ARTICLE | Article in Press

Quantifying controller preference reliability under uncertainty in a rotary inverted pendulum benchmark

Yusuf Coşkun · Lale Özyılmaz · Muhammet Aydın

Received: 22 August 2026 / Accepted: 15 October 2026
© The Author(s) 2026

Abstract
Simulation-based controller comparisons can report a single winner even when score differences are too small to justify a unique engineering preference. We assessed preference reliability for five fixed controllers on a numerically verified nonlinear rotary inverted pendulum model using two independent Latin-hypercube realizations across matched ±10%, ±15%, and ±20% uncertainty bands (30,000 closed-loop simulations). A four-metric score and a prespecified 5% practical margin quantified controller separation. Under the common uncertainty score, PID had the lowest controller robustness index in every band, but the four leading indices remained within 3.87%. In the pooled ±15% domain, all ten paired comparisons were detectable after Holm correction, yet 69.40% of 2,000 scenario blocks lacked a unique practical preference. PID retained rank one in 99.95% of paired-bootstrap replicates but only 63.16% and 43.06% under two weight-variation analyses. Post-review sensitivity analyses showed that the principal practical-separation conclusion was robust to the tested CRI coefficient, realization-specific median normalization, and heavier-tail dependence models, while command delay altered separation among near-tied controllers. These findings are specific to the fixed controller portfolio, score definition, RIP model, and simulated uncertainty domain.

Keywords Rotary inverted pendulum · Controller benchmarking · Uncertainty quantification · Practical indifference · Gaussian process regression · Controller preference

Introduction

Underactuated mechanical systems combine nonlinear coupling, limited actuation and unstable operating regimes, so a controller comparison can depend strongly on the physical parameters, forcing conditions and performance criteria used for evaluation. The rotary inverted pendulum (RIP, or Furuta pendulum) is a compact benchmark for this problem: a motor-driven horizontal arm must regulate an unactuated pendulum about the upright equilibrium while actuator motion and effort remain bounded. Linear, nonlinear, sliding-mode, fuzzy and neural-network-assisted controllers have all been demonstrated on this system[1–15]. Recent studies likewise compare RIP controllers under selected nominal, disturbance or parameter-variation conditions[4,9,11]. Such studies establish controller

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

Published online: 21 October 2026

behaviour for the stated protocol, but a numerical minimum under one protocol does not by itself show that a practically unique controller preference persists when the evaluation criterion or operating conditions change.

Randomized robustness analysis provides one route for examining this dependence. Monte Carlo methods and Latin hypercube sampling (LHS) propagate several uncertain inputs through a closed-loop model and expose the distribution of response metrics rather than a single nominal trajectory[16–19]. For a controller portfolio, applying the same physical scenario to every candidate creates a paired comparison: differences in pendulum regulation, arm motion, actuator effort and control roughness are then attributed to the controller response under that scenario rather than to different random draws. The resulting ranking can nevertheless remain conditional on how those responses are combined. Finite scenario sampling, the engineering weights assigned to the metrics, and the choice of metric set itself are distinct parts of the evaluation protocol and need not produce the same ordering.

Simulation ranking-and-selection addresses a closely related issue. Indifference-zone formulations distinguish a numerically best alternative from alternatives whose performance differs by less than a prespecified practical tolerance[20], whereas indifference-zone-free and fixed-tolerance formulations avoid treating every small numerical difference as decision-relevant[21,22]. Common random numbers strengthen paired comparisons by exposing alternatives to the same stochastic conditions[23]. Ranking and selection with covariates allows the preferred alternative to vary with observable instance descriptors[24], and the broader idea of choosing a method from instance characteristics goes back to the classical algorithm-selection problem[25]. These concepts are relevant to the present comparison, but no new ranking-and-selection procedure or probability guarantee is proposed. The question is instead how often a conventional controller benchmark supports a unique practical preference when the controller portfolio and simulation budget are fixed.

Controller scores vary continuously across the scenario space, so regression models are useful for estimating score differences between simulated points. Gaussian process regression (GPR) supplies a predicted score and predictive dispersion conditional on the scenario descriptors, but this predictive distribution is not treated as physical uncertainty[26–32]. Calibration and coverage diagnostics are therefore considered together with point-prediction accuracy[33,34]. Because every controller is evaluated under the same physical scenario, their prediction errors can also be dependent; joint ordering probabilities cannot be obtained from five marginal models alone[35–37]. The decision rule is allowed to return no unique practical preference when the predicted score separation remains below the prescribed margin[38].

Recent control studies also illustrate why controller comparisons require explicit reporting of architecture-specific design and tuning assumptions. Higher-order model-free control tuned by fictitious reference iterative tuning (FRIT) has been investigated on three-dimensional crane systems[39], while indirect adaptive learning has been used to enhance P-type control through set-point updates[40]. Data-driven FRIT has also been applied to active-disturbance-rejection and model-free control[41], and fuzzy logic control has been integrated with path planning for mobile-robot navigation in dynamic environments[42]. These studies address different plants and objectives and are not direct RIP ranking benchmarks; their relevance here is that controller structures, tuning variables, and intended operating assumptions can differ substantially. Accordingly, the present comparison is restricted to the five fixed controller instances obtained under the declared common protocol and is not interpreted as an architecture-level superiority test.

Accordingly, the analysis focuses on a narrower question than simply identifying the controller with the smallest score: when does the available simulation evidence support a unique practical preference? Five fixed controllers—PID, LQR, SMC, Fuzzy-SMC and ANFIS-SMC—are evaluated on the same numerically verified nonlinear electromechanical RIP model with a common estimator, voltage saturation and paired uncertainty scenarios. The analysis distinguishes criterion dependence, engineering-weight sensitivity and finite-sample variability, then evaluates practical separation with a prespecified relative margin and two independent LHS realizations. Controller-specific GPR score models are finally combined with the residual correlation estimated from training data to obtain joint event probabilities. The contribution is a quantitative assessment of preference reliability for this

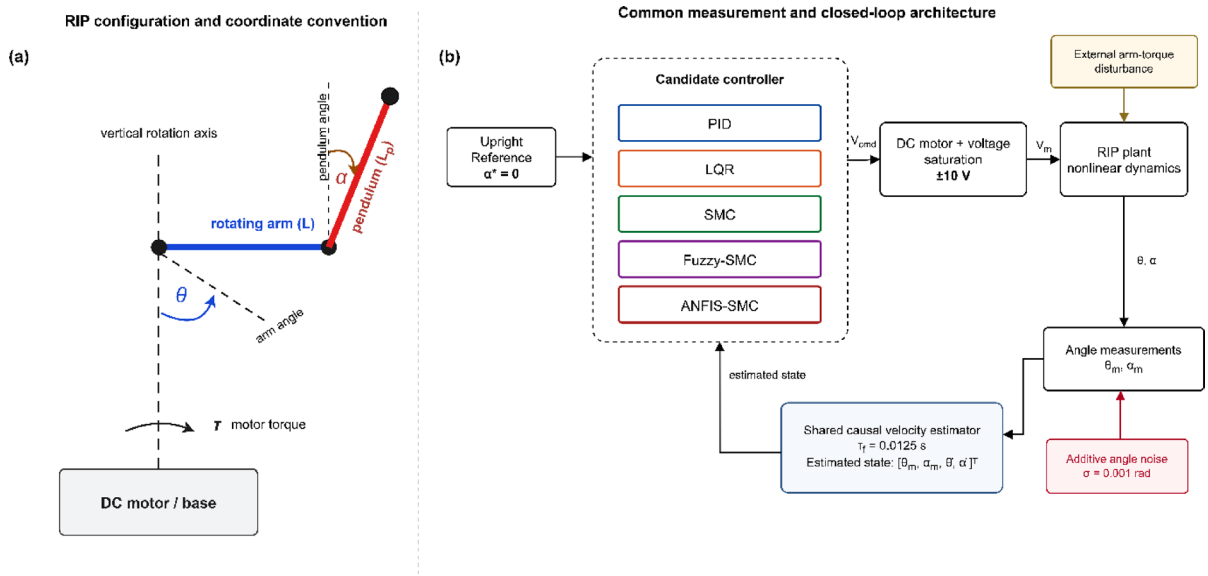


Fig. 1 Rotary inverted pendulum configuration and common measurement/closed-loop architecture. Panel (**a**) defines the coordinate convention, arm and pendulum angles, motor torque and rotation axis. Panel (**b**) shows the common reference, controller portfolio, motor and voltage saturation, nonlinear RIP plant, angle measurements, additive measurement noise, external arm-torque disturbance and shared causal velocity estimator.

Table 1 Core simulation specification and compact numerical-verification summary.

Quantity	Setting/result	Interpretation
Numerical setup	T = 5 s; Δt = 0.002 s; τf = 0.0125 s; actuator limit ± 10 V	Common closed-loop settings
Scenario design	2 independent LHS realizations × 1,000 blocks; bands ± 10%, ± 15%, ± 20%	Same block applied to all five controllers
Uncertain plant quantities	m, Lp, Kt, br; dependent quantities recomputed	Four plant dimensions
Disturbance/noise description	Five disturbance descriptors; Gaussian angle noise σ = 0.001 rad	Five forcing dimensions plus fixed noise level
Energy conservation	Relative drift = 1.608×10^{-11} over 20 s	Conservative-model check
Linearization check	max |ΔA| = 1.5×10^{-12}; max |ΔB| = 3.6×10^{-15}	Analytical vs. finite difference
Local nonlinear–linear agreement	RMS difference at 0.25°: 0.00163° (θ), 0.00195° (α)	Small-perturbation consistency
RK4 convergence	Observed order 3.987–3.997	Time-integration consistency
Closed-loop time-step sensitivity	0.408% αRMS; 0.047% θRMS; 0.133% URMS	Δt = 0.002 vs. 0.001 s

controller portfolio and uncertainty domain; it is not a claim of universal controller superiority or a new ranking-and-selection guarantee.

Results

Controller preference changed with the evaluation criterion before uncertainty was introduced

The five fixed controller designs operate on the same nonlinear plant, causal estimator, voltage-saturation block and disturbance interface (Fig. 1), with the shared simulation and numerical-verification specification summarized in Table 1. They were first tuned under the common five-term objective used in the original controller-design

stage. Under that objective, LQR ranked first (0.572362), followed by PID, ANFIS-SMC, Fuzzy-SMC and SMC. To remove criterion confounding from the subsequent uncertainty comparison, the same fixed controllers were then re-evaluated at the nominal plant point using the four response metrics, pooled-median normalization constants and base weights used in the uncertainty study. This common-score evaluation changed the ordering substantially: ANFIS-SMC gave the lowest nominal score (0.368502), followed by Fuzzy-SMC (0.369421), SMC (0.383970), PID (0.387630) and LQR (0.509571) (Table 2a). Relative to the ANFIS-SMC minimum, the other four scores were 0.25%, 4.20%, 5.19% and 38.3% higher, respectively. The nominal good-controller set contained two controllers at a 2.5% margin, three at 5%, and four at 10%, showing directly that any nominal set-membership statement is conditional on the stated practical tolerance. The corresponding nominal closed-loop trajectories are shown in Fig. 2.

The rank reordering is not interpreted as a change in controller physics caused by uncertainty, because it occurs before uncertainty is introduced. The two criteria answer different questions: the tuning objective includes transient shaping terms such as overshoot and settling time, whereas the common evaluation score summarizes pendulum regulation, sustained arm motion, applied-voltage RMS and control roughness. All five controllers were exposed to the same tuning protocol, but the common score was not the optimization target of any controller. Its values therefore represent a post-tuning comparison of the fixed portfolio rather than the best performance each controller could attain if it were retuned for that score. This criterion dependence is itself relevant to preference reliability because it shows that a controller ranking can change solely through the definition of performance. When the same common score was propagated through uncertainty and summarized by the Controller Robustness Index (CRI), PID ranked first in all three uncertainty bands, while LQR remained fifth (Table 2b).

Most ± 15% scenario blocks did not support a uniquely separated controller

The uncertainty campaign comprised 30,000 closed-loop runs: 2,000 paired scenario blocks, three matched plant-uncertainty bands, and five controllers. All runs completed the prescribed horizon, satisfied the terminal regulation criterion and remained below the ± 10 V actuator limit; the maximum applied voltage was 6.895 V. Here, a scenario block denotes one physical parameter, disturbance and noise realization shared by all five controllers. For pooled ± 15% analyses, 1,000 blocks came from the primary LHS realization and 1,000 from an independently

Table 2 Controller ranking under the original tuning objective, the common nominal evaluation score, and the uncertainty-aware Controller Robustness Index (CRI). The common score uses the same four response metrics, pooled-median normalization constants and base weights as the uncertainty analysis.

