. At ε = 0.30 the calibrating value η□ = 0.050 coincides with the boundary, which is why coverage there (0.948) is

close to nominal. When the same criterion was applied on a grid extended down to 0.01 (Section 3.4), the mean $\hat{\eta}$ at $\varepsilon = 0.30$ fell to 0.010 (with 99.9% of replications at 0.01) and coverage rose to 1.000, so the near-nominal coverage of the adaptive rule at $\varepsilon = 0.30$ in the main analysis depends on the chosen grid and should not be read as evidence of exact calibration.

The coverage gain was purchased at a substantial efficiency cost, visible in Fig. 4. The average credible-interval width of the adaptive rule grew from 0.649 at $\varepsilon = 0$ to 2.200 at $\varepsilon = 0.10$ and 2.258 at $\varepsilon = 0.20$ to 0.30, roughly four to four-and-a-half times the standard-Bayes width of 0.506 for $\varepsilon \geq 0.10$. The adaptive procedure responded to increasing contamination by discounting the entire sample more heavily, which widens the posterior but leaves the location of the posterior mean essentially unchanged: the MSE of the adaptive rule (0.017 at $\varepsilon = 0$, rising to 0.336 at $\varepsilon = 0.30$) tracked standard Bayes at every contamination level, differing only in the third decimal place. In this design, adaptive tempering repaired coverage without repairing point-estimation accuracy.

3.3 Robust Divergence versus Tempering

The DPD generalized posterior produced a materially different pattern from both tempering rules. Its MSE remained low and grew only modestly with contamination: 0.020 at $\varepsilon = 0$ (a small premium over the 0.017 of standard Bayes when robustness is not needed), 0.022 at $\varepsilon = 0.05$, 0.022 at $\varepsilon = 0.10$, 0.028 at $\varepsilon = 0.20$, and 0.032 at $\varepsilon = 0.30$ (Fig. 1). At the highest contamination level this is roughly one tenth of the MSE of standard Bayes (0.338) and of both tempering rules (0.336 to 0.338), and the empirical SD of its posterior mean was 0.178 compared with 0.582. Biases of all four rules were small and showed no systematic pattern across ε (largest absolute bias 0.036; Table 2).

Coverage of the DPD posterior was high at every contamination level (0.996 to 0.999), and with MCSEs of at most 0.002 it lies well above the nominal 0.95. The DPD posterior therefore maintained high but conservative coverage rather than near-nominal coverage. Its average posterior SD (0.222 to 0.267) was about 1.5 times the empirical SD of its posterior mean, because the generalized posterior with unit loss weight is not scaled to match the sampling variability of the DPD estimator (the scaling issue discussed in Section 1). Interval widths of 0.873 to 1.047 were less than half the width of the adaptive-tempering intervals at every $\varepsilon \geq 0.05$ (ratios 0.46 to 0.43), so the DPD posterior reached its conservative coverage at a substantially lower efficiency cost. It achieved a lower MSE than either tempering rule at every contamination level greater than zero, consistent with a mechanism that discounts individual discrepant observations rather than the whole sample uniformly.

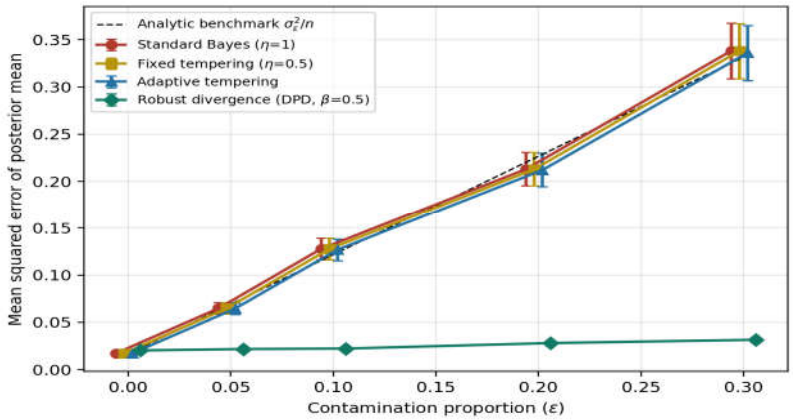

Fig. 1: Mean squared error of the posterior mean as a function of the contamination proportion (ε), for all four updating rules. Error bars are 95% Monte Carlo intervals (±1.96 MCSE); the dashed line is the analytic benchmark σ²_ε/n. Standard Bayes, fixed tempering and adaptive tempering overlap almost exactly.

