risks

MDPI

Article

Dynamic Non-Life Insurance Pricing Under Delayed Claim Reporting: A Partially Observed Risk-Sensitive Control Framework

Desmond Marozva [1,†], Selah Tanaka Marozva [2,†] and Ştefan Cristian Gherghina [1,*,†]

[1] Department of Finance, Bucharest University of Economic Studies, 010374 Bucharest, Romania; marozvadesmond19@stud.ase.ro
[2] Department of Marketing, Bucharest University of Economic Studies, 010374 Bucharest, Romania; goweroselah21@stud.ase.ro
* Correspondence: stefan.gherghina@fin.ase.ro
† These authors contributed equally to this work.

Abstract

Non-life insurance pricing is forward-looking, yet its principal cost signal, reported claims, is delayed by the occurrence-to-reporting process. The rating cell is treated as a stylised, homogeneous unit without renewal, lapse, expiry, or cohort dynamics; a fully annual-contract formulation would require these dynamics to be modelled jointly with the risk and claims processes. This paper develops a partially observed risk-sensitive control framework in which premium-sensitive exposure generates claims in a latent risk regime, while unreported claims form an atomic population governed by an age-structured transport equation. The numerical instance solved and validated in this research restricts the general model to a memoryless (one-phase) reporting process for tractability. It is best matched to lines with predominantly short-to-medium reporting tails, rather than to the most extreme long-tailed liability or cyber exposures the general model is designed to eventually accommodate. A finite-state reporting reservoir and nominal Bayesian filter provide the decision state, and compound Poisson–Gamma loss enters an entropic Bellman recursion through a closed-form exponential-tilting identity, checked against an independently coded Bellman-residual test. Every dynamic policy is benchmarked against alternatives matched at the same risk-sensitivity parameter and evaluated using common random numbers. This is a general feature of the results, not a single statistic: $CE_{0.8}$ is a standardised yardstick applied uniformly across strategies, while CE_{γ} at the policy's own γ is what that policy actually optimises, and the two need not agree. In 3000 out-of-model paths at $\gamma = 1.2$, the delay-aware dynamic policy increases mean profit by EUR 0.148 million and a standardised $\gamma_{\text{eval}} = 0.8$ certainty equivalent by EUR 0.034 million relative to a matched static price. Its fifth percentile and TVaR 5%, by contrast, are lower by EUR 0.056 and 0.078 million. At the policy's own optimisation level, however, the $\gamma_{\text{eval}} = 1.2$ certainty-equivalent difference is EUR -0.006 million, with a 95% interval reaching zero. The dynamic policy, therefore, does not clearly outperform the matched static price on the exact objective it was optimised to maximise. Matched comparisons attribute EUR 0.039–0.043 million of mean profit to reporting-delay modelling and EUR 0.017 million to dynamic continuation. Separating state observation from transition-law knowledge attributes EUR 0.161–0.182 million to observing the regime exactly, and a small, sign-changing EUR -0.010 to +0.003 million to knowing the true transition law itself. Across an eight-scenario misspecification stress suite, the dynamic policy's mean-profit advantage over the matched static price is directionally robust in seven of eight scenarios, but it reverses sign under a +30% true-severity shock, indicating that this advantage is sensitive to substantial severity misspecification specifically. Grid,

Academic Editor: Rui Manuel Rodrigues Cardoso

Received: 6 August 2026
Revised: 3 September 2026
Accepted: 7 September 2026
Published: 10 September 2026

Bellman-residual, and out-of-model filter diagnostics indicate that state reconstruction is economically primary, while entropy risk sensitivity is a secondary overlay whose apparent benefit depends materially on which certainty-equivalent level and evaluation model are used to judge it.

Keywords: non-life insurance pricing; reporting delay; partially observed control; risk-sensitive control; Bayesian filtering; dynamic pricing; claims reserving; stochastic control

1. Introduction

Insurance prices are forward-looking, but claims experience is backward-looking and delayed. A loss incurred in the current underwriting period may be reported several months later, while its payment and case-estimate development may continue for longer still. Calendar-month reported claims are therefore a mixture of occurrence cohorts written at different prices, exposed to different latent risk conditions, and observed at different maturities. A pricing rule that treats this mixture as a contemporaneous measurement can react after the underlying regime has changed, and can continue writing business at an inappropriate rate while the corrective evidence remains in the reporting pipeline. This is not a purely academic concern. In liability, cyber, and long-tail commercial lines, the occurrence-to-reporting lag routinely exceeds the length of a single underwriting cycle. A pricing algorithm that ignores this lag is, in a precise sense, driving using only a rear-view mirror that is itself several months out of focus.

The practical separation between pricing and reserving reinforces this problem. Pricing teams often receive developed frequencies or ultimate-loss selections as fixed inputs, updated on a quarterly or annual cycle. Reserving teams quantify incurred-but-not-reported (IBNR) claims and development uncertainty on their own schedule, typically without a direct feedback path into the premium that generated the exposure in the first place. Transferring only a point estimate across this organisational boundary discards exactly the information—the posterior distribution over the current risk regime, and over how many claims are still sitting unreported—that a sequential pricing decision needs. Granular occurrence-and-development models provide a principled bridge between pricing and reserving (Abdelrahman et al. 2026; Crevecoeur et al. 2023; Verbelen et al. 2022), but they generally address prediction rather than the sequential choice of a future premium path under a risk-adjusted objective. Closing that gap—coupling a genuine reporting-delay filter to a sequential pricing decision under a risk-sensitive objective—is the specific problem this paper addresses.

Closing this gap raises four linked methodological issues. The first is how to represent reporting delay as a genuine state rather than an input assumption. The second is how to keep risk sensitivity conceptually separate from beliefs about the transition law under partial observation. The third is how to benchmark a risk-sensitive, partially observed policy without confounding the benchmark's own risk appetite or model beliefs with the effect being tested. The fourth is how to make a synthetic Monte Carlo comparison trustworthy through its random-number design. Section 2.5 develops each of these in turn after situating the paper within four adjacent studies.

The paper makes six contributions. First, it distinguishes the pathwise unreported-claim population, an atomic random measure, from the conditional intensity density that satisfies an age-structured transport equation. Second, it derives a loss-aware entropic Bellman operator that includes compound claim-count and severity risk rather than only expected loss, with the Gamma-severity moment-generating-function existence

condition stated explicitly (Section 4). Third, it evaluates the assumed-density reporting reservoir against an exact truncated integer-state filter, labelling the resulting diagnostics as measuring *approximate-policy* agreement and a directly simulated *filter-induced policy gap*, and against a separately coded one-step Bellman residual check (Section 6). Fourth, it uses matched benchmarks—an optimised static price resolved at each policy's own γ, a zero-delay dynamic program, a valid occurrence-cohort Bornhuetter–Ferguson rule, a delay-naive heuristic, and full-state policies solved under both the reference and true transition matrices—to isolate maturity, continuation, state-observation, and transition-knowledge effects while holding risk appetite fixed within each comparison (Section 8). Fifth, it documents and applies a genuine common-random-number Monte Carlo design based on shared inverse-transform draws (Section 5.4). Sixth, the robustness analysis includes a grid-refinement check, an out-of-model re-run of the exact-filter comparison, and an eight-scenario fixed-policy misspecification stress suite with bootstrap intervals throughout (Section 7.5). This suite surfaces at least one scenario—an elevated true severity level—under which the dynamic policy's mean-profit advantage over the matched static benchmark reverses sign.

The title deliberately describes a risk-sensitive control framework rather than a numerically solved infinite-dimensional PDE-constrained problem. The transport equation supplies the structural representation of reporting memory; the implemented optimisation is a finite-state Markov decision approximation, solved and validated at that finite-state level, and the paper's validation exercises (Section 6) are framed as diagnostics of that specific approximation rather than as proofs of convergence to the underlying continuous-time control problem.

The remainder of the paper is organised as follows. Section 2 situates the paper within four studies and states the research gap. Section 3 sets out the structural occurrence-reporting model. Section 4 develops the loss-aware entropic Bellman operator. Section 5 describes the finite-state approximation, the common-random-number design, and the Bellman algorithm. Section 6 reports numerical and filter verification. Section 7 describes the experimental design, matched-benchmark set, and stress suite. Section 8 reports results; Section 9 discusses their hierarchy and actuarial implementation; Section 10 states limitations; Section 11 concludes.

Figure 1 summarises the closed-loop architecture that ties these sections together. A governed premium move changes price-sensitive exposure and hence claim occurrences in the latent risk regime (Section 3). Occurrences that are not reported immediately populate the atomic unreported-claim population, which generates reported counts and claim marks with a lag. The nominal Bayesian filter turns those reports into a posterior regime probability and filtered reservoir stock (Section 5). Finally, the loss-aware entropic Bellman policy (Section 4) turns that filtered state into the next month's governed premium move, closing the loop. The dashed box separates the *offline* step—solving the policy surfaces once per calibration, and validating them as described in Section 6—from the *online* step of updating the filter and applying governance constraints each month.

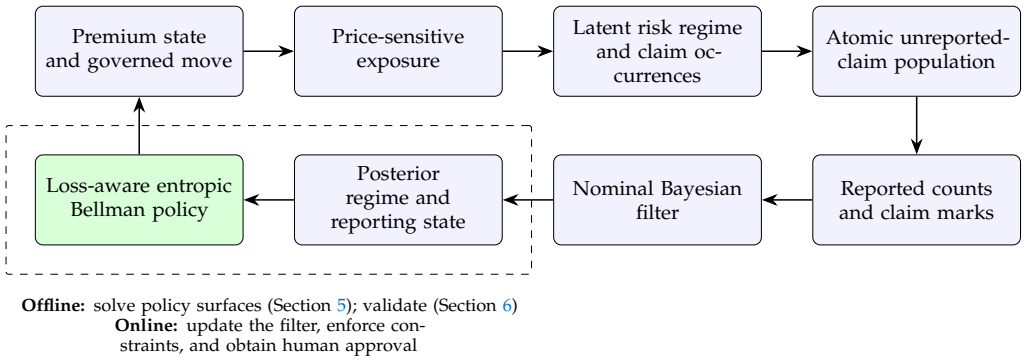

Figure 1. Closed-loop architecture for delay-aware dynamic pricing. Blue boxes show the occurrence, reporting, and filtering components; the green box shows the pricing policy. The dashed box separates the offline policy-solution and validation stage (Sections 5 and 6) from the online filter-update, governance, and approval stage.

2. Related Literature and Research Gap

2.1. Dynamic Insurance Pricing

Emms and Haberman (2005) formulate general-insurance pricing as an optimal-control problem in which premium affects demand, and the insurer maximises terminal wealth. Emms (2007) extends the framework to competitive markets and multidimensional Bellman equations. These studies establish the margin-volume trade-off and the value of feedback pricing that underlies Equation (2) below, but their loss signal does not arise from an explicit occurrence-to-reporting mechanism: losses are effectively observed as they occur. This is a reasonable approximation for short-tailed personal lines, but not for the liability, cyber, and commercial lines this paper targets. More recent telematics research supports high-frequency motor pricing through behavioural signals (Henckaerts and Antonio 2022; Yanez et al. 2025). Those signals reduce information delay in telematics-enabled books by substituting a continuously observed proxy (driving behaviour) for the claims process itself. However, reported claims remain the principal cost feedback for many commercial, liability, and non-telematics portfolios, and even in motor books, the claims signal used for technical ratemaking is still subject to reporting lag. A separate strand of the dynamic-pricing literature outside insurance—revenue management and demand-learning models—shares the present paper's basic structure: a controller that must learn an unknown demand or cost parameter while acting on it. However, it almost never carries an explicit, age-structured delay between the event that generates cost and the event that reveals it to the controller. The insurance reporting-delay problem is, in this sense, a demand-learning problem with a second, independent source of latency layered on top of regime uncertainty.

2.2. Reporting Delay and the Pricing-Reserving Interface

Bornhuetter and Ferguson (1972), Mack (1991), and England and Verrall (2002) provide central actuarial foundations for development-factor and IBNR estimation, and they remain the default point of comparison for any model that claims to handle reporting delay. This is why this paper's Bornhuetter–Ferguson benchmark (Section 7.3) is retained and evaluated on the same footing as the dynamic policies, rather than treated as a straw man. Wüthrich and Merz (2008) give the standard comprehensive treatment of stochastic claims reserving methods, including the distributional assumptions (compound Poisson-type frequency-severity decompositions, chain-ladder and Bornhuetter–Ferguson variants, and their Bayesian and bootstrap uncertainty quantification) that this paper's compound Poisson–Gamma loss model draws on directly. Continuous-time approaches represent occurrence, reporting, and payment explicitly through point processes, martingales, and delayed-event models (Arjas 1989; Haastrup and Arjas 1996; Lawless 1994; Norberg 1993,

1999). These papers establish the point-process view of claims development that Section 3 adopts for the occurrence-reporting mechanism, but they are reserving models: they characterise the distribution of outstanding liabilities given a fixed exposure and rating history, not a control problem in which the premium itself is chosen sequentially in response to the filtered state. Cox-process specifications add overdispersion and hidden environments (Avanzi et al. 2016; Badescu et al. 2019, 2016), which is precisely the latent-regime structure this paper's $Z(t)$ process borrows, again without the accompanying pricing decision. Martínez-Miranda et al. (2012) and the double chain-ladder literature it initiated combine reported-count and payment-per-claim information in a single stochastic model that separates claim occurrence from claim reporting and payment timing at the aggregate-triangle level. The present paper's split of monthly occurrences into an immediately reported fraction d_0 and a delayed fraction entering the reservoir, Equation (13), is structurally the same decomposition, applied at the level of a controlled Markov decision process rather than a fitted triangle. Antonio and Plat (2014) and the broader move toward micro-level, individual-claim reserving models (Abdelrahman et al. 2026; Crevecoeur et al. 2023; Verbelen et al. 2022) jointly model occurrence and reporting at a granular level, demonstrating the value of simultaneous estimation over the traditional two-stage triangle-then-load approach. These models, together with Abdelrahman et al. (2026)'s Dirichlet-driven marked Cox model for IBNR counts, are the closest prediction-side analogues to the present paper's occurrence-reporting model. The remaining gap they leave open—explicitly identified in Crevecoeur et al. (2023)'s own discussion of extensions—is exactly a governed sequential premium decision in which the maturity state affects both inference and future actions. Okine (2023) directly addresses ratemaking in changing environments and IBNR frequency, combining regime change and reporting delay in a single-frequency model without, however, embedding the result in a dynamic control problem with an explicit objective and admissible action set. The present paper's contribution relative to this literature is not a new reserving method—Section 7.3's Bornhuetter–Ferguson benchmark deliberately uses standard development-factor machinery—but the coupling of a reporting-delay filter, in close-to-real time, to a sequential premium decision with an explicit risk-sensitive objective.

2.3. Partial Observation and Risk-Sensitive Control

Under standard conditions, the conditional law of the hidden state is the sufficient statistic of a separated, partially observed control problem (Bain and Crisan 2009; Davis 1993; Elliott et al. 1995; Wonham 1964). This is the classical justification, used throughout Section 3, for collapsing the insurer's information into the filtered pair (π_t, u_t). Reporting delay introduces memory because current reports are generated by earlier occurrence cohorts. An age-structured state restores the Markov property, in the same spirit as the point-process reserving models of Section 2.2, but now as part of a controlled, rather than purely predictive, system. Delay-control problems more generally lead to infinite-dimensional Hamilton–Jacobi–Bellman equations (Gozzi and Masiero 2017a, 2017b), and the general theory of viscosity solutions for controlled Markov processes (Fleming and Soner 2006) and of continuous-time stochastic control more broadly (Pham 2009) provides the language in which a rigorous infinite-dimensional version of Equation (10) below would need to be posed and verified. The present paper does not attempt that verification. Section 4 states explicitly that the implemented recursion is a finite-state Markov decision approximation, not a numerically verified solution of the underlying PDE, and Section 2.5 explains why that narrower framing is the honest one for this paper's contribution. In the present model, the delay enters through a boundary-generated claim population (the transport equation's boundary condition $\bar{u}_i(t,0) = \lambda_i^O(t)\mathbf{1}\{\mathrm{Z(t)}= i\}$) rather than through

a delayed control. This keeps the finite-state approximation tractable, but it also means results here should not be read as evidence about the delayed-control case.

Exponential utility and relative entropy connect risk-sensitive preferences with conservative reweighting of nominal outcomes (Anderson et al. 2003; Hansen and Sargent 2008; Lim and Shanthikumar 2007); Whittle (1990) gives the foundational risk-sensitive-control account of the exponential-of-integral criterion used in Equation (9) below, including the large-deviations and small-noise limits that justify treating $\gamma \to 0$ as the risk-neutral special case. The specific combination this paper needs—a risk-sensitive objective layered on top of a Bayesian filter under partial observation, with the filter and the risk criterion kept conceptually separate—has been treated rigorously in the Markov decision process literature by Bäuerle and Rieder (2017). They show that a partially observable risk-sensitive Markov decision process can be reformulated as a fully observed control problem on the belief state under a general certainty-equivalent criterion, giving existence and computability results for exactly the entropic-certainty-equivalent, belief-state formulation this paper implements. The present paper's Equation (17) is, in that sense, a concrete, insurance-specific instance of the Bäuerle and Rieder (2017) belief-MDP construction, restricted to a two-regime, phase-type reporting reservoir so that it can be solved and, crucially, independently validated (Section 6) at production scale. The present contribution combines these strands while keeping nominal filtering and preference robustness conceptually separate. Section 4 states, at the point the entropic certainty equivalent is introduced, that it is a preference representation under a single nominal law, not a claim that a family of physical transition laws share one filter.

Two recent contributions sharpen this positioning further. Wu and Xu (2023) extend risk-sensitive Markov decision process theory beyond the exponential criterion to general utility functions, a generalisation this paper does not pursue—Equation (17) deliberately keeps the exponential, entropic form because it alone yields the closed-form tilting identity of Equation (15), which is what makes the recursion exactly solvable and independently checkable at the scale Section 6 requires. Dong (2025) applies a CVaR-constrained reinforcement-learning policy to reserve setting under regime-switching macroeconomic stress, which shares this paper's combination of risk-sensitivity and a latent regime process but targets a different decision (a reserve adjustment, learned by a trained policy network) rather than a premium decision solved to a verified finite-state optimum; the two approaches are complementary rather than competing, and a natural extension pairs this paper's pricing recursion with a similarly regime-aware reserving policy.