Panel a. Original tuning objective and nominal common-score evaluation of the fixed controller portfolio.

Controller	Tuning objective	Tuning rank	Nominal common score	Common-score rank	Gap from common minimum
PID	0.584360	2	0.387630	4	5.19%
LQR	0.572362	1	0.509571	5	38.3%
SMC	0.673177	5	0.383970	3	4.20%
Fuzzy-SMC	0.655838	4	0.369421	2	0.25%
ANFIS-SMC	0.655373	3	0.368502	1	0.00%

Panel b. Uncertainty-aware CRI under the same common evaluation score.

Controller	CRI ±10%	CRI ±15%	CRI ±20%	Overall CRI	Overall rank	Bootstrap rank-1	CRI gap, ±10 / ±15 / ±20 (%)
PID	1.194394	1.195591	1.202979	1.196392	1	99.95%	0.00 / 0.00 / 0.00
LQR	1.646985	1.659782	1.668615	1.660685	5	0.00%	37.89 / 38.83 / 38.71
SMC	1.228951	1.224214	1.225298	1.225149	2	0.05%	2.89 / 2.39 / 1.86
Fuzzy-SMC	1.238217	1.232062	1.228954	1.232800	3	0.00%	3.67 / 3.05 / 2.16
ANFIS-SMC	1.240603	1.237038	1.233598	1.237704	4	0.00%	3.87 / 3.47 / 2.55

The common score is a post-tuning evaluation criterion; none of the controller designs was retuned to minimize it. Panel b applies the same 5% practical relevance concept to the CRI point estimates through the bandwise percentage gaps shown in the last column. The table therefore demonstrates criterion dependence and ordering stability of the fixed portfolio without treating a small numerical rank difference as evidence of practical separation.

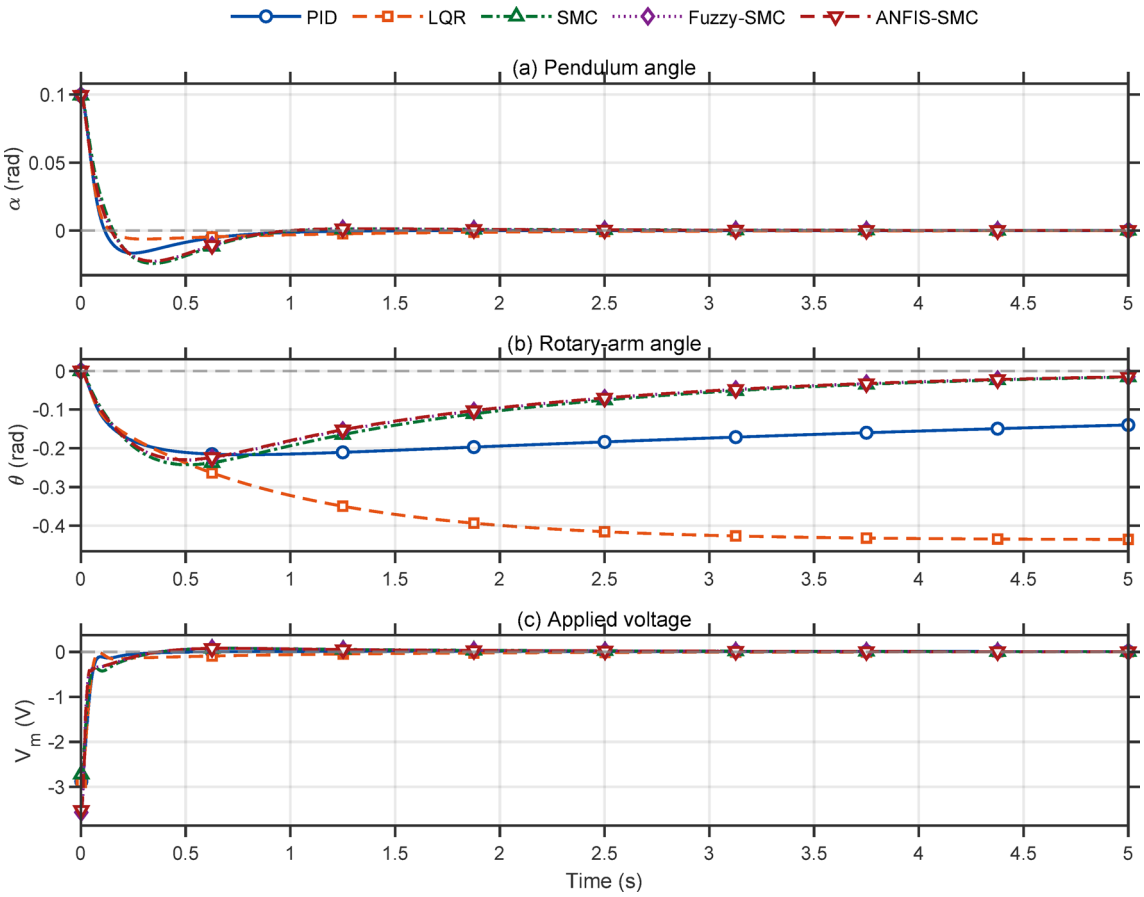


Fig. 2 Nominal closed-loop responses of the five fixed controllers at the common nominal plant point. (**a**) Pendulum angle, (**b**) rotary-arm angle and (**c**) applied motor voltage over the 5 s simulation horizon.

generated replication. The two realizations are pooled only for the $N = 2{,}000$ descriptive and probabilistic analyses; cross-realization tests keep them separate and train on one 1,000-block realization before testing on the other.

At the prespecified 5% practical margin, 1,388 of the 2,000 true simulated score vectors (69.40%) contained at least two controllers within the good-controller set and therefore did not support a uniquely separated preference. The result reproduced closely across the two independent designs: 690/1,000 blocks (69.0%) in the primary realization and 698/1,000 (69.8%) in the replication. Good-set cardinality further showed that 612 blocks (30.60%) contained one controller within the margin, 416 (20.80%) contained two, 640 (32.00%) contained three, and 332 (16.60%) contained four; none contained all five. Increasing the practical margin from 2.5% to 10% increased the fraction of blocks without a unique practical preference from 53.45% to 94.90%, showing that the reported 69.40% is conditional on the stated engineering tolerance rather than a universal property of the controller set (Fig. 3).

Criterion definition, engineering weights and finite scenario sampling produced different forms of ranking variability

Under the fixed base score, the CRI order was unchanged across the ±10%, ±15% and ±20% bands and in the matched overall summary: PID, SMC, Fuzzy-SMC, ANFIS-SMC and LQR (Table 2b). Applying the same 5% practical margin to these aggregate summaries changes the interpretation of that ordering. The four leading controllers differed from the bandwise CRI minimum by at most 3.87%, 3.47% and 2.55% in the ±10%, ±15% and

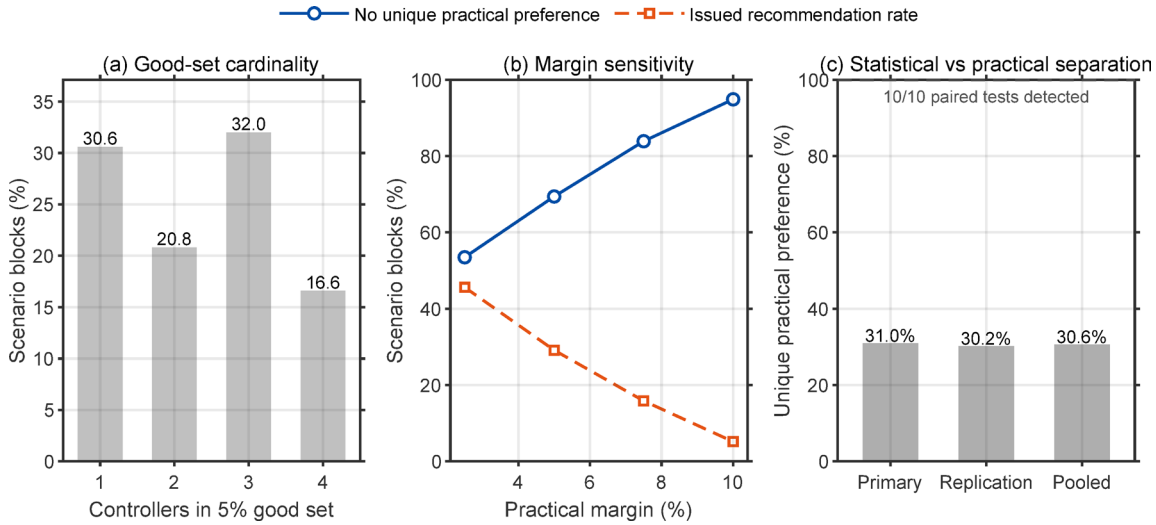


Fig. 3 Practical separation across the ±15% scenario domain. (**a**) Cardinality of the 5% good-controller set, (**b**) sensitivity of the fraction of blocks without a unique practical preference and the issued recommendation rate to the practical margin, and (**c**) comparison of statistical detectability with the fraction of scenario blocks that support a unique practical preference in the primary, replication and pooled datasets.

±20% bands, respectively, whereas LQR was 37.9–38.8% above the minimum. The CRI ordering is therefore internally consistent across uncertainty bands but does not itself establish a practically separated aggregate preference among the four leading designs.

The five paired score distributions differed in the overall $N = 2{,}000$ analysis (Friedman Q = 2895.772, df = 4, asymptotic $P \approx 2.25 \times 10^{-626}$; Kendall's W = 0.362). Two-sided paired Wilcoxon signed-rank tests were Holm-adjusted at $\alpha = 0.05$; the ten adjusted P values are reported in Supplementary Table S1 and ranged from 1.20×10^{-6} to approximately 1.37×10^{-326}. The smallest absolute matched-pairs rank-biserial correlation occurred for PID–Fuzzy-SMC (|r_rb| = 0.1253), whereas the largest occurred for PID–LQR (|r_rb| = 0.9990). These Wilcoxon tests concern the distribution of block-wise paired score differences. By contrast, the paired bootstrap intervals that spanned zero for SMC–Fuzzy-SMC and Fuzzy-SMC–ANFIS-SMC concern sampling uncertainty in the aggregate CRI statistic, median(J) + 0.5IQR(J); the two results address different estimands. Statistical detectability did not imply engineering distinguishability: all ten paired score comparisons were detectable after Holm correction, yet 69.40% of the same scenario blocks contained no uniquely separated controller under the prespecified 5% practical margin. The extreme omnibus P value and W = 0.362 show the same distinction: a large matched simulation sample can make systematic score differences statistically detectable without establishing a practically unique preference.

Band 10 also illustrates why paired-difference statistics and marginal distribution summaries should not be read as interchangeable rankings. The median paired score difference from PID was 0.0221 for Fuzzy-SMC and 0.0384 for SMC, whereas the CRI placed SMC closer to PID. This is not an inconsistency: the median of paired differences is not generally equal to the difference of marginal medians, and CRI additionally includes one-half of the marginal interquartile range. The example therefore reinforces the broader criterion-dependence result rather than supplying a competing controller order. The marginal median(J), IQR(J), and 0.5IQR(J) components underlying every band-specific and overall CRI value are reported in Supplementary Table S10, allowing each CRI value to be reconstructed directly. Supplementary Table S10 also shows that the median and interquartile components order the three SMC-family controllers differently in some contexts; the CRI ordering among these designs therefore reflects the stated balance between central tendency and dispersion rather than a single dominance relation.

Finite-sample stability and preference-weight stability were markedly different under the same score definition. PID retained overall CRI rank one in 99.95% of 2,000 paired bootstrap replicates, whereas its rank-one share fell to 63.16% when the base priority structure was preserved while the remaining weight mass varied, and to 43.06% under unrestricted simplex weights (Fig. 4). The bootstrap result concerns reproducibility of the numerical ordering, not its practical relevance: across the individual bands the other three leading CRIs were only 1.86–3.87% above the minimum, and in the overall summary they were 2.40–3.45% higher, all within the 5% practical margin. Together with the deterministic rank reordering caused by changing the metric set at the nominal point, these analyses isolate three distinct sources of ranking variability: criterion definition, engineering weights and finite scenario sampling. They are not directly commensurate uncertainty measures, but they act at different stages of the evaluation protocol and should not be conflated.

