Erciyes University Journal of Institute Of Science and Technology

Research Article
10.65520/erciyesfen.1994635

Imprint:
Volume: 42(3)
Year: 2026
Page: 419-444

Agah KOZAN[a*]

[a] Dr., Ege University & Lancaster University
agah.kozan@ege.edu.tr

[*] Corresponding Author

Received: 7/15/2026
Accepted: 10/14/2026

Citation:
Agah KOZAN (2026). The Effect of Replication Size and Estimator Choice on Dual Response Surface Optimization under Skewed and Contaminated Data. *Erciyes University Journal of Institute Of Science and Technology*, 42(3), 419-444. https://doi.org/10.65520/erciyesfen.*1994635*

Screened by iThenticate for Authors & Researchers

Except where otherwise noted, content in this article is licensed under a Creative Commons 4.0 International license. Icons by Font Awesome.

The Effect of Replication Size and Estimator Choice on Dual Response Surface Optimization under Skewed and Contaminated Data

Abstract

Within-design-point replication and estimator choice are usually examined separately in dual response surface optimization. This study evaluated their joint influence under clean, skewed, and additively contaminated data using 39 scenarios, 200 Monte Carlo replications, and nested samples at $r \in \{3,5,8,10\}$. The primary analysis fixed a skewness-adjusted location estimator and compared six classical and robust scale estimators. Decision quality was assessed against full-process and baseline-process targets. Increasing r from 3 to 8 reduced mean full-process regret by 66.9% under clean data, 40.5% under skew-only data, and 38.7%, 24.6%, and 22.5% under low, moderate, and high contamination probabilities. In the joint location–scale analysis, the same transition improved full-process regret in 79 of 80 combinations and baseline-process regret in all 80. No estimator was uniformly best. Standard deviation performed best under clean data and often when contamination was treated as part of the operating process. Robust scale estimators performed better in some low- and moderate-contamination settings, especially when preserving the baseline process was the aim. The benefit of additional replication remained consistent across estimator choices, evaluation targets, alternative process geometries, mild response-surface misspecification, finite-sample calibration, and positivity treatments. A printing-process benchmark and a semi-synthetic extension showed the same qualitative pattern under clean and mild-disturbance conditions. Within the fixed 27-point design, replication had the most consistent effect, while the preferred estimator depended on the target and contamination structure. Together, the results provide practical guidance for planning replication and selecting estimators in dual response surface studies.

Keywords: Robust parameter design, dual response surface methodology, replication size, scale estimation, contaminated data

Çarpık ve Kirlenmiş Veriler Altında Tekrar Sayısı ve Tahminci Seçiminin İkili Yanıt Yüzeyi Optimizasyonuna Etkisi

Öz

Deney noktası içindeki tekrar sayısı ve tahminci seçimi, ikili yanıt yüzeyi optimizasyonunda çoğunlukla ayrı ayrı ele alınmaktadır. Bu çalışmada bu iki unsurun temiz, çarpık ve eklemeli olarak kirlenmiş veriler altındaki ortak etkisi, 39 senaryo, 200 Monte Carlo tekrarı ve $r \in \{3,5,8,10\}$ için iç içe örneklemler kullanılarak değerlendirilmiştir. Ana analizde çarpıklığa göre düzeltilmiş bir konum tahmincisi sabit tutulmuş ve altı klasik ve sağlam ölçek tahmincisi karşılaştırılmıştır. Karar kalitesi tam-süreç ve temel-süreç hedeflerine göre değerlendirilmiştir. Tekrar sayısının 3'ten 8'e çıkarılması, tam-süreç için ortalama optimizasyon kaybını temiz veride %66,9, yalnızca

çarpık veride %40,5 ve düşük, orta ve yüksek kirlenme olasılıklarında sırasıyla %38,7, %24,6 ve %22,5 azaltmıştır. Ortak konum–ölçek analizinde aynı geçiş, 80 birleşimin 79'unda tam-süreç, tamamında ise temel-süreç kaybını iyileştirmiştir. Hiçbir tahminci bütün koşullarda üstün değildir. Klasik standart sapma temiz veride ve kirlenmenin operasyonel sürecin parçası sayıldığı birçok durumda en iyi performansı göstermiştir. Sağlam yöntemler ise bazı düşük ve orta kirlenme koşullarında, özellikle temel sürecin korunması amaçlandığında daha iyi sonuç vermiştir. Ek tekrarın yararı; tahminci seçimleri, değerlendirme hedefleri, alternatif süreç geometrileri, hafif yanıt yüzeyi yanlış belirlemesi, küçük örneklem kalibrasyonu ve pozitiflik yaklaşımları altında korunmuştur. Baskı süreci örneği ve yarı-sentetik genişletme, temiz ve hafif bozulma koşullarında aynı nitel örüntüyü göstermiştir. Sabit 27 noktalı tasarım içinde en tutarlı etki tekrar sayısından kaynaklanmış; tercih edilen tahminci ise hedefe ve kirlenme yapısına bağlı kalmıştır. Bulgular, ikili yanıt yüzeyi çalışmalarında tekrar sayısının planlanması ve tahminci seçimi için uygulamaya dönük bir temel sunmaktadır.

Anahtar kelimeler: Sağlam parametre tasarımı, ikili yanıt yüzeyi metodolojisi, tekrar sayısı, ölçek tahmini, kirlenmiş veri

1. INTRODUCTION

Robust parameter design seeks controllable factor settings that achieve a target while limiting sensitivity to uncontrollable variation. Taguchi's parameter-design philosophy provided an influential starting point, although its reliance on signal-to-noise ratios was criticised for obscuring the separate contributions of process location and variability [1, 2]. Response surface methodology provided a model-based alternative [3, 4], and Vining and Myers [5] formalised dual response surface methodology by fitting separate models for process location and dispersion. Subsequent work developed direct loss minimisation, nonlinear-programming formulations, weighted mean-squared-error criteria, alternative solution structures, semiparametric extensions, and distribution-specific designs [6–12].

A dual-response decision depends on the full estimation chain. Location and scale summaries are calculated from replicated observations, response surfaces are fitted to those summaries, and the fitted surfaces enter the loss criterion. Sampling error or bias at the first stage can therefore alter fitted curvature and move the selected operating condition away from the true optimum. This issue is especially important for scale, which is generally less stable than location when within-design-point samples are small [4, 13]. Moreover, the method with the smallest prediction error over the design region need not yield the lowest loss at the selected point.

Skewness, heavy tails, and contamination make this propagation problem more difficult. Classical means and standard deviations can be strongly influenced by a few extreme observations, motivating robust alternatives in parameter-design research [14–16]. Yet robust estimators differ in breakdown behaviour, efficiency, bias, and finite-sample variability [17–23]. Classical standard deviation may remain preferable under clean Gaussian data, whereas robust procedures can be unstable with very few replicates. Comparisons should therefore include both classical and robust benchmarks rather than assume that contamination automatically favours a robust scale estimator.

The scientific interpretation of extreme observations is equally important. Additive contamination may represent transient sensor offsets, calibration glitches, or short-lived disturbances, whereas skewness may arise in non-negative characteristics such as coating thickness, particle size, surface roughness, or cycle time. An extreme observation may be treated either as part of the operating process or as an abnormal departure from an underlying baseline process. These interpretations imply different optimization targets and motivate the two-estimand framework used below.

Relevant research has developed along two largely separate lines. Estimator-focused studies have compared classical and robust summaries under non-normality, asymmetry, high variability, or contamination, generally with fixed replication [14–16, 24–26]. Interval-based methods represented process location using skewness-adjusted confidence-interval information [27–30], while later studies incorporated BS82 and W24 scales and showed that performance depends on asymmetry and the

CC: BY-NC-ND

contamination mechanism [25, 26]. A second line has treated replication as an experimental-design, pure-error, or resource-allocation problem. Pure replication improves error estimation but consumes runs that could be assigned to additional design settings; related work has considered inference-oriented optimal design, cost–precision trade-offs, and prescribed pure-replication structures [31–33]. Tan and Ng [13] studied replication and sample allocation for mean and variance response-surface estimation, but focused on estimation efficiency rather than estimator choice propagated to the final dual-response decision.

To the best of our knowledge, no previous study has jointly examined within-design-point replication size and classical and robust location–scale estimators under skewed and contaminated data. This study traces their combined effects through fitted dual response surfaces to oracle-based decision regret. Its contribution is a unified design–estimation–decision framework that links replication size, estimator choice, and the scientific interpretation of extreme observations to the final operating decision. Three features distinguish the study: replication and estimator choice are varied jointly, performance is evaluated under two scientifically distinct targets, and robustness is assessed at the final decision rather than only at the surface-estimation stage.

Two related modelling issues also require decision-level assessment. An unconstrained polynomial fitted to non-negative scale estimates can predict negative values inside the experimental region; clipping or constrained fitting may restore admissibility without improving the selected decision. Likewise, conclusions under exactly quadratic generating surfaces may be optimistic when the fitted model is mildly misspecified. Global surface accuracy, positivity, and decision regret should therefore be examined jointly rather than treated as interchangeable criteria.

The study uses nested Monte Carlo samples at $r \in \{3,5,8,10\}$. The primary analysis fixes a skewness-adjusted location procedure and compares classical standard deviation, normal-consistent MAD, finite-sample-corrected Q_n, a biweight midvariance-based scale (BWMV), W24, and BS82. Decision quality is evaluated against a full-process target, in which contamination is part of the realised distribution, and a baseline-process target, in which it is treated as an abnormal disturbance. A joint sensitivity analysis then compares four location methods with four representative scale estimators across 39 scenarios.

The study addresses three questions: how much replication improves oracle-based decision quality, whether this pattern remains when location and scale estimators vary jointly, and how estimator preference changes between the full-process and baseline-process targets. Paired nested samples isolate replication effects within a fixed 27-point design, while common decision and diagnostic frameworks distinguish estimator performance, surface accuracy, admissibility, and final regret.

Robustness analyses vary location and scale jointly, change the correctly specified quadratic geometry, introduce mild response-surface misspecification, examine finite-sample calibration and positivity treatments, and assess oracle proximity to the experimental boundary. A final illustration separates decisions from published three-replicate printing-process data [10] from a process-calibrated semi-synthetic replication study under milder disturbances. Accordingly, the study determines which decision-level conclusions remain stable across estimands, estimator choices, modelling assumptions, and process geometries.

