



Calibrated Uncertainty Quantification of LPBF/DMLS SS316L Surface Roughness Using Leave-One-Setting-Out Benchmarking and Reliability Decisions

Valiveti Sivaramakrishna¹, C.V. Subbarao², Sushma Pappula³, Ravikiran Chitalapudi⁴, A. Sunanda⁵ and Venkata Phani Babu Vemuri⁶ ●.

¹Department of Mechanical Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad - 500090, Telangana, India University of, City-Country.

^{2,3}Department of Mechanical Engineering, PACE Institute of Technology & Sciences, Ongole-523272, Andhra Pradesh, India

⁴Department of Mechanical Engineering, MLR Institute of Technology, Hyderabad- 500043, India.

⁵Department of Mechanical Engineering, Sreenidhi Institute of Science and Technology, Yammampet, Hyderabad, India

⁶Department of Civil Engineering, Dadi Institute of Engineering & Technology, Anakapalle, Andhra Pradesh- 531002, India

● These authors contributed equally to this work

DOI: <https://www.doi.org/10.29194/NJES.29030552>

Received: Feb. 23, 2026

Revised: April 24, 2026

Accepted: May 29, 2026

Published: Sep. 20, 2026

Abstract

This study develops a decision-oriented uncertainty quantification methodology for analyzing as-built surface roughness in laser powder bed fusion/direct metal laser sintering (LPBF/DMLS) SS316L. A Taguchi L9 design was used to vary laser power (300-360 W), scan speed (800-1000 mm·s⁻¹), and layer thickness (20-80 μm), producing nine process settings with three independently fabricated specimens per setting, resulting in 27 total specimens. Surface roughness was measured by contact stylus profilometry using arithmetic mean roughness (Ra), root mean square roughness (Rq), and maximum profile height (Rz). The mean roughness varied within narrow ranges, Ra as 5.748-5.952 μm, Rq as 6.673-6.811 μm, and Rz as 28.828-28.892 μm, while within-setting scatter remained non-negligible, particularly for Rz. Probabilistic regression models were evaluated using leave-one-setting-out validation, negative log predictive density, interval coverage, calibration diagnostics, and reliability-driven accept/reject analysis. For Ra and Rq, a low-capacity linear mean model with pooled variance achieved the best predictive density, indicating limited transportable heteroscedastic structure under setting-wise extrapolation. For Rz, a nonlinear mean model with pooled variance performed best. Unregularized two-stage variance learning produced unstable uncertainty estimates, whereas shrinkage regularization improved calibration and reduced spurious setting-dependent variance effects. The decision analysis showed that calibration strongly influences process acceptance, reliability thresholds sharply reduced the number of accepted settings, and shrinkage-stabilized uncertainty produced a conservative and consistent decision frontier. The main contribution of this work is the integration of grouped validation, probabilistic calibration, variance-shrinkage modelling, and reliability-aware decision analysis for surface roughness qualification in LPBF/DMLS SS316L.

Keywords: DMLS, SS316L, Surface Roughness, LOSO, NLPD, ENCE

Corresponding author: Provide the corresponding author information and publisher here. E-mail address: sivaramakrishna_v@vnrvtiet.in

1. Introduction

Laser powder bed fusion (LPBF)/direct metal laser sintering (DMLS) is widely used to fabricate SS316L components requiring corrosion resistance, geometric flexibility, reduced lead time, and functional integration [1]. As-built surface roughness remains a major barrier to LPBF/DMLS component qualification because it can accelerate fatigue initiation, impair sealing performance, and increase dimensional rework. For process qualification, predicting only the average roughness is insufficient; the uncertainty associated with each process setting must also be quantified because manufacturing decisions depend on both accuracy and risk. Uncertainty-aware process improvement is

increasingly important in metal additive manufacturing because physical and operational sources of variability interact strongly, especially when experimental datasets are sparse [2]. In SS316L LPBF/DMLS, data-driven models are increasingly used to relate process parameters to quality outcomes; however, many studies still stop prediction accuracy and do not translate uncertainty into reliability-based acceptance decisions.

For SS316L, changes in scan speed (SS) and laser power (LP) alter microstructure and surface quality through their effects on thermal history and consolidation behaviour [3]. Vertical-wall roughness is particularly sensitive because staircase geometry, contour-track overlap, and side-skin solidification generate surface features that differ from those on top surfaces [4, 5]. Layer thickness (LT) further

complicates surface formation because higher productivity must be balanced against changes in remelting, interlayer bonding, and process stability [6]. Taguchi orthogonal arrays remain useful for screening dominant process factors with a limited number of experiments, particularly in metal processing studies where multi-factor parameter windows must be explored efficiently [7]. Measurement-driven and modelling-assisted DMLS studies have shown the value of data-centric approaches for improving SS316L surface quality [8]. However, uncertainty quantification studies in additive manufacturing show that process variability can be comparable to factor-driven changes; therefore, “best-setting” claims should be interpreted probabilistically when datasets are small and noisy [9]. Recent additive manufacturing study also emphasize that scatter should not be treated only as nuisance noise, because qualification decisions require conservative acceptance under uncertainty [10].