The dispersion coefficient in the descriptive CRI was also varied from lambda = 0 to 1 in increments of 0.25 without retuning the controllers or altering the underlying scenario scores (Supplementary Table S13 and Supplementary Fig. S13). The archived lambda = 0.50 CRI values were reproduced with a maximum absolute difference of 4.89 × 10^−15. PID remained the overall CRI minimum and LQR the maximum at every tested coefficient. Across the same range, the largest overall gap from PID among SMC, Fuzzy-SMC and ANFIS-SMC changed only from 3.37% to 3.51%, so all four leading controllers remained within the 5% practical margin, whereas the LQR gap increased from 34.09% at lambda = 0 to 42.17% at lambda = 1. The overall SMC and Fuzzy-SMC ordering exchanged only near the upper end of the tested range (pairwise crossover lambda = 0.995), and the complete ± 15% band ordering remained unchanged across all five tested coefficients. Thus, the exact ordering among the near-tied leading controllers can depend on the CRI dispersion penalty, but the principal practical-separation interpretation was unchanged over the tested coefficient range.

Score normalization was likewise tested without retuning or resimulation (Supplementary Table S14 and Supplementary Fig. S14). Using medians estimated separately from the primary and replication LHS realizations changed the four normalization constants by no more than 0.70% relative to the locked pooled medians. Both alternatives reproduced the same overall CRI order (PID, SMC, Fuzzy-SMC, ANFIS-SMC, LQR), kept all four leading controllers within 3.43% of the minimum, and yielded 69.50% and 69.35% of ± 15% blocks without a unique 5% preference, respectively, compared with 69.40% under pooled-median scaling. As a stronger stress test, pooled-mean scaling changed the rotary-arm normalization scale by 42.19%; the overall numerical order

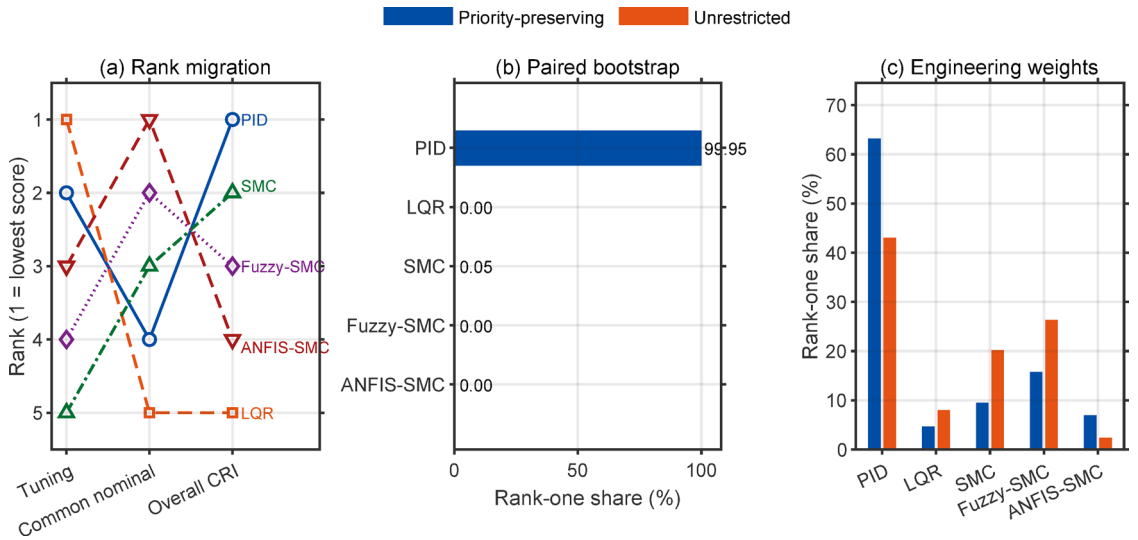


Fig. 4 Ranking variability under changes in the evaluation protocol. (**a**) Rank migration from the nominal tuning objective to the common nominal score and overall CRI, (**b**) paired-bootstrap rank-one frequencies, and (**c**) rank-one shares under priority-preserving and unrestricted engineering-weight variation.

still remained unchanged and 67.50% of ±15% blocks remained non-unique, but the SMC, Fuzzy-SMC and ANFIS-SMC aggregate CRI gaps increased to 5.73%, 6.34% and 6.71%. Thus the rank endpoints and the majority prevalence of scenario-level non-unique preference were stable, whereas the stronger aggregate statement that all four leading CRIs lie inside a 5% band was specific to median-based scaling.

A prescribed command-delay envelope preserved regulation but altered practical separation

A post-review command-delay sensitivity analysis was performed on the pooled±15% domain using the same 2,000 paired scenarios and the same fixed controller instances, without retuning or delay compensation (Supplementary Table S15 and Supplementary Fig. S15). The archived 0 ms baseline was retained and 50,000 new closed-loop runs evaluated command delays of 2, 4, 6, 8 and 10 ms, corresponding to one to five samples at $\Delta t = 0.002$ s. All five controllers completed all 2,000 scenario blocks at every tested delay, so the common all-five-success subset remained 100% throughout. PID remained the minimum-CRI controller and LQR remained fifth at every delay, but the ordering within the SMC-family controllers changed at 4–8 ms. The fraction of blocks without a unique 5% practical preference increased from 69.40% at 0 ms to 70.50%, 71.60%, 73.05% and 74.45% at 2–8 ms, then decreased to 72.00% at 10 ms. At 10 ms, SMC remained only 1.95% above the PID CRI, whereas the Fuzzy-SMC and ANFIS-SMC CRI gaps increased to 7.34% and 8.71%, respectively. Thus, within this prescribed delay envelope, loss of regulation did not occur, but controller separation and the ordering among near-tied designs changed before any feasibility failure was observed.

GPR predictions closely matched the simulated scores, while joint probability estimates were sensitive to cross-controller residual dependence

Controller-specific GPR models were assessed out of fold on the pooled $N = 2{,}000 \pm 15\%$ scenario set. R^2 ranged from 0.992576 to 0.998400 and RMSE from 0.018745 to 0.024006. Pooled empirical predictive-interval coverage was 51.26%, 80.76%, 89.90% and 94.61% at nominal 50%, 80%, 90% and 95% levels, respectively. Standardized residual means were within 0.01 of zero and standard deviations ranged from 1.0037 to 1.0102. Controller-level regression, residual and coverage diagnostics are reported in Supplementary Table S18. Descriptive residual diagnostics showed modest departures from exact Gaussian shape rather than exact normality; the joint score model is treated as a Gaussian approximation whose adequacy is evaluated through residual diagnostics, interval coverage and sensitivity analysis rather than as a verified residual law.

The controller prediction errors were positively dependent. Training-only residual correlations ranged from 0.515 to 0.967, with the largest value between Fuzzy-SMC and ANFIS-SMC. Replacing the estimated correlation structure with independence while holding every marginal GPR prediction fixed changed the probability of no unique practical preference by 0.1123 on average and by as much as 0.4282, changed the six-event maximum-a-posteriori decision in 3.30% of scenario blocks, and produced a mean L1 change of 0.2251 in the six-event probability vector (Fig. 5). These results support the narrower claim that ignoring predictive dependence can materially change the joint decision probabilities for this portfolio; they do not establish that the shared-scenario design is the sole cause of the observed residual correlation.

A post-review dependence-form sensitivity retained the same marginal GPR means and standard deviations and the same training-only, fold-specific residual correlation matrices, but replaced the Gaussian dependence representation with Student-t copulas having $\nu = 8$ and $\nu = 4$ (Supplementary Table S16 and Supplementary Fig. S16). The mean predicted probability of no unique practical preference was 0.700419 under the locked Gaussian model, 0.700486 for $\nu = 8$, and 0.700549 for $\nu = 4$. Relative to the locked Gaussian calculation, the mean absolute change in this probability was 0.003052 and 0.005999, with maxima of 0.02220 and 0.04475, respectively. Six-event MAP and selective-decision labels changed in only 0.05% and 0.10% of blocks. By comparison, replacing the dependence structure with independence produced a mean absolute change of 0.112250, a maximum change of 0.428150, and 3.30% decision disagreement. Selection coverage changed only from 29.15% under the locked

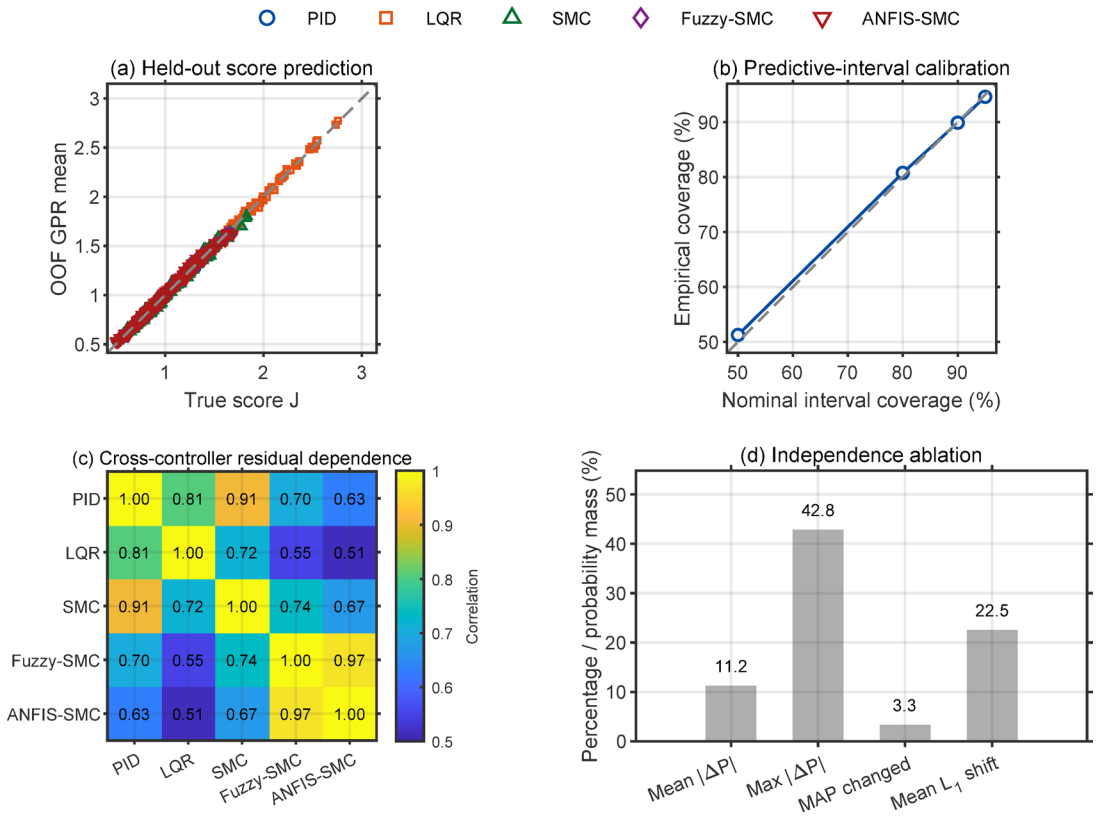


Fig. 5 GPR prediction and cross-controller dependence diagnostics. (**a**) Held-out simulated scores versus out-of-fold GPR means, (**b**) nominal versus empirical predictive-interval coverage, (**c**) cross-controller standardized-residual correlation matrix, and (**d**) effect of replacing the estimated residual correlation structure with independence.

Gaussian model to 29.20% and 29.25% under the two heavier-tail sensitivities, and every selected controller remained the exact simulated minimum in all three dependence-aware calculations. Thus, for the imposed tail forms, the joint decisions were much more sensitive to removing cross-controller dependence than to replacing Gaussian dependence with heavier tails.

Recommendation accuracy and practical separation were evaluated separately

At the 5% margin, the selective rule issued a controller recommendation in 583/2,000 blocks (29.15%) and abstained in 1,417 (70.85%). The recommended controller was the exact simulated minimum in all 583 selected blocks, giving an observed accuracy of 100% with a two-sided 95% Clopper–Pearson interval of 99.37–100%. However, 51 of these 583 selected blocks (8.75%) were in fact non-unique preference cases under the 5% margin. The error in those 51 cases was therefore a declaration of separation rather than identification of the numerical minimum (Table 3).