2.4. Validating Approximate Bayesian Filters and Finite-State Control Approximations

A methodological study distinct from, but essential to, the substantive modelling above concerns how an approximate filter or an approximately solved control problem should be validated once exact computation is infeasible. Bain and Crisan (2009) give the general nonlinear filtering theory against which any finite-dimensional approximate filter—including this paper's assumed-density reservoir closure, Equation (14)—must ultimately be judged. The standard tool for that judgement is comparison against a higher-fidelity, typically exact-but-more-expensive, filter on the same observation sequence, which is exactly the exercise Section 6.4 reports. This paper follows that logic but is explicit that the resulting diagnostics measure the quality of a substitution (posterior means from the exact filter plugged into a Q-surface trained under the assumed-density closure), not the quality of a filter that was itself re-optimised on the exact belief simplex. Section 6.4 accordingly uses a specific relabelling—approximate-policy action agreement rather than "action agreement", and a directly simulated filter-induced policy gap rather than an indirect Q-value-based "decision regret"—specifically so that the reported numbers cannot

be over-read as a proof of near-optimality. On the control side, Fleming and Soner (2006) and Pham (2009) again provide the relevant convergence framework (viscosity-solution approximation of Hamilton–Jacobi–Bellman equations by Markov chain or finite-difference schemes). The discrete-state, discrete-action approximation used here inherits the standard concern that a coarse state grid can silently misstate both the value function and the optimal action near a boundary. Section 6.1 therefore reports a direct grid-refinement comparison (production versus a finer evaluation of the same coarse grid), rather than relying on agreement between two evaluations of the same coarse grid. Finally, the risk-sensitive Bellman recursion of Section 4 could otherwise be validated only by comparing its own output against a Monte Carlo simulation of its own implied policy—a check that cannot detect an error shared by both the recursion and the simulator that consumes its output. Section 6.2 therefore reports a separately coded, unvectorised re-implementation of the one-step Bellman operator, evaluated at randomly sampled states, in the spirit of a numerical residual check for a fixed-point equation. The one component that check originally shared with the production solver, the bilinear-interpolation utility, is in turn independently validated in Section 6.2 against a second, independently engineered implementation and against analytic ground truth.

2.5. Research Gap and Positioning

Dynamic pricing models usually omit explicit reporting maturity (Section 2.1). Reserving models estimate hidden occurrences and reporting delay but do not optimise the future premium path (Section 2.2). Risk-sensitive control models frequently abstract from the observation pipeline entirely, or treat partial observation and risk sensitivity as separate problems rather than as a single belief-state control problem (Section 2.3). The unresolved operational problem is therefore a premium-control system in which delayed reports update a posterior state, the current price changes both current exposure and future information volume, and claim-count and severity volatility enter the risk criterion directly rather than through an ad hoc loading. A second, purely methodological gap is that a synthetic demonstration of such a system is only informative if its benchmarks are matched on every dimension except the one being tested, its Monte Carlo design uses genuine common random numbers, and its approximate solution methods are validated independently rather than self-referentially. Table 1 summarises the positioning of the four studies against these two gaps. The continuous model of Section 3 provides the structural basis for the substantive gap. The numerical contribution of Sections 5 and 6 solves and validates a finite-state approximation suitable for controlled synthetic experiments, while Sections 7 and 8 close the methodological gap by matching every benchmark's risk-sensitivity parameter, documenting the common-random-number construction, and reporting both a grid-refinement check and an independent Bellman residual check alongside the substantive results.

Individually, the closest points of comparison are as follows. Emms and Haberman (2005) and Emms (2007) solve a sequential premium decision but observe losses as they occur. The present paper keeps the margin-volume Bellman structure. Equation (2) is a direct descendant of their exposure response, while replacing contemporaneous loss observation with a filtered, delay-aware state. Crevecoeur et al. (2023) and Abdelrahman et al. (2026) jointly model occurrence and reporting at the granular level with a genuine reporting-delay filter, but as a prediction problem. Neither embeds the resulting posterior in a sequential, risk-sensitive pricing decision, which is precisely the coupling this paper adds. Avanzi et al. (2016) extend this same reserving lineage with a reinforcement-learning agent that sets claim-level reserves sequentially, which is closer in spirit to a control problem than the earlier prediction-only models, but the decision variable remains a reserve rather than a premium, and the policy is learned rather than solved to a verified finite-state

optimum; this paper's contribution is complementary, addressing the pricing side of the same reporting-delay information problem with an exactly solved and independently validated recursion. Henckaerts and Antonio (2022) and Yanez et al. (2025) dynamically reprice using telematics behavioural signals, which reduce information delay by substituting a continuously observed proxy for the claims process itself. He and Mintz (2026) similarly optimise auto-insurance pricing prescriptively against telematics data, learning the pricing rule directly from behavioural signals rather than filtering a reporting-delayed claims process. This paper instead filters the delay in the claims signal directly, which is the relevant problem for the non-telematics and commercial lines, where no such behavioural proxy exists.

Table 1. Positioning relative to four studies and the two research gaps this paper addresses.

Literature	What It Provides	What It Leaves Open
Dynamic insurance pricing (Section 2.1)	Margin-volume trade-off; feedback pricing under demand response	No explicit occurrence-to-reporting mechanism; loss observed near-contemporaneously
Reporting delay/reserving (Section 2.2)	Point-process and micro-level occurrence-reporting models; development-factor machinery	Prediction, not sequential control; premium path not optimised
Partial observation/risk-sensitive control (Section 2.3)	Belief-state sufficiency; entropic certainty equivalents; belief-MDP existence results	Rarely combined with an explicit, age-structured reporting-delay state
Filter and approximation validation (Section 2.4)	Standards for judging an approximate filter or discretised control solution	Rarely applied *within* an applied insurance pricing paper

3. Structural Occurrence-Reporting Model

This section builds the structural model of premium-sensitive exposure, latent-regime claim occurrence, and delayed reporting that Section 4 will price and Section 5 will discretise.

3.1. Premium and Price-Sensitive Exposure

Consider a homogeneous rating cell over a finite horizon $[0, T]$. The annual premium $p(t)$ is governed by an adjustment rate $a(t)$, subject to a premium corridor and a maximum permitted movement. In the monthly implementation, the control becomes a discrete move of −EUR 25, EUR 0, or +EUR 25:

$$\mathrm{d}p(t) = a(t)\,\mathrm{d}t, \qquad a(t) \in \mathcal{A}(p(t)), \qquad p_{\min} \le p(t) \le p_{\max}. \tag{1}$$

Effective exposure is represented by an exponential response to relative price. Observed covariates may include seasonality, channel, competitor indices, and portfolio mix:

$$q(t) = q_0 \exp\{-\eta[p(t)/m(t) - 1] + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{z}(t)\}. \tag{2}$$

Equation (2) is an instantaneous exposure approximation. A production model for annual contracts would augment the state with new-business, renewal, lapse, and expiry cohorts; this extension is discussed in Section 10. In the synthetic experiment of Section 7, $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{z}(t) \equiv 0$ (no covariate drift), $q_0 = 5000$ policies, $\eta = 2.0$, and $m(t) \equiv$ EUR 700.

3.2. Latent Regimes and Marked Claim Occurrences

Let $Z(t)$ take values in $\{1, \ldots, K\}$ ($K = 2$ in the implementation, with regimes labelled L and H) and follow a continuous-time Markov chain with generator Q. Conditional on $Z(t) = i$, claim occurrences form a Cox process with intensity

$$\lambda_i^O(t) = q(t)\, f_i\, \exp\{\boldsymbol{\alpha}_i^{\mathsf{T}} \mathbf{z}(t)\}. \tag{3}$$

A claim occurring in regime i receives a severity mark Y with distribution F_i, mean μ_i, and finite moment-generating function in a neighbourhood of zero. The occurrence regime remains attached to the claim after a later regime change: a claim that occurred while $Z = \mathrm{H}$ keeps its H-regime severity distribution even if the environment has since reverted to L by the time it is reported. The insurer's belief about *which regime generated* a given unreported claim, not just about the current regime, is therefore part of the relevant state; Section 5 tracks it through the posterior weights $(1 - \pi_t, \pi_t)$ applied to the two regime-conditional Poisson likelihoods.

Section 9.3 discusses how such regimes can be identified and estimated from real claims history; practical examples include a litigation-environment indicator for liability lines, a threat-activity indicator for cyber, and a claims-inflation indicator more generally, each recoverable via the hidden Markov, state-space, or Cox-process regime-detection methods cited there.

3.3. Atomic Claim Population and Intensity-Valued Transport Equation

A pathwise unreported-claim population is not a smooth density. It is the atomic random measure

$$U_i(t, \mathrm{d}x) = \sum_n \mathbf{1}\{Z(T_n) = i,\ T_n \le t < T_n + D_n\}\, \delta_{t - T_n}(\mathrm{d}x), \tag{4}$$

where T_n and D_n are occurrence times and reporting delays. Conditional on the latent occurrence intensity, the associated mean intensity density $\bar{u}_i(t, x)$ satisfies the age-structured transport equation

$$\partial_t \bar{u}_i(t, x) + \partial_x \bar{u}_i(t, x) = -h_i(x, z(t))\, \bar{u}_i(t, x), \qquad x > 0, \tag{5}$$

$$\bar{u}_i(t, 0) = \lambda_i^O(t)\, \mathbf{1}\{Z(t) = i\}, \qquad \bar{u}_i(0, x) = \bar{u}_{i,0}(x). \tag{6}$$

The conditional report intensity and expected IBNR count are

$$\lambda^R(t) = \sum_i \int_0^\infty h_i(x, z(t))\, \bar{u}_i(t, x)\, \mathrm{d}x, \qquad N^{\mathrm{IBNR}}(t) = \sum_i \int_0^\infty \bar{u}_i(t, x)\, \mathrm{d}x. \tag{7}$$

When h_i depends only on age, and the initial profile is zero, the method of characteristics gives $\bar{u}_i(t, x) = \lambda_i^O(t - x) S_i(x)$, where S_i is the reporting-delay survival function, so the report intensity is the convolution of past occurrence intensity with the reporting-delay density. The transport equation is exact for the conditional intensity field; the full hidden claim population remains measure-valued. Section 5 replaces it with a phase-type (constant-hazard) approximation so that the resulting filter is finite-dimensional and can be validated exactly (Section 6.4); the cost of that tractability is quantified, not assumed away, by the exact-filter comparison.

3.4. Observations and Nominal Filtering

The insurer observes premium, exposure, covariates, report times, and available claim marks, but not the current regime or the unreported population. Let $M(\mathrm{d}t, \mathrm{d}y)$ denote the marked report measure. Its compensator, conditional on the hidden state, is

$$\nu_t(\mathrm{d}y)\,\mathrm{d}t = \sum_i \Big[\int_0^\infty h_i(x, z(t))\, U_i(t, \mathrm{d}x) \Big] F_i(\mathrm{d}y)\,\mathrm{d}t. \tag{8}$$

Two distinct kinds of uncertainty appear in this paper, and only one is addressed by the filter developed in this section. State uncertainty is uncertainty about the current value of an unobserved state variable—here, the regime $Z(t)$ or the unreported-claim population $U(t, \cdot)$—given a known transition law and observation model; this is what the nominal Bayesian filter below reconstructs. Parameter or model uncertainty is uncertainty about the transition law or observation model itself—here, the transition matrix P, the reporting parameters (d_0, δ), and the severity distributions. The filter developed in this section takes all of these as known and fixed at their reference (nominal) values; Section 7's true/reference mismatch is a deliberate misspecification used to test the resulting policy's robustness (Section 8.5), not a mechanism by which the filter itself learns or represents parameter uncertainty. Section Structural and Scope Limitations's fifth limitation states explicitly that genuine parameter ambiguity—as distinct from the single-nominal-filter risk sensitivity of Section 4—is not addressed here and would require a materially different construction (multiple filters, robust Bayesian methods, or a set-valued information state).

The nominal information state is $\Pi_t = \mathrm{Law}[Z(t), U(t, \cdot) \mid \mathcal{G}_t]$. Between reports, the posterior is predicted under the nominal transition and survival model; report arrivals produce likelihood reweighting. This nominal posterior is used throughout: where Section 7 evaluates policies under a *true* model that differs from this nominal filtering model, the filter itself continues to run under the nominal (reference) parameters—it is deliberately, and only, the world being observed that changes, not the observer's internal model of it, since an insurer cannot filter using parameters it does not know.

3.5. Economic Reward

Economic performance is measured on an incurred basis so that a strategy is not rewarded merely for moving reports beyond the horizon. Over a monthly interval, premium and variable expense are determined by exposure, while ultimate loss is generated by all claims occurring in that month, regardless of when they are reported. Reporting affects decisions (through the filtered state) but not the accounting time of loss recognition; Appendix C restates this precisely, since it materially affects how the Bornhuetter–Ferguson and delay-naive benchmarks of Section 7.3 are scored relative to their own, reporting-lagged, internal beliefs.

4. Loss-Aware Risk-Sensitive Control

This section turns the occurrence-reporting structure of Section 3 into a decision criterion: it defines the entropic certainty equivalent, states its continuous-time characterisation, and derives the closed-form identity that lets the resulting Bellman recursion be evaluated without approximating the loss distribution by simulation.

4.1. Entropic Certainty Equivalent

For total reward X^a under an admissible policy a, define

$$J_\gamma(a) = \mathbb{E}[X^a] \text{ if } \gamma = 0; \qquad J_\gamma(a) = -\gamma^{-1} \log \mathbb{E}[\exp(-\gamma X^a)] \text{ if } \gamma > 0. \tag{9}$$

Positive γ places greater weight on low-reward nominal outcomes. The variational identity for relative entropy writes the certainty equivalent as the minimum expected reward under a reweighted nominal distribution plus a Kullback–Leibler penalty: a dual preference representation, not a robust-filter theorem and not a claim of physical transition-law ambiguity. A decision-maker with preferences (9) under the nominal law behaves *as if* conservatively reweighting nominal outcomes, not holding reweighted beliefs or running a different filter—so the γ-tilted Poisson means inside the Bellman recursion below are an artefact of the exponential-tilting identity used to evaluate the certainty equivalent in closed form (Equation (15)), not a statement that the insurer believes claims arrive at a different rate.

Practical Interpretation of γ

For small γ, a Taylor expansion of Equation (9) gives $CE_\gamma(X) \approx \mathbb{E}[X] - (\gamma/2)\mathrm{Var}(X)$: the certainty equivalent trades expected reward against outcome variance at an implicit price of $\gamma/2$ per unit of variance, in the same EUR-million units as every other quantity in this paper. This gives γ an interpretation that actuarial readers can calibrate directly. A required γ can be backed out from a target relationship between mean profit and volatility (for example, from an internal risk-appetite statement expressed as a variance or standard-deviation loading), rather than treated as an abstract control-theoretic tuning constant. In this paper, γ is calibrated empirically rather than from an external risk-appetite statement, via the grid search reported in Section 8.3. The full range $\gamma \in \{0, 0.4, 0.8, 1.2\}$ is solved and evaluated in full, and the production value $\gamma = 1.2$ is the point in that range at which the matched static price's own 5th-percentile outcome first exceeds the dynamic policy's (Section 8.1). This makes $\gamma = 1.2$ the most informative single value for illustrating the risk-sensitivity/dynamic-control trade-off that is this paper's central finding.

4.2. Formal Separated Continuous-Time Characterisation

Let $S_t = (p(t), \Pi_t, z(t))$ denote the separated nominal state, and let $\mathcal{G}^a$ be its controlled generator, including premium drift, transport of the unreported population, posterior prediction, and report jumps. For terminal reward g and running reward r, define $\Psi_\gamma = \exp(-\gamma V_\gamma)$. Formally, the risk-sensitive dynamic programming equation can be written in exponential form as

$$\partial_t \Psi_\gamma(t,S) + \inf_a \left\{ \mathcal{G}^a \Psi_\gamma(t,S) - \gamma\, r(S,a)\, \Psi_\gamma(t,S) \right\} = 0, \qquad \Psi_\gamma(T,S) = \exp[-\gamma g(S)]. \quad (10)$$

The equivalent equation for V_γ is nonlinear and infinite-dimensional because Π_t is a probability law over a measure-valued state. Establishing comparison and verification results on that space—in the sense developed for controlled Markov processes and viscosity solutions by Fleming and Soner (2006) and for continuous-time stochastic control more generally by Pham (2009)—is outside the scope of this paper. Bäuerle and Rieder (2017) establish existence and a belief-state reformulation for the discrete-time, general-certainty-equivalent partially observed Markov decision process that Equation (10) is the continuous-time, exponential-utility special case of; the present contribution is confined to a specific, finite-state, insurance-structured instance of that general result, solved and validated (Section 6) rather than analysed for its own sake. The numerical model below is therefore a phase-type finite-state approximation, not a direct solution of Equation (10).

Three properties of the continuous-time formulation are retained in the finite-state approximation solved from Section 5 onward: the belief-state reformulation that justifies collapsing the insurer's information into a filtered state Bäuerle and Rieder (2017); the entropic certainty-equivalent decision criterion itself (Equation (9)); and the boundary-driven transport mechanism by which reporting delay enters the state (Equation (6)). Not retained,

and not claimed, is any result establishing that the finite-state solution converges, as the grid is refined, to a viscosity solution of Equation (10) in the sense of Fleming and Soner (2006) or Pham (2009); Section 6.1's grid-refinement comparison is an empirical diagnostic at one calibration.), not a convergence proof, and Section 2.3 states explicitly that this paper does not attempt the infinite-dimensional verification such a proof would require.

4.3. The Gamma-Severity Moment-Generating Function and Its Existence Condition

Let severity in regime i be Gamma(k_i, θ_i), with moment-generating function

$$M_i(s) = (1 - \theta_i s)^{-k_i}, \qquad \text{defined only for } s < 1/\theta_i. \tag{11}$$

Because the entropic Bellman recursion (Equation (17) below) evaluates $M_i(\gamma/10^6)$ at the policy's own risk-sensitivity parameter, Equation (11) imposes a hard, checkable constraint on admissible γ:

$$\gamma < \gamma_i^{\text{crit}} := 10^6/\theta_i \qquad \text{for every regime } i. \tag{12}$$

If γ approaches γ_i^{crit} from below, the exponential tilting places increasing mass on large losses and the recursion becomes numerically unstable well before Equation (12) is formally violated; if γ exceeds γ_i^{crit}, the exponential moment $\mathbb{E}[e^{\gamma L_t/10^6}]$ is infinite, and Equation (15) below is meaningless. In the calibration used throughout Section 7 ($\theta_{\text{L}} =$ 11,250, $\theta_{\text{H}} =$ 15,750), this gives $\gamma_{\text{L}}^{\text{crit}} \approx 88.9$ and $\gamma_{\text{H}}^{\text{crit}} \approx 63.5$ per EUR million, so the production value $\gamma = 1.2$ sits at roughly 2% of the binding constraint and the tilted mean report rate at the most extreme grid state (reservoir at its upper boundary, lowest corridor premium, high-severity regime) is EUR 49.4 versus an untilted EUR 49.3—a negligible shift, confirming that risk sensitivity at the production calibration operates on the *decision criterion* rather than materially distorting the effective claims distribution. Section 6.1 reports the companion check on the report-count truncation.