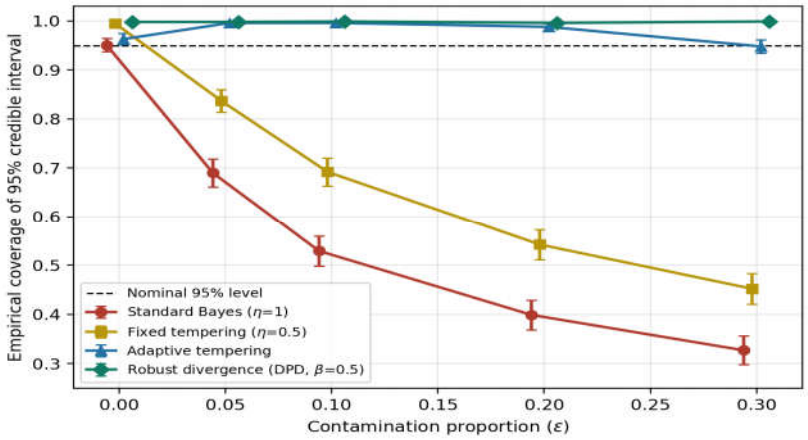

Fig. 2: Empirical coverage of the nominal 95% credible interval as a function of the contamination proportion (ε). The dashed line marks the nominal level. Error bars are 95% Monte Carlo intervals (±1.96 MCSE); differences of a few percentage points may reflect Monte Carlo variation.

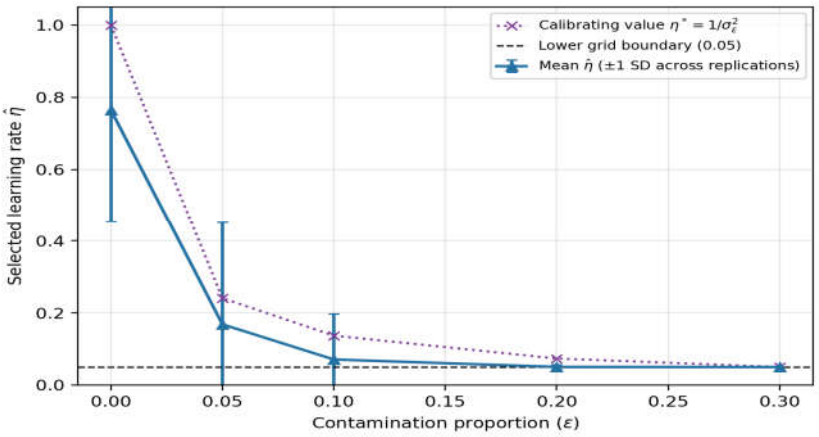

Fig. 3: Mean data-driven learning rate η̂ selected by the adaptive rule (cumulative prequential predictive log-score) as a function of the contamination proportion (ε); error bars show one standard deviation across Monte Carlo replications (truncated at the axis limits). The dashed line marks the lower boundary (0.05) of the search grid and the dotted line the calibrating value η□ = 1/σ²_ε.

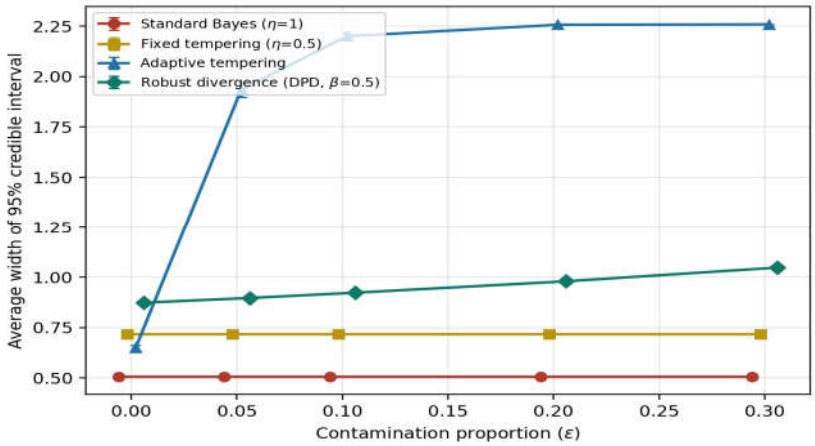

Fig. 4: Average width of the nominal 95% credible interval as a function of the contamination proportion (ε), illustrating the efficiency cost associated with each rule's coverage performance. Error bars are 95% Monte Carlo intervals (±1.96 MCSE) of the mean width.