Small, grouped experiments require validation schemes that reflect practical deployment, where models must predict unseen process settings. Studies on structured data validation show that naive resampling can overestimate model performance when group dependencies are ignored [11]. Because variance learning is unstable when group sizes are small, empirical Bayes shrinkage can stabilize variance estimates by borrowing strength across groups and reducing spurious heteroscedasticity under low replication [5, 12]. Calibration is increasingly treated as a requirement for interpretable uncertainty estimates, rather than as an optional correction applied after prediction [13]. From a decision-support perspective, probabilistic regression can be expressed through downstream quantities such as the probability of exceeding or meeting a specification limit [14]. Probabilistic forecast evaluation considers both error magnitude and distributional accuracy, because point accuracy alone is insufficient for probabilistic decision-making [15]. Conformal methods have also been developed to construct reliable prediction intervals for complex models by targeting interval quality directly [16]. Isotonic recalibration has been used in low signal-to-noise settings, where small-sample models often become overconfident and require post-hoc correction [8, 17]. Multi-task learning strategies that combine mean prediction with residual or auxiliary structure have also been applied in high-stakes optimization, reflecting a broader shift toward modelling both central tendency and dispersion [18].

Large-scale neural regression studies show that probabilistic calibration can remain poor even when mean prediction accuracy appears high [6, 19]. Study on transformed-scale scoring show that evaluation choices can change model ranking, which supports the use of distributional metrics aligned with the decision context [7]. Machine learning studies increasingly view calibration as target-dependent, meaning that the appropriate calibration objective depends on the uncertainty analysis being performed [8]. Active learning studies further show that uncertainty estimates must be empirically credible, because miscalibration can guide sampling toward uninformative or overconfident regions [9]. Treatments of tail calibration showed that getting the extremes right is not automatic and requires diagnostics designed for the tails, not just for the bulk [12]. Probabilistic forecasting studies emphasize that calibration and sharpness must be evaluated together. This trade-off

determines whether uncertainty estimates improve or mislead downstream decisions [19].

Most SS316L LPBF/DMLS roughness studies quantify parameter effects and mean shifts, but few treat within-setting scatter as a primary modelling target under grouped extrapolation. In small Taguchi designs, heteroscedastic variance learning can be confounded by sampling noise, making it difficult to separate true process uncertainty from estimation instability. Even when uncertainty is estimated, it is rarely converted into explicit reliability conditions, such as accept/reject rules that show when feasible settings disappear under stricter qualification thresholds.

The aim of this study is to develop and benchmark uncertainty-aware surrogate models for predicting the as-built surface roughness of LPBF/DMLS-fabricated SS316L under laser power, scan speed, and layer thickness variations. Unlike conventional roughness prediction studies that mainly report point estimates or mean trends, the present work evaluates arithmetic mean roughness (R_a), root mean square roughness (R_q), and maximum profile height (R_z) under a leave-one-setting-out (LOSO) validation strategy, so that each prediction represents generalization to an unseen process condition rather than interpolation among replicated samples. The main contribution is the integration of grouped validation, probabilistic scoring, calibration diagnostics, shrinkage-stabilized heteroscedastic modelling, and reliability-driven accept/reject analysis into a single decision-oriented framework. This enables not only the prediction of roughness magnitude but also the quantification of predictive uncertainty and its practical effect on process qualification decisions.

2. Materials and Methods

2.1 Materials

SS316L is a low-carbon austenitic Cr-Ni-Mo stainless steel widely used in chemically aggressive and chloride-rich environments, including marine and biomedical applications. It offers high corrosion resistance, good weldability, and stable mechanical response [1]. Gas-atomized SS316L powder is commonly used as a benchmark material for dimensional accuracy and surface integrity studies in DMLS. The commercially provided, gas atomized SS316L powder was employed directly in this study. Prior to processing, the powder was maintained in sealed storage under ambient laboratory conditions and then loaded directly into the DMLS system for fabrication. The conceptual schematic of the Electro optical systems (EOS) M 290 DMLS system used in this study is provided as **Figure S1** in the Supplementary file and was adapted from [20].