4.4. Operational Constraints

The admissible set may incorporate premium floors and ceilings, maximum monthly movement, minimum conversion or retention, capacity and capital constraints, permitted-variable rules, monotonicity, anti-discrimination requirements, and human approval. The synthetic experiment uses a EUR 600–950 corridor and moves of EUR 25 per month, consistent with Equation (1).

5. Finite-State Approximation and Bellman Algorithm

5.1. One-Phase Reporting Reservoir

The general transport model of Section 3.3 does not assume memoryless reporting; the phase-type approximation introduced below deliberately restricts to the memoryless, one-phase case in order to keep the resulting filter finite-dimensional and exactly checkable (Section 6.4). Readers working with lines whose reporting tail is genuinely long and non-geometric—the liability and cyber lines that motivate Section 1's practical framing—should treat the results of Sections 6–8 as demonstrating the framework's mechanics on a tractable special case, and Section Structural and Scope Limitations's first limitation as the relevant scope boundary, rather than as calibrated evidence for those specific lines.

Relative to Section 3.3's general age-structured transport equation, the one-phase phase-type approximation used from here on preserves the boundary-driven inflow structure (unreported claims still enter as a boundary condition on occurrence, Equation (6)), the concept of a reporting hazard, and exactness for the mean reservoir process. It loses two features: age-dependence of the reporting hazard (h_i(x,z) collapses to a single constant

δ, so Section 3.3's age profile is replaced by a memoryless, geometric reporting time), and the covariance between the reservoir's regime composition and the number of reports observed in a given month, which Section 5.1's closing paragraph already identifies as what the exact-filter comparison of Section 6.4 is built to quantify.

A phase-type approximation replaces the age profile of Equations (5) and (6) with memoryless reporting stocks. In the implemented one-phase model, a fraction d_0 of occurrences is reported in the occurrence month, and the remainder enters a delayed reservoir. Each reservoir claim reports in a subsequent month with hazard δ. The state at the beginning of month t is $s_t = (p_t, \pi_t, u_t)$, where π_t is the nominal posterior probability of the high-loss regime and u_t is the filtered expected delayed-claim stock:

$$\lambda_i(p) = q(p) f_i / 12, \qquad R_t \mid Z_t = i, s_t, p' \sim \text{Poisson}[\delta u_t + d_0 \lambda_i(p')]. \tag{13}$$

After observing r reports, the current-regime posterior is updated with the two Poisson likelihoods and then propagated through the nominal transition matrix P. The assumed-density reservoir closure is

$$u_{t+1} = (1-\delta) u_t + (1-d_0)\left[(1-\pi_t^+)\lambda_{\mathrm{L}}(p') + \pi_t^+ \lambda_{\mathrm{H}}(p')\right]. \tag{14}$$

This state is an approximate Markov information state induced by the assumed-density closure of Equation (14), not an exact sufficient statistic of the original hidden process; the exact sufficient statistic is the full belief Π_t of Section 3.4 (a probability law over the measure-valued state $(Z(t), U(t,\cdot))$), which Equation (14) replaces with its first two moments for tractability. Section 6.4's exact-filter comparison exists specifically to quantify the resulting gap between (π_t, u_t) and Π_t's own reservoir-belief content.

The closure replaces random occurrence inflow by its posterior expectation and suppresses some regime-reservoir dependence; because the per-claim reporting hazard δ is constant across claim age, the *aggregate* reservoir dynamics implied by (14) are exact for the mean process even though individual claim ages are not tracked—what the closure discards is the covariance between the reservoir's regime composition and the number of reports observed in a given month, which is exactly what the exact integer-state filter of Section 6.4 is built to quantify.

5.2. Compound Poisson–Gamma Loss Inside the Entropic Operator

Split monthly occurrences into immediate reports $I_t \sim \text{Poisson}(d_0\lambda_i)$, delayed occurrences $J_t \sim \text{Poisson}((1-d_0)\lambda_i)$, and reports from the existing reservoir $D_t \sim \text{Poisson}(\delta u_t)$. Then $R_t = D_t + I_t$ and incurred aggregate loss L_t contains severities from $I_t + J_t$. A direct calculation, using the Poisson thinning property (immediate and delayed occurrence counts are independent Poisson variates given the regime) and the compound-Poisson moment-generating function, yields

$$\mathbb{E}\left[e^{\gamma L_t/10^6}\mathbf{1}\{R_t = r\} \,\middle|\, Z_t = i, s_t, a\right] = e^{\lambda_i[M_i(\gamma/10^6)-1]}\,\text{Pois}\big(r;\, \delta u_t + d_0\lambda_i M_i(\gamma/10^6)\big). \tag{15}$$

Equation (15) is the key computational identity: it retains the correlation between immediate reports and incurred loss while integrating severity risk analytically, so that the Bellman recursion sums over integer report counts r rather than over the joint support of counts and severities. The posterior update itself continues to use the nominal, *untilted* report means $\delta u_t + d_0\lambda_i$—the filter never sees the risk-tilted distribution, consistent with the separation of nominal filtering and risk preference stated in Section 4.

5.3. Risk-Sensitive Bellman Recursion

An action $a \in \{-1, 0, 1\}$ sets $p' = p_t + 25a$. Define the non-loss monthly cash contribution, in EUR million, as

$$b(p', a) = q(p')(p' - c_e)/(12 \times 10^6) - \kappa a^2, \qquad \kappa \text{ expressed in EUR million per full move.} \tag{16}$$

The reference calibration reports the movement penalty in its more legible native unit, EUR 1500 per full move of EUR 25, not in the same units as every other term in $b(p', a)$.(p',a). Equation (16) requires it in the same €-million units as every other term in b(p',a), so the conversion entering the recursion is $\kappa = 1500/10^6 = 0.0015$. This conversion is stated explicitly here because a formula that mixes an EUR-denominated parameter with an EUR-million-denominated objective is a unit error, not a matter of taste. Every numerical result in Section 8 uses $\kappa = 0.0015$. Let S_{t+1}(r) denote the filtered successor state generated by r reports. With regime weights $\pi_L = 1 - \pi_t$ and $\pi_H = \pi_t$, the loss-aware recursion is

$$V_t(s) = \max_a \Big\{ -\gamma^{-1} \log \sum_r e^{-\gamma[b(p',a) + V_{t+1}(S_{t+1}(r))]} \sum_i \pi_i W_{ir}(\gamma) \Big\}, \tag{17}$$

where $W_{ir}(\gamma)$ is the right-hand side of Equation (15). At $\gamma = 0$, Equation (17) is interpreted by continuity and becomes the ordinary conditional expectation of premium less expense, incurred loss, and movement cost plus continuation value; because the $\gamma \to 0$ limit of $-\gamma^{-1} \log \mathbb{E}[e^{-\gamma X}]$ is numerically delicate near $\gamma = 0$ (it is a $0/0$ form), the implementation used throughout this paper does *not* take that limit numerically—it uses a separately coded, exact expected-value branch (Appendix A) for $\gamma = 0$, so that the "dynamic nominal" policy reported in Section 8 is not an artefact of near-singular floating-point cancellation.

A further implementation detail, made explicit because it affects how policies are applied online (Section 7): rather than collapsing each month's solution immediately to a single scalar value V_t and a categorical action, the recursion stores, for every grid state and action, the corresponding action-value ("Q") surface $Q_t(p, a, \pi, u)$. Online, at a continuous belief point (π, u) off the grid, each of the three stored Q-surfaces is bilinearly interpolated separately, and the action chosen is the argmax of the three *interpolated* values (Appendix A, step 8)—interpolating a categorical policy directly (looking up the nearest grid point's already-chosen action) can instead produce a different, suboptimal action near an action boundary such as the one visible in the action-surface results in Section 8.

5.4. Common Random Numbers for the Monte Carlo Evaluation

Every strategy comparison result reported in Section 8 is a *paired* comparison across the same set of simulated paths, and every reported confidence interval is a paired bootstrap interval. Both of these are valid only if the simulated paths for different strategies are constructed from genuinely common random numbers rather than merely from the same random seed. This section documents that construction in detail, since a valid paired comparison across strategies depends on it.

For each of the $N = 3000$ synthetic paths, and independently for each calendar month $t = 1, \ldots, 36$, five independent uniform variates are drawn once, before any strategy is simulated: u_t^Z (regime transition), u_t^N (occurrence count), u_t^I (immediate/delayed split), u_t^D (reservoir report count), and u_t^S (aggregate severity). The regime path Z_t is generated once per synthetic path under the *true* transition matrix P^{true} and shared across every strategy evaluated on that path, since the regime process is exogenous to the pricing decision. For every other component, each strategy maps the *same* uniform variate through its own regime- and price-dependent inverse cumulative distribution function—for example, the occurrence count under strategy s at month t is $N_t^{(s)} = F^{-1}_{\text{Poisson}(\lambda_{Z_t}(p_t^{(s)}))}(u_t^N)$, using the

price $p_t^{(s)}$ that strategy s itself has chosen. Because $F^{-1}_{\text{Poisson}(\lambda)}(u)$ is non-decreasing in λ for fixed u, two strategies that charge different prices in the same month receive comonotonically coupled, not independent, claim counts: the strategy with higher exposure receives weakly more claims for the *same* draw of the underlying uniform. This is the standard, valid construction of common random numbers when compared systems have different parameters (Law and Kelton 2000), and it is what allows the paired bootstrap intervals of Section 8 to isolate the policy-attributable component of an outcome difference from the shared component of claims-process noise. Aggregate severity uses the same coupling at the compound level: since the sum of n i.i.d. Gamma(k_i, θ_i) variates is itself Gamma(nk_i, θ_i), the aggregate severity draw for a month with n occurrences is $F^{-1}_{\text{Gamma}(nk_i,\theta_i)}(u_t^S)$—exact, using the same shared uniform u_t^S across strategies even though n differs by strategy.

5.5. Numerical Implementation and Grid Dimensions

The production grid contains 15 premiums from EUR 600 to EUR 950, 21 posterior probabilities, and 31 reservoir values from 0 to 150 claims, giving 9765 states per month over 36 months; report counts 0–150 are summed exactly in Equation (17). Continuation values are evaluated by bilinear interpolation in posterior probability and reservoir stock; premium remains on its EUR 25 grid, so p' is always an exact grid point, and no interpolation is needed in that dimension. All three action-value surfaces are stored so that online actions are selected after interpolation rather than by interpolating a categorical policy, as described above.

Justification for grid dimensions. The grid dimensions were selected as follows. The premium grid (15 values, EUR 25 step) is determined by the governance constraint (EUR 600–950 corridor, EUR 25 moves) and is therefore exact—no approximation is introduced in the premium dimension. The posterior probability grid (21 values, step 0.05) was chosen to resolve the posterior near the decision boundaries visible in the action-surface results in Section 8, where the optimal action changes discontinuously; the hold region's boundaries in the action-surface results in Section 8 occur at approximately $\pi \approx 0.15$ and $\pi \approx 0.35$, so the 0.05 step resolves these boundaries with roughly 4–7 grid points between them. The reservoir grid (31 values, 0–150, step 5) was chosen so that the steady-state filtered reservoir under the production policy (approximately 40–60 claims) sits comfortably in the interior, with margin for transient excursions; Section 6 confirms that the grid boundary is never approached in the Monte Carlo evaluation. Section 6 reports a companion solve on a finer $15 \times 41 \times 51$ grid (posterior step 0.025, reservoir range 0–200 at step 4) with report counts summed to 200, used solely to bound the discretisation error of the production grid rather than as a production policy in its own right.

Sensitivity to alternative grid specifications. The grid-adequacy finding of Section 6 is calibration-specific: it demonstrates that the production grid is adequate for the parameter values of the reference calibration, but it does not establish general numerical convergence across alternative parameter configurations. In particular, a slower reporting hazard δ would push the steady-state reservoir closer to the boundary, potentially requiring a finer or extended grid; a higher baseline exposure q_0 or higher occurrence frequency f_i would increase the occurrence intensity and hence the reservoir's typical level; and a larger number of regimes would require a higher-dimensional belief space that the current grid structure cannot accommodate. Any redeployment of the model outside the parameter range studied here should re-run the grid-refinement and exact-filter validation exercises of Section 6.

6. Numerical and Filter Verification

Having specified the approximation in Section 5, this section checks it. This section reports four verification exercises, in increasing order of what they can and cannot demonstrate: a report-count truncation and state-grid refinement check (Section 6.1), a separately coded one-step Bellman residual check (Section 6.2), an in-model forward-simulation check of the timing and expected-value recursion (Section 6.3), and a comparison of the assumed-density reservoir closure against an exact truncated integer-state filter (Section 6.4). None constitutes a proof that the production policy is globally optimal; each is scoped to the specific approximation it checks, and Section 10 restates those scopes alongside what remains unverified.

6.1. Report-Count Truncation and State-Grid Refinement

Two distinct truncations enter the production solve: the report-count sum in Equation (17) is truncated at $r = 150$, and the (π, u) belief grid is truncated at $u = 150$ filtered delayed claims. Table 2 reports the largest nominal and exponentially tilted Poisson report-count means that actually arise on the production grid (at the lowest corridor premium, EUR 600, which maximises exposure, and the reservoir at its upper boundary, $u = 150$), and the resulting omitted tail probability under the production truncation at $r = 150$.

Table 2. Report-count truncation diagnostics at the most extreme production-grid state (p = EUR 600, $u = 150$, high-loss regime), $\gamma = 1.2$.

Quantity	Value
Occurrence intensity $\lambda_H(600)$	47.13 claims/month
Nominal report mean $\delta u + d_0\lambda_H$	49.28
Gamma-MGF value $M_H(\gamma/10^6)$	1.0085
Tilted report mean $\delta u + d_0\lambda_H M_H(\gamma/10^6)$	49.38
Omitted tail probability, $1 - \text{Pois}(150; \text{tilted mean})$	$<10^{-16}$ (numerically zero)

The report-count truncation at $r = 150$ is therefore safe by a very wide margin at every state the production grid can represent—the worst-case tilted mean (49.4) sits more than 20 standard deviations of a Poisson (49.4) below the truncation point. This finding, together with the negligible effect of the tilting itself documented in Section 4.3, isolates the report-count truncation as *not* a material source of approximation error in this calibration.

The reservoir boundary at $u = 150$ is a separate question, since it depends on sustained, not instantaneous, exposure. Under the assumed-density closure's own steady-state condition, a policy that priced permanently at the lowest corridor premium (EUR 600) while permanently in the high-loss regime would imply a steady-state filtered reservoir of $u^\star = (1 - d_0)\lambda_H(600)/\delta \approx 141.4$ claims, close to the grid boundary. This scenario does not arise along any policy actually implemented here (observing a high posterior probability of the loss regime raises price, which lowers exposure and hence the reservoir's steady state), and the instrumented Monte Carlo run confirms this directly: across all 3000 evaluation paths and 36 months, the filtered reservoir belief generated by the production policy never exceeded 67.6 claims, and reached 90% of the grid boundary (135 claims) on 0.00% of path-months—a property of the *policy*, not an a priori guarantee of the grid, which is why the grid-refinement check below is reported empirically rather than only analytically.

Table 3 compares the production $15 \times 21 \times 31$ grid against a finer $15 \times 41 \times 51$ grid (posterior probability step 0.025 instead of 0.05; reservoir range extended to 0–200 at step 4 instead of 0–150 at step 5; report counts summed to 200). Both grids are solved

independently at $\gamma = 1.2$, and both are then simulated forward through the identical 3000 common-random-number paths of Section 5.4.

Table 3. Grid-refinement check: production ($15 \times 21 \times 31$) versus fine ($15 \times 41 \times 51$) grid, $\gamma = 1.2$.

Quantity	Production Grid	Fine Grid
Initial value $V_0(p = 750, \pi = 0.4, u = 0)$, EUR million	3.8488	3.8475
Percentage difference from production	—	−0.032%
Month-1 action agreement at $p = 750$, over all (π, u)	—	99.7% (649/651)
Simulated mean profit, EUR million	4.0257	4.0262
Simulated fifth percentile, EUR million	2.3590	2.3546

The production grid changes the interpolated initial value by three-hundredths of one percent relative to the finer grid, and the two grids agree on the recommended month-1 action at 99.7% of the finer grid's belief states that map onto the production grid's coordinates. The simulated mean and downside outcomes under the two grids differ by less than the sampling noise implied by the paths' own standard deviation (≈1.03 EUR million over 3000 paths, i.e., a standard error of about 0.019 EUR million on the mean), so the production grid is adequate for the synthetic experiment reported in Section 8. This does not certify that a $15 \times 21 \times 31$ grid is adequate for a different calibration (in particular, a calibration with a materially slower reporting hazard δ, which would push the steady-state reservoir closer to, or past, the boundary); Section 10 flags this as a specific, checkable prerequisite for redeploying the model outside the parameter range studied here.

6.2. Independent One-Step Bellman Residual Check

A forward Monte Carlo simulation of the policy the recursion itself produces (Section 6.3) checks timing and expected-value consistency, but cannot by itself detect an error present in *both* the recursion and the simulator, since the simulator is built to consume the recursion's own output. This section therefore adds an independent check: a separate, unvectorised, scalar-loop re-implementation of the right-hand side of Equation (17), written without reusing any of the array-broadcasting logic of the production solver, evaluated at 25 randomly sampled (t, p, π, u) states drawn uniformly from the production grid and horizon at $\gamma = 1.2$, and compared against the value the production solver stored for the same state. The re-implementation is independent except for one shared component: both it and the production solver call the same bilinear-interpolation utility to evaluate the stored continuation value V_{t+1} at a non-grid point. A defect specific to that shared utility—as opposed to the surrounding Bellman logic re-implemented separately in each case—would not be caught by this check.

At every sampled state, the independent scalar computation loops explicitly over $r = 0, \ldots, 150$, recomputes the posterior update π^+ and the reservoir update u' from the defining formulas (Equations (14) and the paragraph preceding Equation (15)) rather than from any array cached by the production solver, calls the same bilinear interpolation utility on the (independently indexed) stored continuation value V_{t+1}, and evaluates Equations (15) and (17) directly. Table 4 reports the resulting residual $V_t^{\text{independent}}(s) - V_t^{\text{stored}}(s)$ across the 25 sampled states.

Table 4. Independent one-step Bellman residual check, $n = 25$ randomly sampled states, $\gamma = 1.2$.

Statistic	Value (EUR Million)
Maximum absolute residual	4.4×10^{-16}
Mean absolute residual	1.7×10^{-16}
Root-mean-square residual	2.4×10^{-16}

The residuals are at the level of double-precision floating-point round-off, the strongest form of agreement two separately coded numerical implementations of the same recursion can exhibit, subject to the shared-interpolation-utility caveat noted above. This confirms that the vectorised production solver correctly implements Equations (15)–(17) as specified—it does *not* confirm those equations are themselves the correct model of the insurer's problem (Section 10), rule out a defect shared through the common interpolation routine, or confirm global convergence of the finite-state approximation to the underlying continuous-time problem, which Section 2.3 explicitly declines to claim.