The abstained set shows the distinction more directly. Among the 1,417 abstained blocks, the raw predicted controller was the exact simulated minimum in 1,092 (77.06%), but if a choice had been forced it still lay inside the 5% good-controller set in 1,416/1,417 blocks (99.93%; 95% CI 99.61–100%). Across all 2,000 blocks, the forced raw prediction was inside the good-controller set in 1,999 cases (99.95%; 95% CI 99.72–100%). Thus, the abstention output is not primarily a mechanism for preventing a poor controller choice in this dataset; it communicates that the simulated evidence often supports several practically near-equivalent alternatives rather than one uniquely separated controller. Consistent with that interpretation, 1,337 of the 1,417 abstained blocks (94.35%) did not support a unique practical preference.

Table 3 Recommendation and empirical good-controller-set performance at the prespecified 5% practical margin (N = 2,000 scenario blocks).

Decision subset	N	Exact-minimum accuracy	Predicted controller in 5% good set	No unique practical preference
Selected	583	100% [99.37–100]	100% [99.37–100]	8.75% (51/583)
Abstained	1417	77.06% [74.79–79.23]	99.93% [99.61–100]*	94.35% (1337/1417)
All blocks	2000	83.75% [82.06–85.34]	99.95% [99.72–100]*	69.40% (1388/2000)

*For abstained/all rows, "predicted controller" denotes the raw-best controller that would have been returned if a choice had been forced. The high good-set rate does not make abstention redundant; abstention reports whether the scores support a unique practical preference, whereas a forced choice only identifies an acceptable controller.

Same-domain replication did not require pooling the two realizations. Training the probabilistic models on the primary 1,000-block realization and testing on the replication gave 93.00% six-event agreement; reversing the direction gave 93.20%. Selection coverage was 28.0% and 29.4%, and the selected controller was the exact simulated minimum in every selected block in both directions. The close agreement supports reproducibility within the prescribed ± 15% scenario domain, but they do not test extrapolation beyond that domain.

Discussion

The results do not identify one controller as universally preferable. Instead, the support for a controller preference depends on how performance is defined and on whether numerical differences exceed a practical relevance threshold. The criterion change at the nominal point produced a large rank reordering even before uncertainty was introduced: LQR moved from first under the tuning objective to fifth under the common evaluation score, while ANFIS-SMC moved from third to first. The mechanism is visible in the metric definitions. The tuning objective rewards transient quantities such as overshoot and settling time together with pendulum error, voltage effort and control roughness, whereas the common score replaces the transient-shaping terms with sustained rotary-arm RMS motion. LQR combined the lowest nominal pendulum RMS with the largest nominal rotary-arm RMS, so adding arm motion as an explicit evaluation quantity penalized a trade-off that the tuning objective did not represent in the same way. This is both a finding and a limitation: the common score was deliberately used to compare the fixed controllers under one engineering evaluation protocol, but none was retuned to minimize it. The resulting values describe the selected portfolio, not each architecture's attainable optimum under that criterion.

Practical distinguishability changes the interpretation of these rankings. At the nominal point, Fuzzy-SMC, SMC and PID were 0.25%, 4.20% and 5.19% above the ANFIS-SMC minimum, whereas LQR was 38.3% higher; correspondingly, the nominal good set changed from two to three to four controllers as the practical margin increased from 2.5% to 5% to 10%. The aggregate CRI summaries led to the same qualitative separation pattern: PID had the numerical minimum in each uncertainty band, but SMC, Fuzzy-SMC and ANFIS-SMC all remained within 3.87% of it, while LQR remained approximately 38% higher. Across the ± 15% domain, 69.40% of true simulated score vectors contained at least two controllers within the 5% good-controller set, and this prevalence reproduced closely in the two independently randomized LHS realizations. Thus the nominal point, the aggregate uncertainty summaries and the scenario-block analysis converge on the same practical conclusion: a numerical rank can be well defined while the leading controllers remain insufficiently separated for a unique engineering preference. The good-controller set exposes this distinction without converting the chosen margin into a formal equivalence test or a statistical significance threshold.

Ranking variability arose from several different sources. Changing the metric set reshuffled the deterministic nominal order; varying the engineering weights redistributed rank-one shares; and paired resampling of the finite scenario set rarely displaced the base-weight CRI winner. These are different mechanisms rather than competing estimates of the same probability. In particular, the 99.95% bootstrap rank-one frequency indicates that the numerical CRI order is reproducible under paired resampling, but the CRI gaps among the four leading designs

remain inside the 5% practical margin; ordering stability therefore does not establish practical separation. The paired significance tests add another distinction between statistical detectability and practical separation: all ten score comparisons were detectable with $N = 2{,}000$ after Holm adjustment, but most scenario blocks still lacked a controller separated by the 5% engineering margin. Increasing sample size can sharpen inference about small score differences, but it cannot make those differences practically important or remove sensitivity to the evaluation criterion and engineering priorities. The added CRI-coefficient sensitivity further separates aggregate rank details from the broader engineering interpretation: the PID and LQR endpoints and the 5% leading-set conclusion were stable for lambda between 0 and 1, whereas a small number of crossovers occurred among SMC, Fuzzy-SMC and ANFIS-SMC. Accordingly, no architecture-level conclusion is based on the exact internal ordering of these near-tied controllers.

The normalization analysis further separates robust conclusions from criterion-dependent details. Realization-specific median scaling produced essentially the same normalization constants, aggregate ordering and ± 15% non-unique-preference prevalence as the pooled-median analysis. The pooled-mean stress test was intentionally more disruptive because the rotary-arm RMS scale increased by 42.19%; even then, PID and LQR remained the numerical CRI endpoints and 67.50% of scenario blocks still lacked a unique 5% preference. However, mean scaling moved the three SMC-family aggregate gaps just beyond the 5% threshold. Accordingly, the manuscript does not treat membership of all four leading aggregate CRIs inside the 5% band as invariant to the normalization statistic; the more general distinction between numerical ordering and frequent scenario-level practical non-separation is better supported across the tested alternatives.

The command-delay sensitivity adds a related implementation-level qualification. A uniform 0–10 ms command delay, inserted after the fixed controller output and before physical voltage saturation, produced no dynamic or terminal-regulation failures in the evaluated ± 15% scenarios, and PID and LQR remained the numerical CRI endpoints. However, the SMC-family ordering changed at intermediate delays, the prevalence of non-unique practical preference reached 74.45% at 8 ms, and at 10 ms the aggregate Fuzzy-SMC and ANFIS-SMC CRI gaps exceeded the 5% margin while the SMC gap remained below it. These results show that practical-separation conclusions can change before outright loss of regulation. They should not be interpreted as a hardware latency specification: the tested delays form a prescribed sampled-data sensitivity envelope, the controllers were not retuned or redesigned for delay, and separate sensor, computation, communication and actuator dynamics were not identified from hardware.

The selective rule mainly reports whether the predicted difference is practically meaningful; it does not materially improve identification of the minimum-score controller. A forced raw prediction remained within the 5% good-controller set in 99.95% of all scenario blocks, including 99.93% of the blocks on which the rule abstained. Abstention is therefore interpreted as a statement about practical separation: in 94.35% of abstained blocks, the simulated scores placed more than one controller inside the margin. Conversely, 51 of the 583 selected blocks (8.75%) were not uniquely separated even though the selected controller was the exact numerical minimum. These are errors in classifying separation, not errors in identifying the minimum-score controller. Because the decision is based on predicted score distributions rather than the unobserved true score vector, the discrepancy may reflect surrogate mean error, marginal calibration, the Gaussian dependence approximation, or a combination of these effects. The available analyses do not isolate a single cause. Its practical value in this dataset is therefore communicative rather than corrective: it flags when the evidence does not justify presenting a single numerically best prediction as a practically separated controller, even though a forced choice will usually remain inside the simulated good-controller set.

The probabilistic results remain conditional on an approximate joint dependence model. Marginal prediction accuracy and interval coverage were close to their nominal targets, but standardized residuals showed modest departures from exact Gaussian shape and the locked model summarizes dependence through a fold-level linear correlation matrix. The post-review tail-form sensitivity provides a targeted robustness check: holding the marginal GPR predictions and fold-specific correlation matrices fixed, Student-t copulas with $\nu = 8$ and $\nu = 4$ changed the mean probability of no unique practical preference by only 0.003052 and 0.005999 on average, and altered

the six-event MAP decision in 0.05% and 0.10% of blocks. In contrast, discarding dependence entirely changed the probability by 0.112250 on average and altered the decision in 3.30% of blocks. Within these tested alternatives, the downstream conclusions were therefore much more sensitive to ignoring cross-controller dependence than to the imposed heavier-tail forms. This analysis does not validate a Student-t residual law: the copulas were stress-test models, not fitted alternatives, and neither they nor the locked Gaussian representation capture arbitrary nonlinear, state-dependent, or scenario-dependent dependence. The positive residual dependence also partly reflects the composition of the controller portfolio, particularly the closely related Fuzzy-SMC and ANFIS-SMC structures. Hardware experiments, identified friction/backlash effects, correlated physical parameter distributions and external-domain testing remain necessary before transferring the reported preference behavior to a physical RIP or another underactuated platform.

Additional physical effects could change both absolute performance and the apparent separation among controllers. The present model uses linear viscous-friction terms and does not represent Coulomb/static friction, speed-, frequency- or temperature-dependent friction, backlash, gearbox compliance, actuator dead zones, encoder quantization, current-driver dynamics, parameter dependence, or non-Gaussian sensor errors. The post-review command-delay study addresses only a uniform sampled-data delay placed between the fixed controller command and the physical saturation block; it does not identify or separately model sensor-acquisition, computation, communication or actuator delays, nor does it establish a safe hardware-delay limit. Recent delay-aware control studies likewise show that delay can materially alter closed-loop design and performance[43–45]. Physical transfer therefore requires identified delay distributions together with friction, backlash, driver and sensing characteristics of the intended hardware.

Generality is also intentionally limited. The investigation concerns one nonlinear rotary inverted pendulum model, one five-controller portfolio, one tuning protocol and one prescribed simulated uncertainty domain. The reported values—including the 69.40% frequency of non-unique practical preference at the 5% margin—are therefore case-study results for the evaluated ± 15% RIP domain and must not be interpreted as generic frequencies for controller comparisons, other plants or other uncertainty models. Additional benchmark systems and hardware experiments would be required to test whether the same preference-reliability patterns generalize beyond this case study.

Conclusion

This case study shows that a stable numerical controller ordering need not imply a uniquely meaningful engineering preference. For the fixed five-controller RIP portfolio and prescribed simulated uncertainty domain, statistical detectability coexisted with frequent practical near-ties, while ranking sensitivity depended on the evaluation criterion, engineering weights, score normalization and implementation delay. The additional revision analyses showed that the main scenario-level non-unique-preference finding was stable under the tested CRI dispersion coefficients, realization-specific median normalizations and heavier-tail dependence representations, but the ordering and separation of near-tied controllers could change under command delay. These results should therefore be read as evidence about preference reliability within this defined benchmark, not as an architecture-level ranking or a general frequency for controller comparisons. Hardware validation and additional benchmarks are needed before broader transfer.

Methods

Simulation design and scenario-block definition

Five fixed controller designs were evaluated on one nonlinear rotary inverted pendulum (RIP) model using the same measurement, estimator and actuator assumptions. Two independently randomized Latin-hypercube sampling (LHS) realizations supplied paired scenario blocks: within each block, all five controllers received the same plant parameters, disturbance realization and measurement-noise sequence. The two 1,000-block realizations were pooled only for same-domain $N=2{,}000$ analyses; cross-realization evaluations kept them separate for A→B and B→A testing. The analysis sequence separates nominal controller selection, uncertainty propagation, common-score evaluation, surrogate modeling and the later practical-margin decision analysis.

All nominal and uncertainty simulations used the same plant initial condition, $x(0) = [0, 0.0994838, 0, 0]^T$, corresponding to $\theta(0)=0$ rad, $\alpha(0)=5.7°$ and zero initial angular velocities. At the first sample, the shared causal differentiator initialized both velocity estimates to zero and stored the current measured angles as its previous-position states. The PID integral state was initialized to zero ($I\alpha(0)=0$), and all controller memory states were reset before each independent objective evaluation or scenario simulation. These initialization settings are now summarized explicitly in Supplementary Table S11.