3.4 Sensitivity to Tuning Parameters and to the Adaptive Criterion

A modest sensitivity analysis examined whether the ranking of the methods depends on the single tuning values used in the main comparison. Table 4 reports the fixed power $\eta \in \{0.25, 0.50, 0.75\}$ and the DPD parameter $\beta \in \{0.25, 0.50, 0.75\}$, and Fig. 5 summarises coverage and MSE.

Table 4: Sensitivity of fixed tempering and the DPD posterior to their tuning parameters

ε	Method	MSE	Post. SD	Coverage95 (MCSE)	CI Width
0.00	Fixed tempering, η = 0.25	0.017	0.258	1.000 (0.000)	1.012
	Fixed tempering, η = 0.50	0.017	0.183	0.995 (0.002)	0.716
	Fixed tempering, η = 0.75	0.017	0.149	0.972 (0.005)	0.584
	DPD, β = 0.25	0.018	0.172	0.988 (0.003)	0.674
	DPD, β = 0.50	0.020	0.222	0.998 (0.001)	0.873
	DPD, β = 0.75	0.023	0.284	0.999 (0.001)	1.116
0.05	Fixed tempering, η = 0.25	0.065	0.258	0.955 (0.007)	1.012
	Fixed tempering, η = 0.50	0.065	0.183	0.837 (0.012)	0.716
	Fixed tempering, η = 0.75	0.065	0.149	0.753 (0.014)	0.584
	DPD, β = 0.25	0.020	0.176	0.986 (0.004)	0.691
	DPD, β = 0.50	0.022	0.228	0.998 (0.001)	0.897
	DPD, β = 0.75	0.024	0.292	1.000 (0.000)	1.148
0.10	Fixed tempering, η = 0.25	0.128	0.258	0.844 (0.011)	1.012
	Fixed tempering, η = 0.50	0.128	0.183	0.692 (0.015)	0.716
	Fixed tempering, η = 0.75	0.128	0.149	0.588 (0.016)	0.584
	DPD, β = 0.25	0.021	0.181	0.984 (0.004)	0.710
	DPD, β = 0.50	0.022	0.235	0.999 (0.001)	0.923
	DPD, β = 0.75	0.025	0.301	1.000 (0.000)	1.183
0.20	Fixed tempering, η = 0.25	0.213	0.258	0.727 (0.014)	1.012
	Fixed tempering, η = 0.50	0.213	0.183	0.542 (0.016)	0.716
	Fixed tempering, η = 0.75	0.213	0.149	0.461 (0.016)	0.584
	DPD, β = 0.25	0.027	0.192	0.977 (0.005)	0.752
	DPD, β = 0.50	0.028	0.249	0.996 (0.002)	0.979
	DPD, β = 0.75	0.030	0.320	1.000 (0.000)	1.260
0.30	Fixed tempering, η = 0.25	0.338	0.258	0.620 (0.015)	1.012
	Fixed tempering, η = 0.50	0.338	0.183	0.452 (0.016)	0.716
	Fixed tempering, η = 0.75	0.338	0.149	0.361 (0.015)	0.584
	DPD, β = 0.25	0.031	0.205	0.977 (0.005)	0.803
	DPD, β = 0.50	0.032	0.267	0.999 (0.001)	1.047
	DPD, β = 0.75	0.034	0.343	1.000 (0.000)	1.349

All rules were applied to the same simulated samples as Table 2. DPD intervals are posterior quantile intervals. Coverage is given with its Monte Carlo standard error in parentheses.

Fixed tempering. Varying the fixed exponent changed the width of the interval but not the point estimate: the MSE was identical to that of standard Bayes to three decimals at every ε (Table 4). Smaller exponents gave higher coverage, but for $\varepsilon \geq 0.10$ none of the three fixed values restored coverage (best value 0.844, obtained with $\eta = 0.25$ at $\varepsilon = 0.10$), and only $\eta = 0.25$ at $\varepsilon = 0.05$ (0.955 (0.007)) came close to nominal under contamination. The ranking of fixed tempering below adaptive tempering in coverage therefore does not depend on the choice $\eta = 0.5$. DPD posterior. Coverage stayed at or above 0.977 for every β and every ε, and increased with β (from 0.977–0.988 at $\beta = 0.25$ to 0.999–1.000 at $\beta = 0.75$), while interval width also increased with β. The MSE ranged from 0.018 to 0.034 across all β and ε, with smaller β slightly more efficient in this symmetric design, and was below the MSE of every tempering rule for all $\varepsilon > 0$ and all three β values. For $\varepsilon \geq 0.05$ the DPD interval widths (0.691 to 1.349) were narrower than those of adaptive tempering (1.935 to 2.259). The qualitative ranking reported in Section 3.3 therefore does not depend on the single choices $\eta = 0.5$ and $\beta = 0.5$.