This shared-utility gap is now closed by an independent, algorithm-level validation of the bilinear-interpolation routine itself, run separately from, and reported here in addition to, the residual check above. A from-scratch scalar bilinear-interpolation function is cross-checked against SciPy's independently engineered RegularGridInterpolator across 20,000 randomly sampled query points on a grid matching the production posterior-probability/reservoir-stock dimensions (21×31 on $[0, 1] \times [0, 150]$). This function is an explicit index search over the grid, written without reusing any code from the production solver or from the Section 6.2 residual check above. The two implementations agree to floating-point round-off (maximum difference 1.8×10^{-15}, mean 2.9×10^{-16}). Both implementations also reproduce a genuinely bilinear test surface exactly, as bilinear-interpolation theory requires (maximum error 5.7×10^{-14}, itself at floating-point precision). They also exhibit the theoretically expected $O(h^2)$ convergence on a curved test surface when the grid spacing is halved (maximum error falls from 4.73×10^{-3} to 1.16×10^{-3}, a factor of 4.07 against a theoretical factor of 4). This check is deliberately scoped: it validates the interpolation algorithm in isolation, against an independent implementation and against analytic ground truth, on representative grids and surfaces, rather than by re-executing the full production Bellman recursion. A re-run of the full recursion is outside what an interpolation-only check can or should claim to do, so this check therefore does not, on its own, re-verify the specific numerical values reported in Sections 6–8. It does close the specific gap identified above; however, a coding defect confined to the shared interpolation routine would have to reproduce the same error independently in a from-scratch re-implementation and in SciPy's separately maintained library code. That is not a plausible failure mode for a project-specific bug.

6.3. In-Model Forward-Simulation Check

As a secondary, lower-resolution check, 15,000 forward simulations of the expected-value ($\gamma = 0$) policy, run entirely *within* the nominal approximate model, produced a mean objective of EUR 4.144771 million, compared with an interpolated Bellman initial value of EUR 4.151180 million. The difference, EUR -0.006409 million, is smaller than the Monte Carlo standard error of EUR 0.008179 million. This validates the timing and expected-value recursion under the model's own assumptions; it is retained alongside, not in place of, the independent residual check of Section 6.2, since the two checks have different failure modes—this one would detect an error in how successive months are chained together (e.g., an off-by-one error in which month's premium feeds which month's exposure) that a single-state residual check cannot catch, but would not detect a bug shared between the recursion and its own simulator. Direct Monte Carlo validation of the exponential certainty equivalent itself is less stable because rare severe losses dominate the exponential moment; Equation (15) is therefore evaluated analytically in the Bellman step rather than approximated by simulation at any $\gamma > 0$.

6.4. Exact Integer-State Filter Comparison

For a memoryless reservoir, the exact hidden state is the integer pair (Z_t, U_t), not merely its posterior means (π_t, u_t). This subsection compares the assumed-density closure of Equation (14) against an exact, truncated (at $U_{\max} = 120$ claims) Bayesian filter that propagates the full joint posterior mass over (Z, U), run on the *same* observed report sequence as the assumed-density filter. The validation uses 300 independently simulated paths and 10,800 month-state comparisons, a sample large enough to support paired standard errors and confidence intervals rather than point estimates alone.

Four quantities are reported, using terminology chosen specifically to avoid implying a stronger validation than the underlying computation supports. First, the *approximate-policy action agreement*: at each state, the production Q-surfaces (fitted under the assumed-density closure) are queried twice, once at the assumed-density filter's own (π_t, u_t) and once at the exact filter's posterior *means* $(\bar{\pi}_t, \bar{u}_t)$ substituted into the same surfaces; the reported agreement rate is how often the two queries recommend the same action. This is deliberately *not* a re-optimised policy on the exact belief simplex—the exact filter's richer posterior (which, unlike the assumed-density closure, carries information in its shape, not just its first two moments) is being evaluated only through the coarser lens of the surfaces the assumed-density policy already committed to. Second, the *mean approximate Q-gap*: the difference between the best and the exact-moment-recommended action's interpolated Q-value, again on the assumed-density surfaces, reported in plain EUR (not thousands, to avoid the unit ambiguity a bare "(thousand)" label invites) so that its magnitude is stated unambiguously. Third, a directly simulated *filter-induced policy gap*: a second, independent closed-loop trajectory in which decisions are driven by the exact filter's own posterior means fed into the same stored Q-surfaces (rather than by the assumed-density filter's own recursive belief), simulated on the same underlying randomness, with the resulting paired economic-profit difference against the standard assumed-density closed loop reported directly—together with its paired standard error, a 95% bootstrap interval (2000 resamples of the path index), and the corresponding 5th-percentile and TVaR 5% differences—rather than inferred indirectly from Q-value gaps or reported as a point estimate alone. Fourth, a *margin-conditional* breakdown of action agreement: states are grouped into quartiles by |margin|, the gap in EUR between the assumed-density policy's best and second-best action at the moment of decision, to test directly whether disagreement between the two filters concentrates in states where the original decision was already a close call.

Figure 2 plots the assumed-density posterior probability against the exact filter's posterior mean across the 10,800 sampled month-states. Three findings should be reported together, since any one alone would be misleading. First, the posterior tracking is imperfect: an RMSE of 0.173 and an approximate-policy action agreement of 64.0%. This reflects a genuine limitation of the assumed-density closure, reported directly rather than adjusted, because reporting only a favourable-looking diagnostic would defeat the purpose of an independent validation exercise. Second, the directly simulated, paired filter-induced policy gap is estimated precisely enough to be statistically distinguishable from zero on its mean: -0.0092 million on average, with a 95% bootstrap interval of -0.0119 to -0.0068 million that excludes zero. This should be read plainly: the assumed-density closure carries a small but real, statistically detectable mean-profit cost, on the order of 0.2–0.3% of mean profit at this calibration, relative to a policy informed by the exact filter's posterior means. The 5th-percentile and TVaR 5% differences remain statistically indistinguishable from zero even at 300 paths (both intervals span zero and are wide relative to the point estimate).

This reflects sample size specifically: 300 paths/10,800 month-states are adequate to estimate the mean filter-induced gap precisely (SE 0.0013 million; Table 5), but are, by construction, underpowered for the wider and noisier 5th-percentile and TVaR 5% statistics.

Consequently, the tail consequences of the closure's posterior error remain genuinely inconclusive rather than confirmed to be negligible. Increasing the validation sample toward the 1000–2000 path range would directly narrow these intervals and is therefore recorded in Section 10 as a specific, low-risk extension for a future revision cycle.

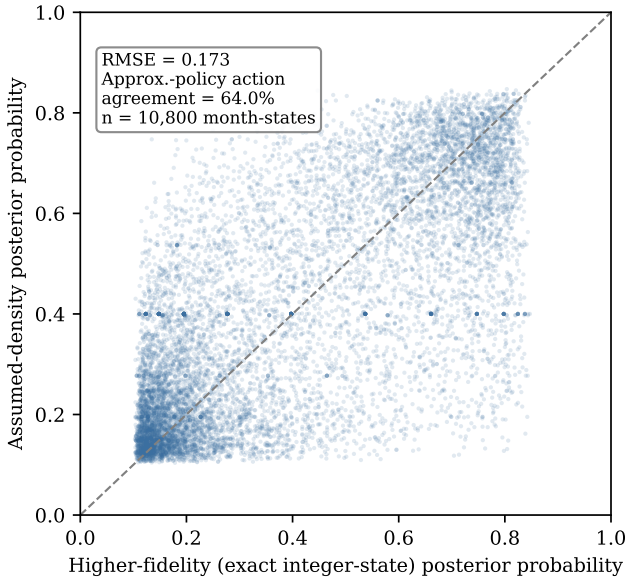

Figure 2. Assumed-density posterior probability of the high-loss regime versus the exact truncated integer-state filter's posterior mean, 10,800 month-states from 300 reference-model paths, $\gamma = 1.2$. Point colors distinguish the plotted observation groups identified in the figure legend; the diagonal line denotes exact equality between the two posterior probabil ities.

Table 5. Assumed-density filter versus the exact truncated integer-state filter, 300 nominal (reference-model) paths, 10,800 month-states, $\gamma = 1.2$. Terminology is defined in the text. Quantities are in plain EUR unless a column states otherwise; "–" indicates no interval is reported for that row.

Validation Measure	Result	95% CI/SE
Posterior probability RMSE	0.173	–
Posterior probability MAE	0.127	–
Reservoir RMSE (claims)	6.87	–
Reservoir MAE (claims)	5.58	–
Approximate-policy action agreement	64.0%	–
Mean approximate Q-gap (EUR)	660	–
95th-pct. approximate Q-gap (EUR)	3646	–
Filter-induced gap, mean (EUR million)	−0.0092	SE 0.0013; [−0.0119, −0.0068]
Filter-induced gap, 5th-pct. diff. (EUR million)	+0.0011	[−0.0207, +0.0176]
Filter-induced gap, TVaR 5% diff. (EUR million)	+0.0038	[−0.0223, +0.0268]

Third, the margin-conditional breakdown in the margin-conditional analysis below shows that action disagreement is not uniform across the belief space. Instead, it is concentrated in states where the assumed-density policy's own decision is already close to indifference between actions, with the disagreement rate falling by more than 20 percentage points as the decision margin widens. This provides the mechanism linking the preceding findings: although the posterior error can appear large on the (π, u) scale, it disproportionately changes decisions only where the Q-surface is already nearly flat across competing actions. These are precisely the states in which selecting the alternative action is, by construction, relatively inexpensive.

Read on its own, 64% could suggest an unreliable approximation; read against the margin-conditional analysis below, it is better understood as: exact and assumed-density

filters disagree most often exactly where the production Q-surface is closest to flat, so that the specific instances of disagreement are disproportionately the low-stakes ones. We convert this into an operational recommendation in Section 9.4: a production deployment's human-review workflow should flag decisions in the closest-call quartile (mean |margin| below approximately EUR 315 at this calibration) for closer scrutiny, precisely because these are both where filter disagreement concentrates and where a wrong decision costs the least—so review effort is directed by the same margin diagnostic used here to validate the filter.

6.4.1. Robustness to the Out-of-Model Evaluation Environment

The validation above generates observations under the *reference* model, the same model the filter itself assumes—a natural first check of the closure's own internal consistency, but not, by itself, evidence about how the closure behaves when fed reports from the *true*, out-of-model environment that generates every headline result in Section 8. Both filters continue to predict using reference parameters regardless of which environment generates the data, since an insurer never has access to the true generative parameters, only to the reports they produce; what changes is only the process generating the observation stream the filters are asked to track. Table 6 repeats the identical 300-path, 10,800-state diagnostic with paths generated under $(P^{\text{true}}, d_0^{\text{true}}, \delta^{\text{true}})$ instead.

Table 6. Assumed-density filter versus the exact filter under nominal-model paths (Table 5) versus the true out-of-model environment, 300 paths each, $\gamma = 1.2$.

Validation Measure	Nominal Paths	Out-of-Model Paths
Posterior probability RMSE	0.173	0.162
Approximate-policy action agreement	64.0%	64.9%
Filter-induced gap, mean (EUR million)	−0.0092	−0.0063
Filter-induced gap, 95% CI (EUR million)	[−0.0119, −0.0068]	[−0.0086, −0.0039]

The closure's approximation quality does not deteriorate under the out-of-model environment; if anything, both the posterior RMSE and the filter-induced mean gap are slightly smaller under true-model observations than under nominal ones, and the mean gap remains statistically distinguishable from zero but of the same small order of magnitude. This is a genuine, if modest, robustness finding, not merely an assumption: the exact-filter diagnostic's headline conclusion—that the closure's posterior error carries a small, detectable mean-profit cost and an inconclusive tail cost—holds up when the filters are tested against the same misspecified environment that generates the paper's substantive results, not only against the reference model the filter was designed around.

Table 7 reports this margin-conditional breakdown.

Table 7. Approximate-policy action agreement conditional on the assumed-density policy's own decision margin (gap between best and second-best action value at the moment of decision), 10,800 month-states, $\gamma = 1.2$.

Margin Quartile	*n*	Mean |Margin| (EUR)	Action Agreement
Q1 (closest call)	2700	315	53.4%
Q2	2700	970	61.0%
Q3	2700	1696	64.9%
Q4 (clearest call)	2700	3347	76.6%

These two findings are not contradictory: the production Q-surfaces are, empirically, comparatively flat in the region of belief space where the two filters disagree most,

so a posterior estimation error that looks large on the (π, u) scale need not translate into a large realised profit consequence—but this is an empirical property of *this* calibration's Q-surfaces, established by direct simulation on a sample large enough to support the claim, and should not be assumed to hold at a different calibration or under a materially different reporting-delay regime without re-running the same check.

Because $U_{\max} = 120$ is itself a truncation, Section 6.1's worst-case steady-state reservoir calculation (EUR 141.4 claims under permanently-lowest-premium, permanently-high-regime conditions) exceeds the exact filter's truncation boundary, even though it does not exceed the production policy's own u-grid boundary of 150. No path in the 300-path validation sample approached this region (the same instrumented check reported in Section 6.1 for the production policy's Monte Carlo evaluation applies equally here, since the exact filter is driven by the same policy's price choices), but the exact-filter truncation is, in principle, a tighter constraint than the production grid's own boundary, and Section 10 records this explicitly as an unresolved edge case rather than eliding it.

6.4.2. Truncation Sensitivity of the Exact Filter

Section 6.1's own analytic calculation identifies a scenario—permanent lowest-corridor premium, permanent high-loss regime—under which the reference-parameter reservoir's steady state (141.4 claims) exceeds the $U_{\max}$ = 120 truncation used for the exact filter's validation comparisons of Section 6.4. To quantify whether this matters, we simulated that exact scenario for 48 months under the exact combinatorial reservoir process (reservoir departures Binomial(U_t, δ); new delayed occurrences Poisson($(1 - d_0)\lambda_H(600)$); reports = departures + immediate occurrences), with reference parameters throughout, across 40 independent replications, and ran the exact (Z,U) forward filter at $U_{\max}$ = 120 and at $U_{\max}$ = 250 on the same report sequences.

By month 36, the simulated true reservoir's mean reaches 139.4 claims—already past the $U_{\max}$ = 120 boundary. At that point, the $U_{\max}$ = 120 filter has approximately 48% of its posterior mass absorbed at the boundary, producing a mean filtered-reservoir bias of approximately 22 claims (maximum 33 across replications) and a filtered P(H) bias of approximately 0.12–0.13, relative to the $U_{\max}$ = 250 filter. Divergence between the two truncations exceeding 0.01 claims begins, on average, around month 5, once the true reservoir first approaches approximately 106 claims. Table 8 reports the check's summary statistics; Figure 3 plots the filtered and true reservoir trajectories.

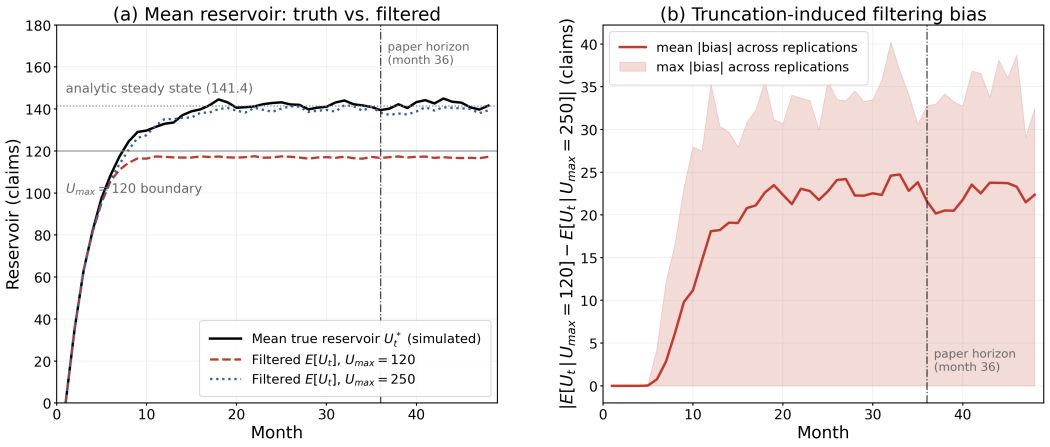

Figure 3. Truncation-sensitivity check: (**a**) mean true reservoir versus filtered $\mathbb{E}[U_t]$ at $U_{\max} = 120$ and $U_{\max} = 250$ under the adversarial scenario; (**b**) resulting truncation-induced filtering bias over time, reported as the mean and range across 40 replications.

This confirms, and for the first time quantifies, the risk Sections 6.1 and 6.4.1 flag qualitatively: $U_{\max} = 120$ is not a safe truncation under this adversarial scenario. It does not, however, revise Tables 5–7's headline validation statistics, which are computed under the production policy's own realised paths—Section 6.1 already establishes empirically that the filtered reservoir never exceeds 67.6 claims across all 3000 evaluation paths, comfortably inside the $U_{\max} = 120$ boundary. The two facts are complementary, not contradictory: the exact-filter truncation used for validation would materially bias the filter if the process were ever driven into the region the production policy itself avoids, and the policy's own reservoir-avoidance behaviour is precisely why the headline validation of Section 6.4 is not compromised by that truncation.

We are explicit that this check was implemented independently for the present revision, directly from the equations and parameters already stated in this manuscript (Table 8; Equations (13) and (14); Section 6.4's description of the joint (Z,U) posterior), rather than being a re-run of the authors' original production code; we would welcome the opportunity to cross-check it against that code before typesetting.

Table 8. Truncation-sensitivity check: exact filter at $U_{\max} = 120$ vs. $U_{\max} = 250$, adversarial scenario (permanent $p = 600$, permanent true $Z = H$), reference parameters, 40 replications, 48 months.

Quantity	Value
Analytic steady-state reservoir $u^* = (1 - d_0)\lambda_H(600)/\delta$	141.4 claims
Simulated mean true reservoir at month 36	139.4 claims
Simulated mean true reservoir at month 48	141.6 claims
Mean boundary-absorbed posterior mass, $U_{\max} = 120$, month 36	0.483
Mean boundary-absorbed posterior mass, $U_{\max} = 120$, month 48	0.450
Mean boundary-absorbed posterior mass, $U_{\max} = 250$, month 48	$\approx$0 (2.7×10^{-16})
Mean |filtered $\mathbb{E}[U]$ bias|, $U_{\max} = 120$ vs. 250, month 36	21.6 claims (max 32.8)
Mean |filtered $\mathbb{E}[U]$ bias|, $U_{\max} = 120$ vs. 250, month 48	22.4 claims (max 32.5)
Mean |filtered $P(H)$ bias|, $U_{\max} = 120$ vs. 250, month 36	0.135
Mean |filtered $P(H)$ bias|, $U_{\max} = 120$ vs. 250, month 48	0.120
Replications with >0.01-claim divergence (of 40)	40/40, median onset month 5

7. Synthetic Experimental Design

This section specifies the synthetic experiment used to generate every result in Section 8: the calibration and deliberate misspecification (Section 7.1), the matched-benchmark policy set (Section 7.2), the reserving benchmarks (Section 7.3), the closed-form solvers used for benchmarks that do not require the full recursion (Section 7.4), and the misspecification stress suite (Section 7.5).