Section 2 describes the methods, Section 3 presents and discusses the results, and Section 4 summarises the conclusions. Supporting numerical results are provided in Appendix A.

2. MATERIALS AND METHODS

The study consists of a primary simulation analysis, a joint location–scale sensitivity analysis, targeted assessments of numerical and modelling assumptions, and a printing-process illustration combining published observations with a process-calibrated semi-synthetic extension. The primary analysis isolated the combined effects of replication size and scale-estimator choice by holding the location estimator fixed, whereas the additional analyses examined the robustness and practical

CC: BY-NC-ND

interpretation of the resulting decision patterns.

2.1. Dual response surface framework

Let Y denote the quality characteristic and let $\mathbf{x} = (x_1, x_2, x_3)^\top$ denote the controllable factors. At each design point, r replicated responses are observed. A location summary and a scale summary are calculated pointwise and fitted separately using the same second-order polynomial basis [4, 5]:

$$\hat{\mu}(\mathbf{x}) = \beta_0 + \sum_{j=1}^{3} \beta_j x_j + \sum_{j=1}^{3} \beta_{jj} x_j^2 + \sum_{1 \le j < k \le 3} \beta_{jk} x_j x_k, \tag{1}$$

$$\hat{\sigma}(\mathbf{x}) = \gamma_0 + \sum_{j=1}^{3} \gamma_j x_j + \sum_{j=1}^{3} \gamma_{jj} x_j^2 + \sum_{1 \le j < k \le 3} \gamma_{jk} x_j x_k. \tag{2}$$

Both surfaces were fitted by ordinary least squares in the primary analysis. The same polynomial basis was used for all pointwise estimator and replication-size comparisons so that the resulting differences were not confounded with changes in response-surface structure. The scale summaries were modelled directly on their original scale because the true scale functions in the primary and additional exactly quadratic process geometries belonged to the fitted second-order model class. This choice provided a correctly specified benchmark for evaluating first-stage scale estimation, although an unconstrained polynomial could predict negative scale values within the experimental region. Such predictions were retained as diagnostics rather than silently removed, and their consequences were examined separately through clipping and positivity-constrained sensitivity analyses.

2.2. Location estimators

2.2.1. Primary skewness-adjusted location estimator

The primary analysis used the midpoint of Johnson's skewness-adjusted interval [27]. This procedure was selected as a common location benchmark rather than because it was expected to be uniformly optimal. First, it provides a skewness-adjusted location summary consistent with the interval-based parameter-design literature [28–30]. Second, holding one location rule fixed allowed the primary experiment to isolate the effects of replication size and scale-estimator choice.

The Johnson midpoint adjusts the sample mean using an estimated third central moment. It should therefore be interpreted as a skewness-adjusted procedure rather than as a high-breakdown robust location estimator. In particular, its third-moment correction can be sensitive to extreme observations and unstable when the within-design-point sample is very small. The Johnson midpoint was consequently not assumed to be uniformly preferable, and the full joint sensitivity analysis repeated all 39 scenarios using the sample mean, sample median, and Huber M-location as alternatives.

For the r responses at design point i, let $\bar{y}_i$, s_i, and $\hat{\mu}_{3i}$ denote the sample mean, sample standard deviation, and sample third central moment, respectively, where

$$\hat{\mu}_{3i} = \frac{1}{r} \sum_{j=1}^{r} \left(y_{ij} - \bar{y}_i\right)^3.$$

The pointwise location summary was

$$\tilde{\mu}_{J,i} = \bar{y}_i + \frac{\hat{\mu}_{3i}}{6 r s_i^2}. \tag{3}$$

CC: BY-NC-ND

When $s_i = 0$, $\bar{y}_i$ was used. Because Equation (3) is the midpoint of the skewness-adjusted interval, its value does not depend on the confidence coefficient.

2.2.2. Location sensitivity methods

The joint sensitivity analysis also considered the sample mean, sample median, and a Huber M-location estimator [17]. At each design point i, the Huber procedure was initialised at the sample median. Standardised residuals were calculated using the normal-consistent MAD,

$$S_i = 1.482602\,\mathrm{median}_j\left|y_{ij} - \mathrm{median}_\ell(y_{i\ell})\right|.$$

For the current location value T_i, the residuals, weights, and updated location were

$$u_{ij} = \frac{y_{ij} - T_i}{S_i}, \qquad w_{ij} = \begin{cases} 1, & |u_{ij}| \le 1.345, \\ \dfrac{1.345}{|u_{ij}|}, & |u_{ij}| > 1.345, \end{cases} \qquad T_{i,\mathrm{new}} = \frac{\sum_{j=1}^{r} w_{ij}\, y_{ij}}{\sum_{j=1}^{r} w_{ij}}.$$

Iteration stopped when the relative change in location was below 10^{-10} or after 100 iterations.

2.3. Scale estimators

Six pointwise scale estimators were compared. At a generic design point, the design-point index is suppressed below and the r replicated responses are denoted by $y_1, \ldots, y_r$. A common numerical safeguard of $\varepsilon = 0.02$ was applied when a pointwise scale estimate was non-finite, undefined, non-positive, or smaller than the bound. In such cases, the value supplied to the response-surface fit was set to ε, and activation of this rule was recorded as a diagnostic. This safeguard operated only on the pointwise scale summaries; it did not constrain the subsequently fitted polynomial scale surface to remain positive throughout the experimental region.

The estimators were implemented using the published normal-consistency constants and tuning constants stated below. Only Q_n included an explicitly r-specific finite-sample correction. The other procedures retained their literature-based implementations, and no post hoc simulation-derived correction was imposed in the primary analysis. Consequently, especially at $r = 3$ and $r = 5$, the comparisons reflect not only robustness–efficiency and sampling-variability differences, but also the native finite-sample calibration of the implemented estimators. This choice treated the published procedures as the primary objects of comparison rather than forcing them to have identical normal-model means. Their finite-sample behaviour was characterised separately, and the complete primary analysis was repeated with r-specific normal-model mean calibration as a sensitivity assessment.

2.3.1. Classical standard deviation and MAD

The classical sample standard deviation was

$$\hat{\sigma}_{SD} = \left\{\frac{1}{r-1}\sum_{i=1}^{r}(y_i - \bar{y})^2\right\}^{1/2}.$$

The normal-consistent median absolute deviation was

$$\hat{\sigma}_{MAD} = 1.482602\,\mathrm{median}_i\left\{\left|y_i - \mathrm{median}_j(y_j)\right|\right\}.$$

The multiplier provides asymptotic normal consistency and should not be interpreted as an exact finite-sample unbiasedness correction at each investigated value of r.

2.3.2. Finite-sample-corrected Q_n

The Q_n estimator is based on pairwise absolute differences and combines a high breakdown point with good Gaussian efficiency [20]. Let

CC: BY-NC-ND

$$h = \left\lfloor \frac{r}{2} \right\rfloor + 1, \qquad k = \binom{h}{2},$$

and let

$$\{|y_i - y_j| : i < j\}_{(k)}$$

denote the kth order statistic of the pairwise absolute differences. The estimator was

$$\hat{\sigma}_{Q_n} = 2.21914\, d_r \{|y_i - y_j| : i < j\}_{(k)},$$

where d_r is the finite-sample correction corresponding to the investigated replication sizes $r \in \{3,5,8,10\}$. The exact correction values are listed in Appendix Table A1.

2.3.3. W24 estimator

For the median-based initial location and scale

$$T_0 = \text{median}_i(y_i), \qquad S_0 = \text{median}_i|y_i - T_0|,$$

define

$$z_i = \frac{y_i - T_0}{2.4 S_0}.$$

Following Gross [34] and its dual-response application [26], the W24 estimator was

$$\hat{\sigma}_{W24} = 2.4 S_0 \left[r \frac{\sum_{|z_i|<\pi} \sin^2(z_i)}{\left\{\sum_{|z_i|<\pi} \cos(z_i)\right\}^2} \right]^{1/2}.$$

If the initial scale or the resulting expression was undefined or non-positive, the common numerical safeguard described above was applied.

2.3.4. Biweight midvariance-based scale

Using the median as centre, define

$$\text{MAD} = \text{median}_i \left| y_i - \text{median}_j(y_j) \right|, \qquad u_i = \frac{y_i - \text{median}_j(y_j)}{9\text{MAD}}.$$

The standard-deviation form of the biweight midvariance was [35, 36]

$$\hat{\sigma}_{BWMV} = \left[r \frac{\sum_{|u_i|\le 1} \{y_i - \text{median}_j(y_j)\}^2 (1 - u_i^2)^4}{\left\{\sum_{|u_i|\le 1} (1 - u_i^2)(1 - 5u_i^2)\right\}^2} \right]^{1/2}.$$

The common numerical safeguard was used when the initial MAD or the resulting scale expression was undefined or non-positive.

2.3.5. BS82 M-scale estimator

For

$$z_i = \frac{y_i - T_0}{8.2 S_0},$$

define

$$\psi(z) = z(1 - z^2)^2, \qquad \psi'(z) = 1 - 6z^2 + 5z^4,$$

for $|z| \le 1$, and set both functions to zero outside this interval. The BS82 scale used in earlier robust dual-response work [25, 37] was

CC: BY-NC-ND

$$\hat{\sigma}_{BS82} = 8.2S_0\left[r\frac{\sum_{i=1}^{r}\psi^2(z_i)}{\{\sum_{i=1}^{r}\psi'(z_i)\}^2}\right]^{1/2}.$$

If the initial scale or the resulting expression was undefined or non-positive, the common numerical safeguard was applied.

2.4. Nominal-the-best criterion and evaluation targets

For target τ, the estimated nominal-the-best loss followed Lin and Tu [7]:

$$\hat{L}(\mathbf{x}) = \{\hat{\mu}(\mathbf{x}) - \tau\}^2 + \hat{\sigma}^2(\mathbf{x}), \qquad \hat{\mathbf{x}}^* = \arg\min_{\mathbf{x}\in[-1,1]^3}\hat{L}(\mathbf{x}). \tag{4}$$

The unweighted form avoided an additional application-specific parameter. The operating condition $\hat{\mathbf{x}}^*$ was selected once from this fitted loss and evaluated against both oracle targets.