Nonlinear electromechanical model and common measurement architecture

The RIP consists of a motor-driven horizontal arm with angle θ and an unactuated pendulum with angle α, where $\alpha=0$ is upright. The RIP configuration and coordinate convention are shown in Fig. 1a. The simulation state is defined in Eq. (1), and the Euler–Lagrange dynamics with $q=[\theta,\alpha]^T$ are written in Eq. (2).

$$x = \begin{bmatrix} \theta & \alpha & \dot{\theta} & \dot{\alpha} \end{bmatrix}^T \tag{1}$$

$$M(q)\ddot{q} + c(q,\dot{q}) + D\dot{q} + g(q) = \begin{bmatrix} \tau_m + \tau_d \\ 0 \end{bmatrix} \tag{2}$$

The inertia terms are given in Eqs. (3)–(4), and the velocity-dependent, gravitational, and viscous-damping terms in Eqs. (5)–(7). Symbols and nominal mechanical parameters are listed in Supplementary Table S2.

$$M_{11} = J_a + mL^2 + ml^2\sin^2\alpha \tag{3}$$

$$M_{12} = M_{21} = -mLl\cos\alpha, M_{22} = J_p + ml^2 \tag{4}$$

$$c_1 = 2ml^2\sin\alpha\cos\alpha\dot{\theta}\dot{\alpha} + mLl\sin\alpha\dot{\alpha}^2 \tag{5}$$

$$c_2 = -ml^2\sin\alpha\cos\alpha\dot{\theta}^2 \tag{6}$$

$$g(q) = \begin{bmatrix} 0 \\ -mgl\sin\alpha \end{bmatrix}, D\dot{q} = \begin{bmatrix} b_r\dot{\theta} \\ b_p\dot{\alpha} \end{bmatrix} \tag{7}$$

Motor current is represented quasi-steadily because $La/Ra \approx 6.9\times10^{-5}$ s is much shorter than the 0.002 s simulation step. Eqs (8)–(10) define the motor torque relation, effective motor coefficients, and explicit ±10 V voltage saturation. The saturated voltage drives the nonlinear plant, while external disturbance torque enters the actuated arm channel.

$$\tau_m = K_{eff}\left(V_m - K_{back}\dot{\theta}\right) \tag{8}$$

$$K_{eff} = \frac{\eta_g K_g \eta_m K_t}{R_a}, K_{back} = K_g K_e \tag{9}$$

$$V_m = sat\,(V_{cmd}, -V_{max}, V_{max}) \tag{10}$$

Supplementary Table S2 gives the nominal parameters used consistently for plant simulation, controller synthesis, and uncertainty mapping; they are reference simulation values rather than parameters identified from new hardware experiments.

For reproducibility, the complete nominal parameter set is stated here in addition to Supplementary Table S2: pendulum mass m = 0.127 kg; rotary-arm length L = 0.216 m; pendulum full length Lp = 0.337 m; pendulum centre-of-mass distance l = 0.1685 m; arm/motor inertia Ja = 0.0028 kg·m²; pendulum centre-of-mass inertia Jp = 1.202×10^{-3} kg·m²; arm and pendulum viscous-friction coefficients br = bp = 0.0024 N·m·s/rad; gravitational acceleration g = 9.81 m/s²; armature resistance Ra = 2.6 Ω; armature inductance La = 1.8×10^{-4} H; torque and back-EMF constants Kt = Ke = 7.68×10^{-3} in their respective SI units; gear ratio Kg = 70; motor efficiency ηm = 0.69; gearbox efficiency ηg = 0.90; and actuator voltage limit Vmax = 10 V. These are reference simulation parameters rather than values identified from a new physical RIP unit.

All controllers use the same closed-loop interface shown in Fig. 1b. Measured θ and α are used directly, while θ and $\dot{\alpha}$ are estimated by the same first-order causal differentiator defined below; true angular velocities are not supplied to any controller.

$$\hat{v}_k = a\hat{v}_{k-1} + b\,(q_{m,k} - q_{m,k-1}), a = \frac{\tau_f}{\tau_f + \Delta t}, b = \frac{1}{\tau_f + \Delta t}$$

The estimator uses $\tau f = 0.0125$ s and $\Delta t = 0.002$ s, with zero initial velocity and the current angle as the previous-position state. During noisy evaluations, independent zero-mean Gaussian noise with $\sigma = 0.001$ rad is added to both angle channels before differentiation. The same noise realization, estimator, and ± 10 V actuator saturation are shared by all controllers within each paired scenario.

At each completed control interval, the estimated state is first supplied to the active controller to obtain Vcmd. The common saturation block is then applied once, before motor-torque evaluation, so that Vm = sat(Vcmd) is the voltage entering Eqs. (8)–(9) and the nonlinear plant. For PID, the previous interval's saturation state is also used by the conditional-integration rule in Eq. (13); the remaining controllers contain no integral state and therefore require no integral anti-windup state.

Numerical verification

Model verification addressed analytical and numerical consistency rather than experimental validation. The checks comprised energy conservation, analytical-versus-finite-difference linearization, local nonlinear–linear agreement, RK4 convergence, and closed-loop timestep sensitivity (Table 1).

The verification strategy also situates the formulation against established Furuta-model practice. The nonlinear model follows an Euler–Lagrange formulation and upright-equilibrium Jacobian linearization; the checks below address analytical consistency and numerical convergence. They do not constitute physical validation.

In the unforced frictionless model, the maximum relative energy drift over 20 s was 1.608×10^{-11}. Maximum discrepancies between the analytical and finite-difference linearization were 1.5×10^{-12} for A, 3.6×10^{-15} for B, and 7.1×10^{-14} for matched eigenvalues. The nonlinear–linear trajectory discrepancy decreased with perturbation size; at 0.25°, RMS differences were 0.00163° for θ and 0.00195° for α. RK4 convergence orders of 3.987–3.997 were consistent with fourth-order behavior.

A demanding Fuzzy-SMC condition with $\alpha\,(0) = 0.5$ rad was used to compare $\Delta t = 0.002$ and 0.001 s. The smaller step changed $\alpha\,RMS$, $\theta\,RMS$, and applied-voltage RMS by 0.408%, 0.047%, and 0.133%, respectively, with unchanged 0.133% saturation fraction; $\Delta t = 0.002$ s was therefore used in the comparison.

Fixed controller portfolio and nominal tuning

The candidate set comprises PID, LQR, SMC, Fuzzy-SMC, and ANFIS-SMC. All designs use the estimated state in Eq. (11), are synthesized or tuned on the nominal plant, and are kept unchanged during uncertainty propagation; only the plant and prescribed scenario inputs vary.

$$\hat{x}_k = \left[\theta_{m,k}, \alpha_{m,k}, \hat{\dot{\theta}}_k, \hat{\dot{\alpha}}_k\right]^T \tag{11}$$

PID with conditional-integration anti-windup

The PID command in Eq. (12) regulates α with additional proportional feedback from θ and $\dot{\theta}$. Its integral state is updated once per completed control interval according to Eq. (13); integration is suspended when the previous command is saturated and the accumulated error would drive it farther into saturation. No additional anti-windup gain is introduced.

$$V_{cmd} = -\left(K_p\alpha_m + K_i I_\alpha + K_d\hat{\dot{\alpha}} + K_\theta\theta_m + K_{\dot{\theta}}\hat{\dot{\theta}}\right) \tag{12}$$

$$I_{\alpha,k} = \begin{cases} I_{\alpha,k-1}, & F_{k-1}, \\ I_{\alpha,k-1} + \alpha_{m,k-1}\Delta t, & otherwise. \end{cases} \tag{13}$$

Linear quadratic regulator

LQR is synthesized from the continuous-time upright linearization and applies the state-feedback law in Eq. (14). No integral state is added. The weighting matrices are parameterized and selected under the common tuning protocol in the controller-tuning procedure.

For completeness, the upright-equilibrium matrices used for LQR synthesis are stated explicitly. With $a = Ja + mL^2$, $d = Jp + ml^2$, $b = -mLl$, $h = br + Kback$, and $\Delta = ad - b^2$, the Jacobian of Eqs. (2)–(10) for the state order $x = [\theta, \alpha, \dot{\theta}, \dot{\alpha}]^T$ gives the following continuous-time matrices:

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & -\frac{bmgl}{\Delta} & -\frac{dh}{\Delta} & \frac{bb_p}{\Delta} \\ 0 & \frac{amgl}{\Delta} & \frac{bh}{\Delta} & -\frac{ab_p}{\Delta} \end{bmatrix}, B = \begin{bmatrix} 0 \\ 0 \\ \frac{dK_{eff}}{\Delta} \\ -\frac{bK_{eff}}{\Delta} \end{bmatrix}.$$

The displayed matrices make the local linear model transparent and also define the A and B used in the finite-difference verification of Numerical verification. The linear model is used only for LQR synthesis; every nominal and uncertain closed-loop performance result is generated with the nonlinear plant.

$$V_{cmd} = -K\hat{x}, K = lqr(A, B, Q, R) \tag{14}$$

Sliding-mode controller family

SMC, Fuzzy-SMC, and ANFIS-SMC share the same nonlinear voltage-based structure. The sliding variable is defined in Eq. (15) using C obtained from the nominal upright linearization with $Qs = diag(1{,}100{,}0.5{,}5)$ and $Rs = 1$. The nominal equivalent voltage and boundary-layer reaching law are given in Eqs. (16)–(17). Over the verified region $|\alpha| \leq 0.60$ rad and $|\dot{\theta}|, |\dot{\alpha}| \leq 8$ rad/s, $\gamma = Cg(x; p0)$ remained positive. Equation (18) gives the corresponding local nominal sliding dynamics; this is not a global stability claim for the estimator-active saturated system.

$$\dot{x} = f(x; p_0) + g(x; p_0)V, s = C\hat{x}, \gamma = Cg(\hat{x}; p_0) \tag{15}$$

$$V_{eq} = -\frac{Cf(\hat{x}; p_0)}{Cg(\hat{x}; p_0)} \tag{16}$$

$$V_{cmd} = V_{eq} - \eta sat\left(\frac{s}{\phi}\right) - ks \tag{17}$$

$$\dot{s} = -\gamma\left[\eta sat\left(\frac{s}{\phi}\right) + ks\right) \tag{18}$$

Fuzzy-SMC replaces the fixed reaching gain η with a Mamdani scheduler. The normalized sliding-surface magnitude and causal surface-rate estimate are defined in Eq. (19), and Eq. (20) maps them to η. Small, Medium, and Large triangular membership functions are used for both inputs and the normalized output; min–max inference and centroid defuzzification are applied. Supplementary Table S3 gives the 3×3 rule base.

$$z_s = \min\left(\frac{|s)}{\phi}, 1\right), z_r = \min\left(\frac{\tau_f|\hat{s})}{\phi}, 1\right) \tag{19}$$

$$\eta = \eta_{\max} y_F(z_s, z_r), 0 \le y_F \le 1 \tag{20}$$

ANFIS-SMC uses the same SMC structure and normalized inputs, replacing only the Mamdani scheduling surface by the ANFIS mapping in Eq. (21). The selected model has five Gaussian premise functions per input, 25 rules, and linear Sugeno consequents. It approximates the Fuzzy-SMC scheduling surface rather than plant dynamics or the complete controller; $\eta\, max$ is shared with Fuzzy-SMC, while k and φ are tuned separately.

$$\eta = \eta_{max} y_A(z_s, z_r), 0 \le y_A \le 1 \tag{21}$$

Controller tuning and ANFIS selection

All controllers were tuned before uncertainty propagation using the same nominal plant, estimator, actuator saturation, $\alpha(0) = 5.7^\circ$, $T = 4$ s, $\Delta t = 0.002$ s, and $\tau f = 0.0125$ s. Measurement noise and external disturbance were disabled during tuning, and controller internal states were reset between objective evaluations.

The common five-term objective in Eq. (22) combines pendulum RMS error, applied-voltage RMS, overshoot, settling time, and chattering with controller-independent normalizers. Candidates that failed, exceeded $|\alpha| = 1.5$ rad, or did not remain within $|\alpha| < 0.05$ rad during the final 0.40 s received $J = 10^6$. This objective is used only for nominal parameter selection and is distinct from the downstream scenario score in the common-score definition.

For the nominal tuning metrics, overshoot was the positive excess of the maximum absolute pendulum angle over its initial absolute value; settling time was the last sample time at which the absolute pendulum angle exceeded 0.02 rad (zero if no such sample occurred); and the tuning chattering index was the mean absolute applied-voltage increment normalized by the maximum absolute applied voltage, with a 10^{-6} V denominator floor.

$$J = 0.40\frac{e_{RMS}}{0.012} + 0.20\frac{u_{RMS}}{0.3} + 0.10\frac{OS}{0.03} + 0.15\frac{t_s}{2.0} + 0.15\frac{CI}{0.001} \tag{22}$$

Bounded genetic-algorithm searches used 150 objective evaluations per independently tuned parameter per seed and five pseudorandom seeds. Supplementary Table S4 reports the controller-specific variables, bounds, and resulting evaluation budgets. LQR searches only $q2$ and $q4$ in logarithmic coordinates; the SMC-family bounds include $k \ge 0$ and $\varphi \le 5$.