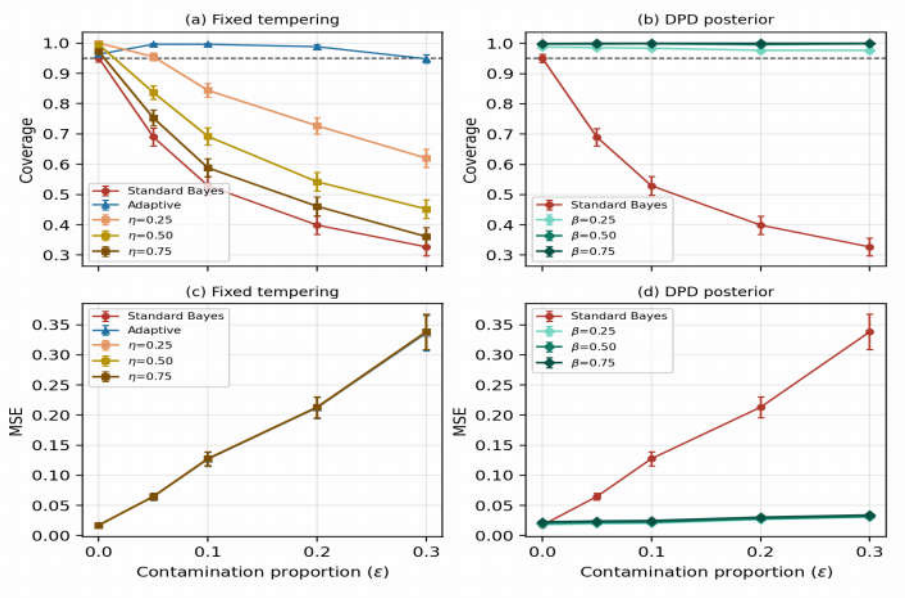

Fig. 5: Sensitivity of coverage (panels a and b) and MSE (panels c and d) to the fixed tempering exponent η (left) and the DPD tuning parameter β (right). Standard Bayes and, in panels (a) and (c), adaptive tempering are shown for reference. Error bars are 95% Monte Carlo intervals (±1.96 MCSE).

The choice of the adaptive criterion mattered far more than the choice of fixed tuning values. Table 5 compares four implementations of the prequential idea on the same samples: the criterion of Eq. (3), the same criterion with a random observation order in every replication, and the I-log and R-log variants of Grünwald and van Ommen [4], and reports the effect of extending the grid down to 0.01.

Table 5: Adaptive tempering under alternative prequential criteria

ε	Criterion	Mean η̂ (SD)	% at bound	Coverage95 (MCSE)	CI Width
0.00	Predictive log-score, Eq. (3)	0.765 (0.310)	0.0	0.963 (0.006)	0.649
	Predictive, random order	0.764 (0.309)	0.0	0.970 (0.005)	0.647
	I-log	0.517 (0.463)	42.3	0.977 (0.005)	1.323
	R-log	1.000 (0.000)	0.0	0.951 (0.007)	0.506
	Predictive, grid to 0.01	0.765 (0.310)	0.0	0.963	0.649
0.05	Predictive log-score, Eq. (3)	0.168 (0.283)	75.6	0.996 (0.002)	1.935
	Predictive, random order	0.173 (0.289)	76.1	0.996 (0.002)	1.931
	I-log	0.390 (0.448)	57.7	0.909 (0.009)	1.580
	R-log	0.996 (0.060)	0.3	0.692 (0.015)	0.512
	Predictive, grid to 0.01	0.144 (0.294)	45.1	0.996	3.435
0.10	Predictive log-score, Eq. (3)	0.071 (0.127)	95.5	0.996 (0.002)	2.200
	Predictive, random order	0.069 (0.117)	95.2	0.997 (0.002)	2.200
	I-log	0.326 (0.423)	64.9	0.880 (0.010)	1.698
	R-log	0.994 (0.072)	0.4	0.532 (0.016)	0.515
	Predictive, grid to 0.01	0.036 (0.133)	78.9	0.997	4.510
0.20	Predictive log-score, Eq. (3)	0.050 (0.007)	99.8	0.988 (0.003)	2.258
	Predictive, random order	0.050 (0.005)	99.9	0.988 (0.003)	2.258
	I-log	0.200 (0.342)	82.0	0.895 (0.010)	1.964
	R-log	0.987 (0.106)	0.9	0.407 (0.016)	0.526
	Predictive, grid to 0.01	0.011 (0.008)	98.1	1.000	4.984
0.30	Predictive log-score, Eq. (3)	0.050 (0.000)	100.0	0.948 (0.007)	2.259
	Predictive, random order	0.050 (0.000)	100.0	0.948 (0.007)	2.259
	I-log	0.144 (0.281)	88.3	0.880 (0.010)	2.071
	R-log	0.977 (0.141)	1.0	0.340 (0.015)	0.536
	Predictive, grid to 0.01	0.010 (0.000)	99.9	1.000	5.017