The full-process target treated contamination as part of the realised operating distribution:

$$\begin{aligned} L_c(\mathbf{x}) &= \{\mu_c(\mathbf{x}) - \tau\}^2 + \sigma_c^2(\mathbf{x}), \quad \mathbf{x}_c^* = \arg\min_{\mathbf{x}\in[-1,1]^3} L_c(\mathbf{x}), \\ R_c &= L_c(\hat{\mathbf{x}}^*) - L_c(\mathbf{x}_c^*). \end{aligned} \tag{5}$$

The baseline-process target treated contamination as an abnormal disturbance:

$$\begin{aligned} L_0(\mathbf{x}) &= \{\mu_0(\mathbf{x}) - \tau\}^2 + \sigma_0^2(\mathbf{x}), \quad \mathbf{x}_0^* = \arg\min_{\mathbf{x}\in[-1,1]^3} L_0(\mathbf{x}), \\ R_0 &= L_0(\hat{\mathbf{x}}^*) - L_0(\mathbf{x}_0^*). \end{aligned} \tag{6}$$

The two regrets answer different scientific questions and were reported separately. Apart from optimization tolerance, each is non-negative because its oracle minimises the corresponding loss over the same region.

2.5. Simulation design

The coded region was $[-1,1]^3$. A full 3^3 factorial at levels -1, 0, and 1 provided 27 design points for the 10-coefficient second-order model. This transparent three-factor setting provided broad geometric coverage while keeping the factorial sensitivity analyses computationally manageable, thereby offering a controlled benchmark for isolating replication and estimator effects.

The baseline mean and scale functions were

$$\begin{aligned} \mu_0(\mathbf{x}) &= 50 + 4x_1 - 3x_2 - 6x_3 + 2x_1x_2 - 1.5x_2x_3 + 2.5x_1^2, \\ \sigma_0(\mathbf{x}) &= 6\{1 + 0.25x_1^2 + 0.35(x_2+1)^2 + 0.15(x_3+1)^2\}. \end{aligned} \tag{7}$$

The target was $\tau = 50$, and the scale function was positive throughout the region and belonged to the fitted model class. Responses were generated as

$$Y(\mathbf{x}) = \mu_0(\mathbf{x}) + \sigma_0(\mathbf{x})U, \qquad U = E_g + CO. \tag{8}$$

Here $C \sim \text{Bernoulli}(c)$. For $g = 0$, $E_g = Z$ with $Z \sim N(0,1)$; for $g > 0$,

$$E_g = \frac{\exp(gZ) - \exp(g^2/2)}{\{\exp(2g^2) - \exp(g^2)\}^{1/2}},$$

so $\mathrm{E}(E_g) = 0$ and $\mathrm{Var}(E_g) = 1$. Thus g changed asymmetry while preserving the first two moments; $g = 0.60$ and 0.90 correspond to standardised third moments of approximately 2.26 and

CC: BY-NC-ND

4.75.

For one-sided contamination, $O = k + 0.8V$; for two-sided contamination, $O = S(k + 0.8V)$, where $V \sim N(0,1)$ and S independently equals -1 or $+1$ with probability $1/2$. All random components were independent. Contamination was additive on the standardised-error scale, so k represented displacement relative to the local baseline scale. The exact moments were

$$\begin{aligned} \mathrm{E}(U) &= ck, \quad \mathrm{Var}(U) = 1 + c(0.8)^2 + c(1-c)k^2, && \text{one-sided,} \\ \mathrm{E}(U) &= 0, \quad \mathrm{Var}(U) = 1 + c\{k^2 + (0.8)^2\}, && \text{two-sided.} \end{aligned} \tag{9}$$

Consequently,

$$\mu_c(\mathbf{x}) = \mu_0(\mathbf{x}) + \sigma_0(\mathbf{x})\mathrm{E}(U), \qquad \sigma_c(\mathbf{x}) = \sigma_0(\mathbf{x})\{\mathrm{Var}(U)\}^{1/2}. \tag{10}$$

The 39 scenarios comprised one clean control, two skew-only controls, and 36 contaminated settings obtained by crossing three probabilities, two magnitudes, three asymmetry values, and two directions. Each used 200 Monte Carlo replications; Table 1 summarises the factorial structure.

Table 1. Structure of the 39 primary simulation scenarios.

Scenario class	Contamination probability	Outlier magnitude	Asymmetry strength	Direction	Number
Clean	0	0	0.00	Not applicable	1
Skew-only	0	0	0.60, 0.90	Not applicable	2
Low contamination probability	0.05	12, 25	0.00, 0.60, 0.90	One- or two-sided	12
Moderate contamination probability	0.15	12, 25	0.00, 0.60, 0.90	One- or two-sided	12
High contamination probability	0.25	12, 25	0.00, 0.60, 0.90	One- or two-sided	12
Total					39

The labels low, moderate, and high refer only to c, not total distributional impact. Because k is standardised, $k = 12$ and 25 are intentionally strong stress tests: the implied standard-deviation multiplier ranges from 2.81 to 12.55 and the one-sided mean shift ck from 0.60 to 6.25 (Appendix Table A6). The levels were informed by earlier W24 screening ranges [26]; a milder $c = 0.05$, $k = 3$ condition was examined in the printing-process illustration.

2.6. Nested replication-size comparison

For each scenario, design point, and Monte Carlo replication, ten responses were generated first. Analyses at $r \in \{3,5,8,10\}$ used the first 3, 5, 8, and 10 observations, so smaller samples were nested within larger ones. The same samples were supplied to every estimator. Replication-size and estimator contrasts were therefore paired, reducing irrelevant Monte Carlo variation; this dependence was retained in the tests and bootstrap procedures.

2.7. Performance measures

Full-process regret R_c was primary and baseline-process regret R_0 secondary. For $t \in \{c, 0\}$, total regret was decomposed as

CC: BY-NC-ND

$$\begin{aligned} R_t &= R_{\mu,t} + R_{\sigma,t}, \\ R_{\mu,t} &= \{\mu_t(\hat{\mathbf{x}}^*) - \tau\}^2 - \{\mu_t(\mathbf{x}_t^*) - \tau\}^2, \\ R_{\sigma,t} &= \sigma_t^2(\hat{\mathbf{x}}^*) - \sigma_t^2(\mathbf{x}_t^*). \end{aligned} \tag{11}$$

Either component may be negative because the oracle minimises their sum; total regret is non-negative apart from numerical tolerance.

Location- and scale-surface RMSEs were calculated on a diagnostic grid separate from optimization. For example,

$$\mathrm{RMSE}_{\sigma,t} = \left[\frac{1}{G}\sum_{g=1}^{G}\{\hat{\sigma}(\mathbf{x}_g) - \sigma_t(\mathbf{x}_g)\}^2\right]^{1/2}. \tag{12}$$

Additional diagnostics included oracle distance, floor activation, negative fitted-scale rates on the grid and at the selected optimum, boundary frequency, and optimization failures. These measures separated pointwise stability, global fit, admissibility, and final decision quality; regret remained the principal criterion.

2.8. Statistical comparison procedure

For each scenario, replication size, estimator, and outcome, means, medians, standard deviations, and relevant interquartile ranges were calculated. A normal-approximation 95% Monte Carlo error interval for a mean was $\overline{Z} \pm 1.96 s_Z/\sqrt{M}$; this quantifies simulation error rather than uncertainty for a real process.

Five paired replication contrasts were prespecified: 5 versus 3, 8 versus 3, 10 versus 3, 8 versus 5, and 10 versus 8. Two-sided Wilcoxon signed-rank tests excluded zero differences, with Holm adjustment within each scenario–estimator–outcome family [38, 39]. The six scale estimators generated 15 paired comparisons per scenario, replication size, and outcome; Holm adjustment covered those 15. Mean and median differences, win rates, and rank-biserial effects accompanied the tests. Counts of significant results across scenarios were descriptive rather than one global family-wise test.

For the joint location–scale analysis, mean, median, and Huber locations were compared separately with Johnson after averaging over scenarios within class and Monte Carlo replication. Holm adjustment covered the three contrasts in each class–replication–scale–outcome family, giving 80 families per outcome.

Primary class summaries were likewise formed within Monte Carlo replication by averaging over relevant scenarios and the six scale estimators. If $\overline{R}_{c,r}$ denotes the class mean, absolute and relative reductions per added replicate were

$$\begin{aligned} G_{c,a\to b} &= \frac{\overline{R}_{c,a} - \overline{R}_{c,b}}{b-a}, \\ G^{\%}_{c,a\to b} &= 100\,\frac{\overline{R}_{c,a} - \overline{R}_{c,b}}{(b-a)\overline{R}_{c,a}}. \end{aligned} \tag{13}$$

The win rate was the proportion of paired summaries with lower regret at the larger r. Paired-bootstrap 95% intervals used 5000 resamples of the Monte Carlo index, with the same indices applied to all contributing scenarios, estimators, and replication sizes [40]. Main and joint analyses used 200 replications and targeted checks 100; conclusions combined intervals, paired tests, effect summaries, and win rates [41].

CC: BY-NC-ND

2.9. Sensitivity analyses

Complementary analyses tested whether the replication conclusion depended on location–scale choice, the numerical floor, process geometry, model specification, positivity treatment, finite-sample calibration, or oracle boundary proximity. Unless stated otherwise, the nested samples and the full-process and baseline-process targets were retained.

2.9.1. Joint location–scale analysis

All 39 scenarios and 200 replications were repeated in a 4×4 comparison of Johnson, mean, median, and Huber locations with SD, finite-sample-corrected Q_n, BWMV, and W24 scales. The 499,200 combinations used the same samples, seeds, targets, second-order basis, and optimization procedure as the primary analysis. The purpose was to evaluate whether the replication pattern remained stable when both first-stage choices varied jointly.