For Fuzzy-SMC, $\eta\, max \le 10/0.87$ keeps the scheduled reaching gain within the common 10 V actuator range.

The search dimension and therefore the total objective-evaluation budget differed across controller parameterizations. Per seed, PID optimized five variables and used 750 objective evaluations; LQR optimized two variables and used 300; SMC and Fuzzy-SMC each optimized three variables and used 450; and ANFIS-SMC

optimized two post-training variables and used 300, with five seeds used for every controller. The corresponding totals across five seeds were 3,750, 1,500, 2,250, 2,250 and 1,500 objective evaluations, respectively. The explicit bounds were PID: Kp∈[10,200], Ki∈[0,5], Kd∈[0.5,20], Kθ∈[−5,0], Kθ̇∈[−5,0]; LQR: log10(q2)∈[0,4] and log10(q4)∈[−3,2] with q1 = q3 = 10^{-6} and $R = 1$ fixed; SMC: η∈[0.01,10], k∈[0,3], ϕ∈[0.05,5]; Fuzzy-SMC: ηmax∈[0.01,10/0.87], k∈[0,3], ϕ∈[0.05,5]; and ANFIS-SMC: k∈[0,3], ϕ∈[0.05,5], with ηmax inherited from Fuzzy-SMC. The common protocol therefore equalized the evaluation allocation per independently optimized parameter and the number of random seeds, but it did not impose an equal total search budget or equal structural flexibility across architectures. This distinction is important: the reported rankings compare the selected fixed controller instances, and differences in parameterization or tuning flexibility may contribute to their numerical ordering. Supplementary Tables S4, S5 and S12 provide the detailed search specification, selected parameters and fairness summary.

ANFIS architecture selection compared 3 × 3, 4 × 4, and 5 × 5 Gaussian-premise models with linear or constant Sugeno consequents on the common deterministic Fuzzy-SMC scheduling surface. A 5 × 5 model with linear consequents and 25 rules was selected before the reserved holdout was examined.

The reference scheduling surface was sampled on a deterministic 51 × 51 grid over the normalized unit-square input domain, yielding 2,601 scheduler targets. A deterministic spatial split produced 1,822 training, 389 validation, and 390 reserved test samples. Candidate FISs were initialized by grid partitioning and trained for up to 100 epochs with MATLAB's hybrid ANFIS optimization; validation-selected checkpoints were compared on an independent 80 × 80 one-third-cell-offset grid, and the 5 × 5 linear-consequent model was selected before the 390-point test holdout was evaluated.

The selected ANFIS was evaluated on the 390-point holdout; only k and φ were subsequently tuned, while $\eta\,max$ was inherited from Fuzzy-SMC. Supplementary Table S5 reports the selected controller parameters.

For each controller, the lowest independently re-evaluated objective across five seeds defined the parameter set used in all subsequent analyses. The nominal objective order was LQR, PID, ANFIS-SMC, Fuzzy-SMC, and SMC.

Nominal objective-weight sensitivity was evaluated without retuning. The selected controllers were rescored using 5,000 local multiplicative perturbations of the base weights, 5,000 Dirichlet(1,1,1,1,1) vectors, and deterministic equal-weight or 70%-emphasis cases. The resulting rank changes measure preference sensitivity of the nominal comparison; they do not represent re-optimization under alternative objectives.

Paired uncertainty design and Monte Carlo propagation

Uncertainty was propagated with two independently randomized $N = 1000$ LHS designs in nine dimensions: four plant quantities and five disturbance descriptors. The same unit-hypercube quantiles were mapped to matched ± 10%, ± 15%, and ± 20% plant-uncertainty bands so that controller and band comparisons used common scenario realizations.

Equation (23) maps the first four dimensions to pendulum mass m, pendulum length Lp, motor torque constant Kt, and arm viscous friction br. Dependent quantities l, Jp, and $Keff$ are recomputed from each sampled plant; Supplementary Table S6 gives the complete mapping.

$$p_{i,j}^{(b)} = p_{0,j}\left[1 + b\left(2u_{i,j} - 1\right)\right) \tag{23}$$

The remaining five LHS dimensions define the composite arm disturbance in Eq. (24): sinusoidal amplitude and frequency, bias torque, and the amplitude and onset of a 0.05 s rectangular impulse. Measurement noise is separate from the LHS because its standard deviation is fixed at 0.001 rad; Eq. (25) defines the two noisy angle channels. For each scenario, the same disturbance and two-channel noise realization are used across all five controllers and three uncertainty bands.

$$\tau_{d,i}(t) = A_{\sin,i}\sin(\omega_i t) + \tau_{bias,i} + A_{imp,i}\left[H\left(t - t_{imp,i}\right) - H\left(t - t_{imp,i} - 0.05\right)\right) \quad (24)$$

$$q_{m,k} = q_k + \epsilon_k, \epsilon_k \sim N\left(0, \sigma_q^2 I_2\right), \sigma_q = 0.001\text{rad} \quad (25)$$

LHS realizations A and B use independent random streams (LHS seeds 42001 and 52001; noise seed bases 240000 and 340000). Each realization therefore contains 15,000 controller–band simulations. Their concatenation gives $N = 2000$ distinct paired scenario blocks; it is not treated as a single formal $N = 2000$ LHS design.

The ±15% band was prespecified for probabilistic score modeling; ±10% and ±20% were used for uncertainty-band sensitivity. This avoids treating the three matched band realizations from one scenario as independent learning rows.

These three bands are sensitivity envelopes rather than asserted universal manufacturing tolerances. Because the nominal values are reference simulation parameters rather than parameters identified from one physical unit, application-specific bounds should ultimately be obtained from component specifications, identification uncertainty, and operating data for the intended hardware.

All Monte Carlo runs used $T = 5$ s, $\Delta t = 0.002$ s, $\tau f = 0.0125$ s, and $[\theta, \alpha, \dot{\theta}, \dot{\alpha}]^T = [0, 5.7^\circ, 0, 0]^T$. A run failed if $|\alpha| > 1.5$ rad; successful regulation required completion of the horizon and $max|\alpha| < 0.10$ rad over the final 0.5 s.

Post-review command-delay sensitivity

To assess one implementation nonideality without changing the submitted controller designs, a post-review command-delay sensitivity analysis was performed on the pooled ± 15% scenario domain. The controller command was delayed by d = 0, 1, 2, 3, 4 or 5 fixed simulation samples, corresponding to 0, 2, 4, 6, 8 or 10 ms at Δt = 0.002 s. The delay was inserted after evaluation of the unchanged fixed controller and before the common ± 10 V physical saturation block; command prehistory was set to zero. No controller was retuned, no delay compensation was introduced, and the PID anti-windup logic was left unchanged. The archived 0 ms results served as the baseline, while the five nonzero delays required 50,000 new closed-loop runs (2,000 scenarios × 5 controllers × 5 delays). Before these runs were interpreted, the delay-capable simulator was required to reproduce six zero-delay sentinel scenarios across all five controllers; the maximum metric difference was $8.44 \times 10^{-}15$. Success rate, CRI and 5% good-controller-set summaries were then evaluated at each delay. Because every controller remained successful in every tested block, the practical-separation calculations used the full 2,000-scenario paired set at all delay values. This analysis is a prescribed command-delay sensitivity envelope rather than a measurement of hardware latency.

Performance metrics, common score and practical margin

Four continuous metrics represent pendulum regulation, rotary-arm motion, actuator effort, and control roughness: α_{RMS}, θ_{RMS}, applied-voltage RMS, and the dimensionless chattering index in Eqs. (26)–(29). Success and saturation fraction are feasibility indicators rather than score terms.

$$\alpha_{RMS} = \sqrt{\frac{1}{N_x}\sum_{k=1}^{N_x}\alpha_k^2}. \quad (26)$$

$$\theta_{RMS} = \sqrt{\frac{1}{N_x}\sum_{k=1}^{N_x}\theta_k^2}. \quad (27)$$

$$U_{RMS} = \sqrt{\frac{1}{N_u}\sum_{k=1}^{N_u}V_{m,k}^2}. \quad (28)$$

$$CI = \frac{\frac{1}{N_u-1}\sum_{k=2}^{N_u}|V_{m,k}-V_{m,k-1})}{max\left(\underset{k}{max}|V_{m,k}),10^{-6}\right)}. \tag{29}$$

Each metric is normalized by the pooled median over the completed 30,000 Monte Carlo records. Supplementary Table S7 reports these scales and the base weights, which are used unchanged in the later regression and cross-realization analyses. Pooled-median scaling was selected to place metrics with different physical units on comparable dimensionless scales while reducing sensitivity to extreme response values relative to mean-based scaling; the same fixed scales were then used for all controllers and downstream analyses.

A post-review normalization sensitivity analysis then recomputed the same fixed score using four alternative sets of scaling constants: the locked pooled median, the median estimated from LHS realization A only, the median estimated from realization B only, and the pooled mean as a secondary stress test. The controller responses, four metric weights, CRI coefficient (0.5), and 5% practical margin were held fixed, and no controller was retuned. The realization-specific medians test whether the reported score depends materially on which independent LHS design supplies the normalization constants, whereas pooled-mean scaling intentionally tests a less robust summary statistic that is more sensitive to skewed or extreme responses. This sensitivity analysis was applied to the archived Monte Carlo records only; it did not refit the downstream GPR models.

Equation (30) combines the normalized metrics into the preference-dependent scenario score J, where lower is better. This downstream score differs from the nominal tuning objective in Eq. (22) in both metric set and normalization.

$$J_{i,c}^{(b)} = 0.45\frac{\alpha_{RMS,i,c}^{(b)}}{s_\alpha} + 0.20\frac{\theta_{RMS,i,c}^{(b)}}{s_\theta} + 0.20\frac{U_{RMS,i,c}^{(b)}}{s_U} + 0.15\frac{CI_{i,c}^{(b)}}{s_{CI}}. \tag{30}$$

For descriptive controller summaries, Eq. (31) defines the Controller Robustness Index (CRI) as median(J)+0.5IQR(J). The regression models predict individual scenario scores J, not CRI. For the matched three-band summary, the three normalized metric vectors are first averaged within each scenario and controller before J is evaluated, preserving the $N=2000$ paired scenario-block structure. The coefficient 0.5 was fixed to apply a moderate dispersion penalty: the median represents central performance, whereas one-half of the IQR penalizes broad scenario-to-scenario variability without weighting dispersion as strongly as the full IQR. Because this coefficient is an engineering summary choice rather than a uniquely determined constant, the major revision evaluates the stability of the aggregate conclusions over alternative IQR coefficients and reports the results in the revised Supplementary Information. To test the dependence of this descriptive summary on the chosen dispersion penalty, a post-review sensitivity analysis recomputed CRI_lambda = median(J) + lambda IQR(J) for lambda = 0, 0.25, 0.50, 0.75, and 1.00 from the archived scenario scores, without controller retuning, resimulation, or any change to the scenario-score definition. The lambda = 0.50 calculation was required to reproduce the locked CRI values before alternative coefficients were interpreted.

$$\mathrm{CRI}_c = median_i\left(J_{i,c}\right) + 0.5IQR_i\left(J_{i,c}\right). \tag{31}$$

The practical-margin analysis uses the same scenario score. For scenario block i, the good-controller set contains every controller whose score is within the relative margin δ of the block minimum, with the primary value fixed at δ = 0.05 before the decision analysis. A block is classified as having no unique practical preference when the good-controller set contains at least two controllers. The 2.5%, 7.5% and 10% margins are sensitivity analyses; δ is an engineering relevance threshold and not a formal equivalence-test boundary. The 5% primary margin was fixed before the downstream probabilistic decision analysis as an interpretable relative engineering tolerance: score differences smaller than this threshold were treated as insufficient, by themselves, to justify a unique practical preference. It is not claimed to be a universal physical tolerance for RIP control. The pre-existing 2.5%, 7.5% and 10% sensitivity analyses are therefore used to show explicitly which conclusions depend on this choice.

CRI preference sensitivity uses two sets of 5,000 score-weight vectors (seed 94101): $w\alpha \sim U(0.30,0.60)$ with the remaining mass distributed by Dirichlet(1,1,1), and unrestricted Dirichlet(1,1,1,1) weights. Sampling uncertainty is assessed with 2,000 paired scenario-block bootstrap replicates (seed 95103), using common resampled indices across controllers and bands to obtain percentile 95% intervals and rank frequencies.