η̂ is the selected learning rate (mean, with SD in parentheses); "% at bound" is the percentage of replications with η̂ equal to the lower grid value (0.05, or 0.01 for the extended grid). Coverage is given with its Monte Carlo standard error in parentheses.

Reordering the observations at random left the results essentially unchanged (mean $\hat{\eta}$ 0.050 versus 0.050 at $\varepsilon = 0.30$, and coverage within 0.007 at every ε), and no exact ties occurred. The I-log criterion selected highly variable learning rates (for example mean $\hat{\eta}$ 0.517 at $\varepsilon = 0$, with 42% of replications at 0.05 and 47% at 1.00) and under-covered under contamination (0.880 to 0.909 for $\varepsilon \geq 0.05$). The R-log criterion stayed at $\hat{\eta} \approx 1$ and therefore did not protect coverage at all (coverage 0.340 at $\varepsilon = 0.30$). Only the full predictive log-score used in the main analysis restored coverage in this setting. Its results should therefore be regarded as characterising one prequential rule rather than the SafeBayes family. Finally, extending the grid to 0.01 moved the mean $\hat{\eta}$ at $\varepsilon \geq 0.10$ further down (0.036, 0.011 and 0.010 at $\varepsilon = 0.10$, 0.20 and 0.30), with coverage of 1.000 or more at $\varepsilon \geq 0.20$ and intervals up to 5.017 wide, confirming that the criterion is boundary-driven at high contamination.

3.5 Interval Construction and Numerical Checks for the DPD Posterior

Because the DPD posterior is available numerically, its 95% intervals were computed directly from the 2.5th and 97.5th posterior percentiles. Table 6 compares them with the interval mean ± 1.96 SD. The two constructions were essentially the same: coverage differed by at most 0.001 and average widths by at most 0.003 at $\beta = 0.5$, and by at most 0.001 in coverage for the other β values. The posterior is therefore close to symmetric, and the conservative coverage of the DPD posterior does not come from the choice of interval. The numerical accuracy of the grid was also checked. The largest normalised posterior density at the grid end points was 6.0×10-13 for $\beta = 0.5$ and 6.6×10-7 over all β, so no appreciable posterior mass was truncated. Recomputing all 1000 replications for each ε on a finer and wider grid (12001 points on [−10, 14]) changed posterior means by at most 2.6×10-12, posterior SDs by at most 7.5×10-11, the 2.5% and 97.5% quantiles by at most 2.9×10-5 and the log normalising constant by at most 5.6×10-13. The reported DPD results are therefore stable with respect to the grid.

Table 6: DPD posterior ($\beta = 0.5$): interval construction and grid sensitivity

ε	Cov. quantile	Cov. mean±1.96 SD	Width quantile	Width mean±1.96 SD	|Δ mean|	|Δ SD|	|Δ quantile|	End-point density
0.00	0.998	0.997	0.873	0.872	1.3×10-15	1.3×10-14	2.9×10-5	1.6×10-24
0.05	0.998	0.998	0.897	0.895	1.3×10-15	1.9×10-14	2.9×10-5	8.4×10-22
0.10	0.999	0.999	0.923	0.921	1.3×10-15	1.3×10-14	2.9×10-5	4.7×10-20
0.20	0.996	0.996	0.979	0.977	1.3×10-15	2.2×10-14	2.7×10-5	8.9×10-17
0.30	0.999	0.999	1.047	1.045	2.6×10-12	7.5×10-11	2.8×10-5	6.0×10-13

Coverage and width are given for the quantile interval and for mean ± 1.96 SD. The last four columns are maximum absolute differences over all replications between the main grid (4001 points on [−6, 10]) and a finer, wider grid (12001 points on [−10, 14]); the final column is the largest normalised posterior density at the end points of the main grid.