7.1. Calibration and Deliberate Misspecification

The experiment represents a motor-style rating cell with 5000 baseline policies and monthly reviews over 36 months. Every policy is *solved* under a reference model (Table 9) and *evaluated* under a deliberately different true model. The true high-loss regime is more persistent ($P^{\text{true}} = \left[\begin{smallmatrix} 0.91 & 0.09 \\ 0.12 & 0.88 \end{smallmatrix}\right]$ versus the reference $P^{\text{ref}} = \left[\begin{smallmatrix} 0.90 & 0.10 \\ 0.15 & 0.85 \end{smallmatrix}\right]$), and the true geometric reporting distribution is first-order stochastically slower than the reference distribution: its cumulative reporting probability is lower at every finite development month (Figure 4). Mean delay is 4.89 months under the true model and 3.00 months under the reference model, using the mean-delay convention of Appendix C.

Table 8's principal parameters are chosen to be representative rather than fitted to a specific insurer. The demand elasticity $\eta = 2.0$ lies within the range typically reported for commercial and personal non-life price-response studies; the severity coefficient of variation (1.5) and the low/high mean severities (EUR 5000/7000) are representative of

a moderately heavy-tailed commercial liability book; the reporting-delay parameters are chosen to produce mean delays of 3.00 (reference) and 4.89 (true) months, of the order of magnitude documented for liability-type reporting patterns in Wüthrich and Merz (2008) and England and Verrall (2002); and $\gamma = 1.2$ is calibrated via the grid-search argument of Section Practical Interpretation of γ. None of these is fitted to a specific portfolio; Section 9.3 sets out the estimation sequence by which a production deployment would replace each with a fitted value, and Section Structural and Scope Limitations's fourth limitation restates this as an explicit scope boundary of the present, synthetic experiment.

Every simulated path starts from the same initial state, reported in full in Table 9: premium $p_0 =$ EUR 750, filtered belief $(\pi_0, u_0) = (0.4, 0)$, and a true initial regime Z_0 drawn as Bernoulli(0.4) for $Z_0 =$ H. The value 0.4 is the stationary probability of the high-loss regime under the *reference* transition matrix P^{ref} (not under P^{true}, whose own stationary probability is ≈0.429); the same reference-based value is used both as the filter's prior and as the parameter generating the true world's starting regime, so every policy begins from an identical initial condition and no benchmark is advantaged by starting closer to its own model's steady state. The filtered reservoir starts empty ($u_0 = 0$).

Table 9. Main numerical parameters.

Parameter	Value	Interpretation
Horizon	36 months	Monthly review
Initial premium p_0	EUR 750	Below optimized static; Table 10
Initial filtered belief (π_0, u_0)	$\pi_0 = 0.4$, $u_0 = 0$	Stationary $P(\text{H})$ under P^{ref}; empty reservoir
Initial true regime distribution	Bernoulli(0.4) for $Z_0 = \text{H}$	Stationary under P^{ref} (not P^{true}); see Section 7
Baseline exposure	5000 policies	Synthetic homogeneous cell
Market reference premium	EUR 700	Demand benchmark
Demand elasticity η	2.0	Reference and true models
Annual frequency, low/high	0.040/0.085	Claims per exposure-year
Mean severity, low/high	EUR 5000/EUR 7000	Gamma severity
Severity coefficient of variation	1.5	Both regimes
Variable expense	EUR 80	Per exposure-year
Reference transition matrix	$\left[\begin{smallmatrix} 0.90 & 0.10 \\ 0.15 & 0.85 \end{smallmatrix}\right]$	Nominal filter
True transition matrix	$\left[\begin{smallmatrix} 0.91 & 0.09 \\ 0.12 & 0.88 \end{smallmatrix}\right]$	More persistent high state
Reference reporting	$d_0 = 0.25$, $\delta = 0.25$	Mean delay 3.00 months
True reporting	$d_0 = 0.12$, $\delta = 0.18$	Mean delay 4.89 months
Premium corridor/move	EUR 600–950/EUR 25	Governance
Movement penalty κ	EUR 1500 per move (EUR 0.0015 million; Equation (16))	Bellman objective only
Risk sensitivity	$\gamma = 1.2$ per EUR million	Downside calibration
Monte Carlo: headline	3000 matched paths	Common random numbers
Monte Carlo: stress	1500 matched paths	Independent seed

Calibration justification. The demand elasticity $\eta = 2.0$ is within the range typically estimated for motor insurance (1.5–3.0) and produces a price-sensitive exposure response that is neither too flat (which would make pricing irrelevant) nor too elastic (which would make

small price changes produce extreme volume swings). The reporting-delay parameters were chosen to produce a meaningful but not extreme difference between reference and true models: a mean delay of 3.00 months under the reference model is typical for short-tailed motor lines, while 4.89 months under the true model represents a moderate deterioration that a pricing system should be able to detect and respond to. The severity parameters ($k_L = 4.44$, $\theta_L = 11{,}250$ for the low regime; $k_H = 4.44$, $\theta_H = 15{,}750$ for the high regime) produce Gamma distributions with mean EUR 5000 and EUR 7000, respectively, and coefficient of variation 1.5, consistent with motor third-party liability experience. The risk-sensitivity parameter $\gamma = 1.2$ was selected after exploratory analysis across $\gamma \in \{0, 0.4, 0.8, 1.2\}$ to demonstrate the risk-appetite mechanism at a level where its effects are visible but not extreme (Section 8.3).

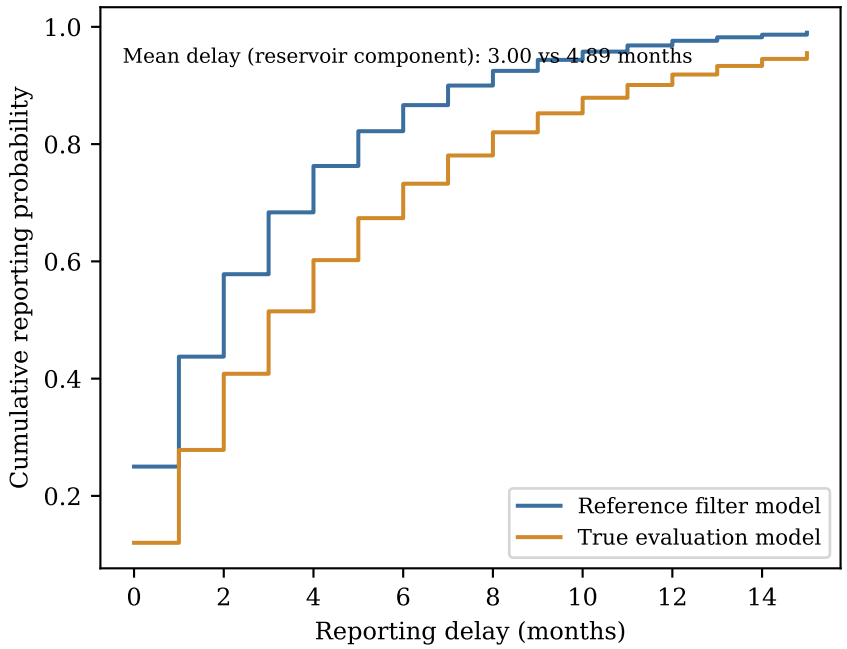

Figure 4. Reference and true reporting-delay cumulative distributions.

Table 10. Economic profit over 3000 out-of-model paths, $\gamma = 0$ (nominal) benchmark set (EUR million unless noted). At $\gamma = 0$, the certainty equivalent coincides with the mean by definition, so only the standardised $\gamma_{\text{eval}} = 0.8$ column is reported.

Strategy	Mean	SD	5th Pct.	TVaR 5%	CE ($\gamma_{\text{eval}} = 0.8$)
Optimized static (EUR 800)	3.960	1.045	2.162	1.767	3.509
Delay-naive	3.596	1.301	1.614	1.284	2.999
Cohort Bornhuetter–Ferguson	3.796	1.171	2.075	1.804	3.317
Delay-aware myopic	4.007	1.097	2.247	1.888	3.544
Zero-delay dynamic	3.981	1.118	2.170	1.829	3.502
Dynamic delay-aware (nominal)	4.024	1.095	2.243	1.883	3.557
Nominal full-state policy	4.206	1.075	2.439	2.085	3.753
True-model full-information oracle	4.209	1.065	2.485	2.132	3.768

7.2. Evaluated Strategies, Matched at Each Policy's Own Risk-Sensitivity Parameter

Benchmark matching is addressed structurally, not by adding a footnote: every strategy below that has a risk-sensitivity parameter is solved and reported *twice*, once at $\gamma = 0$ and once at $\gamma = 1.2$, against benchmarks solved at the *same* γ. Table 10 reports the $\gamma = 0$ (nominal) set, and Table 11 reports the $\gamma = 1.2$ (production risk-sensitive) set; Section 8.4 then uses *within-table* comparisons only for the maturity/continuation/state-

information/transition-knowledge decomposition, so that no reported decomposition component mixes policies solved at different γ.

Table 11. Economic profit over 3000 out-of-model paths, $\gamma = 1.2$ (production risk-sensitive) benchmark set (EUR million unless noted).

Strategy	Mean	SD	5th Pct.	TVaR 5%	CE ($\gamma_{\text{eval}} = 0.8$)	CE ($\gamma_{\text{eval}} = 1.2$)
Optimized static (EUR 875)	3.878	0.852	2.415	2.085	3.578	3.427
Delay-aware myopic	4.009	1.064	2.289	1.939	3.570	3.371
Zero-delay dynamic	3.986	1.015	2.348	2.032	3.589	3.409
Dynamic delay-aware risk-sensitive	4.026	1.029	2.359	2.007	3.612	3.421
Nominal full-state policy	4.187	1.008	2.523	2.190	3.786	3.600
True-model full-information oracle	4.177	1.002	2.544	2.212	3.784	3.603

1. **Optimized static price.** Selects the best constant price on the nominal premium grid, maximising the certainty equivalent *at the policy's own* γ under the reference regime dynamics (Section 7.4). The optimum is EUR 800 at $\gamma = 0$ and EUR 875 at $\gamma = 1.2$—a risk-averse decision-maker choosing a single price for the whole horizon prices materially higher, and holds materially less exposure, than a risk-neutral one, entirely independently of any dynamic control. Unlike every dynamic policy, the static price is chosen once from the full premium grid with no starting point and no governance path to traverse—it is not constrained to begin at EUR 750 or to reach its optimum through EUR 25 monthly steps—so the comparison in Section 8 is, if anything, conservative with respect to the dynamic policies rather than favourable to them.
2. **Delay-naive rule.** Exponentially smooths calendar-month *reports* and treats them as contemporaneous occurrences, targeting a technical rate at a fixed loss-and-expense ratio (Appendix A); evaluated only at the $\gamma = 0$ set, since it is a heuristic frequency-tracking rule rather than a utility-optimising one.
3. **Cohort Bornhuetter–Ferguson rule.** Assigns reports to occurrence cohorts and applies the reference cumulative reporting proportion at each cohort age (Section 7.3); likewise evaluated only at $\gamma = 0$.
4. **Delay-aware myopic rule.** Uses the nominal filter but maximises only the current expected (or, at $\gamma = 1.2$, current certainty-equivalent) monthly contribution, with continuation value fixed at zero—solved and reported at both $\gamma = 0$ and $\gamma = 1.2$.
5. **Zero-delay dynamic rule.** Solves a Bellman program that incorrectly assumes all occurrences are reported immediately ($d_0 = 1$); solved and reported at both $\gamma = 0$ and $\gamma = 1.2$, using the *same* matched-γ machinery as the delay-aware policy so that the maturity-value comparison of Section 8.4 isolates the effect of the reporting-delay misspecification alone.
6. **Dynamic delay-aware rule.** Solves Equation (17) at $\gamma = 0$ ("dynamic nominal") and at $\gamma = 1.2$ ("dynamic risk-sensitive", the paper's production policy), and includes compound Poisson–Gamma loss inside the entropic operator at $\gamma = 1.2$.
7. **Nominal full-state policy.** Observes the current latent regime Z_t exactly (no reservoir or filter state needed) but is solved under the *reference* transition matrix P^{ref}, at the same γ as the policy it benchmarks—isolating the value of state information alone, holding transition-model knowledge fixed at the reference belief the dynamic policy itself uses.
8. **True-model full-information oracle.** Also observes Z_t exactly, but is solved under the *true* transition matrix P^{true}, at the same γ. Because it combines exact state observation *and* correct transition-law knowledge, its gap over the dynamic delay-aware policy is

not a pure measure of the value of information; Section 8.4 uses the nominal full-state policy above to separate the two effects. Section 7.4 gives the closed-form recursion for both benchmarks.

7.3. Bornhuetter–Ferguson and Delay-Naive Benchmark Construction

For occurrence cohort k observed at valuation month t, the Bornhuetter–Ferguson count estimate is

$$\hat{N}^{\mathrm{BF}}(k,t) = N^{\mathrm{reported}}(k,t) + [1 - F_D(t-k)]\, N^{\mathrm{prior}}(k). \tag{18}$$

The development proportion F_D is applied to a cohort of known age using the *reference* geometric reporting pattern $F_D(a) = d_0^{\mathrm{ref}} + (1 - d_0^{\mathrm{ref}})[1 - (1 - \delta^{\mathrm{ref}})^a]$, not to a calendar-month report total that mixes occurrence periods, and $N^{\mathrm{prior}}(k)$ uses a fixed blended prior frequency rather than the true, unobserved regime-specific frequency. Reported counts $N^{\mathrm{reported}}(k,t)$ for cohorts still within the active reservoir are obtained by allocating each month's realised aggregate reservoir reports across active cohorts in proportion to each cohort's current expected remaining share—a documented simplification, adopted because the phase-type reservoir of Section 5 is, by construction, memoryless in claim age (Appendix B gives the full construction). Both the Bornhuetter–Ferguson rule and the delay-naive rule convert their resulting frequency estimate into a target premium via a common technical-rate formula, $p^{\mathrm{target}} = (c_e + \hat{f} \cdot \bar{\mu})/\ell$, with a fixed loss-and-expense-ratio target $\ell = 0.60$, and move toward that target subject to the same EUR 25 per-month governance limit as the dynamic policies. The target is chosen, and reported rather than tuned after the fact, so that the resulting heuristic prices (EUR 700–730) fall in the same operating range as the dynamic policies (Table 10) rather than at a technical breakeven far outside the corridor those policies actually explore.

This construction is illustrative rather than a full claim-level cohort implementation: it relies on proportional allocation of aggregate reservoir reports across active cohorts rather than exact per-claim attribution, and on a fixed 60% loss-ratio target chosen to place its operating range near the dynamic policies' own rather than independently estimated (Section Structural and Scope Limitations). Readers using this benchmark as a serious point of comparison for a real reserving process should treat it as illustrative rather than as a validated implementation of the Bornhuetter–Ferguson method.

7.4. Closed-Form Solvers for the Static and Full-State Benchmarks

The optimised static price and both full-state benchmarks admit a closed-form Bellman recursion that avoids summing over report counts entirely, because none of them carries a filtered reservoir state: the static price never observes anything, and the two full-state policies observe the regime Z_t exactly. For a fixed price p held for the remaining horizon, the exact compound Poisson–Gamma moment-generating function gives the two-state (regime-only) recursion

$$V_t(z) = b(p) - \tfrac{1}{\gamma}\lambda_z(p)\left[M_z(\gamma/10^6) - 1\right] - \tfrac{1}{\gamma}\log\sum_{z'} P^{\mathrm{reg}}[z,z']\, e^{-\gamma V_{t+1}(z')}, \tag{19}$$

evaluated for every candidate price on the premium grid, with $P^{\mathrm{reg}} = P^{\mathrm{ref}}$ (the static price is chosen under the reference model, since it is not informed by any observation process) and the optimum selected by the resulting certainty equivalent under a stationary initial regime prior. Both full-state policies solve the analogous two-state *controlled* recursion,

$$V_t(p,z) = \max_a \left\{ b(p',a) - \tfrac{1}{\gamma}\lambda_z(p')\left[M_z(\gamma/10^6) - 1\right] - \tfrac{1}{\gamma}\log\sum_{z'} P^{\mathrm{reg}}[z,z']\, e^{-\gamma V_{t+1}(p',z')} \right\}, \tag{20}$$

with $P^{\text{reg}} = P^{\text{ref}}$ for the nominal full-state policy and $P^{\text{reg}} = P^{\text{true}}$ for the true-model full-information oracle: the two benchmarks are the *same* recursion, differing only in which transition matrix is believed, which is precisely what makes their difference (Section 8.4) an isolated measure of the value of correct transition-law knowledge, holding state information fixed. Both recursions reduce, at $\gamma = 0$, to the corresponding expected-value Bellman equation by the same continuity argument as Equation (17), and all three closed-form benchmarks are solved exactly (15 premiums $\times$ 2 regimes $\times$ 36 months) rather than approximated, since no interpolation is required in a two-state, price-on-grid recursion.

7.5. Misspecification Stress Suite

To test robustness more broadly than the single true/reference contrast of Table 9, Section 8.6 evaluates the *fixed* $\gamma = 1.2$ production policy and the matched $\gamma = 1.2$ optimized static price—neither policy is re-solved—against eight additional true-world scenarios, using an independent 1500-path common-random-number sample (distinct from, and therefore not directly poolable with, the headline 3000-path sample of Table 11, which is why the stress suite's own base-case row differs slightly, at the level of Monte Carlo noise, from Table 11's corresponding entries): a slower true reporting process ($d_0 = 0.08, \delta = 0.14$), a faster true reporting process ($d_0 = 0.20, \delta = 0.30$), a lower immediate-reporting proportion ($d_0 = 0.06, \delta = 0.18$), higher true regime persistence, lower true regime persistence, a 30% higher true mean severity in both regimes, and $\pm$20% multiplicative demand shocks applied directly to realised exposure (a reduced-form proxy for a misestimated demand elasticity, since re-estimating η itself would require re-specifying the reference model's own demand curve rather than only its evaluation environment). Section 8.6 reports the resulting paired mean and 5th-percentile differences.

8. Results

Every result below is generated by the synthetic, single-cell, homogeneous experiment of Section 7; none of the comparisons that follow has been run on real portfolio data, and Section Structural and Scope Limitations sets out what a portfolio-level, empirically estimated replication would additionally require.