2.2 Experimental design and DMLS fabrication conditions

Specimens were fabricated on a commercial DMLS system (EOS M 290) under an argon atmosphere, with the build environment controlled to maintain a low oxygen level throughout processing (chamber $O_2 < 0.1\%$). A Taguchi L9 orthogonal array was used to examine the effects of the main energy-input parameters on surface roughness. Laser power ($LP = 300, 330, 360$ W), scan speed ($SS = 800, 900, 1000$ mm·s⁻¹), and layer thickness ($LT = 20, 40, 80$ μm) were each studied at three levels, yielding nine unique LP-SS-LT settings. For each setting, three specimens were manufactured to assess repeatability and estimate within-setting uncertainty (total $n = 27$ samples). The complete experimental matrix and specimen-to-setting mapping are provided in Table 1. All other process parameters were kept constant to minimize confounding effects. Hatch spacing was fixed at 0.11 mm and a stripe scanning strategy with 67° inter layer rotation was employed with stripe width of 5 mm. Each layer was processed using a single contour pass with a 0.12 mm contour offset. The build plate was preheated to 80 °C to reduce thermal gradients and mitigate distortion, and shielding was maintained via controlled argon flow (3.5 L·min⁻¹). Powder was

deposited using a blade-type recoater with a polymer edge to promote uniform layer spreading. A single benchmark sample geometry (Figure (1)) [2] was used for all builds to eliminate part-to-part geometric variation. The geometry consists of an external base with four outer fillet faces (OF1-OF4) and a raised central boss incorporating internal fillet faces (IF1-IF4) and a circular internal feature (C1). For consistent feature comparisons, surface regions were defined as a priori solid regions (S1, S2), and elliptical features

(E1, E2). All samples were built using the same orientation and a consistent placement strategy, with parts positioned away from the build plate perimeter to minimize location dependent effects associated with local flow fields and edge thermal conditions. These controls ensured that roughness variations were mainly associated with the prescribed LP-SS-LT settings rather than geometry or build orientation. A photograph of the 27 printed specimens is provided as Figure S2 in the Supplementary file and was adapted from [2].

Table 1. The experimental matrix and specimen-to-setting mapping

Sample No.	Layer thickness (μm)	Laser power (W)	Scan speed (mm/s)	Sample No.	Layer thickness (μm)	Laser power (W)	Scan speed (mm/s)
1	20	300	800	15	80	330	900
2	20	300	800	16	20	330	1000
3	20	300	800	17	20	330	1000
4	40	300	900	18	20	330	1000
5	40	300	900	19	80	360	800
6	40	300	900	20	80	360	800
7	80	300	1000	21	80	360	800
8	80	300	1000	22	20	360	900
9	80	300	1000	23	20	360	900
10	40	330	800	24	20	360	900
11	40	330	800	25	40	360	1000
12	40	330	800	26	40	360	1000
13	80	330	900	27	40	360	1000
14	80	330	900	-	-	-	-

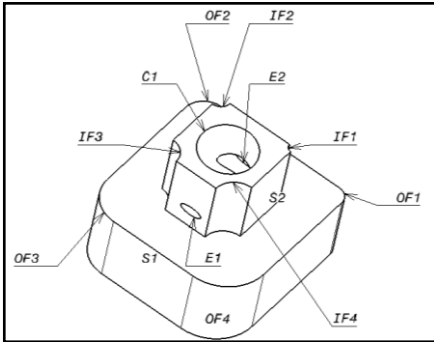


Figure 1. The benchmark sample geometry used for fabricating 27 samples [2]

2.3 Surface roughness measurement protocol

Surface roughness of the as-built SS316L specimens was measured using a contact stylus profilometer (Mitutoyo SurfTest SJ-210, Mitutoyo Corp., Japan). The utilized device possesses a diamond stylus (with a nominal tip radius of 2 μm, with a tip angle of 60°) and provides a high vertical resolution (0.002 μm) within a measurement range of 360 μm. The parameters of profile acquisition and filtering were in accordance with ISO 4287/ISO 4288. A cut-off length of 0.8 mm and evaluation length of 4.0 mm were used for all measurements. Measurements were taken, without post-processing, of the vertical wall surfaces to capture the as-built surface state generated by the LP-SS-LT parameter combinations. Five traces per specimen were recorded from uniformly spaced locations on the chosen wall area following a consistent traverse direction coincident with the build direction. The average of five traces for each specimen was calculated for arithmetic mean roughness (Ra), root mean square roughness (Rq) and maximum height of the profile or peak-valley extreme metric (Rz) to diminish the sensitivity for local heterogeneity. Mean values from three replicates produced independently were reported for each LP-SS-

LT configuration. The characterization of DMLS surface morphology by means of Ra, Rq and Rz together is superior to the use of a single roughness metric. Ra reflects the average size of an irregularity's amplitude whereas Rq is more sensitive to larger deviations. Rz measures extreme peak-valleys that link to melt-pool instability, partially fused particles, balling, and stairs that layer-wise step over.