4.0 Discussion

The central pattern in these results is a separation between two things that are often discussed as though they were a single problem: getting the location of the posterior right, and getting its spread right. In this design the contaminating component shares the mean of the bulk of the data, so contamination does not shift the estimator; it inflates the variance of each observation from 1 to 19.9 at $\varepsilon = 0.30$, and the estimator inherits that variance. The MSE of the posterior mean equals its squared bias plus its sampling variance (Eq. 7), so MSE rose sharply because the sampling variability of the posterior mean increased (empirical SD from 0.131 to 0.582), in line with the analytic benchmark of Table 3. The failure of the standard-Bayes posterior SD to expand (it stayed at 0.129) is a different matter: it explains the poor coverage of the 95% interval, not the increase in MSE. Adaptive tempering corrected the second failure without touching the first: a uniform power exponent multiplies the precision contribution of every observation by the same factor, so it can inflate the posterior variance until the interval regains coverage, but the posterior mean remains a weighted average of the prior mean and the sample mean, with weights that differ only through the (weak) prior, so the MSE tracked that of standard Bayes. The DPD posterior lowered the MSE because it reduced the sampling variability of the point estimator (empirical SD 0.178 against 0.582 at $\varepsilon = 0.30$) by

down-weighting observations far from the current location, not because its posterior SD tracked the contamination: its average posterior SD (0.222 to 0.267) rose only modestly with ε and remained about 1.5 times the empirical SD of the estimator.

These results are consistent with, and help to make concrete for one setting, several claims that have previously been stated more abstractly. Grünwald and van Ommen [4] motivate tempering through predictive Kullback-Leibler risk and the repair of posterior consistency, not through the accuracy of a point estimator under symmetric contamination; the present results show that, in a symmetric design of this kind, tempering does not improve point-estimation MSE relative to standard Bayes, which is compatible with their framing. The finding of Wu and Martin [7] that no single learning-rate selection method uniformly resolves severe misspecification is echoed here in a narrow form: the outcome depended strongly on which prequential criterion was used (Table 5), with the predictive log-score restoring coverage, the I-log criterion under-covering, and the R-log criterion leaving coverage unprotected. Even the successful criterion resolved the coverage problem but not the accuracy problem, which suggests that "resolving misspecification" is at least two separable achievements and that evaluations of adaptive Bayesian methods should report both. The results for the DPD posterior echo the qualitative claim of Jewson et al. [10] that targeting a divergence chosen for tail robustness protects inference against tail misspecification; the present study adds a side-by-side quantification for one design, an MSE gap of roughly 11-fold at $\varepsilon = 0.30$.

The practical implication, restricted to this simulation, is that the choice of updating rule should be driven by the quantity a practitioner ultimately needs. If only a credible interval with at least nominal coverage is needed, both adaptive tempering and the DPD posterior delivered it, the DPD posterior with intervals less than half as wide for $\varepsilon \geq 0.05$. If a point estimate is also reported or acted upon, tempering alone should not be relied upon: an analyst who observes that an adaptive procedure has selected a small learning rate might reasonably, but incorrectly, take this as evidence that the point estimate has become more trustworthy, whereas in this design the point estimate was essentially that of standard Bayes and only the reported uncertainty grew. The results also distinguish restoration of coverage from exact calibration. The oracle tempering rule with $\eta = \eta$□ is close to exactly calibrated (Table 3), whereas adaptive tempering overcovered at moderate contamination and matched nominal coverage at $\varepsilon = 0.30$ only because the grid boundary coincided with η□. The DPD posterior overcovered in every condition (0.996 to 0.999), which is a safe but inefficient property; a calibrated scaling of the generalized posterior in the spirit of Syring and Martin [9] would be needed to bring its coverage to the nominal level.