GPR score models and joint residual dependence

For each controller, a separate GPR model predicts the ±15% scenario score J from the nine descriptors in Supplementary Table S6, using the combined $N = 2000$ scenarios from LHS realizations A and B. Equation (32) defines the marginal predictive distribution and Eq. (33) the ARD squared-exponential covariance with a constant basis.

Predictor standardization, hyperparameter estimation, and the Gaussian response-noise standard deviation are learned within each training set only; the noise standard deviation is bounded below by $max[10^{-8}, 10^{-6}SD(Jtrain)]$. Supplementary Table S8 summarizes the model specification. This surrogate noise term is not interpreted as a physical aleatoric/epistemic decomposition.

$$J_c \mid x, D_{tr} \sim N\left(\mu_c(x), \sigma_c^2(x)\right) \tag{32}$$

$$k_c(x, x') = \sigma_{f,c}^2 exp\left[-\frac{1}{2}\sum_{j=1}^{9}\frac{(x_j - x'_j)^2}{\ell_{c,j}^2}\right) \tag{33}$$

Marginal performance is evaluated with five realization-balanced OOF folds, each holding out 200 scenarios from LHS realization A and 200 from LHS realization B. Reported metrics include RMSE, MAE,, RMSE/IQR, and realization-specific 95% interval coverage, width, and negative log predictive density (NLPD). The resulting controller-level regression and calibration diagnostics are summarized in Supplementary Table S18.

Predictive uncertainty was assessed with central 50%, 80%, 90%, and 95% intervals, standardized residuals [Eq. (34)], NLPD [Eq. (35)], and Gaussian CRPS [Eq. (36)]. Coverage is interpreted together with interval width and proper scores; no post hoc variance scaling or conformal adjustment is applied, and no coverage threshold is used as a calibration certificate.

$$z_{i,c} = \frac{J_{i,c} - \mu_c(x_i)}{\sigma_c(x_i)} \tag{34}$$

$$NLPD_{i,c} = \frac{1}{2}ln\left(2\pi\,\sigma_c^2(x_i)\right) + \frac{\left[J_{i,c} - \mu_c(x_i)\right)^2}{2\sigma_c^2(x_i)} \tag{35}$$

$$CRPS_{i,c} = \sigma_c(x_i)\left[z_{i,c}\left(2\Phi\left(z_{i,c}\right) - 1\right) + 2\phi\left(z_{i,c}\right) - \frac{1}{\sqrt{\pi}}\right) \tag{36}$$

As a deterministic diagnostic, OOF predictive means are also assessed for minimum-score top-one/top-two recovery, score regret, and agreement with the same 5% six-event labeling rule used later. These diagnostics are distinct from the joint probabilistic recommendation.

Standardized out-of-fold residuals were examined using their mean, standard deviation, skewness, excess kurtosis and Q–Q diagnostics. Exact Gaussianity was not used as an acceptance criterion: the joint probability model is explicitly treated as a Gaussian approximation, and its adequacy is assessed together with predictive-interval coverage, proper scores and the dependence-sensitivity analysis.

Within each outer fold, cross-controller dependence is estimated only from its 1600 training scenarios. Five inner OOF fits produce standardized residuals for the five controller scores, from which the correlation matrix Rf is calculated by Eq. (37). The matrix is symmetrized, renormalized to unit diagonal, and assigned an eigenvalue floor of 10^{-8} before Cholesky factorization; no outer-test score contributes to this estimate.

$$\hat{R}^{(f)} = Corr\left(Z_{OOF}^{(f)}\right) \tag{37}$$

For each held-out scenario, Eq. (38) combines the five marginal GPR means and standard deviations with the training-derived Cholesky factor to form a correlation-aware Gaussian approximation of the joint score distribution. The construction preserves the marginal Gaussian predictions but is not an exact multivariate generative model. It assumes that standardized-residual dependence is adequately summarized by an approximately scenario-invariant linear correlation within each outer fold and that Gaussian dependence is suitable for the downstream probability calculations. In particular, this construction does not model scenario-dependent correlation, nonlinear copula structure, or tail-specific dependence; such features could alter near-tie event probabilities when controller scores become jointly extreme. The revision therefore treats the correlation model as a pragmatic dependence approximation rather than a physical law and adds a targeted dependence-form sensitivity analysis in the Supplementary Information.

$$\tilde{J}_i^{(m)} = \mu_i + diag\left(\sigma_i\right) L_f \varepsilon^{(m)}, \varepsilon^{(m)} \sim N\left(0, I_5\right), L_f L_f^T = \hat{R}^{(f)} \tag{38}$$

Event probabilities were estimated with 20,000 antithetic Gaussian draws per scenario. Draw-count and independent-seed sensitivity assess numerical stability; an identity-correlation calculation isolates the effect of ignoring residual dependence. Supplementary Table S9 summarizes these settings. The fixed random-stream definitions used in this and the preceding analyses are consolidated in Supplementary Table S17.

To probe tail-form sensitivity without changing the marginal GPR predictions, a post-review analysis retained each held-out scenario's GPR mean and standard deviation and each outer fold's training-only correlation matrix, then replaced the Gaussian copula with Student-t copulas with $\nu = 8$ and $\nu = 4$[46]. Gaussian marginal predictions were restored by probability-integral transformation, so the sensitivity altered the joint tail dependence while preserving the fitted univariate predictive distributions. The same 5% practical margin and 20,000 integration draws per scenario were used. These Student-t copulas were deliberately imposed stress-test models rather than fitted residual laws; the calculation therefore probes heavier joint tails but does not establish robustness to arbitrary nonlinear or scenario-dependent dependence (Supplementary Table S16 and Supplementary Fig. S16).

Decision rule, cross-realization evaluation and deterministic benchmark

For each joint score draw, Eq. (39) defines the relative separation between the two smallest scores. A non-unique preference event occurs when this separation is below the practical margin δ; the primary $\delta = 0.05$ is prespecified and is not a formal equivalence test. Equation (40) defines each controller's minimum-score probability, the probability of a controller-specific unique preference, and the probability of no unique preference; the five controller-specific events and the non-unique preference event are mutually exclusive and sum to one.

$$g^{(m)} = \frac{\tilde{J}_{(2)}^{(m)} - \tilde{J}_{(1)}^{(m)}}{\max\left(|\tilde{J}_{(1)}^{(m)}|, \int\right)} \tag{39}$$

$$\begin{aligned} P_c^{min} &= Pr\left(c = arg\min_j \tilde{J}_j\right), \\ P_c^{clear} &= Pr\left(c = arg\min_j \tilde{J}_j, g \geq \delta\right), P_{NW} = Pr\left(g < \delta\right) \end{aligned} \tag{40}$$

Equation (41) defines the implemented two-stage recommendation. If the non-unique preference event has the largest six-event probability, the model abstains; otherwise the recommended controller is the one with the largest minimum-score probability. No additional confidence threshold is imposed.

$$\hat{r}_i = \begin{cases} NW, & P_{i,NW} > \max_c P_{i,c}^{clear} \\ \arg\max_c P_{i,c}^{\min}, & otherwise. \end{cases} \quad (41)$$

For decision-level evaluation, the six-event MAP label is defined as the highest-probability event among the five controller-specific unique preference events and the non-unique preference event; this diagnostic is distinct from the two-stage selective rule in Eq. (41).

The six-event vector is assessed with multiclass Brier score and log loss, while the binary probability of no unique preference uses Brier score, log loss, and 10 equal-width reliability bins with expected calibration error. Log-loss probabilities are floored at $0.5/Ndraws$, and six-event entropy is normalized by log(6). Because some controller-specific unique preference events are rare or absent, calibration claims are limited to the aggregate six-event and binary diagnostics for no unique preference.

Sensitivity is evaluated at practical margins of 2.5%, 7.5%, and 10%, at 5,000/10,000/20,000 integration draws, and with an independent integration seed. These calculations modify only the probability calculation.

For the empirical good-selection check, the predicted controller is additionally checked against the true 5% good-controller set computed from the simulated score vector. These frequencies describe empirical behavior in the evaluated sample and are not probability-of-correct-selection or probability-of-good-selection guarantees in the ranking-and-selection sense.

Same-domain cross-realization assessment trains the five GPR models and the residual-dependence estimate on one $N = 1000$ LHS realization and evaluates them on the other, then reverses the roles. Predictors, $\pm 15\%$ domain, score definition, 5% practical margin, and 20,000-draw integration remain unchanged. This assessment is conditional on the fixed global score definition in the common-score definition; the normalization constants are not re-estimated within either cross-realization direction.

Dependence is estimated by five-fold OOF predictions within the training realization before the marginal GPR models are refitted on that realization. The independent realization is then used for predictive accuracy, interval coverage, proper scores, event probabilities, reliability, regret, and selective recommendation. This tests independent-design reproducibility within the prescribed domain; it is not external, hardware, or out-of-distribution validation.

After evaluation, five GPR models are fitted to all $N = 2000$ scenarios; the associated dependence matrix is estimated from full OOF standardized residuals rather than in-sample errors. Predictor ranges are stored to restrict use to the observed feature domain unless separate validation supports extrapolation.

A deterministic RBF-SVR model is fitted separately for each controller using the same nine predictors and score response. Each model uses MATLAB fitrsvm with a Gaussian kernel, automatic kernel-scale selection, and predictor standardization; BoxConstraint and epsilon are left at the function defaults. The same realization-balanced OOF partitions and the two cross-realization directions are used, with no hyperparameter optimization on held-out folds; the benchmark provides point predictions only.

GPR means and RBF-SVR predictions are compared on identical held-out scenarios using joint RMSE/MAE, minimum-score top-one/top-two recovery, regret, 5% six-event agreement, F1 for the non-unique preference class, and balanced accuracy. Paired differences use 5,000 scenario-level bootstrap replicates. The benchmark separates point-prediction quality from the additional predictive distributions, joint event probabilities based on the estimated residual dependence, and abstention provided by the probabilistic method.

Statistical analysis and reproducibility

Statistical comparisons use the $N = 2000$ matched scenario blocks as the analysis units, with controller comparisons paired within scenario. Friedman testing with Kendall's W is followed by two-sided Wilcoxon signed-rank tests for all controller pairs with Holm correction; median paired differences and matched-pairs rank-biserial correlations quantify effect size.

The family-wise error rate for the ten pairwise Wilcoxon comparisons was controlled at $\alpha = 0.05$ using Holm adjustment. All Wilcoxon tests were two-sided. Actual asymptotic P values, N values and effect sizes for each uncertainty context are reported in Supplementary Table S1; the Friedman P values were recomputed at high precision where the original double-precision output underflowed to zero. Friedman and Wilcoxon procedures are non-parametric and no normality assumption is invoked for those controller comparisons. CRI sampling uncertainty uses 2,000 paired scenario-block bootstrap replicates with common resampled indices across controllers and bands.

The numerical analysis was performed in MATLAB R2023b Update 7 (23.2.0.2515942). Random seeds for controller tuning, LHS generation, measurement-noise generation, weight sensitivity, bootstrap resampling, GPR fitting and probability integration were fixed. The principal seeds and stream definitions are listed in Supplementary Table S17 and the complete computational records are stored with the archived analysis outputs. The complete MATLAB code and data package, including the four sensitivity analyses added during revision, is publicly archived on Zenodo[47].

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-76797-y.

Author contributions Y.C.: conceptualization, methodology, software, formal analysis, investigation, data curation, visualization, writing—original draft, writing—review and editing. L.Ö.: conceptualization, methodology, supervision, writing—review and editing. M.A.: conceptualization, methodology, supervision, writing—review and editing. All authors reviewed and approved the manuscript.

Data availability The data, MATLAB code, fitted model files, figure-generation scripts, required archived inputs and stored outputs supporting the revised study are publicly available in the updated Zenodo record cited as reference[47](https://doi.org/10.5281/zenodo.23263340). The public package includes the complete materials supporting the four analyses added during revision: CRI dispersion-coefficient sensitivity, normalization sensitivity, command-delay sensitivity and dependence/tail-form sensitivity.