2.4 Dataset structure

The dataset consists of nine LP-SS-LT settings created via Taguchi L9 design and three independently fabricated samples were created per setting (in total 27 samples). The surface roughness responses of each specimen are taken as (Ra, Rq, Rz). Let $y_{g,r}^{(m)}$ denote the observed roughness for metric $m \in (Ra, Rq, Rz)$, at process setting (group) $g \in (1, \dots, 9)$ and replicate $r \in (1, 2, 3)$. The predictor vector is shown in eq. (1). The predictor vector in question is represented in eq. (1). To facilitate a meaningful comparison of heteroscedasticity and predictive reliability across responses, each roughness metric m is modelled independently using an identical modelling pipeline.

$$X_g = [LP_g, SS_g, LT_g]^T \quad \dots(1)$$

2.5 Grouped validation protocol (Leave-One-Setting-Out, LOSO)

A grouped cross-validation strategy was adopted, whereby every fold holds out all replicates associated with one LP-SS-LT setting, to ensure the evaluation quantifies generalization to unseen process settings. Specifically, LOSO defines nine folds at fold k , the test set is $\mathcal{T}_k = (g = k, r = 1, 2, 3)$, and the training set is the complement $D \setminus \mathcal{T}_k$. This validation design directly probes whether a surrogate can deliver both accurate mean prediction and credible uncertainty when transferred to a new parameter setting. This is consistent with established practice for group structured data where within-group observations are not independent for generalization claims [3]. All

reported performance metrics were aggregated across folds, and uncertainty diagnostics were computed from out-of-fold predictive distributions, ensuring that every prediction used for benchmarking is out-of-sample with respect to its setting.

2.6 Two-stage heteroscedastic surrogate modelling

2.6.1 Stage 1: mean model

For each metric m , the first stage fits a low-capacity mean function $\mu^{(m)}(X)$ mapping LP-SS-LT to the conditional expectation of roughness, as shown in eq. (2). $\mu^{(m)}(\cdot)$ is restricted to a transparent functional class, to avoid variance inflation and unstable fold-to-fold behaviour.

$$y^{(m)} = \mu^{(m)}(X) + \epsilon^{(m)}(X), \mathbb{E}(\epsilon^{(m)}(X)|X) = 0 \quad \dots(2)$$

2.6.2 Stage 2: Variance model trained on setting-wise residual variance

After fitting the mean model $\mu^{(m)}(\cdot)$ on the training split in each LOSO fold, residuals are computed for each training observation:

$$e_{g,r}^{(m)} = y_{g,r}^{(m)} - \bar{\mu}^{(m)}(X_g) \quad \dots(3)$$

Because each process setting g contains 3 replicates, a setting-wise residual variance from the training data was estimated.

$$s_g^{2(m)} = \frac{1}{n_g-1} \sum_{r=1}^{n_g} (e_{g,r}^{(m)} - \bar{e}_{g,r}^{(m)})^2 \quad \dots(4)$$

The variance learning target was constructed as the log-variance,

$$z_g^{(m)} = \log(\tilde{s}_g^{2(m)} + \epsilon) \quad \dots(5)$$

where $\tilde{s}_g^{2(m)}$ is the shrinkage stabilized variance and ϵ is a small constant for numerical stability. The variance model $h^{(m)}(\cdot)$ is trained to map $X_g \mapsto z_g^{(m)}$ and predictions are mapped back to variance by

$$\hat{\sigma}^{2(m)}(X) = \exp(h^{(m)}(X)) \quad \dots(6)$$

2.6.3 Predictive distribution

Predictive distribution of each metric was approximated as Gaussian, which enables closed form likelihood scoring, prediction intervals, and probability of specification computations, using eq. (7).

$$y^{(m)}|X \sim \mathcal{N}(\hat{\mu}^{(m)}(X), \sigma^{2(m)}(X)) \quad \dots(7)$$

2.6.4 Variance shrinkage for stability

Because each setting has only three replicates, setting-wise empirical variances are noisy. To stabilize training targets for the variance stage and mitigate fold instability, an empirical Bayes shrinkage was introduced that pools setting variances toward a global value with strength governed by $\nu 0$, as shown in eq. (8).

$$\tilde{s}_g^{2(m)} = \frac{\nu 0 s_0^{2(m)} + (n_g - 1) s_g^{2(m)}}{\nu 0 + (n_g - 1)}, n_g = 3 \quad \dots(8)$$

where $s_g^{2(m)}$ is the sample variance of residuals within setting g (computed on training data only) and $s_0^{2(m)}$ is the pooled residual variance over all training points.