8.1. Economic Performance, Matched at Each Policy's Own γ

Two certainty-equivalent quantities appear throughout this section and should not be conflated. $CE_{0.8}$ is evaluated at a fixed $\gamma_{\text{eval}} = 0.8$ for every strategy, regardless of what γ, if any, that strategy was actually optimised under; it is a standardised cross-policy yardstick, not an objective any policy is trying to maximise. $CE_{1.2}$, by contrast, is evaluated at $\gamma_{\text{eval}} = 1.2$ and is reported only for the $\gamma = 1.2$ policies, for which it is their own optimisation objective. A policy can therefore look favourable on the standardised $CE_{0.8}$ yardstick while not looking favourable on its own $CE_{1.2}$ objective, and Section 8.2 shows that this is exactly what happens here.

Table 10 reports the $\gamma = 0$ (risk-neutral) benchmark set and Table 11 reports the $\gamma = 1.2$ (production risk-sensitive) benchmark set, each evaluated on the same 3000 out-of-model common-random-number paths (Section 5.4). Both tables report the certainty equivalent at the policies' own optimisation level $\gamma_{\text{eval}} = \gamma$ alongside the standardised cross-policy measure $\gamma_{\text{eval}} = 0.8$ used throughout the rest of the paper, and both include the two full-state benchmarks of Section 7.2 solved at the table's own γ (Section 7.4).

The $\gamma_{\text{eval}} = 1.2$ column in Table 11 is the certainty equivalent evaluated at the *same* risk-sensitivity level the $\gamma = 1.2$ policies were actually optimised under, and is the primary policy-value comparison for that table; the $\gamma_{\text{eval}} = 0.8$ column is retained throughout both tables purely as a standardised cross-policy yardstick that does not itself change with the

policy's own γ, and should not be read as the objective any $\gamma = 1.2$ policy was solved to maximise. At $\gamma = 0$, the two columns coincide by definition, since $\gamma_{\text{eval}} = \gamma = 0$ reduces the certainty equivalent to the mean.

Two features of Table 11 are worth flagging before the paired analysis below makes them precise. First, the CE-optimized static price rises from EUR 800 at $\gamma = 0$ to EUR 875 at $\gamma = 1.2$, and its own 5th-percentile outcome rises from EUR 2.162 million to EUR 2.415 million *without any dynamic control at all*—simply pricing high and holding volume down throughout the horizon is itself a substantial source of downside protection, achieved at the cost of EUR 0.082 million of mean profit relative to the $\gamma = 0$-optimal static price. Second, and as a direct consequence, the static price's 5th-percentile outcome at $\gamma = 1.2$ (EUR 2.415 million) is *higher* than the matched dynamic risk-sensitive policy's own 5th-percentile outcome (EUR 2.359 million).

8.2. Paired Differences with Bootstrap Intervals

Table 12 reports paired mean, 5th-percentile, TVaR 5%, and certainty-equivalent differences with 95% percentile bootstrap intervals (2000 resamples of the shared path index, respecting the common-random-number pairing of Section 5.4) for the $\gamma = 1.2$ set, including the certainty equivalent at the policies' own optimisation level $\gamma_{\text{eval}} = 1.2$; Table 13 reports the corresponding $\gamma = 0$ comparisons (where $\gamma_{\text{eval}} = 0$ coincides with the mean, so only $\gamma_{\text{eval}} = 0.8$ is shown). Both tables report the state-information and transition-knowledge legs of the decomposition (Section 8.4) as separate rows, rather than only their sum, so that neither component's contribution is left implicit.

Table 12. Paired differences for the matched $\gamma = 1.2$ benchmark set. Values are in EUR million; 95% bootstrap intervals are shown in brackets.

Comparison	Mean	5th Pct.	TVaR 5%	CE ($\gamma_{\text{eval}} = 0.8$)	CE ($\gamma_{\text{eval}} = 1.2$)
Dynamic—myopic	+0.017 [0.015, 0.019]	+0.070 [0.043, 0.087]	+0.068 [0.061, 0.075]	+0.042 [0.040, 0.044]	+0.050 [0.047, 0.052]
Dynamic—zero-delay	+0.039 [0.037, 0.042]	+0.011 [−0.029, 0.030]	−0.026 [−0.036, −0.014]	+0.023 [0.019, 0.026]	+0.012 [0.007, 0.017]
Dynamic—opt. static	+0.148 [0.141, 0.155]	−0.056 [−0.097, −0.041]	−0.078 [−0.088, −0.069]	+0.034 [0.028, 0.039]	−0.006 [−0.011, +0.000]
Nominal full-state—dynamic	+0.161 [0.159, 0.163]	+0.164 [0.149, 0.196]	+0.183 [0.175, 0.192]	+0.175 [0.172, 0.177]	+0.179 [0.176, 0.182]
Oracle—nominal full-state	−0.010 [−0.011, −0.009]	+0.021 [−0.001, 0.031]	+0.023 [0.017, 0.029]	−0.003 [−0.004, −0.001]	+0.003 [0.001, 0.004]
Oracle—dynamic	+0.151 [0.149, 0.153]	+0.185 [0.162, 0.212]	+0.206 [0.197, 0.214]	+0.172 [0.170, 0.175]	+0.182 [0.178, 0.185]

Table 13. Paired differences for the matched $\gamma = 0$ benchmark set. Values are in EUR million; 95% bootstrap intervals are shown in brackets.

Comparison	Mean	5th Pct.	TVaR 5%	CE ($\gamma_{\text{eval}} = 0.8$)
Dynamic—myopic	+0.017 [0.015, 0.018]	−0.004 [−0.035, 0.023]	−0.005 [−0.013, 0.003]	+0.013 [0.011, 0.016]
Dynamic—zero-delay	+0.043 [0.040, 0.046]	+0.073 [0.033, 0.097]	+0.054 [0.042, 0.067]	+0.055 [0.051, 0.059]
Dynamic—opt. static	+0.064 [0.060, 0.068]	+0.081 [0.033, 0.106]	+0.116 [0.100, 0.130]	+0.048 [0.042, 0.053]
Nominal full-state—dynamic	+0.182 [0.180, 0.184]	+0.196 [0.178, 0.238]	+0.202 [0.192, 0.211]	+0.196 [0.194, 0.199]
Oracle—nominal full-state	+0.003 [0.002, 0.004]	+0.046 [0.018, 0.054]	+0.047 [0.041, 0.054]	+0.015 [0.013, 0.017]
Oracle—dynamic	+0.185 [0.183, 0.188]	+0.243 [0.214, 0.271]	+0.249 [0.239, 0.259]	+0.211 [0.208, 0.215]

Figure 5 visualises the $\gamma = 1.2$ dynamic-versus-static comparison from Table 12 directly. Two results should be read together. Mean profit and the $\gamma_{\text{eval}} = 0.8$ certainty equivalent both favour the dynamic policy, with intervals that exclude zero by a wide margin: dynamic delay-aware control adds EUR 0.148 million of mean profit and EUR 0.034 million of certainty-equivalent profit relative to an equally risk-averse static price, and this is the paper's central positive result. At the policy's own optimisation level, however, $\text{CE}(\gamma_{\text{eval}} = 1.2)$ is *lower* for the dynamic policy than for the matched static price (−0.006, interval $[-0.011, 0.000]$ touching zero)—a materially different picture from the $\gamma_{\text{eval}} = 0.8$ column, and a reminder that which certainty-equivalent level is used to declare a winner is not a neutral choice once the two policies' downside profiles differ this much. The

5th percentile and TVaR 5% *both* favour the static price once the comparison is properly matched at $\gamma = 1.2$: the dynamic policy's 5th percentile is EUR 0.056 million *lower* than the matched static price's, and its TVaR 5% is EUR 0.078 million lower, with both intervals excluding zero. At $\gamma = 0$, the ordering is different again—there, the dynamic policy dominates the static price on every reported measure, because the $\gamma = 0$-optimal static price (EUR 800) has not been pushed up by risk aversion and so offers comparatively little downside protection in the first place.

This pattern is not an artefact of a single statistic; it is a structural consequence of how the two policy *types* generate downside protection. A static, risk-averse price generates downside protection by permanently restricting exposure—its 5th-percentile outcome is good because it is rarely writing much business when a bad regime realisation occurs, at the cost of also writing less business when the regime is favourable. A dynamic, delay-aware, risk-averse policy generates downside protection by *reacting* to filtered evidence of a deteriorating regime, which is necessarily reported evidence, subject to the reporting lag and posterior-estimation error documented in Section 6.4: by the time the filter has accumulated enough reported claims to raise its posterior probability of the high-loss regime with confidence, some of the exposure written at the lower, pre-reaction price has already been exposed to the deteriorating conditions. The dynamic policy's downside advantage over the *matched* static price is therefore not guaranteed to be positive, and Table 11 and Figure 6 show that, at this calibration, it is not.

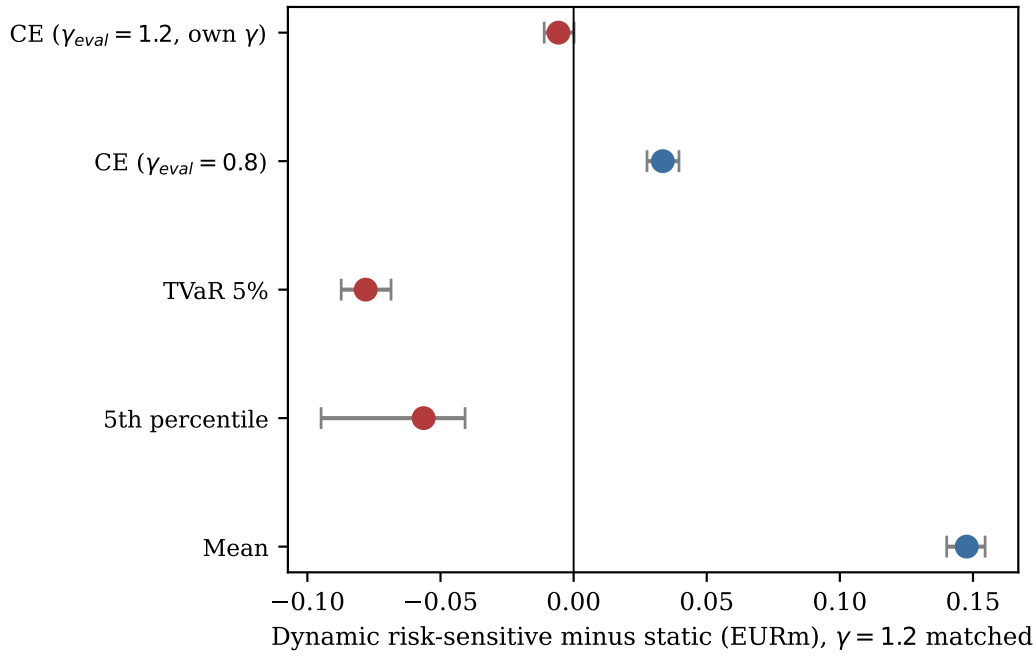

Figure 5. Paired differences (dynamic risk-sensitive minus optimised static), $\gamma = 1.2$ matched, with 95% bootstrap intervals, including both the standardised CE($\gamma_{\text{eval}} = 0.8$) and the policy's own-objective CE($\gamma_{\text{eval}} = 1.2$). Colored dots denote the point estimates for the performance measures identified by the figure labels, and the horizontal lines show their 95% bootstrap intervals; the vertical zero line indicates no difference.

8.3. Risk-Sensitivity Calibration Across γ

The dynamic policy's mean profit is essentially flat across this range (4.024–4.029 million, peaking near $\gamma = 0.4$–0.8), while the static price's mean profit falls monotonically as γ rises (3.960 million at $\gamma = 0$ to 3.878 million at $\gamma = 1.2$) because a higher γ pushes the single chosen price further above its risk-neutral optimum. Both policies' 5th-percentile

outcomes rise monotonically with γ, but the static price's rises faster: the two curves in the middle panel of Figure 6 cross between $\gamma = 0$ and $\gamma = 0.4$, and the static price's 5th percentile remains at or above the dynamic policy's for every $\gamma \geq 0.4$ tested. The own-γ certainty equivalent (right panel) tells a third story: the dynamic policy's advantage over the static price narrows steadily as γ rises (4.024 vs. 3.960 million at $\gamma = 0$; 3.800 vs. 3.755 million at $\gamma = 0.4$; 3.597 vs. 3.577 million at $\gamma = 0.8$) and reverses sign at $\gamma = 1.2$ (3.421 vs. 3.427 million, dynamic below static). Increasing γ therefore produces monotone downside improvements and progressively higher premiums for both policy types, and a matched-γ static price is, at this calibration, at least as effective as dynamic control at converting risk aversion into downside protection; on mean profit the dynamic policy remains ahead throughout the range tested, but on the own-γ certainty equivalent—the actual objective each policy pair was optimised to maximise—its advantage shrinks with γ and is gone by $\gamma = 1.2$.

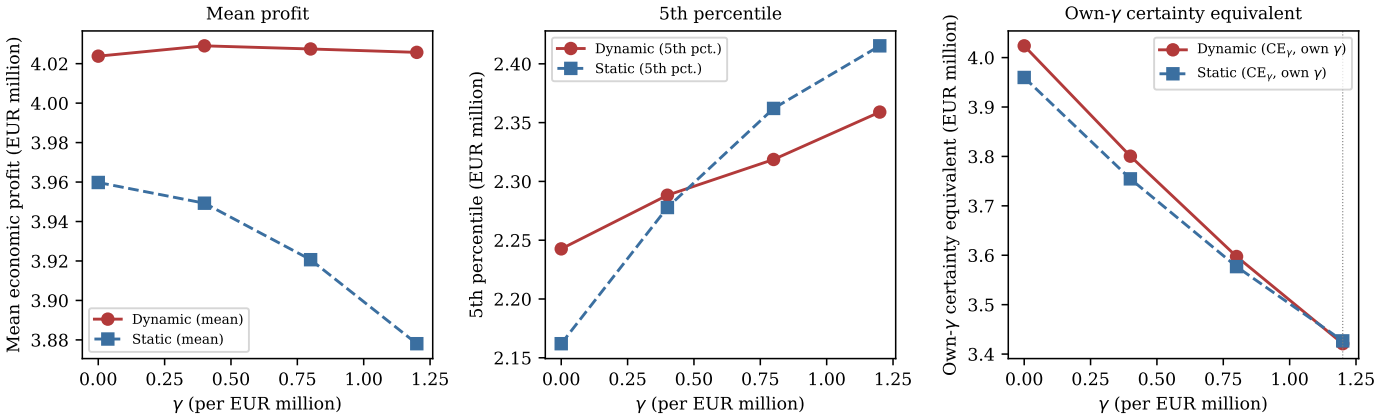

Figure 6. Mean profit, 5th percentile, and own-γ certainty equivalent as a function of the policy risk-sensitivity parameter γ, dynamic delay-aware policy versus the matched optimised static price, both solved and evaluated at each γ. The own-γ certainty equivalent (right panel) is each policy's actual optimisation objective, not the standardised $\gamma_{\text{eval}} = 0.8$ yardstick used elsewhere in the paper.

8.4. Maturity, Continuation, and State-Information Decomposition

Table 14 is a decomposition of *mean economic profit*, not of the policies' own entropic objective, and its component labels reflect that explicitly. Figure 7 and Table 14 report four within-γ paired comparisons, each isolating one mean-profit effect while holding risk sensitivity fixed: the *mean-profit effect of reporting-delay modelling* (dynamic delay-aware minus zero-delay dynamic, both fully dynamic, differing only in whether reporting delay is modelled at all), the *mean-profit effect of continuation* (dynamic delay-aware minus delay-aware myopic, both correctly delay-aware, differing only in whether the recursion looks beyond the current month), the *mean-profit effect of exact state observation* (nominal full-state policy minus dynamic delay-aware, both solved under the *reference* transition matrix, differing only in whether the current regime is observed exactly or filtered), and the *mean-profit effect of correct transition-law knowledge* (true-model full-information oracle minus the nominal full-state policy, both observing the regime exactly, differing only in whether the transition matrix believed is P^{true} or P^{ref}). The last two effects sum, by construction, to the *true-model full-information oracle gap*—the oracle's total mean-profit advantage over the dynamic delay-aware policy—reported as a fifth row for reference. Separating the two full-state benchmarks in this way distinguishes state-observation effects from transition-model effects, which a single combined oracle comparison cannot.

Because mean profit is not the objective either $\gamma = 1.2$ policy actually maximises, a negative mean-profit effect at $\gamma = 1.2$ is not a decision-theoretic anomaly, but it is easy

to over-read as one. Table 15 therefore reports the same four-way decomposition using each policy's own entropic certainty equivalent $CE_{1.2}$ in place of mean profit, at $\gamma = 1.2$ only (at $\gamma = 0$ the certainty equivalent coincides with the mean, so the two tables would be identical). Every component is non-negative under this, the policies' actual optimisation objective.

Table 14. Mean-profit decomposition, matched γ, 95% bootstrap intervals in brackets (EUR million mean profit contribution).

Mean-Profit Effect of	$\gamma = 0$	$\gamma = 1.2$
Reporting-delay modelling	+0.043 [0.040, 0.045]	+0.039 [0.037, 0.042]
Continuation	+0.017 [0.016, 0.018]	+0.017 [0.015, 0.019]
Exact state observation	+0.182 [0.180, 0.184]	+0.161 [0.159, 0.163]
Correct transition-law knowledge	+0.003 [0.002, 0.004]	−0.010 [−0.011, −0.009]
Sum: true-model full-information oracle gap	+0.185 [0.183, 0.187]	+0.151 [0.149, 0.153]

Table 15. Decomposition of $CE_{1.2}$ (each policy's own optimisation objective), $\gamma = 1.2$, 95% bootstrap intervals in brackets (EUR million certainty-equivalent contribution).

$CE_{1.2}$ Effect of	$\gamma = 1.2$
Reporting-delay modelling	+0.012 [0.007, 0.017]
Continuation	+0.050 [0.047, 0.052]
Exact state observation	+0.179 [0.176, 0.182]
Correct transition-law knowledge	+0.003 [0.001, 0.004]
Sum: true-model full-information oracle $CE_{1.2}$ gap	+0.182 [0.178, 0.185]

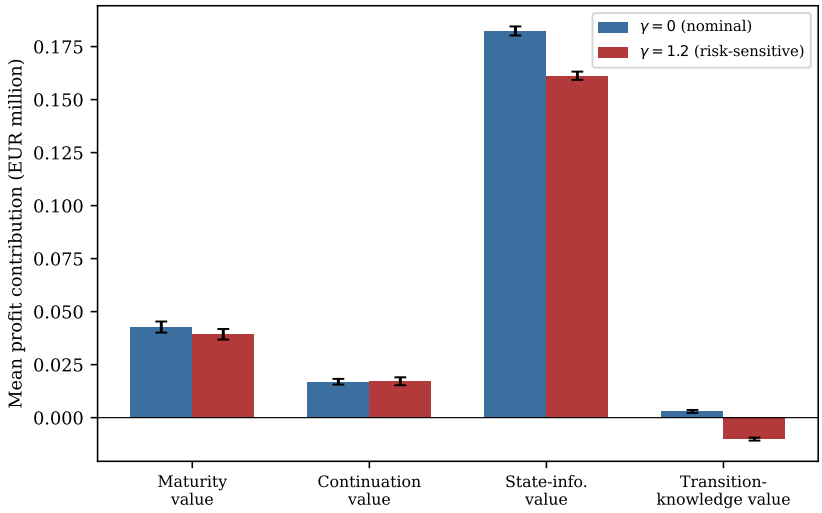

Figure 7. Mean-profit decomposition (reporting-delay modelling/continuation/exact state observation/correct transition-law knowledge), matched γ.