2.9.2. Numerical-floor sensitivity

The pointwise scale floor was varied over $\varepsilon \in \{10^{-6}, 0.01, 0.02\}$ in six representative scenarios with 100 replications. The run used seed 20260708, a 15^3 optimization grid, refinement from the three best finite points, and a separate 13^3 diagnostic grid. This check assessed whether small-r behaviour was created by the default numerical safeguard; it did not impose positivity on the fitted polynomial.

2.9.3. Additional exactly quadratic process geometry

Generalisability across a second correctly specified quadratic system was examined using

$$\begin{aligned} \mu_{0,B}(\mathbf{x}) &= 50 - 5x_1 + 4x_2 + 3x_3 - 2x_1x_2 + 2.5x_1x_3 + 1.5x_2^2 - 2x_3^2, \\ \sigma_{0,B}(\mathbf{x}) &= 5\{1 + 0.30(x_1 + 1)^2 + 0.25(x_2 - 1)^2 + 0.35x_3^2\}. \end{aligned} \tag{14}$$

The full scenario grid was repeated with 100 replications, the same nested r values, Johnson location, and W24, BWMV, and BS82 scales, using seed 20260708 and the standard optimization and diagnostic grids. This changed coefficient geometry while retaining correct second-order specification.

2.9.4. Response-surface misspecification

A separate stress test fitted the same second-order basis to true functions containing omitted third-order structure:

$$\begin{aligned} \mu_M(\mathbf{x}) &= \mu_0(\mathbf{x}) + 1.50x_1x_2x_3 + 1.20x_1^2x_2 - 1.00x_2^2x_3 + 0.80x_1x_3^2, \\ \sigma_M(\mathbf{x}) &= \sigma_0(\mathbf{x})\exp\{0.08x_1x_2x_3 + 0.07x_1^2x_3 - 0.06x_2^2x_1 + 0.05x_2x_3^2\}. \end{aligned} \tag{15}$$

The exponential multiplier preserved positive true scale. All 39 scenarios used 100 replications, Johnson, median, and Huber locations, and SD, Q_n, BWMV, and W24 scales. Oracle losses used the complete true functions and a 31^3 grid with 12 refinement starts. Unlike the alternative geometry, this analysis deliberately introduced model-form error.

2.9.5. Grid-constrained positivity analysis

Unconstrained OLS was compared with post-fit clipping,

$$\hat{\sigma}_{\text{clip}}(\mathbf{x}) = \max\{\hat{\sigma}_{\text{OLS}}(\mathbf{x}), 0.02\},$$

and the quadratic programme

$$\min_{\boldsymbol{\gamma}} \| \mathbf{X}\boldsymbol{\gamma} - \mathbf{s} \|_2^2 \quad \text{subject to} \quad \mathbf{X}_G\boldsymbol{\gamma} \geq 0.02, \tag{16}$$

where $\mathbf{X}_G$ was evaluated on an 11^3 constraint grid. Because this guarantees positivity only at

CC: BY-NC-ND

those points, fits were re-evaluated on the separate diagnostic grid and at the continuous optimum. Twelve representative scenarios used 100 replications, $r = 3,8,10$, Johnson and median locations, and SD, Q_n, BWMV, and W24. CVXPY used OSQP with CLARABEL fallback; solutions required maximum violation at most 10^{-6}. Comparisons included negative predictions, RMSE, and regret.

2.9.6. Finite-sample calibration and normal-model mean calibration sensitivity

Each scale estimator was evaluated on 10^6 independent $N(0,1)$ samples at every r, using the primary constants and safeguards. The complete primary experiment was then repeated after dividing each pointwise estimate by its corresponding normal-model Monte Carlo mean. This forced unit mean only under the reference model and did not imply unbiasedness under skewness or contamination. The paired published-versus-calibrated comparison contained 374,400 decision results generated from the same observations and settings.

2.9.7. Oracle-boundary assessment

The 39 scenarios yielded 13 distinct oracle structures. Each exact-process optimum from the standard 15^3 grid and three starts was checked with a 31^3 grid and 12 starts. For $\mathbf{x}^*$, the coordinate distance to the boundary was

$$d_\partial = \min_{j=1,2,3}\{1 - |x_j^*|\},$$

with $d_\partial \leq 10^{-3}$ used as a descriptive boundary classification. This assessed whether frequent fitted boundary solutions merely reflected boundary oracle optima.

2.10. Printing-process benchmark illustration

The benchmark of Ding et al. [10] contains three coded factors in a complete 3^3 factorial, three responses at each of 27 points, and target 500. The illustration had two stages. First, the 81 published observations were analysed with Johnson location and all six scales. Because the physical process truth and oracle are unknown, this stage reported fitted decisions and diagnostics rather than common oracle regret; estimator-specific fitted losses were not used for external ranking. Data-integrity checks verified the factorial structure, source values, point summaries, and published rounded coefficients, after which mean and standard-deviation surfaces were refitted using unrounded OLS coefficients.

Second, the refitted surfaces were treated as a benchmark geometry rather than exact physical truth. Clean, mild one-sided, and mild two-sided conditions were studied; disturbed cases used $c = 0.05$ and $k = 3$. Each used 200 replications, nested $r \in \{3,5,8,10\}$, Johnson location, all six scales, and both targets. Because the semi-synthetic truth was known, oracle regret was available. Standard oracle searches were checked with a 31^3 grid and 12 starts. This stage extended the simulation framework to a process geometry calibrated from published industrial observations.

2.11. Computational reproducibility

Analyses used Python with NumPy, pandas, SciPy, and Matplotlib; positivity fitting additionally used CVXPY with OSQP and CLARABEL. Seeds, scenario definitions, estimator settings, optimization settings, and metadata were recorded. Unless otherwise stated, operating conditions used a bounded 15^3 grid followed by L-BFGS-B refinement from the three best finite points; diagnostics used a separate 13^3 grid, and designated oracle checks used a 31^3 grid with 12 starts.

Failed refinement did not delete a replication: the best finite grid solution was retained and logged. QP fits required maximum constraint violation at most 10^{-6}. Detailed scripts, scenario-level outputs, optimization logs, and machine-readable results are available from the corresponding author upon reasonable request.

2.12. Ethical Approval

Ethical approval is not required for this study.

CC: BY-NC-ND

3. RESULTS AND DISCUSSION

3.1. Computational completeness and overall replication effect

The primary analysis comprised 187,200 fitted combinations. Four local refinements failed (0.0021%); the best finite grid solution was retained and no replication was removed.

Increasing replication reduced class-level mean full-process regret in every scenario class (Table 2; Figure 1). From $r = 3$ to $r = 8$, mean regret fell by 66.9% under clean data, 40.5% under skew-only data, and 38.7%, 24.6%, and 22.5% at the low, moderate, and high contamination probabilities. Corresponding baseline-process reductions were 66.9%, 40.5%, 32.0%, 16.9%, and 13.8%; detailed mean results are provided in Appendix Tables A2 and A3. Medians also decreased monotonically, so the mean pattern was not driven only by extreme values.

Table 2. Mean full-process optimization regret by scenario class and replication size. Parentheses contain normal-approximation 95% Monte Carlo error intervals based on 200 replicate-level summaries averaged over the six scale estimators.

Scenario class	$r = 3$	$r = 5$	$r = 8$	$r = 10$
Clean	64.66 (56.29–73.03)	38.65 (33.51–43.80)	21.39 (18.87–23.90)	19.39 (16.97–21.80)
Skew-only	53.49 (48.96–58.01)	39.97 (36.80–43.14)	31.83 (29.05–34.61)	27.93 (25.72–30.14)
Low contamination probability	1265.68 (1215.37–1316.00)	963.18 (921.11–1005.25)	775.86 (746.15–805.56)	738.06 (707.12–769.00)
Moderate contamination probability	4173.21 (4031.06–4315.35)	3527.51 (3374.93–3680.09)	3147.88 (3011.61–3284.15)	2660.42 (2532.13–2788.71)
High contamination probability	6677.90 (6424.07–6931.74)	5965.96 (5720.25–6211.66)	5173.11 (4954.59–5391.63)	4619.18 (4408.89–4829.48)

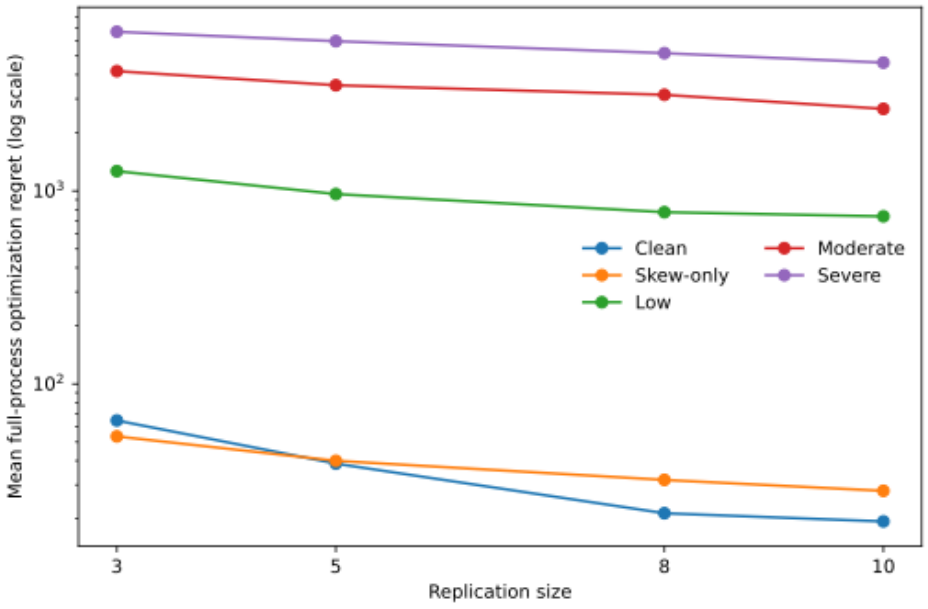

Figure 1. Mean full-process optimization regret by replication size and scenario class, averaged over the six scale estimators. The vertical axis is logarithmic.

CC: BY-NC-ND

Location RMSE also improved with replication. Nevertheless, the variance component dominated mean full-process regret under contamination: at $r = 3$ it represented about 99.0%, 94.3%, and 89.3% of total regret in the classes with low, moderate, and high contamination probabilities. These are class-mean contributions; the two components need not be separately non-negative.