In each LOSO fold, $\tilde{s}_g^{2(m)}$ is computed on the training split only and then used to form the Stage-2 regression target $z_g^{(m)}$ in Eq. (5). When $\nu 0 = 0$, $\tilde{s}_g^{2(m)} = s_g^{2(m)}$ and M3 reduces to an unshrunk setting variance learner. This moderate variance method is a standard for variance estimation under small sample sizes and has a clear Bayesian interpretation via an inverse χ^2 prior on variances [5]. In

practice, $\nu 0$ is treated as a sensitivity/robustness parameter, where $\nu 0 = 0$ corresponds to no shrinkage, while larger $\nu 0$ increases pooling.

2.7 Benchmarking baselines and model capacity control

To ensure that any apparent heteroscedasticity is not an artefact of over-parameterized learners, all models were evaluated under the identical LOSO protocol using a controlled capacity benchmark suite. The benchmark includes both homoscedastic and heteroscedastic formulations so that improvements can be attributed to meaningful variance learning rather than to increased mean function flexibility alone. Because the experimental design comprises only nine settings (with three replicates per setting), model complexity was restricted a priori to promote small sample stability and to reduce the risk of spuriously overconfident uncertainty estimates, an established failure mode in probabilistic regression when data are limited [6]. Model M0 denotes a low-capacity linear mean model paired with a pooled (homoscedastic) Gaussian variance estimated from training residuals. Model M1 denotes a nonlinear mean model (Gradient boosting regressor based on histogram) paired with the same pooled Gaussian variance, isolating the benefit of mean function nonlinearity under a fixed uncertainty model. Model M2 is quantile baseline with Gaussian reconstruction for NLPD. M2 predicts conditional quantiles $\hat{q}_r(X)$ directly. To enable comparison under density-based scores (Negative log predictive density; NLPD), an approximate Gaussian distribution was constructed by matching the median and the central 90% interval implied by $\hat{q}_{0.05}(x)$ and $\hat{q}_{0.95}(x)$ denote the predicted 5th and 95th conditional quantiles. Then,

$$\hat{\mu}(x) = \hat{q}_{0.05}(x) \quad \dots(9)$$

and the predictive standard deviation is estimated by matching the central 90% Gaussian interval,

$$\hat{\sigma}(x) = \frac{\hat{q}_{0.95}(x) - \hat{q}_{0.05}(x)}{2z_{0.95}}, z_{0.95} = 1.645 \quad \dots(10)$$

This yields the approximate predictive density, which is then used to compute NLPD consistently with the density-based models.

$$p(y|x) = \mathcal{N}(\hat{\mu}(x), \hat{\sigma}^2(x)) \quad \dots(11)$$

Model M3 denotes a two-stage heteroscedastic surrogate in which Stage-2 learns a setting-wise variance map $X_g \mapsto \hat{\sigma}^2(X_g)$ using the log of the shrinkage-stabilized residual variance $\log(\tilde{s}_g^{2(m)} + \epsilon)$ as the regression target, with $\nu 0 \in \{0, 2, 5, 10\}$. This structured benchmark enables a controlled comparison between linear versus nonlinear mean response, pooled versus input dependent variance, and distribution free interval estimation versus parametric probabilistic scoring, all under the same grouped generalization test.

2.8 Probabilistic and point-prediction evaluation metrics

2.8.1 Point prediction accuracy

The mean accuracy was assessed using out-of-fold error criteria, such as Root mean square error (RMSE), as shown in eq. (12).

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{\mu}(x_i))^2} \quad \dots(12)$$

w2.8.2 Negative log predictive density

Probabilistic accuracy was quantified using the NLPD, a scoring rule for distribution forecasts, as shown in eq. (13).

$$NLPD = \frac{1}{N} \sum_{i=1}^N \left[\frac{1}{2} \log(2\pi\hat{\sigma}^2(x_i)) + \frac{(y_i - \hat{\mu}(x_i))^2}{2\hat{\sigma}^2(x_i)} \right] \quad \dots(13)$$

Lower NLPD indicates better-calibrated, sharper probabilistic predictions in aggregate and is a standard metric in probabilistic forecast evaluation [7].