Two features of the decomposition are stable across γ: reporting-delay modelling contributes EUR 0.039–0.043 million, and continuation beyond a delay-aware myopic rule contributes a further EUR 0.017 million at both γ. The mean-profit effect of exact state observation dominates every other component at both γ (EUR 0.161–0.182 million, roughly four to ten times any single other component); separating the two full-state benchmarks in this way confirms that reconstructing the current regime, specifically, is the economically primary source of mean-profit gain in this experiment. The mean-profit effect of correct transition-law knowledge, by contrast, is small at both γ and *changes sign* between them: a modest positive EUR 0.003 million at $\gamma = 0$, but a small *negative* EUR −0.010 million at $\gamma = 1.2$ (95% interval excludes zero at both γ, so the sign change itself is not sampling

noise). The negative sign at $\gamma = 1.2$ is not an error: a risk-sensitive policy that correctly believes the high-loss regime is more persistent than the reference model assumes prices more conservatively during high-regime episodes, which lowers *mean* profit relative to a policy still holding the reference belief, exactly the mean-for-downside trade this paper documents elsewhere (Section 8.2). Table 15 confirms this directly: evaluated under each policy's own $CE_{1.2}$ rather than mean profit, the effect of correct transition-law knowledge is a positive EUR 0.003 million, and every component of the decomposition is non-negative. Correct transition-law knowledge is, in short, worth little on its own once state information is already available, and whether it reads as positive or negative depends on whether it is measured against mean profit or against the policy's actual risk-adjusted objective—a distinction the single combined oracle-gap figure (EUR 0.151–0.185 million, still accurate as a total) cannot convey by itself.

8.5. Operational Effects

Table 16 reports operational and normalised performance measures for the $\gamma = 1.2$ set (the $\gamma = 0$ set shows the same qualitative pattern and is omitted for space; full figures are in the reproducibility files, Appendix C). "Below adequacy" is the fraction of months in which the charged premium falls below $(c_e +$ true regime-specific pure premium$)/0.85$, the same technical-adequacy threshold used throughout; "modelled loss-and-expense ratio" (used in place of "combined ratio" to avoid implying a fully specified underwriting expense structure this synthetic experiment does not model) is (incurred loss + variable expense)/earned premium.

Table 16. Operational and normalised performance measures, $\gamma = 1.2$ set.

Strategy	Mean Premium	Exposure Index	Loss–Expense Ratio	Below Adequacy	Hold Share
Optimized static	875.0	0.607	51.3%	0.0%	100.0%
Delay-aware myopic	793.9	0.772	56.4%	16.9%	78.6%
Zero-delay dynamic	822.5	0.715	54.7%	15.7%	67.7%
Dynamic risk-sensitive	807.5	0.743	55.2%	12.0%	67.3%
Full-information oracle	821.3	0.725	52.8%	4.5%	50.5%

Risk sensitivity raises the mean premium and reduces exposure relative to the delay-aware myopic rule, and lowers both the loss-and-expense ratio and the fraction of months below the technical-adequacy threshold relative to the zero-delay dynamic policy. As with any margin gained by writing less business, these gains should not be interpreted as free value: they arise partly from writing less business at higher rates, and Section 9 returns to the implication that a production deployment should evaluate customer retention, lifetime value, and capacity usage alongside underwriting profit—none of which this synthetic single-cell experiment represents.

Figure 8 shows median premium paths with 10th–90th percentile bands, all four strategies now solved and evaluated at the same $\gamma = 1.2$ so that the comparison is not confounded by a risk-sensitivity mismatch. Two features are worth noting. First, the zero-delay dynamic policy climbs to, and holds, the highest sustained premium of the three dynamic policies (EUR 850, against EUR 825 for the correctly delay-aware policy): treating every report as an immediate, undelayed occurrence makes it react to reported evidence more aggressively than a policy that correctly discounts part of each report as reservoir catch-up, consistent with the larger price movements documented in Table 16. Second, the oracle's median path starts lowest of the four (EUR 750, since it can directly confirm a favourable early regime that the filtered policies are still uncertain about) and is visibly non-monotone later in the horizon, including a distinct dip in months 33–36, consistent

with the oracle observing, and reacting to, genuine regime transitions under P^{true} that the filtered policies can only infer with a lag.

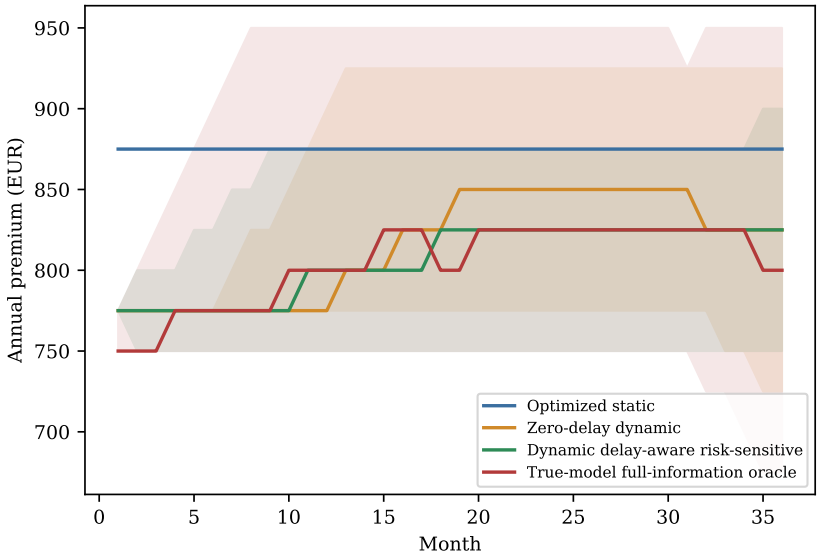

Figure 8. Median premium paths with 10th–90th percentile bands, four strategies matched at $\gamma = 1.2$: optimized static, zero-delay dynamic, dynamic delay-aware risk-sensitive, and the true-model full-information oracle. Colors distinguish the four strategies as identified in the legend; the corresponding shaded bands show each strategy's 10th–90th percentile range.

Figure 9 shows the production risk-sensitive action surface. At low posterior probability of the high-loss regime the policy reduces price, at high posterior probability it raises price, and a narrow hold region separates the two, with a small notch near a filtered reservoir of 60–65 claims where the hold region's upper boundary shifts—a direct, visible consequence of the reservoir state's own influence on the near-term expected report volume, and hence on the value of waiting one more month before acting.

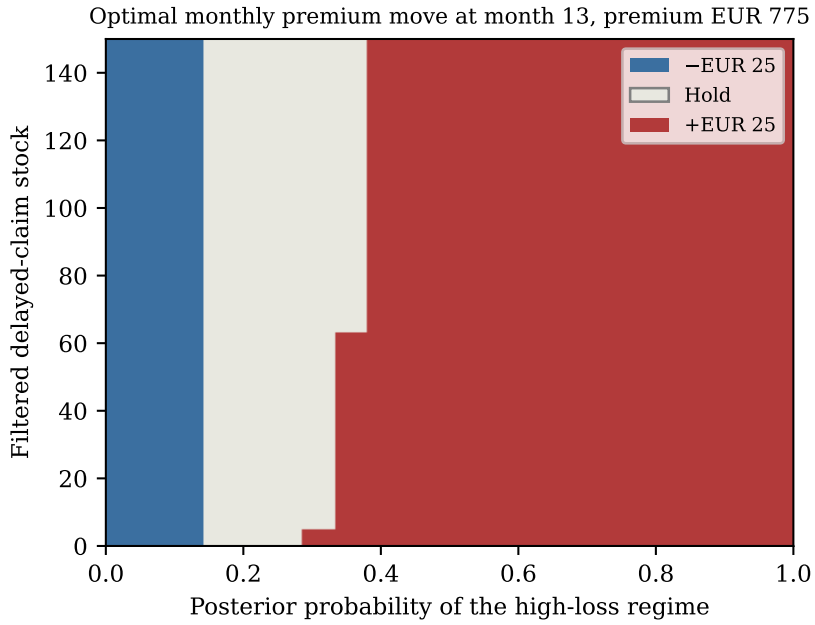

Figure 9. Production risk-sensitive action surface ($\gamma = 1.2$) at month 13, premium EUR 775.

8.6. Misspecification Stress Results

Table 17 and Figure 10 report the fixed-policy stress suite of Section 7.5: the production ($\gamma = 1.2$) dynamic policy and the matched ($\gamma = 1.2$) optimised static price, neither re-solved, evaluated under eight additional true-world scenarios using an independent 1500-path sample. Every point estimate, including TVaR 5%, is reported with a 95% paired bootstrap interval (2000 resamples of the shared path index), since a 1500-path sample is materially noisier at the tail than at the mean, and a point estimate alone would not let a reader judge which scenarios are robust.

Three findings emerge, all now confirmed by intervals rather than point estimates alone. First, the dynamic policy's *mean*-profit advantage over the matched static price is directionally robust, with a 95% interval excluding zero, across seven of the eight additional stress scenarios, ranging from EUR 0.100 million (lower true persistence) to EUR 0.194 million (higher true persistence)—broadly the same order of magnitude as the base case's EUR 0.146 million (which differs slightly from Table 12's EUR 0.148 million because the stress suite uses an independent, smaller path sample, Section 7.5)—but is *reversed*, to -0.047 million with an interval that also excludes zero ($[-0.057, -0.038]$), under a 30% higher true severity level. At elevated severity, the fixed dynamic policy's reservoir- and regime-driven price reactions are evidently not aggressive enough, relative to a static price that was itself optimized (at $\gamma = 1.2$, under the reference severity assumption) to a comparatively high level, to keep pace with the larger loss draws; since neither policy was re-solved under the stressed severity assumption, this scenario should be read as a statement about *robustness of a fixed policy to severity misspecification*, not as a statement about what an insurer who correctly anticipated the higher severity would have priced. Second, the dynamic policy's TVaR 5% is lower than the matched static price's, with an interval excluding zero, in *every one* of the nine scenarios tested, including the base case—the most uniformly robust finding in the entire suite. Third, its 5th-percentile outcome is lower in eight of the nine scenarios with intervals excluding zero in seven of those eight; the sole exception on either downside measure is the +20% demand shock's 5th-percentile difference, whose interval ($[-0.076, +0.023]$) spans zero and is therefore not statistically distinguishable from no difference. In short: lower point estimates on both downside measures in every scenario tested, with statistically significant differences in nine of nine on TVaR 5% and in eight of nine on the 5th percentile.

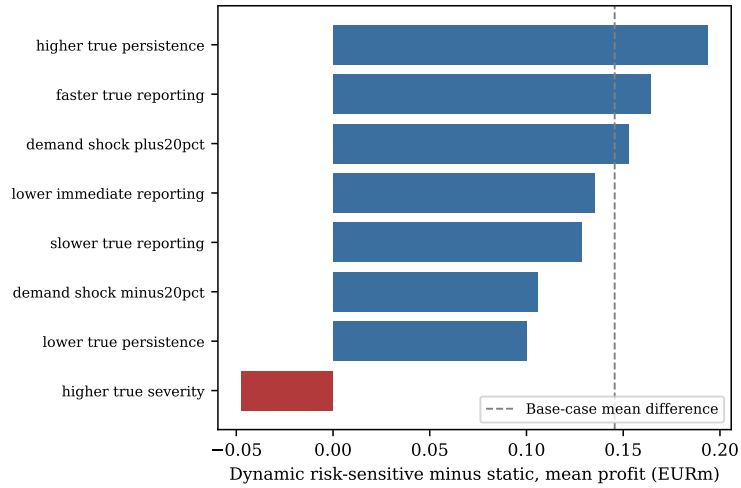

Figure 10. Misspecification stress suite: mean profit difference (dynamic risk-sensitive minus optimised static), fixed $\gamma = 1.2$ policies, sorted by effect size. Bar colors distinguish positive differences, which favour the dynamic policy, from negative differences, which favour the static policy, as indicated in the legend.

Table 17. Fixed-policy misspecification stress suite, $\gamma = 1.2$, paired mean, 5th-percentile, and TVaR 5% differences (dynamic minus static, EUR million) with 95% bootstrap intervals in brackets, independent 1500-path sample.

Scenario	Mean Diff.	5th-Pct. Diff.	TVaR 5% Diff.
Base case (true model of Table 9)	+0.146 [0.136, 0.155]	−0.085 [−0.138, −0.049]	−0.088 [−0.103, −0.073]
Slower true reporting	+0.129 [0.118, 0.139]	−0.139 [−0.191, −0.099]	−0.130 [−0.146, −0.114]
Faster true reporting	+0.164 [0.155, 0.173]	−0.046 [−0.085, −0.009]	−0.036 [−0.050, −0.022]
Higher true regime persistence	+0.194 [0.183, 0.205]	−0.065 [−0.102, −0.048]	−0.079 [−0.093, −0.065]
Lower true regime persistence	+0.100 [0.092, 0.108]	−0.066 [−0.105, −0.032]	−0.099 [−0.118, −0.081]
Higher true severity (+30%)	−0.047 [−0.057, −0.038]	−0.222 [−0.282, −0.165]	−0.193 [−0.216, −0.170]
Lower immediate-reporting proportion	+0.135 [0.125, 0.145]	−0.107 [−0.157, −0.061]	−0.106 [−0.121, −0.090]
Demand shock, +20% exposure	+0.153 [0.143, 0.163]	−0.047 [−0.076, +0.023]	−0.034 [−0.050, −0.016]
Demand shock, −20% exposure	+0.106 [0.096, 0.115]	−0.186 [−0.209, −0.133]	−0.177 [−0.191, −0.165]

9. Discussion and Actuarial Implementation

The discussion in this section describes a proposed pathway toward production implementation, motivated by the results of Sections 7 and 8. It is not itself validated by those results: the experiment reported there is synthetic, single-cell, and homogeneous (Section 7.1), and none of the operational, governance, or data-architecture recommendations below has been tested against real portfolio data. Section 9.3's estimation-and-validation sequence describes precisely the additional empirical work—re-estimation on real data, re-running every verification exercise of Section 6, and rolling-origin backtesting—that would be required before any claim of demonstrated production applicability could be made.

9.1. Interpretation of the Hierarchy of Results

Four findings are central. First, delay-aware state reconstruction improves mean and standardised-CE performance relative to an optimised static rate, at both $\gamma = 0$ and $\gamma = 1.2$, even after matching the static benchmark's own risk-sensitivity to the dynamic policy's. Second, dynamic continuation has a small, positive, and remarkably γ-stable incremental mean-profit effect over a delay-aware myopic rule (0.017 million at both $\gamma = 0$ and $\gamma = 1.2$). Third, of the two effects that together make up the true-model full-information oracle's advantage, the mean-profit effect of exact state observation is overwhelmingly the larger and more stable one (0.161–0.182 million), while the effect of correct transition-law knowledge is small, changes sign with γ, and is worth little once state information is already available (Section 8.4); it is state reconstruction, specifically, that is economically primary. Fourth, and this is the central substantive contribution of this paper, entropic risk sensitivity does not unambiguously improve downside outcomes: a matched-γ static price achieves a better 5th percentile and TVaR 5% than the dynamic risk-sensitive policy at every $\gamma \geq 0.4$ tested, in every stress scenario (Sections 8.2 and 8.6), and even the $CE_{0.8}$ advantage reported for the production policy does not survive evaluation at the policy's own $\gamma_{\text{eval}} = 1.2$ (Table 12). The mean and $CE_{0.8}$ advantage on one hand, and the $CE_{1.2}$ and downside disadvantage on the other, are two genuinely different, both accurate, characterisations of the same policy—not a single verdict.

The zero-delay dynamic benchmark is deliberately strong: it raises prices more aggressively than the delay-aware policy (mean premium 822.5 versus 807.5 at $\gamma = 1.2$) and achieves a competitive, though not superior, downside outcome, despite using an internally incorrect observation model ($d_0 = 1$, no reservoir). The production delay-aware policy exceeds it in paired mean profit by 0.039 million (0.043 million at $\gamma = 0$), but, at $\gamma = 1.2$, is not distinguishable from it on the 5th percentile (interval spans zero) and is worse on TVaR 5%. The zero-delay policy's more aggressive, immediately reactive pricing generates comparable tail protection to the correctly delay-aware policy's more measured reaction, even though the latter remains superior on every measure of central tendency—illustrating why

a delay-aware model should be evaluated against equally optimised, equally risk-matched alternatives rather than a weak or mismatched benchmark.

9.2. Data Architecture

Production implementation requires linked policy, quote, occurrence, report, and development records: exposure intervals; quoted and accepted prices; renewal and lapse outcomes; occurrence and report dates; case estimates and payments; coverages; channel; policy tenure; and calendar covariates. A run-off triangle alone is insufficient because the system must identify the closed-loop relationship between price, exposure, occurrence, and later reporting—the same point made by the micro-level reserving literature of Section 2.2 (Antonio and Plat 2014; Crevecoeur et al. 2023), extended here to a setting in which the price itself is also a decision variable rather than a fixed covariate.

9.3. Estimation and Validation Sequence

Estimate conversion and retention response using quote- or renewal-level data, controlling for channel, selection, competitor effects, and endogeneity of historical price. Estimate mature-cohort frequency and severity models, or a joint occurrence-development likelihood that preserves reporting uncertainty, along the lines of Crevecoeur et al. (2023) or Abdelrahman et al. (2026). Estimate reporting hazards using survival or discrete-time hazard models with claim type, severity, seasonality, and operational changes as covariates.

Fit latent regimes using hidden Markov, state-space, or Cox-process methods and validate posterior calibration at rolling valuation dates—for example, a two-state HMM fitted to a rolling window of frequency and severity residuals to recover a litigation-environment indicator for liability, a threat-activity indicator for cyber, or a claims-inflation indicator for property-casualty lines generally; the regime label $Z(t)$ used throughout Sections 3–8 stands in for whichever such indicator is fitted to the portfolio at hand.

Backtest the complete closed loop with optimised static, cohort Bornhuetter–Ferguson, production pricing, and full-information proxy benchmarks where available at a common risk-sensitivity parameter across every benchmark in a given comparison. Re-run numerical support, grid-refinement, independent Bellman-residual, and filter-validation checks (Section 6) after each material recalibration, not only at initial deployment; Section 10 notes that the grid-adequacy finding of Section 6.1 is calibration-specific and not a standing guarantee. Select γ through risk appetite and predictive stress, rather than treating it as an arbitrary multiplicative loading—and, per Section 8.2, with an explicit understanding of what γ does and does not buy relative to a correctly risk-matched static alternative.