Several aspects of the study's design limit how far its conclusions can be generalised. First, only one Gaussian location model with known variance, one sample size ($n = 60$) and one symmetric scale-contamination mechanism were examined, so the findings should not be taken as statements about model misspecification in general. The finding that tempering cannot change the point estimate is in part a direct algebraic consequence of the conjugate normal updating equations; in non-conjugate or nonlinear models, such as generalized linear or hierarchical models, a uniform learning rate could also shift the posterior location. Second, the mean-preserving contamination is a relatively favourable case for the bias of standard Bayes, and asymmetric or location-shifting contamination would probably add substantial bias for standard Bayes and fixed tempering and narrow, though probably not eliminate, the accuracy gap in favour of the DPD posterior. Third, the DPD parameter β was not adaptively selected, and Yonekura and Sugasawa [21] showed that this is a nontrivial problem; the sensitivity analysis covered only three values of β and three fixed values of η, and the adaptive tempering rule used one prequential criterion on a coarse grid with a lower bound that proved binding at high contamination. Fourth, the known-variance assumption removes uncertainty about the scale parameter, which is typically present in applied work. Fifth, $R = 1000$ replications give Monte Carlo standard errors of up to 0.016 for coverage, so differences of one or two percentage points should not be interpreted. Finally, no out-of-sample predictive score and no computational benchmark was evaluated, and this study therefore makes no claim about predictive performance or computational cost. Natural extensions are asymmetric and heavier-tailed contamination, several sample sizes, unknown variance, regression and other non-conjugate models, calibrated or adaptively tuned DPD posteriors, and predictive evaluation.

5.0 Conclusion

This study compared standard Bayesian updating, fixed and adaptive power tempering, and a density-power-divergence generalized posterior in a Gaussian location model with known variance (n = 60) under symmetric scale contamination of up to 30%. The principal finding is that repairing coverage and repairing accuracy are separable. At $\varepsilon = 0.30$, standard Bayes had an MSE of 0.338 and coverage of 0.327; adaptive tempering restored coverage to 0.948 but left the MSE at 0.336 with intervals 4.5 times as wide; the DPD posterior ($\beta = 0.5$) reduced the MSE to 0.032 and gave coverage of 0.999 with intervals of width 1.047 against 2.259.

Within this simulation, the increase in MSE under contamination reflects the greater sampling variability of the posterior mean, whereas the loss of coverage reflects a posterior spread that did not respond to the contamination. Tempering changes spread but not location, while the DPD posterior reduces the variability of the estimator. Both restored coverage, but often conservatively rather than exactly. The ranking of the methods was unchanged for fixed $\eta \in \{0.25, 0.50, 0.75\}$ and $\beta \in \{0.25, 0.50, 0.75\}$, but it depended strongly on the adaptive criterion: the I-log and R-log variants did not restore coverage here.

These findings are limited to a single Gaussian location model with known variance, one sample size, one symmetric contamination mechanism, a fixed β, a lower learning-rate grid bound that was binding at high contamination, and 1000 replications; predictive performance and computational cost were not evaluated. Future work should examine asymmetric and heavier-tailed contamination, several sample sizes, unknown variance, non-conjugate models, calibrated or adaptively tuned DPD posteriors, and predictive scores.

Data and Code Availability

All simulation code, the random seed, package versions, replicate-level results, summary tables and the code that generates Figs 1 to 5 are provided as supplementary material (Supplementary_Code.zip). Running python run_simulation.py with the default settings (seed 20260929, R = 1000) reproduces every number reported in Tables 2 to 6.

References

1. Bissiri, P. G., Holmes, C. C., & Walker, S. G. (2016). A general framework for updating belief distributions. Journal of the Royal Statistical Society: Series B, Vol. 78, No. 5, pp. 1103–1130. https://doi.org/10.1111/rssb.12158
2. Kleijn, B. J. K., & van der Vaart, A. W. (2012). The Bernstein–von Mises theorem under misspecification. Electronic Journal of Statistics, No. 6, pp. 354–381. https://doi.org/10.1214/12-EJS675
3. Lyddon, S. P., Holmes, C. C., & Walker, S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. Biometrika, Vol. 106, No. 2, pp. 465–478. https://doi.org/10.1093/biomet/asz006
4. Grünwald, P., & van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. Bayesian Analysis, Vol. 12, No. 4, pp. 1069–1103. https://doi.org/10.1214/17-BA1085
5. Owhadi, H., Scovel, C., & Sullivan, T. (2015). Brittleness of Bayesian inference under finite information in a continuous world. Electronic Journal of Statistics, Vol. 9, No. 1, pp. 1–79. https://doi.org/10.1214/15-EJS989
6. Holmes, C. C., & Walker, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. Biometrika, Vol. 104, No. 2, pp. 497–503. https://doi.org/10.1093/biomet/asx013
7. Wu, P.-S., & Martin, R. (2023). A comparison of learning rate selection methods in generalized Bayesian inference. Bayesian Analysis, Vol. 18, No. 1, pp. 105–132. https://doi.org/10.1214/21-BA1302
8. Heide, R., Kirichenko, A., Grünwald, P., & Mehta, N. (2020). Safe-Bayesian generalized linear