2.8.3 Prediction interval coverage and width

The prediction interval (PI) for a nominal $100(1-\alpha)\%$ is estimated using eq. (14). The coverage is computed using eq. (15).

$$PI_{1-\alpha}(X) = \left[\hat{\mu}(x_i) - z_{1-\frac{\alpha}{2}} \hat{\sigma}(x_i), \hat{\mu}(x_i) + z_{1-\frac{\alpha}{2}} \hat{\sigma}(x_i) \right] \dots (14)$$

$$PICP_{1-\alpha} = \frac{1}{N} \sum_{i=1}^N \mathbb{I} y_i \in PI_{1-\alpha}(x_i) \dots (15)$$

The sharpness is summarized through the mean prediction interval width (MPI), using eq. (16). This pair explicitly operationalizes the calibration sharpness trade-off that highlights uncertainty evaluation [8].

$$MPIW_{1-\alpha} = \frac{1}{N} \sum_{i=1}^N \left(2z_{1-\frac{\alpha}{2}} \hat{\sigma}(x_i) \right) \dots (16)$$

2.9 Heteroscedasticity diagnostics under LOSO

Three complementary diagnostics were used, all computed from out-of-fold predictions, to ascertain if uncertainty varies systematically with LP-SS-LT in a way that is stable out-of-sample.

2.9.1 Variance-error agreement (uncertainty calibration)

The degree of calibration of uncertainty was measured by means of a regression analogue of calibration error, which uses the difference between predicted uncertainty and realized squared error. A pragmatic solution is Expected Normalized Calibration Error (ENCE) that bins the predictions by $\hat{\sigma}$ and compares the root mean predicted variance to RMSE within bins [9]. A lower ENCE means that the predicted uncertainty matches up with what has happened.

2.9.2 Evidence of heteroscedastic structure across settings

For each setting g , empirical out-of-fold error dispersion was compared against mean predicted uncertainty. A metric is considered genuinely heteroscedastic under LOSO if settings with larger predicted $\hat{\sigma}$ systematically exhibit larger realised squared errors or wider empirical residual spreads, and if this relationship persists across $\nu 0$ values rather than collapsing under mild shrinkage [5].

2.9.3 Stability to shrinkage as a robustness check

Because very small groups can induce spurious heteroscedastic fits, $\nu 0 \in \{0, 2, 5, 10\}$ is treated as a robustness sweep. A heteroscedastic conclusion is regarded as strong only when NLPD improves relative to homoscedastic baselines, PI coverage remains near nominal without excessive widening, and the setting-wise uncertainty pattern is stable as $\nu 0$ increases.

2.10 Decision-utility analysis and specification scenarios

2.10.1 Specification based decision scenarios

Uncertainty is ultimately valuable when it changes decisions. Therefore, the probabilistic surrogate is additionally evaluated for its ability to support accept/reject decisions under a roughness limit τ . For a fitted Gaussian predictive distribution, the probability that roughness meets the limit is shown in eq. (17).

$$p_{acc}(X) = \mathbb{P}(y \leq \tau | X) = \Phi \left(\frac{\tau - \hat{\mu}(X)}{\hat{\sigma}(X)} \right) \dots (17)$$

where $\Phi(\cdot)$ is the standard normal Cumulative Distribution Function (CDF). This turns each setting into a decision with an explicit reliability statement, consistent with probabilistic regression approaches that target decision functionals such as threshold probabilities [10].

2.10.2 Percentile specifications as scenarios

External engineering limits are often context-dependent such as component function, post-processing allowance, and inspection standard. When a universally agreed τ is not available, the

percentile-based specification is considered as the primary scenario and treat reliability thresholds as sensitivity analysis. In the present study, q90 spec scenario is taken as the primary scenario. Define τ as the 90th percentile of the observed roughness distribution for each metric, computed from the dataset, using eq. (18). This yields a transparent, stringent but attainable benchmark and avoids arbitrary limits. τ sets the roughness limits and r sets how confident the model must be $P((y \leq \tau | X) \geq r)$

$$\tau^{(m)} = \text{Quantile}_{0.90} \left(\left\{ y_i^{(m)} \right\}_{i=1}^N \right) \dots (18)$$

Reliability frontier was considered as the secondary scenario (0.80/0.85) for sensitivity. if $p_{acc}(X) \geq r$ for $r \in \{0.80, 0.85\}$. This evaluates the sensitivity of the conclusions to varying risk tolerance, avoiding the assumption that any single reliability threshold r is universally applicable.

2.10.3 Utility reporting

For each metric and LOSO fold, decisions are summarized by predicted acceptance probability, observed pass/fail against τ , and decision outcomes aggregated across folds. This reporting explicitly penalises overconfident models that approve settings with underestimated uncertainty, aligning with the studies on reliability and sharpness as joint requirements for decision grade uncertainty [12].