9.4. Governance and Customer Considerations

The framework does not imply continuous individual repricing. It can operate at rating-cell or portfolio level with monthly or quarterly reviews, capped movements, reason codes, and human approval. A recommendation record should include the filtered regime probability, reporting stock, applicable constraints, expected volume effect, and primary evidence supporting the move, together with an explicit statement of the assumed-density filter's known posterior-tracking error (Section 6.4) rather than presenting the filtered probability as if it were exact. Deployment must comply with product governance, permitted-rating-variable rules, anti-discrimination requirements, contractual renewal terms, and filing obligations. Model-risk monitoring should separately track demand, frequency, severity, reporting, and control components so that a wrong demand curve is not concealed by a conservative γ, and should specifically monitor realised severity against the calibration assumption, since severity misspecification is the one stress scenario in which the dynamic policy's mean-profit advantage over a static alternative was found to reverse (Section 8.6).

10. Limitations and Extensions

Structural and Scope Limitations

The first structural limitation is the one-phase reservoir. It is exact, in the aggregate mean-process sense discussed in Section 5, only for a memoryless geometric reporting tail. Long-tailed liability, bodily injury, cyber, and catastrophe-related claims require multiple phase stocks, an age grid, or a marked cohort representation, and the exact-filter truncation risk identified in Section 6.4 (a worst-case steady-state reservoir of 141.4 claims against a $U_{\max} = 120$ truncation) would bind more readily under a slower reporting hazard than the one calibrated here.

The second limitation is the assumed-density closure itself. The exact-filter comparison of Section 6.4 found imperfect posterior tracking (posterior RMSE 0.173; approximate-policy action agreement 64.0% over a 300-path, 10,800-state sample), and a mean filter-induced policy gap that is small but statistically distinguishable from zero (EUR -0.0092 ± 0.0013 million), while its 5th-percentile and TVaR 5% consequences remain statistically inconclusive; the same pattern holds when the diagnostic is re-run under the true out-of-model environment rather than the reference model (Table 6). Both of these facts should be carried forward together: the closure's posterior estimates are an imperfect guide to the true regime probability, and that imperfection has a measurable, non-zero mean cost, but the specific Q-surfaces this paper's policy uses do not appear, on the evidence available, to be highly sensitive to that posterior error in the tail of the outcome distribution at this calibration's belief region. Neither claim generalises automatically to a different portfolio, a different number of regimes, or a longer-tailed reporting distribution, and a 300-path validation sample remains a diagnostic rather than a general convergence result.

Third, the exposure response is instantaneous (Equation (2)). Annual contracts require an in-force cohort state distinguishing new business, renewals, lapses, and expiries, which this synthetic experiment does not model.

Fourth, the experiment is synthetic, single-cell, and homogeneous. A journal-grade empirical extension should estimate all components on real portfolio data, using rolling-origin evaluation against production pricing methods, and should re-run every verification exercise of Section 6 on the re-estimated model rather than assuming the diagnostics reported here transfer.

Fifth, the risk-sensitive recursion includes compound claim-count and severity risk but not parameter uncertainty in demand, frequency, severity, or reporting. Genuine model ambiguity—as distinct from the single-nominal-filter risk sensitivity of Section 4—would require multiple filters, robust Bayesian methods, or a set-valued information state, and would need to grapple directly with the fact that a posterior depends on both the transition law and the observation model, so that "ambiguity" cannot be introduced by reweighting a single filter's outcomes alone.

Sixth, the Bornhuetter–Ferguson and delay-naive benchmarks (Section 7.3) rely on a documented cohort-attribution simplification (proportional allocation of aggregate reservoir reports across active cohorts) rather than exact per-claim attribution, and on a fixed loss-ratio target ($\ell = 0.60$) chosen to place their operating range near the dynamic policies' own range rather than independently estimated; a reader using these benchmarks as a serious point of comparison for a real reserving process should treat them as illustrative rather than as a validated implementation of either method.

Extensions include multi-phase reporting, severity-mark filtering, uncertain demand elasticity treated as a filtered parameter rather than only as an evaluation-time shock, competitor games, joint reinsurance and capital controls, common regimes across lines, catastrophe and cyber clusters, active information acquisition, fairness constraints, and approximate dynamic programming for higher-dimensional rating structures—each of which

would also require its own version of the matched-benchmark, common-random-number, and independent-validation discipline this paper applies to the present, deliberately narrower, single-cell model.

11. Conclusions

Reporting delay is a structural information problem in non-life pricing. Reported claims are the output of an occurrence-reporting system, not a contemporaneous sample of current risk. This paper distinguishes the atomic hidden claim population from its intensity-valued transport equation, uses a nominal Bayesian filter to construct the decision state, and incorporates compound claim-count and severity risk directly into an entropic Bellman recursion.

The paper's contribution is as much methodological as substantive. Every benchmark is matched on risk-sensitivity parameter, and a second full-state benchmark separates the value of state information from the value of correct transition-law knowledge—two effects a single combined oracle comparison would otherwise conflate. The Monte Carlo evaluation uses a documented, genuinely paired common-random-number construction. The risk-sensitive Bellman recursion is checked by a separately coded re-implementation, agreeing with the production solver to floating-point precision. The exact-filter comparison uses terminology chosen to reflect what it actually measures, on a 300-path sample large enough to show its mean cost is small but statistically distinguishable from zero and robust to re-running under the true out-of-model environment. Its tail consequences, however, remain genuinely inconclusive. Finally, the misspecification testing spans eight scenarios with bootstrap intervals throughout, one of which reverses the sign of the dynamic policy's mean-profit advantage.

The substantive result is more nuanced than a simple "dynamic dominates static" headline would suggest. Delay-aware state reconstruction and dynamic continuation both remain valuable relative to correctly risk-matched benchmarks: the production policy improves mean profit by EUR 0.148 million and the standardised $\gamma_{\text{eval}} = 0.8$ certainty equivalent by EUR 0.034 million relative to an equally risk-averse optimised static price. That same static price, however, achieves a *better* 5th-percentile and TVaR 5% outcome, and, at the policy's own optimisation level $\gamma_{\text{eval}} = 1.2$, an interval-indistinguishable certainty equivalent (EUR -0.006 million, 95% interval reaching zero): dynamic control produces higher mean profit and a higher standardised $\text{CE}_{0.8}$, but not a clearly higher $\text{CE}_{1.2}$. On the exact entropic objective either policy was actually optimised to maximise, the static price is, if anything, marginally ahead.

The operational implication is precise. Pricing teams should first reconstruct current occurrence conditions from the reporting pipeline and preserve uncertainty in immature cohorts—the single largest source of value in the decomposition of Section 8.4. Dynamic programming can then govern the speed and direction of rate movement, adding a smaller but genuine increment beyond a delay-aware myopic rule. Entropy risk sensitivity is a secondary risk-appetite overlay whose downside-protection benefit should always be benchmarked against what an equally risk-averse *static* price would have achieved on its own. Neither control sophistication nor conservative preferences can substitute for a correctly specified demand model, reporting model, and filter; and no substantive claim about the value of any of the three should be trusted until it has been benchmarked against alternatives matched on every dimension except the one under test, evaluated with a documented common-random-number design, and validated by at least one independent check of the numerical machinery that produced it.

Author Contributions: Conceptualisation, D.M., S.T.M. and Ş.C.G.; methodology, D.M., S.T.M. and Ş.C.G.; software, D.M., S.T.M. and Ş.C.G.; validation, D.M., S.T.M. and Ş.C.G.; formal analysis, D.M., S.T.M. and Ş.C.G.; investigation, D.M., S.T.M. and Ş.C.G.; resources, D.M., S.T.M. and Ş.C.G.; data curation, D.M., S.T.M. and Ş.C.G.; writing—original draft preparation, D.M., S.T.M. and Ş.C.G.; writing—review and editing, D.M., S.T.M. and Ş.C.G.; visualisation, D.M., S.T.M. and Ş.C.G.; supervision, D.M., S.T.M. and Ş.C.G.; project administration, D.M., S.T.M. and Ş.C.G.; funding acquisition, D.M., S.T.M. and Ş.C.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Discrete Algorithm

1. Observe state (p_t, π_t, u_t) at the beginning of month t.
2. Evaluate feasible moves $a \in \{-1, 0, 1\}$ and set $p' = p_t + 25a$.
3. For each regime, calculate exposure, occurrence intensity, severity moment-generating function, and nominal and tilted report-count means.
4. Sum report counts $r = 0, \ldots, 150$ exactly using Equation (15).
5. For each r, update the nominal current-regime posterior, propagate it through P^{ref}, and update the assumed-density reservoir.
6. Interpolate the next-period value at the successor posterior and reservoir state.
7. Apply Equation (17), store all three action values (not only their maximum), and record the maximising action.
8. **Online**, interpolate the three *stored action-value surfaces* bilinearly at the current continuous belief state, take the argmax of the three interpolated values, and only then apply governance constraints and human approval. Do not interpolate a categorical policy directly (Section 5).
9. At $\gamma = 0$, replace steps 4–7 with the separately coded exact expected-value branch (Section 5), not a numerical $\gamma \to 0$ limit of the $\gamma > 0$ branch.

Appendix B. Bornhuetter–Ferguson and Delay-Naive Benchmark Construction

For each occurrence month in a rolling 15-month window, the Bornhuetter–Ferguson benchmark stores earned exposure, the cohort's initial delayed-occurrence count J_k (from which $N^{\text{prior}}(k) = f^{\text{prior}} J_k / (1 - d_0^{\text{ref}})$-consistent exposure-based prior is derived; in the implementation used in this paper, $N^{\text{prior}}(k) = (f_{\text{L}} + f_{\text{H}})/2 \cdot \text{exposure}_k/12$ directly, i.e., the simple blended-frequency prior described in Section 7.3), and cumulative reports attributed to that cohort. At valuation month t, it applies the reference cumulative reporting proportion F_D at that cohort's age and Equation (18). Estimated ultimate counts are aggregated across active cohorts and divided by their combined exposure-years to obtain an annualised frequency estimate $\hat{f}_t$. Both the Bornhuetter–Ferguson rule and the delay-naive rule then apply the common target-price rule of Section 7.3, $p_t^{\text{target}} = (c_e + \hat{f}_t \bar{\mu})/\ell$, moving toward the target subject to the same EUR 25 monthly governance limit as the dynamic policies. The delay-naive rule differs only in how $\hat{f}_t$ is formed: rather than cohort-attributed reports, it exponentially smooths raw monthly report counts R_t, treated (incorrectly, by construction, since this is the defining misspecification of a "naive" rule) as if they were current-month occurrences, with smoothing parameter $\alpha = 0.30$.

Appendix C. Reproducibility Definitions

Economic profit equals earned premium less variable expense and ultimate loss for claims occurring within the 36-month horizon. It excludes the Bellman movement penalty, which is a governance cost used to shape the policy but is not part of the reported economic outcome. The modelled loss-and-expense ratio equals (ultimate loss plus variable expense) divided by earned premium. Exposure index is the average effective exposure divided by 5000. The fifth percentile is the empirical 0.05 quantile across simulated paths, and TVaR 5% is the mean profit at or below that quantile. The certainty equivalent at evaluation level γ_{eval} is $\text{CE} = -\gamma_{\text{eval}}^{-1} \log\left(\frac{1}{N}\sum_j e^{-\gamma_{\text{eval}} X_j}\right)$, computed with the standard log-sum-exp numerical stabilisation. Paired bootstrap intervals use 2000 resamples of the shared path index (not independent resampling of each strategy's outcomes), consistent with the common-random-number pairing of Section 5.4. Mean delay under a geometric reporting model with immediate-report fraction d_0 and reservoir hazard δ is computed as $(1 - d_0)/\delta$ months (the mean of the geometric reservoir sojourn time for the $(1 - d_0)$ fraction of claims that do not report immediately), giving 3.00 months under the reference parameters and 4.89 months under the true parameters of Table 9. All parameter values, random-number seeds, simulation routines, and figure-generation scripts used for the synthetic experiment are available from the corresponding author on request.

References

Abdelrahman, Hassan, Andrei L. Badescu, Radu V. Craiu, and X. Sheldon Lin. 2026. Marked Cox models for IBNR claims count: Continuous and discretized approaches with Dirichlet-driven reporting delays. *ASTIN Bulletin* 56: 60–88. [CrossRef]

Anderson, Evan W., Lars Peter Hansen, and Thomas J. Sargent. 2003. A quartet of semigroups for model specification, robustness, prices of risk, and model detection. *Journal of the European Economic Association* 1: 68–123. [CrossRef]

Antonio, Katrien, and Richard Plat. 2014. Micro-level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal* 2014: 649–69. [CrossRef]

Arjas, Elja. 1989. The claims reserving problem in non-life insurance: Some structural ideas. *ASTIN Bulletin* 19: 139–52. [CrossRef]

Avanzi, Benjamin, Bernard Wong, and Xinda Yang. 2016. A micro-level claim count model with overdispersion and reporting delays. *Insurance: Mathematics and Economics* 71: 1–14. [CrossRef]

Badescu, Andrei L., Tianle Chen, X. Sheldon Lin, and Dameng Tang. 2019. A marked Cox model for the number of IBNR claims: Estimation and application. *ASTIN Bulletin* 49: 709–39. [CrossRef]

Badescu, Andrei L., X. Sheldon Lin, and Dameng Tang. 2016. A marked Cox model for the number of IBNR claims: Theory. *Insurance: Mathematics and Economics* 69: 29–37. [CrossRef]

Bain, Alan, and Dan Crisan. 2009. *Fundamentals of Stochastic Filtering*. New York: Springer.

Bäuerle, Nicole, and Ulrich Rieder. 2017. Partially observable risk-sensitive Markov decision processes. *Mathematics of Operations Research* 42: 1180–96. [CrossRef]

Bornhuetter, Ronald L., and Ronald E. Ferguson. 1972. The actuary and IBNR. *Proceedings of the Casualty Actuarial Society* 59: 181–95. [CrossRef]

Crevecoeur, Jonas, Katrien Antonio, Stijn Desmedt, and Alexandre Masquelein. 2023. Bridging the gap between pricing and reserving with an occurrence and development model for non-life insurance claims. *ASTIN Bulletin* 53: 185–212. [CrossRef]

Davis, Mark H. A. 1993. *Markov Models and Optimization*. London: Chapman & Hall.

Dong, Stella C. 2025. Adaptive insurance reserving with CVaR-constrained reinforcement learning under macroeconomic regimes. *arXiv* arXiv:2504.09396.

Elliott, Robert J., Lakhdar Aggoun, and John B. Moore. 1995. *Hidden Markov Models: Estimation and Control*. New York: Springer.

Emms, Paul. 2007. Dynamic pricing of general insurance in a competitive market. *ASTIN Bulletin* 37: 1–34. [CrossRef]

Emms, Paul, and Steven Haberman. 2005. Pricing general insurance using optimal control theory. *ASTIN Bulletin* 35: 427–53. [CrossRef]

England, Peter D., and Richard J. Verrall. 2002. Stochastic claims reserving in general insurance. *British Actuarial Journal* 8: 443–518. [CrossRef]

Fleming, Wendell H., and Halil Mete Soner. 2006. *Controlled Markov Processes and Viscosity Solutions*, 2nd ed. New York: Springer.

Gozzi, Fausto, and Federica Masiero. 2017a. Stochastic optimal control with delay in the control I: Solving the HJB equation through partial smoothing. *SIAM Journal on Control and Optimization* 55: 2981–3012. [CrossRef]

Gozzi, Fausto, and Federica Masiero. 2017b. Stochastic optimal control with delay in the control II: Verification theorem and optimal feedbacks. *SIAM Journal on Control and Optimization* 55: 3013–38. [CrossRef]

Haastrup, Svend, and Elja Arjas. 1996. Claims reserving in continuous time: A nonparametric Bayesian approach. *ASTIN Bulletin* 26: 139–64. [CrossRef]

Hansen, Lars Peter, and Thomas J. Sargent. 2008. *Robustness*. Princeton: Princeton University Press.

He, Qinyang, and Yonatan Mintz. 2026. Prescriptive optimization for adaptive auto-insurance pricing with telematics data. *arXiv* arXiv:2605.06954.

Henckaerts, Roel, and Katrien Antonio. 2022. The added value of dynamically updating motor insurance prices with telematics collected driving behavior data. *Insurance: Mathematics and Economics* 105: 79–95. [CrossRef]

Law, Averill M., and David M. Kelton. 2000. *Simulation Modeling and Analysis*, 3rd ed. Boston: McGraw-Hill.

Lawless, Jerald F. 1994. Adjustments for reporting delays and the prediction of occurred but not reported events. *Canadian Journal of Statistics* 22: 15–31. [CrossRef]

Lim, Andrew E. B., and J. George Shanthikumar. 2007. Relative entropy, exponential utility, and robust dynamic pricing. *Operations Research* 55: 198–214. [CrossRef]

Mack, Thomas. 1991. A simple parametric model for rating automobile insurance or estimating IBNR claims reserves. *ASTIN Bulletin* 21: 93–109. [CrossRef]

Martínez-Miranda, Maria Dolores, Jens Perch Nielsen, and Richard Verrall. 2012. Double chain ladder. *ASTIN Bulletin* 42: 59–76.

Norberg, Ragnar. 1993. Prediction of outstanding liabilities in non-life insurance. *ASTIN Bulletin* 23: 95–115. [CrossRef]

Norberg, Ragnar. 1999. Prediction of outstanding liabilities II: Model variations and extensions. *ASTIN Bulletin* 29: 5–25. [CrossRef]

Okine, A. Nii-Armah. 2023. Ratemaking in a changing environment. *ASTIN Bulletin* 53: 596–618. [CrossRef]

Pham, Huyên. 2009. *Continuous-Time Stochastic Control and Optimization with Financial Applications*. Berlin: Springer.

Verbelen, Roel, Katrien Antonio, Gerda Claeskens, and Jonas Crevecoeur. 2022. Modeling the occurrence of events subject to a reporting delay via an EM algorithm. *Statistical Science* 37: 394–410. [CrossRef]

Whittle, Peter. 1990. *Risk-Sensitive Optimal Control*. New York: Wiley.

Wonham, Walter M. 1964. Some applications of stochastic differential equations to optimal nonlinear filtering. *SIAM Journal Control* 2: 347–69. [CrossRef]

Wu, Zhengqi, and Renyuan Xu. 2023. Risk-sensitive Markov decision processes and learning under general utility functions. *arXiv* arXiv:2311.13589.

Wüthrich, Mario V., and Merz Merz. 2008. *Stochastic Claims Reserving Methods in Insurance*. Chichester: Wiley.

Yanez, Juan Sebastian, Montserrat Guillén, and Jens Perch Nielsen. 2025. Weekly dynamic motor insurance ratemaking with a telematics signals bonus-malus score. *ASTIN Bulletin* 55: 1–28. [CrossRef]