regression. In Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. Vol. 108, pp. 2623–2633. PMLR.

9. Syring, N., & Martin, R. (2019). Calibrating general posterior credible regions. Biometrika, Vol. 106, No. 2, pp. 479–486. https://doi.org/10.1093/biomet/asy054
10. Jewson, J., Smith, J. Q., & Holmes, C. (2018). Principles of Bayesian inference using general divergence criteria. Entropy, Vol. 20, No.6, pp. 442. https://doi.org/10.3390/e20060442
11. Knoblauch, J., Jewson, J., & Damoulas, T. (2022). An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. Journal of Machine Learning Research, Vol. 23, No. 132, pp. 1–109.
12. Basu, A., Harris, I. R., Hjort, N. L., & Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. Biometrika, Vol. 85, No. 3, pp. 549–559. https://doi.org/10.1093/biomet/85.3.549
13. Zellner, A. (1988). Optimal information processing and Bayes' theorem. The American Statistician, Vol. 42, No. 4, pp. 278–280. https://doi.org/10.1080/00031305.1988.10475585
14. Barron, A. R., & Cover, T. M. (1991). Minimum complexity density estimation. IEEE Transactions on Information Theory, Vol. 37, No. 4, pp. 1034–1054. https://doi.org/10.1109/18.86996
15. Walker, S. G., & Hjort, N. L. (2001). On Bayesian consistency. Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 63, No. 4, pp. 811–821. https://doi.org/10.1111/1467-9868.00314
16. Dawid, A. P. (1984). Statistical theory: The prequential approach. Journal of the Royal Statistical Society: Series A, Vol. 147, No. 2, pp. 278–292. https://doi.org/10.2307/2981683
17. Matsubara, T., Knoblauch, J., Briol, F.-X., & Oates, C. J. (2022). Robust generalised Bayesian inference for intractable likelihoods. Journal of the Royal Statistical Society: Series B, Vol. 84, No. 3, pp. 997–1022. https://doi.org/10.1111/rssb.12500
18. Huggins, J. H., & Miller, J. W. (2021). Using bagged posteriors for robust inference and model criticism. arXiv. https://arxiv.org/abs/1912.07104
19. Miller, J. W., & Dunson, D. B. (2019). Robust Bayesian inference via coarsening. Journal of the American Statistical Association, Vol. 114, No. 527, pp. 1113–1125. https://doi.org/10.1080/01621459.2018.1469995
20. Wenzel, F., Roth, K., Veeling, B. S., Świątkowski, J., Tran, L., Mandt, S., et al. (2020). How good is the Bayes posterior in deep neural networks really? In Proceedings of the 37th International Conference on Machine Learning (Vol. 119). PMLR.
21. Yonekura, S., & Sugasawa, S. (2023). Adaptation of the tuning parameter in general Bayesian inference with robust divergence. Statistics and Computing, Vol. 33, Article 39. https://doi.org/10.1007/s11222-023-10205-x
22. Ghosh, A., & Basu, A. (2016). Robust Bayes estimation using the density power divergence. Annals of the Institute of Statistical Mathematics, Vol. 68, No. 2, pp. 413–437. https://doi.org/10.1007/s10463-014-0499-0
23. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D. et al. (2020). Array programming with NumPy. Nature, Vol. 585, pp. 357–362. https://doi.org/10.1038/s41586-020-2649-2
24. Virtanen, P., Gommers, R., Oliphant, T. E., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods, Vol. 17, pp. 261–272. https://doi.org/10.1038/s41592-019-0686-2
25. Hunter, J. D. (2007). Matplotlib: a 2D graphics environment. Computing in Science & Engineering, Vol. 9, No. 3, pp. 90–95. https://doi.org/10.1109/MCSE.2007.55