2.11 Implementation details

All analyses were implemented in Python (version 3.10) using standard scientific and Machine learning libraries, including NumPy for numerical operations, pandas for data handling and tabulation, and scikit-learn for regression modelling, grouped cross-validation and model evaluation to ensure that identical preprocessing, fold assignments, and scoring routines were applied consistently across all the response metrics (Ra, Rq, and Rz). All three replicate specimens associated with a given Setting ID were held out together as a single test fold, while the remaining eight settings (24 specimens) formed the corresponding training fold. This grouping prevented information leakage between train and test splits arising from replicate measurements produced under the same process conditions. For homoscedastic (pooled variance) baselines, the residual variance was estimated within each training fold only. Specifically, after fitting the mean model on the training split, the pooled variance $\hat{\sigma}_{0,(-k)}^2$ was computed from the training residuals and then used unchanged when scoring predictions on the held-out setting. This fold-wise estimation ensured that probabilistic scores and prediction intervals were strictly out-of-sample with respect to the held-out process setting. All stochastic components were controlled using a fixed random seed (random state=42), and the same seed values were reused across metrics and across model variants. Deterministic models were unaffected by seed choice but were nonetheless executed under the same controlled pipeline for uniformity. All intermediate and final outputs including out-of-fold predictions ($\hat{\mu}, \hat{\sigma}$), probabilistic benchmark summaries (RMSE, NLPD, coverage, and interval width), and decision utility tables were evaluated.

All equations presented in Sections 2.4-2.10 were used directly in the computational workflow: Eqs. (1)-(8) define the predictor structure, mean model, residual variance estimation, predictive distribution, and shrinkage-stabilized uncertainty model; Eqs. (9)-(11) define the quantile-to-Gaussian reconstruction used for the M2 baseline; Eqs. (12)-(16) define the point and probabilistic evaluation metrics; and the final specification-probability equations Eqs. (17)-(18) define the reliability-driven accept/reject decision analysis.

3. Results

Within this constrained EOS parameter set with fixed contour/hatch strategy, vertical-wall roughness is weakly sensitive to LP-SS-LT in mean, and the main challenge becomes scatter and decision-grade uncertainty.

3.1 Descriptive statistics and variability across settings

Across the nine settings ($n=3$), the mean roughness varied in a relatively narrow band, $R_a=5.748\text{--}5.952\text{ }\mu\text{m}$, $R_q=6.673\text{--}6.811\text{ }\mu\text{m}$, and $R_z=28.828\text{--}28.892\text{ }\mu\text{m}$ (Table 2). The within-setting scatter (SD across three replicates) was non-negligible relative to the between-setting spread, particularly for R_z , where the range of means ($0.064\text{ }\mu\text{m}$) was comparable to the typical within-setting standard deviation (SD) ($0.038\text{--}0.087\text{ }\mu\text{m}$). For R_a and R_q , the between-setting spread remained larger than the typical replicate scatter, but the overlap across settings was substantial (R_a SD:

$0.024\text{--}0.145\text{ }\mu\text{m}$; R_q SD: $0.007\text{--}0.077\text{ }\mu\text{m}$). These results indicate that, within the investigated parameter window, surface texture is not purely deterministic-by-setting but stochastic, replicate variability remains a component of the signal that subsequent modeling must respect. Although the three replicate specimens within each Taguchi L9 setting were fabricated under identical nominal LP-SS-LT conditions, they produced slightly different R_a , R_q , and R_z values because of inherent LPBF/DMLS process variability, local powder bed heterogeneity, melt pool instability, spatter adhesion, and measurement location effects. Therefore, replication was essential not to create new process settings, but to estimate within setting scatter, report mean $\pm$ standard deviation and support the uncertainty quantification and reliability-decision framework. A single specimen per setting would not have allowed estimation of setting-wise variance or calibration of predictive uncertainty.

Table 2. Setting level surface roughness statistics (mean $\pm$ SD, $n=3$ per setting)

Setting	LP (W)	SS (mm/s)	LT (μm)	R_a (μm)	R_q (μm)	R_z (μm)
1	300	800	20	5.7603 ± 0.145	6.6853 ± 0.006	28.8497 ± 0.051
2	300	900	40	5.7730 ± 0.076	6.7363 ± 0.048	28.8843 ± 0.087
3	300	1000	80	5.9220 ± 0.023	6.7993 ± 0.066	28.8400 ± 0.055
4	330	800	40	5.9520 ± 0.068	6.8077 ± 0.073	28.8917 ± 0.041
5	330	900	80	5.7483 ± 0.125	6.6733 ± 0.004	28.8377 ± 0.054
6	330	1000	20	5.7610 ± 0.056	6.7243 ± 0.048	28.8723 ± 0.081
7	360	800	80	5.9100 ± 0.013	6.7873 ± 0.061	28.8280 ± 0.052
8	360	900	20	5.9400 ± 0.068	6.7957 ± 0.072	28.8797 ± 0.044
9	360	1000	40	5.9193 ± 0.036	6.8110 ± 0.077	28.8783 ± 0.037