Numerical difficulties were concentrated at small r. Floor activation declined from 0.293% at $r = 3$ to zero at $r = 8$ and 10; negative diagnostic-grid predictions fell from 8.00% to 3.08%, and negative fitted scale at the selected optimum from 33.1% to 14.5%. Thus additional replication improved both decision performance and surface behaviour, although it did not eliminate inadmissible polynomial predictions (Appendix Table A4).

The improvement with increasing replication size is consistent with greater stability of the pointwise location and, particularly, scale estimates. At very small replication sizes, scale estimates are more variable, and this variability propagates to the fitted dispersion surface and ultimately to the selected operating condition. As replication size increased, location RMSE improved, floor activation became negligible, and negative fitted-scale predictions became less frequent. These patterns suggest that the decision-level benefit of replication arises from more stable first-stage estimation being propagated through response-surface fitting to the final optimization decision.

3.2. Marginal return and scenario heterogeneity

Marginal gains varied across transitions and scenario classes (Table 3; Figure 2). Clean and low-contamination settings showed clear diminishing relative returns: the reduction per added replicate fell from 20.1% to 4.7% and from 12.0% to 2.4%, respectively. The pattern was less regular under skew-only, moderate, and high contamination probabilities, so diminishing returns were not universal.

Table 3. Relative mean full-process regret reduction per added replicate. Each cell reports the percentage reduction and its paired-bootstrap 95% interval; the percentage of paired replicate-level summaries favouring the larger replication size is shown after the semicolon.

Scenario class	$r = 3 \rightarrow 5$	$r = 5 \rightarrow 8$	$r = 8 \rightarrow 10$
Clean	20.1 (15.1–24.5); 68.5	14.9 (12.0–17.6); 74.5	4.7 (0.9–8.3); 62.0
Skew-only	12.6 (8.5–16.4); 67.5	6.8 (4.5–9.0); 66.5	6.1 (3.0–9.1); 57.0
Low contamination probability	12.0 (10.0–13.8); 80.0	6.5 (5.3–7.6); 77.0	2.4 (0.7–4.1); 61.0
Moderate contamination probability	7.7 (5.8–9.7); 72.5	3.6 (2.2–5.0); 62.5	7.7 (5.8–9.6); 75.5
High contamination probability	5.3 (3.2–7.4); 59.5	4.4 (2.9–5.8); 69.0	5.4 (3.1–7.4); 64.5

CC: BY-NC-ND

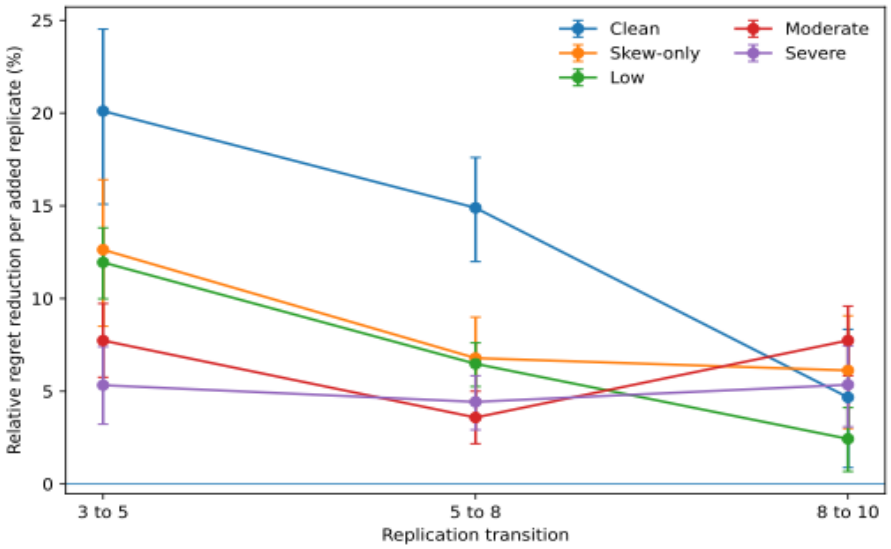

Figure 2. Relative mean full-process regret reduction per added replicate. Error bars are paired-bootstrap 95% intervals based on 5000 resamples of the Monte Carlo replication index.

All class-mean gains were positive and their bootstrap intervals excluded zero. Win rates of 57.0%–80.0% showed that the overall benefit was strong, although its magnitude varied across paired experiments. Relative and absolute gains also differed because regret scales varied. For $r = 8 \rightarrow 10$, the per-replicate relative reductions were 4.7% under clean and 5.4% under high contamination, whereas absolute reductions were 1.00 and 276.96.

Broad probability classes concealed further heterogeneity. The total $r = 8 \rightarrow 10$ reduction ranged from 0.5% for 5% one-sided contamination with magnitude 25 to 24.8% for 15% two-sided contamination with magnitude 25. Thus $r = 8$ is only an intermediate investigated level. Increasing the replication size from 8 to 10 required 54 additional runs, making this transition particularly relevant for future fixed-budget comparisons with additional or relocated design points.

3.3. Scale-estimator choice and evaluation target

Estimator rankings changed with scenario class, replication size, and target (Table 4). Entries are descriptive class-mean winners, not claims of significance against every alternative; detailed regret results are provided in Appendix Tables A2–A3.

Table 4. Scale estimator with the lowest class-level mean outcome. Within each cell, entries from left to right correspond to $r = 3,5,8,$ and 10. Rankings are descriptive; estimator-specific regret results are provided in Appendix Tables A2–A3.

Scenario class	Full-process regret	Baseline-process regret	Full-process scale RMSE
Clean	SD, SD, SD, SD	SD, SD, SD, SD	SD, SD, SD, SD
Skew-only	SD, SD, Q_n, Q_n	SD, SD, Q_n, Q_n	SD, SD, SD, SD
Low contamination probability	BWMV, BS82, BWMV, BWMV	BS82, Q_n, BS82, BWMV	SD, SD, SD, SD
Moderate contamination probability	SD, SD, SD, SD	SD, Q_n, SD, SD	SD, SD, SD, SD
High contamination probability	SD, SD, SD, SD	SD, SD, SD, SD	SD, SD, SD, SD

CC: BY-NC-ND

SD had the smallest full-process scale RMSE in all 156 scenario–replication cells but the lowest mean full-process regret in only 85. Global dispersion-surface accuracy and local decision quality were therefore non-equivalent. At the low contamination probability at $r = 10$, BWMV regret was 616.83 versus 1241.85 for SD, although its scale RMSE was 38.81 versus 20.27. Conversely, at the high contamination probability at $r = 10$, SD regret was 1715.75, while resistant estimators ranged from 4855.73 to 5357.54.

This result does not imply that SD is always superior. When contamination belongs to the full process, its added second-moment variation is part of the target estimated by SD, whereas resistant scales deliberately downweight it. Under clean data, SD led regret and RMSE at every r; under skew-only data, Q_n became the regret winner at $r = 8$ and 10 while SD retained the lowest RMSE.

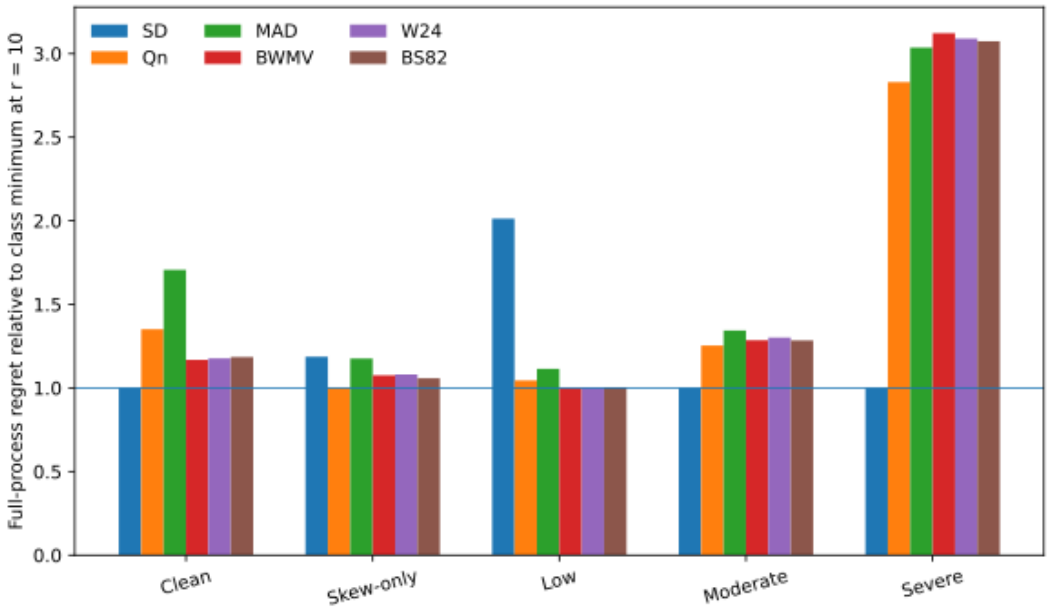

Figure 3. Mean full-process regret at $r = 10$, normalised by the minimum within each scenario class. A value of one denotes the lowest class-level mean regret. The normalisation facilitates within-class comparison and should not be used to compare absolute regret magnitudes across scenario classes.

The baseline-process target made robust scale estimators more competitive, especially at the low contamination probability, but did not produce a general reversal. SD still had the lowest mean baseline regret in 74 of 156 cells, compared with 28 for Q_n and 23 for BWMV. In the primary analysis, scale changed while Johnson location remained fixed, so contamination could still affect the fitted location surface; small samples and multiple contaminated observations also limited recovery of the baseline optimum.

No scale estimator was uniformly preferable. Selection should reflect the process distribution, replication size, and whether extremes belong to routine performance or represent an abnormal disturbance. These findings extend fixed-replication robust dual-response comparisons [14–16, 25, 26] and complement the estimation-efficiency focus of Tan and Ng [13] by propagating first-stage behaviour to oracle decision regret.

3.4. Joint location–scale sensitivity

The joint analysis comprised 499,200 combinations. Sixteen refinements failed (0.0032%); grid solutions were retained, and Johnson-based results reproduced the corresponding primary outputs.