Code availability All bespoke MATLAB code used for nonlinear RIP simulation, controller tuning and verification, uncertainty propagation, score analysis, GPR and dependence calculations, figure generation, and the four revision sensitivity analyses is publicly available in the updated Zenodo record[47](https://doi.org/10.5281/zenodo.23263340), together with the corresponding stored outputs required to reproduce the revised results.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Furuta, K., Yamakita, M. & Kobayashi, S. Swing-up control of inverted pendulum using pseudo-state feedback. *Proc. Inst. Mech. Eng. Part. I J. Syst. Control Eng.* **206**, 263–269. https://doi.org/10.1243/PIME_PROC_1992_206_341_02 (1992).
2. Hamza, M. F. et al. Current development on using rotary inverted pendulum as a benchmark for testing linear and nonlinear control algorithms. *Mech. Syst. Signal Process.* **116**, 347–369. https://doi.org/10.1016/j.ymssp.2018.06.054 (2019).

3. Cazzolato, B. S. & Prime, Z. On the dynamics of the Furuta pendulum. *J. Control Sci. Eng.* 528341. https://doi.org/10.1155/2011/528341 (2011).
4. Nguyen, T. V. A., Dao, Q. T. & Bui, N. T. Optimized fuzzy logic and sliding mode control for stability and disturbance rejection in rotary inverted pendulum. *Sci. Rep.* **14**, 31116. https://doi.org/10.1038/s41598-024-82471-y (2024).
5. Soydemir, M. U., Şahin, S., Kocaoğlu, A., Bulucu, P. & Güzeliş, C. Online learning of stable robust adaptive controllers design based on data-dependent feedback linearization with application to rotary inverted pendulum. *Neural Comput. Appl.* **36**, 10881–10896. https://doi.org/10.1007/s00521-024-09621-1 (2024).
6. Ben Hazem, Z., Fotuhi, M. J. & Bingül, Z. Development of a fuzzy-LQR and fuzzy-LQG stability control for a double link rotary inverted pendulum. *J. Frankl. Inst.* **357**, 10529–10556. https://doi.org/10.1016/j.jfranklin.2020.08.030 (2020).
7. El-Sousy, F. F. M., Alattas, K. A., Mofid, O., Mobayen, S. & Fekih, A. Robust adaptive super-twisting sliding mode stability control of underactuated rotational inverted pendulum with experimental validation. *IEEE Access* **10**, 100857–100866. https://doi.org/10.1109/ACCESS.2022.3208412 (2022).
8. Nagarajan, A. & Victoire, A. A. Optimization reinforced PID-sliding mode controller for rotary inverted pendulum. *IEEE Access* **11**, 24420–24430. https://doi.org/10.1109/ACCESS.2023.3254591 (2023).
9. Pham, D. B., Dao, Q. T., Bui, N. T. & Nguyen, T. V. A. Robust-optimal control of rotary inverted pendulum control through fuzzy descriptor-based techniques. *Sci. Rep.* **14**, 5593. https://doi.org/10.1038/s41598-024-56202-2 (2024).
10. Gupta, N. & Dewan, L. Adaptive neural network-based sliding mode control of rotary inverted pendulum system. *J. Control Decis.* **12**, 819–828. https://doi.org/10.1080/23307706.2024.2310666 (2025).
11. Shalbafian, A. & Amiri, F. Nonlinear control design for stabilization and tracking of a rotary inverted pendulum. *Sci. Rep.* https://doi.org/10.1038/s41598-026-57277-9 (2026).
12. Chawla, I. & Singla, A. Real-time control of a rotary inverted pendulum using robust LQR-based ANFIS controller. *Int. J. Nonlinear Sci. Numer. Simul.* **19**, 379–389. https://doi.org/10.1515/ijnsns-2017-0139 (2018).
13. Nghi, H. V., Nhien, D. P. & Ba, D. X. A LQR neural network control approach for fast stabilizing rotary inverted pendulums. *Int. J. Precis. Eng. Manuf.* **23**, 45–56. https://doi.org/10.1007/s12541-021-00606-x (2022).
14. Bajodah, A. H. & Ansari, U. Rotary inverted pendulum control using robust generalized dynamic inversion with adaptive neural estimation. *IFAC-PapersOnLine* **56**, 10570–10575. https://doi.org/10.1016/j.ifacol.2023.10.675 (2023).
15. Ben Hazem, Z. Study of Q-learning and deep Q-network learning control for a rotary inverted pendulum system. *Discov Appl. Sci.* **6**, 49. https://doi.org/10.1007/s42452-024-05690-y (2024).
16. Ray, L. R. & Stengel, R. F. A Monte Carlo approach to the analysis of control system robustness. *Automatica* **29**, 229–236. https://doi.org/10.1016/0005-1098(93)90187-X (1993).
17. Calafiore, G. C., Dabbene, F. & Tempo, R. Research on probabilistic methods for control system design. *Automatica* **47**, 1279–1293. https://doi.org/10.1016/j.automatica.2011.02.029 (2011).
18. Helton, J. C. & Davis, F. J. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliab. Eng. Syst. Saf.* **81**, 23–69. https://doi.org/10.1016/S0951-8320(03)00058-9 (2003).
19. McKay, M. D., Beckman, R. J. & Conover, W. J. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21**, 239–245. https://doi.org/10.1080/00401706.1979.10489755 (1979).
20. Kim, S. H. & Nelson, B. L. A fully sequential procedure for indifference-zone selection in simulation. *ACM Trans. Model. Comput. Simul.* **11**, 251–273. https://doi.org/10.1145/502109.502111 (2001).
21. Fan, W., Hong, L. J. & Nelson, B. L. Indifference-zone-free selection of the best. *Oper. Res.* **64**, 1499–1514. https://doi.org/10.1287/opre.2016.1530 (2016).
22. Eckman, D. J. & Henderson, S. G. Fixed-confidence, fixed-tolerance guarantees for ranking-and-selection procedures. *ACM Trans. Model. Comput. Simul.* **31**, 1–33, 7. https://doi.org/10.1145/3432754 (2021).
23. Nelson, B. L. & Matejcik, F. J. Using common random numbers for indifference-zone selection and multiple comparisons in simulation. *Manag. Sci.* **41**, 1935–1945. https://doi.org/10.1287/mnsc.41.12.1935 (1995).
24. Shen, H., Hong, L. J. & Zhang, X. Ranking and selection with covariates for personalized decision making. *INFORMS J. Comput.* **33**, 1500–1519. https://doi.org/10.1287/ijoc.2020.1009 (2021).
25. Rice, J. R. The Algorithm Selection Problem. *Adv. Comput.* **15**, 65–118. https://doi.org/10.1016/S0065-2458(08)60520-3 (1976).
26. Del Corso, G., Colantonio, S. & Caudai, C. Shedding light on uncertainties in machine learning: Formal derivation and optimal model selection. *J. Frankl. Inst.* **362**, 107548. https://doi.org/10.1016/j.jfranklin.2025.107548 (2025).
27. Tyralis, H. & Papacharalampous, G. A review of predictive uncertainty estimation with machine learning. *Artif. Intell. Rev.* **57**, 94. https://doi.org/10.1007/s10462-023-10698-8 (2024).
28. Psaros, A. F., Meng, X., Zou, Z., Guo, L. & Karniadakis, G. E. Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *J. Comput. Phys.* **477**, 111902. https://doi.org/10.1016/j.jcp.2022.111902 (2023).
29. Hofman, P., Sale, Y. & Hüllermeier, E. Uncertainty Quantification for Machine Learning: One Size Does Not Fit All. *Proc. AAAI Conf. Artif. Intell.* **40**, 21726–21734. https://doi.org/10.1609/aaai.v40i26.39323 (2026).
30. Gao, R., Yin, M., McInerney, J. & Kallus, N. Adjusting regression models for conditional uncertainty calibration. *Mach. Learn.* **113**, 8347–8370. https://doi.org/10.1007/s10994-024-06627-7 (2024).

31. Landgraf, D. et al. Probabilistic prediction methods for nonlinear systems with application to stochastic model predictive control. *Annu. Rev. Control*. **56**, 100905. https://doi.org/10.1016/j.arcontrol.2023.100905 (2023).
32. Scampicchio, A., Arcari, E., Lahr, A. & Zeilinger, M. N. Gaussian processes for dynamics learning in model predictive control. *Annu. Rev. Control*. **60**, 101034. https://doi.org/10.1016/j.arcontrol.2025.101034 (2025).
33. Sluijterman, L., Cator, E. & Heskes, T. How to evaluate uncertainty estimates in machine learning for regression?. *Neural Netw.* **173**, 106203. https://doi.org/10.1016/j.neunet.2024.106203 (2024).
34. Iwata, T. & Kumagai, A. Meta-learning to calibrate Gaussian processes with deep kernels for regression uncertainty estimation. *Neurocomputing* **579**, 127441. https://doi.org/10.1016/j.neucom.2024.127441 (2024).
35. Allen, S., Ziegel, J. & Ginsbourger, D. Assessing the calibration of multivariate probabilistic forecasts. *Q. J. R. Meteorol. Soc.* **150**, 1315–1335. https://doi.org/10.1002/qj.4647 (2024).
36. Grothe, O., Kächele, F. & Krüger, F. From point forecasts to multivariate probabilistic forecasts: The Schaake shuffle for day-ahead electricity price forecasting. *Energy Econ.* **120**, 106602. https://doi.org/10.1016/j.eneco.2023.106602 (2023).
37. Berrisch, J. & Ziel, F. Multivariate probabilistic CRPS learning with an application to day-ahead electricity prices. *Int. J. Forecast.* **40**, 1568–1586. https://doi.org/10.1016/j.ijforecast.2024.01.005 (2024).
38. Hendrickx, K., Perini, L., Van der Plas, D., Meert, W. & Davis, J. Machine learning with a reject option: A survey. *Mach. Learn.* **113**, 3073–3110. https://doi.org/10.1007/s10994-024-06534-x (2024).
39. Roman, R. C., Precup, R. E. & Petriu, E. M. Higher-order model-free control tuned by fictitious reference iterative tuning for 3D cranes. *Rom J. Inf. Sci. Technol.* **28**, 89–102. https://doi.org/10.59277/ROMJIST.2025.1.08 (2025).
40. Chi, R., Li, H., Shen, D., Hou, Z. & Huang, B. Enhanced P-type control: Indirect adaptive learning from set-point updates. *IEEE Trans. Autom. Control* **68**, 1600–1613. https://doi.org/10.1109/TAC.2022.3154347 (2023).
41. Roman, R.-C., Precup, R.-E. & Petriu, E. M. Active disturbance rejection control and model-free control tuned via fictitious reference iterative tuning. *Facta Univ. Ser. Mech. Eng.* **23**, 183–196. https://doi.org/10.22190/FUME250520021R (2025).
42. Hentout, A., Maoudj, A. & Kouider, A. Shortest path planning and efficient fuzzy logic control of mobile robots in indoor static and dynamic environments. *Rom J. Inf. Sci. Technol.* **27**, 21–36. https://doi.org/10.59277/ROMJIST.2024.1.02 (2024).
43. Yi, H., Chen, Z., He, Y., Liu, C. & Su, H. Active disturbance rejection control synthesis for industrial time-delayed process: an observation reconfiguration perspective. *ISA Trans.* **170**, 333–347. https://doi.org/10.1016/j.isatra.2026.01.022 (2026).
44. Cai, J., Zhang, L. & Gao, Q. Two-stage optimization of multi-channel delayed feedforward-feedback control for an active quasi-zero-stiffness vibration isolation system. *Mech. Syst. Signal. Process.* **258**, 114679. https://doi.org/10.1016/j.ymssp.2026.114679 (2026).
45. Cai, J., Zhang, L. & Gao, Q. Physics-informed neural network for spectrum design in delayed vibration control. *Int. J. Mech. Sci.* **325**, 111840. https://doi.org/10.1016/j.ijmecsci.2026.111840 (2026).
46. Demarta, S. & McNeil, A. J. The t copula and related copulas. *Int. Stat. Rev.* **73**, 111–129. https://doi.org/10.1111/j.1751-5823.2005.tb00254.x (2005).
47. Coşkun, Y. Probabilistic Controller Selection for a Rotary Inverted Pendulum under Uncertainty. *Zenodo* https://doi.org/10.5281/zenodo.23263340 (2026).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Yusuf Coşkun[1] · Lale Özyılmaz[2] · Muhammet Aydın[3]

✉ Yusuf Coşkun
yusuf.coskun@istun.edu.tr

[1] Department of Mechatronics Engineering, Faculty of Engineering and Natural Sciences, Istanbul Health and Technology University, Istanbul, Türkiye, Turkey

[2] Department of Avionics, Faculty of Applied Sciences, Istanbul Gelişim University, Istanbul, Türkiye, Turkey

[3] Department of Mechatronics Engineering, Faculty of Engineering, Fırat University, Elazığ, Türkiye, Turkey