From $r = 3$ to $r = 8$, class-mean full-process regret decreased in 79 of 80 location–scale combinations, with positive bootstrap intervals in 78; baseline regret decreased with positive intervals in all 80. The $r = 8 \rightarrow 10$ transition was less uniform: mean full- and baseline-process regret decreased

CC: BY-NC-ND

in 79 and 76 combinations, respectively, with median total reductions of 11.6% and 9.3%.

Location choice materially affected rankings. Against Johnson, mean, median, and Huber produced significantly lower full-process regret in 15, 28, and 27 of 80 comparisons; Johnson was better in 8, 12, and 12. Under the baseline-process target, the corresponding improvements were 19, 44, and 42, while Johnson was better in 7, 7, and 6. Many comparisons were not significant, so no location procedure dominated.

This target difference has a direct interpretation. Resistance can protect the uncontaminated location when extremes are abnormal, but may suppress a one-sided mean shift that belongs to the full-process target. Descriptive best combinations likewise varied: median with robust scales often led baseline regret at low and moderate contamination probabilities, whereas SD remained the scale component of the best combinations at the high contamination probability. The replication direction was therefore more consistent than the preferred location–scale estimator combination.

3.5. Targeted stress tests: numerical floor, process geometry, model misspecification, and positivity

The targeted checks used 100 Monte Carlo replications and addressed complementary assumptions. Their main decision-level results are summarised in Table 5.

Table 5. Decision-level summary of the targeted stress tests.

Analysis	Evidence	Decision-level implication
Numerical scale floor	Across the representative scenarios, the maximum change attributable to varying the pointwise floor was 0.1918% for mean regret and 0.0007% for scale-surface RMSE.	The default value $\varepsilon = 0.02$ did not materially create the small-r findings.
Alternative exactly quadratic geometry	The class-level full-process regret reduction from $r = 3$ to $r = 8$ ranged from 11.3% to 59.0%. The $r = 3 \rightarrow 8$transition was the most consistently favourable, whereas the incremental gain from $r = 8 \rightarrow 10$ varied across classes.	The main replication gain extended to a second correctly specified quadratic coefficient structure. The evidence therefore identifies $r = 3 \rightarrow 8$ as the more stable transition across the two quadratic geometries.
Mild model misspecification	From $r = 3$ to $r = 8$, mean full-process and baseline-process regret decreased in all 60 tested location–scale combinations. Bootstrap intervals were entirely positive in 57 of 60 full-process and all 60 baseline-process combinations.	The principal replication pattern did not require the true mean and scale surfaces to belong exactly to the fitted second-order model class.
Grid-constrained positivity	QP reduced the mean diagnostic-grid negative-scale rate from 4.850% to 0.042% and improved full-process scale RMSE in 222 of 288 paired comparisons, but did not consistently reduce decision regret.	Decision regret remained largely unchanged despite improved admissibility and global scale-surface accuracy, demonstrating that surface diagnostics and decision quality are distinct criteria.

CC: BY-NC-ND

Changing the numerical floor had negligible effects, indicating that small-r behaviour came from estimator and surface instability rather than the chosen bound. Under the second correctly specified quadratic geometry, $r = 3 \rightarrow 8$ reduced full-process regret in every class by 11.3%–59.0%, while $r = 8 \rightarrow 10$ was not uniformly favourable (Appendix Table A5).

With omitted third-order structure, mean regret decreased from $r = 3$ to $r = 8$ in all 60 tested combinations under both targets; bootstrap intervals were positive in 57 full-process and all 60 baseline comparisons. For $r = 8 \rightarrow 10$, all mean changes were positive but only 40 and 39 intervals were positive, confirming weaker incremental evidence beyond eight.

All 14,400 QP fits met the feasibility rule. QP reduced the diagnostic-grid negative-scale rate from 4.850% to 0.042% and improved scale RMSE in 222 of 288 comparisons, but aggregate full- and baseline-process regret increased from 2466.38 to 2605.51 and from 61.31 to 65.80. Clipping reduced aggregate full-process regret to 2423.10 but barely changed baseline regret and left the raw polynomial unchanged. Thus positivity and global RMSE improvements did not guarantee better decisions; unconstrained OLS remained the common primary specification.

3.6. Finite-sample calibration and oracle-boundary assessments

Native small-sample calibration differed markedly. Under $N(0,1)$ at $r = 3$, mean scale estimates ranged from 0.620 for BS82 to 1.000 for corrected Q_n; SD, MAD, BWMV, and W24 had means 0.886, 0.673, 0.627, and 0.624. Mean calibration did not equal accuracy: Q_n RMSE was 0.825 versus 0.476 for SD, which had the lowest normal-model RMSE at every r (Appendix Table A7, Panel A).

Normal-model mean calibration preserved the principal result. The $r = 3 \rightarrow 8$ transition reduced both regrets in all 60 class–estimator–target combinations, and calibrated class reductions closely matched the published implementations. The descriptive winner changed in only 2 of 40 class–replication–target cells. Calibration improved neither regret nor surface accuracy uniformly and equalised only the $N(0,1)$ mean, not bias or variance under skewness and contamination. Nine refinements failed versus four in the primary implementation, with grid fallbacks retained (Appendix Table A7, Panel B).

The 39 scenarios yielded 13 oracle structures. Standard and dense searches agreed to within 4.09×10^{-12} in loss, and none was classified as a boundary optimum at tolerance 10^{-3}. The closest, for high two-sided contamination with $c = 0.25$ and $k = 25$, lay 0.002871 from the boundary. Frequent fitted boundary solutions therefore mainly reflected estimation and surface uncertainty, although near-boundary geometry could amplify small fitting errors.

3.7. Printing-process illustration

Data checks confirmed the 3^3 design, 81 published responses, point summaries, and rounded coefficients; the refitted standard-deviation surface remained positive on the validation grid. In the observed $r = 3$ data, every estimator selected $\hat{x}_1 = 1$, but the other coordinates differed, with maximum pairwise distance 0.205. Because each method supplied its own fitted loss, these observed decisions were not ranked by common true regret. Floor activation was 7.4% for SD and 29.6% for the other scales; negative-grid rates were zero for SD and 12.4%–19.3% otherwise (Appendix Table A8, Panel A).

The semi-synthetic stage produced all 14,400 results without optimization failure, and dense oracle checks agreed within 2.3×10^{-12}. From $r = 3$ to $r = 8$, full-process regret fell by 69.5%, 70.4%, and 63.5% under clean, mild one-sided, and mild two-sided conditions; baseline reductions were 69.5%, 66.2%, and 62.3%, with all bootstrap intervals positive (Appendix Table A8, Panel B). This extended the replication pattern to milder disturbances using a process-calibrated benchmark derived from published industrial data.

SD had the lowest mean regret in 22 of 24 condition–replication–target cells; BWMV led only the mild one-sided $r = 10$ cells. Boundary solutions occurred in 88.2% of fitted decisions, consistent with all three benchmark oracle optima satisfying $x_1^* = 1$. Thus boundary selection may reflect either

CC: BY-NC-ND

estimation uncertainty or genuine process geometry. Even under mild contamination, robust-scale performance remained dependent on the evaluation target and process geometry.

3.8. Practical implications

Application should begin by deciding whether extreme observations belong to routine process performance or represent an abnormal disturbance. That estimand, together with credible assumptions about frequency, magnitude, direction, and asymmetry, should guide process-specific comparison of location–scale estimator combinations rather than automatic selection of a robust method.

Replication should be planned explicitly. Within the fixed 27-point design, increasing replication from 3 to 8 requires 135 additional runs, while increasing it from 8 to 10 requires another 54. These additional runs should be weighed against time, material, labour, equipment, operational-loss costs, and the corresponding gains in decision quality. The decision-quality estimates reported here provide the quantitative basis for application-specific economic evaluation of replication.

Under clean, approximately Gaussian conditions, SD was a strong benchmark. Robust scale estimators were more competitive in selected settings with low and moderate contamination probabilities, especially for baseline preservation, but could downweight variation deliberately included in the full-process target. Location and scale should be chosen jointly. Regret should also be reported with floor activation, negative predictions, RMSE, boundary behaviour, and optimization diagnostics; no single surface diagnostic substituted for evaluation against the scientific target.

3.9. Study scope and future directions

By fixing the 27 design points and using nested samples, the study isolated the effect of within-design-point replication on the final operating decision. This design enabled direct paired comparisons across four replication sizes. Future work can extend the framework to fixed-budget comparisons involving additional, relocated, or unequally replicated design points and explicit economic costs.

Generalisability was examined through two quadratic geometries, mild response-surface misspecification, two evaluation targets, strong contamination stress tests, a milder benchmark disturbance, finite-sample calibration, and positivity treatments. These checks support the stability of the main replication finding within the investigated framework. Further extensions could consider additional targets, weighted losses, alternative positive scale models, tuning constants, and broader low-magnitude disturbance regimes.

The printing-process analysis linked the simulations to published industrial observations, while the higher-replication component remained semi-synthetic so that oracle regret could be evaluated. Prospective application-specific studies would provide the next step in translating the observed decision-quality gains into operational and economic recommendations.

4. CONCLUSION

Within the fixed 27-point design, replication size had a more consistent influence on decision quality than estimator identity. The principal replication gain remained strongly supported across the full-process and baseline-process targets, joint location–scale estimator choices, an alternative quadratic geometry, mild response-surface misspecification, finite-sample calibration, positivity treatments, and a process-calibrated printing benchmark. The study therefore contributes a decision-focused framework showing how replication gains propagate through fitted dual response surfaces to the final operating decision.

Evidence for the r=3→8 transition was especially stable, whereas the incremental benefit from r=8→10 was more scenario-dependent. Estimator performance was target-dependent: SD was strongest under clean data and often when extreme outcomes belonged to full-process variability, whereas resistant estimators were more competitive in selected low- and moderate-contamination settings aimed at preserving the baseline process. The finding that improved positivity or global RMSE did not necessarily reduce regret further demonstrates the importance of evaluating methods at the

CC: BY-NC-ND

decision level.

Taken together, the results support replication size as a central design lever and the joint selection of location and scale estimators according to the process distribution and the scientific meaning of extreme observations. The framework also provides a basis for future fixed-budget and application-specific economic evaluations.

Peer-review: External, Independent.

Acknowledgements:

-

Declarations:

1. Statement of Originality:

This work is original.

2. Author Contributions:

Concept: AK; ***Conceptualization:*** AK; ***Literature Search:*** AK; ***Data Collection:*** AK; ***Data Processing:*** AK; ***Analysis:*** AK; ***Writing – original draft:*** AK; ***Writing – review & editing:*** AK.

3. Ethics approval:

Not applicable.

4. Funding/Support:

This work has not received any funding or support.

5. Competing Interests:

The author declares no competing interests.

6. GenAI Usage Statement:

No GenAI tools were used at any stage of the study.

7. Sustainable Development Goals:

REFERENCES

[1] Genichi Taguchi. Introduction to Quality Engineering: Designing Quality into Products and Processes. Asian Productivity Organization, Tokyo, 1986. ISBN 9789283310839.

[2] George E. P. Box. Discussion of "off-line quality control, parameter design, and the Taguchi method". Journal of Quality Technology, 17(4):189–190, 1985. doi: 10.1080/00224065.1985.11978965.

[3] George E. P. Box and K. B. Wilson. On the experimental attainment of optimum conditions. Journal of the Royal Statistical Society: Series B (Methodological), 13(1):1–38, 1951. doi: 10.1111/j.2517-6161.1951.tb00067.x.

[4] Raymond H. Myers, Douglas C. Montgomery, and Christine M. Anderson-Cook. Response Surface Methodology: Process and Product Optimization Using Designed Experiments. John Wiley &

CC: BY-NC-ND

Sons, Hoboken, NJ, 4th edition, 2016. ISBN 9781118916018.

[5] G. Geoffrey Vining and Raymond H. Myers. Combining Taguchi and response surface philosophies: A dual response approach. Journal of Quality Technology, 22(1):38–45, 1990. doi: 10.1080/00224065.1990.11979204.

[6] Enrique Del Castillo and Douglas C. Montgomery. A nonlinear programming solution to the dual response problem. Journal of Quality Technology, 25(3):199–204, 1993. doi: 10.1080/00224065.1993.11979454.

[7] Dennis K. J. Lin and Wanzhu Tu. Dual response surface optimization. Journal of Quality Technology, 27(1):34–39, 1995. doi: 10.1080/00224065.1995.11979556.

[8] Karen A. F. Copeland and Peter R. Nelson. Dual response optimization via direct function minimization. Journal of Quality Technology, 28(3):331–336, 1996. doi: 10.1080/00224065.1996.11979683.

[9] Enrique Del Castillo, Shu-Kai Fan, and John Semple. Optimization of dual response systems: A comprehensive procedure for degenerate and nondegenerate problems. European Journal of Operational Research, 112(1):174–186, 1999. doi: 10.1016/S0377-2217(97)00382-2.

[10] Rui Ding, Dennis K. J. Lin, and Duan Wei. Dual-response surface optimization: A weighted MSE approach. Quality Engineering, 16(3):377–385, 2004. doi: 10.1081/QEN-120027940.

[11] Stephanie M. Pickle, Timothy J. Robinson, Jeffrey B. Birch, and Christine M. Anderson-Cook. A semi-parametric approach to robust parameter design. Journal of Statistical Planning and Inference, 138(1):114–131, 2008. doi: 10.1016/j.jspi.2007.05.018.

[12] Sarah E. Burke, Douglas C. Montgomery, Christine M. Anderson-Cook, and Rachel T. Silvestrini. Optimal design of experiments for dual-response systems with normal, binary, and Poisson distributions. Military Operations Research, 28(2):109–134, 2023. URL: https://www.jstor.org/stable/27227017.

[13] Matthias H. Y. Tan and Szu Hui Ng. Estimation of the mean and variance response surfaces when the means and variances of the noise variables are unknown. IIE Transactions, 41(11):942–956, 2009. doi: 10.1080/07408170902735418.

[14] Chanseok Park and Byung Rae Cho. Development of robust design under contaminated and non-normal data. Quality Engineering, 15(3):463–469, 2003. doi: 10.1081/QEN-120018045.

[15] Seong Beom Lee, Chanseok Park, and Byung Rae Cho. Development of a highly efficient and resistant robust design. International Journal of Production Research, 45(1):157–167, 2007. doi: 10.1080/00207540600649202.

[16] Gregory L. Boylan and Byung Rae Cho. Comparative studies on the high-variability embedded robust parameter design from the perspective of estimators. Computers & Industrial Engineering, 64(1):442–452, 2013. doi: 10.1016/j.cie.2012.10.012.

[17] Peter J. Huber. Robust estimation of a location parameter. The Annals of Mathematical Statistics, 35(1):73–101, 1964. doi: 10.1214/aoms/1177703732.

[18] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. Robust Statistics: The Approach Based on Influence Functions. John Wiley & Sons, New York, 1986. ISBN 9780471829218. doi: 10.1002/9781118186435.

[19] David A. Lax. Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. Journal of the American Statistical Association, 80(391):736–741, 1985. doi: 10.1080/01621459.1985.10478177.

[20] Peter J. Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. Journal

CC: BY-NC-ND

of the American Statistical Association, 88(424):1273–1283, 1993. doi: 10.1080/01621459.1993.10476408.

[21] Peter J. Rousseeuw and Sabine Verboven. Robust estimation in very small samples. Computational Statistics & Data Analysis, 40(4):741–758, 2002. doi: 10.1016/S0167-9473(02)00078-6.

[22] John A. Randal. A reinvestigation of robust scale estimation in finite samples. Computational Statistics & Data Analysis, 52(11):5014–5021, 2008. doi: 10.1016/j.csda.2008.04.016.

[23] Chanseok Park, Haewon Kim, and Min Wang. Investigation of finite-sample properties of robust location and scale estimators. Communications in Statistics—Simulation and Computation, 51(5):2619–2645, 2022. doi: 10.1080/03610918.2019.1699114.

[24] Gregory L. Boylan and Byung Rae Cho. Studies on the effects of estimator selection in robust parameter design under asymmetric conditions. Quality and Reliability Engineering International, 29(4):571–582, 2013. doi: 10.1002/qre.1406.

[25] Elif Kozan. A new robust estimation approach to dual response surface models. Duzce University Journal of Science and Technology, 13(3):1089–1097, 2025. doi: 10.29130/dubited.1537921.

[26] Agah Kozan. An interval-based robust response surface design using the W24 estimator for asymmetric and contaminated data. Symmetry, 18(4):644, 2026. doi: 10.3390/sym18040644.

[27] Norman J. Johnson. Modified t tests and confidence intervals for asymmetrical populations. Journal of the American Statistical Association, 73(363):536–544, 1978. doi: 10.1080/01621459.1978.10480051.

[28] Melis Zeybek. Confidence interval based quality improvement for non-normal responses. Celal Bayar University Journal of Science, 15(2):199–204, 2019. doi: 10.18466/cbayarfbe.518736.

[29] Melis Zeybek. Interval robust design under contaminated and non-normal data. Communications in Statistics—Theory and Methods, 49(22):5406–5418, 2020. doi: 10.1080/03610926.2019.1710198.

[30] Linhan Ouyang, Yizhong Ma, Jai-Hyun Byun, Jianjun Wang, and Yiliu Tu. An interval approach to robust design with parameter uncertainty. International Journal of Production Research, 54(11):3201–3215, 2016. doi: 10.1080/00207543.2015.1078920.

[31] Steven G. Gilmour and Luzia A. Trinca. Optimum design of experiments for statistical inference. Journal of the Royal Statistical Society: Series C (Applied Statistics), 61(3):345–401, 2012. doi: 10.1111/j.1467-9876.2011.01000.x.

[32] Gregory L. Boylan, Paul L. Goethals, and Byung Rae Cho. Robust parameter design in resource-constrained environments: An investigation of trade-offs between costs and precision within variable processes. Applied Mathematical Modelling, 37(4):2394–2416, 2013. doi: 10.1016/j.apm.2012.05.017.

[33] Stephen J. Walsh, Thomsen B. Bolton, and John J. Borkowski. Generating optimal designs with user-specified pure replication structure. Quality and Reliability Engineering International, 40(4):2019–2047, 2024. doi: 10.1002/qre.3506.

[34] Alan M. Gross. Confidence interval robustness with long-tailed symmetric distributions. Journal of the American Statistical Association, 71(354):409–416, 1976. doi: 10.1080/01621459.1976.10480359.

[35] David C. Hoaglin, Frederick Mosteller, and John W. Tukey, editors. Understanding Robust and Exploratory Data Analysis. John Wiley & Sons, New York, 1983. ISBN 9780471097778.

[36] Timothy C. Beers, Kevin Flynn, and Karl Gebhardt. Measures of location and scale for velocities in clusters of galaxies—a robust approach. The Astronomical Journal, 100:32–46, 1990. doi:

CC: BY-NC-ND

10.1086/115487.

[37] Alan M. Gross. Confidence intervals for bisquare regression estimates. Journal of the American Statistical Association, 72(358):341–354, 1977. doi: 10.1080/01621459.1977.10481001.

[38] Frank Wilcoxon. Individual comparisons by ranking methods. Biometrics Bulletin, 1(6):80–83, 1945. doi: 10.2307/3001968.

[39] Sture Holm. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2):65–70, 1979. URL: https://www.jstor.org/stable/4615733.

[40] Bradley Efron and Robert J. Tibshirani. An Introduction to the Bootstrap. Chapman & Hall, New York, 1993. ISBN 9780412042317. doi: 10.1201/9780429246593.

[41] Tim P. Morris, Ian R. White, and Michael J. Crowther. Using simulation studies to evaluate statistical methods. Statistics in Medicine, 38(11):2074–2102, 2019. doi: 10.1002/sim.8086.

❋ ❋ ❋

APPENDIX A. SUPPORTING NUMERICAL RESULTS

Table A1. Finite-sample corrections used for Q_n.

r	d_r	$2.21914 d_r$
3	0.99365	2.20505
5	0.84401	1.87298
8	0.66993	1.48667
10	0.72014	1.59809

Table A2. Mean full-process regret by scenario class, replication size, and scale estimator.

Scenario class	r	SD	MAD	Q_n	BWMV	W24	BS82
Clean	3	44.35	73.55	72.10	64.64	66.47	66.85
	5	27.26	49.27	41.51	37.14	37.52	39.23
	8	18.14	26.39	23.57	19.88	20.13	20.20
	10	15.33	26.16	20.72	17.91	18.05	18.17
Skew-only	3	45.84	55.90	57.78	52.96	54.66	53.80
	5	37.74	41.08	39.25	39.61	41.41	40.74
	8	34.39	32.19	28.63	31.36	32.23	32.17
	10	30.24	29.97	25.47	27.40	27.53	26.96
Low	3	1435.87	1260.26	1275.50	1200.39	1218.19	1203.90
	5	1356.58	928.63	888.76	873.69	866.46	864.97
	8	1281.77	718.43	675.19	658.75	661.66	659.34
	10	1241.85	687.44	644.51	616.83	617.99	619.74
Moderate	3	4067.48	4309.83	4152.53	4274.24	4095.75	4139.41
	5	3380.87	3588.50	3477.49	3596.96	3586.52	3534.70
	8	2593.57	3387.46	3117.31	3288.24	3244.24	3256.44
	10	2137.11	2871.71	2680.05	2747.23	2781.63	2744.77
High	3	5796.90	6803.69	6700.11	6957.51	6916.08	6893.14

CC: BY-NC-ND

Scenario class	r	SD	MAD	Q_n	BWMV	W24	BS82
	5	4063.04	6354.56	6252.67	6413.30	6343.34	6368.83
	8	2129.98	5843.64	5775.31	5755.19	5787.70	5746.83
	10	1715.75	5209.38	4855.73	5357.54	5302.01	5274.70

Note. Values are Monte Carlo means. Low, moderate, and high refer to contamination probability.

Table A3. Mean baseline-process regret by scenario class, replication size, and scale estimator.

Scenario class	r	SD	MAD	Q_n	BWMV	W24	BS82
Clean	3	44.35	73.55	72.10	64.64	66.47	66.85
	5	27.26	49.27	41.51	37.14	37.52	39.23
	8	18.14	26.39	23.57	19.88	20.13	20.20
	10	15.33	26.16	20.72	17.91	18.05	18.17
Skew-only	3	45.84	55.90	57.78	52.96	54.66	53.80
	5	37.74	41.08	39.25	39.61	41.41	40.74
	8	34.39	32.19	28.63	31.36	32.23	32.17
	10	30.24	29.97	25.47	27.40	27.53	26.96
Low	3	80.96	80.52	81.20	77.95	78.14	76.90
	5	75.17	65.09	62.87	63.52	63.06	62.92
	8	72.64	52.29	49.95	49.76	49.52	49.48
	10	73.03	49.84	48.26	46.22	46.25	46.42
Moderate	3	87.24	92.39	89.93	93.40	90.85	91.88
	5	82.51	82.49	79.36	85.88	85.36	85.13
	8	69.42	77.39	72.19	78.72	77.86	77.80
	10	64.01	69.75	66.97	69.94	70.69	70.33
High	3	85.37	89.76	88.88	91.88	92.06	93.17
	5	70.67	86.75	85.00	90.19	90.10	89.69
	8	54.49	83.76	78.98	83.01	83.09	83.25
	10	50.84	74.51	70.00	76.27	75.95	75.42

Note. Values are Monte Carlo means. Low, moderate, and high refer to contamination probability.

Table A4. Numerical and optimization diagnostics averaged over the six primary scale estimators.

Class	r	Floor activation (%)	Negative grid (%)	Negative at optimum (%)	Boundary (%)	Failure (%)
Clean	3	0.275	0.91	6.17	75.17	0.0000
	5	0.000	0.08	0.58	70.67	0.0000
	8	0.000	0.00	0.00	60.75	0.0000
	10	0.000	0.00	0.00	57.25	0.0000
Skew-only	3	0.384	2.74	18.71	63.92	0.0000

CC: BY-NC-ND

Class	r	Floor activation (%)	Negative grid (%)	Negative at optimum (%)	Boundary (%)	Failure (%)
	5	0.000	0.47	4.79	71.75	0.0000
	8	0.000	0.07	1.13	71.04	0.0000
	10	0.000	0.04	0.46	66.33	0.0000
Low	3	0.374	6.68	32.17	56.12	0.0000
	5	0.004	2.31	14.53	70.40	0.0000
	8	0.000	0.83	6.74	73.78	0.0000
	10	0.000	0.42	4.65	74.81	0.0000
Moderate	3	0.289	9.40	36.24	52.49	0.0000
	5	0.001	6.37	26.92	64.59	0.0000
	8	0.000	4.60	19.44	73.06	0.0000
	10	0.000	2.64	11.41	77.24	0.0000
High	3	0.202	9.40	35.67	56.69	0.0139
	5	0.002	7.89	31.42	64.67	0.0000
	8	0.000	7.39	34.33	67.74	0.0069
	10	0.000	6.94	30.87	68.56	0.0069

Note. Rates are descriptive averages over the six estimator-specific class summaries.

Table A5. Mean full-process regret under the alternative exactly quadratic process geometry.

Scenario class	$r = 3$	$r = 5$	$r = 8$	$r = 10$	$r = 3 \rightarrow 8$ reduction	$r = 3 \rightarrow 10$ reduction
Clean	49.63	31.89	20.37	17.17	59.0%	65.4%
Skew-only	36.30	27.57	21.44	22.29	40.9%	38.6%
Low	1012.79	742.24	626.84	584.68	38.1%	42.3%
Moderate	3605.66	3132.63	3051.36	2529.58	15.4%	29.8%
High	6005.70	5390.15	5325.20	5076.95	11.3%	15.5%

Note. Values are replicate-level class means averaged over W24, BWMV, and BS82. Low, moderate, and high refer to contamination probability.

Table A6. Implied standardized mean shifts and standard-deviation multipliers for the contaminated scenarios.

c	k	One-sided $\mathrm{E}(U)$	One-sided $\{\mathrm{Var}(U)\}^{1/2}$	Two-sided $\{\mathrm{Var}(U)\}^{1/2}$
0.05	12	0.60	2.806	2.869
0.05	25	1.25	5.543	5.682
0.15	12	1.80	4.411	4.764
0.15	25	3.75	8.988	9.739
0.25	12	3.00	5.307	6.096
0.25	25	6.25	10.879	12.546

CC: BY-NC-ND

Note. For two-sided contamination, E(U) = 0 in every row. The labels low, moderate, and high refer only to c = 0.05, 0.15, and 0.25, respectively.

Table A7. Finite-sample calibration and normal-model mean-calibration sensitivity.
Panel A. Finite-sample calibration under N(0,1): Monte Carlo mean [RMSE], based on 10^6 samples per replication size.

Estimator	$r = 3$	$r = 5$	$r = 8$	$r = 10$
SD	0.886 [0.476]	0.940 [0.347]	0.965 [0.264]	0.973 [0.234]
MAD	0.673 [0.644]	0.821 [0.512]	0.887 [0.383]	0.913 [0.348]
Q_n	1.000 [0.825]	0.999 [0.537]	1.000 [0.344]	1.000 [0.300]
BWMV	0.627 [0.587]	0.816 [0.417]	0.903 [0.295]	0.928 [0.257]
W24	0.624 [0.591]	0.814 [0.423]	0.903 [0.299]	0.929 [0.260]
BS82	0.620 [0.594]	0.812 [0.426]	0.902 [0.301]	0.929 [0.262]

Panel B. r = 3 → 8 relative regret reduction under the published and calibrated implementations.

Scenario class	Full-process: published	Full-process: calibrated	Baseline-process: published	Baseline-process: calibrated
Clean	66.9%	66.5%	66.9%	66.5%
Skew-only	40.5%	40.2%	40.5%	40.2%
Low	38.7%	39.7%	32.0%	32.6%
Moderate	24.6%	25.1%	16.9%	17.6%
High	22.5%	22.8%	13.8%	13.8%

Note. Panel A characterizes the native finite-sample behavior of the published implementations. In Panel B, values are class-level means averaged over the six scale estimators.

Table A8. Printing-process benchmark results.
Panel A. Selected settings from the observed three-replicate data.

Estimator	$\hat{x}_1$	$\hat{x}_2$	$\hat{x}_3$	$\hat{\mu}(\hat{\mathbf{x}})$	$\hat{\sigma}(\hat{\mathbf{x}})$	Floor at design points (%)	Negative diagnostic grid (%)
SD	1.000	0.079	-0.252	494.665	44.573	7.4	0.0
MAD	1.000	-0.078	-0.122	496.633	18.756	29.6	19.3
Q_n	1.000	-0.088	-0.132	492.894	26.897	29.6	19.3
BWMV	1.000	-0.002	-0.185	496.065	24.741	29.6	12.4
W24	1.000	-0.003	-0.184	496.112	24.591	29.6	12.9
BS82	1.000	-0.002	-0.185	496.128	24.721	29.6	13.5

Panel B. r = 3 → 8 regret reduction in the semi-synthetic illustration.

Condition	Full R_3	Full R_8	Full reduction (95% bootstrap CI)	Baseline R_3	Baseline R_8	Baseline reduction (95% bootstrap CI)
Clean	1260.17	383.95	69.5% (62.7, 75.2)	1260.17	383.95	69.5% (62.8, 75.2)
Mild one-sided	1904.58	563.65	70.4% (64.7, 75.1)	1451.67	491.35	66.2% (59.8, 71.3)
Mild two-	1505.47	549.99	63.5% (55.5,	1155.52	435.29	62.3% (54.5,

CC: BY-NC-ND

Condition	Full R_3	Full R_8	Full reduction (95% bootstrap CI)	Baseline R_3	Baseline R_8	Baseline reduction (95% bootstrap CI)
sided			70.1)			69.0)

Note. In Panel A, all six optimizations succeeded and all selected settings had $\hat{x}_1 = 1$. In Panel B, regrets were averaged over the six scale estimators within each Monte Carlo replication; intervals used 5000 paired bootstrap resamples.

CC: BY-NC-ND