These magnitudes and trends are consistent with prior LPBF studies on stainless steels, which reported that roughness arises from coupled melt track morphology, partially fused particles/ spatter attachment, and scan vector dynamics [13]. The parameter changes do not always translate into large monotonic shifts when contour/strategy and orientation are held fixed. In particular, study by [14] on vertical SS316L surfaces showed that roughness mechanisms can switch with energy input, producing sensitivity that is highly contingent on the specific process window and surface class being interrogated. Study on LPBF by Chowdhury et al., [15] emphasized that apparent weak parameter effects can emerge with a relatively stable processing corridor and uncontrolled stochastic phenomena (powder scale heterogeneity, spatter redeposition, local thermal fluctuations) contribute variance comparable to factor-driven shifts. R_a / R_q exhibited modest setting level movement, suggesting that LP-SS-LT influence average texture but only weakly within this L9 array. R_z was effectively dominated by replicate scatter rather than by setting-to-setting separation, implying that extreme peak/valley behavior is either intrinsically less sensitive to LP-SS-LT or sensitive but masked by stochastic effects at the three-replicate resolution. This is precisely the region where uncertainty aware evaluation under LOSO is justified. If within-setting variability competes with between-setting differences, the credibility of any best setting claim could be expressed probabilistically rather than as a single point optimum [15].

3.2 LOSO probabilistic benchmarking

Across the nine LP-SS-LT settings, the LOSO benchmark provided a generalization test, where each fold holds out an entire parameter setting, so every reported score reflects true extrapolation across process conditions rather than interpolation among replicates. In this context, probabilistic skill was prioritized with NLPD as a scoring rule for distributional forecasts together with empirical interval coverage (Cov80/Cov95) and corresponding widths (Width80/ Width95). This diagnoses calibration sharpness trade-offs rather than point accuracy alone [16]. This is in line with small setting LPBF datasets where overconfident uncertainty models can appear accurate by RMSE yet fail under distributional metrics, and where reliability should be interpreted as conditional on process space rather than as a global average [8]. Table 3 shows that the simplest pooled variance baseline (M0) attains the best (lowest) NLPD for R_a (-0.392) and R_q (-1.006). This indicates that, under LOSO, these metrics demonstrate a stable mean response with comparatively weak residual heteroscedasticity at the setting level. For R_z , the best NLPD was obtained by the nonlinear mean model with pooled variance (M1, NLPD = -1.489), implying that R_z benefited more from flexible mean structure than from elaborate variance learning in the present data region. The quantile baseline (M2) achieved the lowest RMSE for R_a and R_z yet produced extremely poor NLPD for R_z (NLPD = 3.434), reinforcing that point error optimality does not guarantee credible uncertainty when evaluated by proper distributional scores [17].

Table 3. LOSO probabilistic benchmarking results

Metric	Model	RMSE	NLPD	Cov80	Cov95	Width80	Width95
R _a	M0	0.116	-0.392	0.593	0.889	0.218	0.334
	M1	0.176	1.023	0.407	0.556	0.208	0.235
	M2	0.108	-0.120	0.630	0.630	0.179	0.168
	M3 $\nu 0 = 0$	0.116	0.692	0.593	0.667	0.204	0.313
	M3 $\nu 0 = 2$	0.116	-0.266	0.593	0.889	0.218	0.333
	M3 $\nu 0 = 5$	0.116	-0.346	0.593	0.889	0.220	0.336
	M3 $\nu 0 = 10$	0.116	-0.372	0.593	0.889	0.220	0.336
R _q	M0	0.076	-1.006	0.630	0.852	0.147	0.228
	M1	0.106	-0.020	0.333	0.593	0.149	0.228
	M2	0.081	0.310	0.519	0.519	0.111	0.122
	M3 $\nu 0 = 0$	0.076	0.723	0.370	0.741	0.101	0.155
	M3 $\nu 0 = 2$	0.076	-0.945	0.556	0.889	0.143	0.218
	M3 $\nu 0 = 5$	0.076	-0.985	0.593	0.852	0.147	0.224
	M3 $\nu 0 = 10$	0.076	-0.997	0.630	0.852	0.148	0.226
R _z	M0	0.053	-1.461	0.741	0.963	0.154	0.236
	M1	0.052	-1.490	0.741	1.000	0.156	0.237
	M2	0.050	3.435	0.407	0.407	0.050	0.040
	M3 $\nu 0 = 0$	0.053	-1.479	0.741	0.963	0.146	0.224
	M3 $\nu 0 = 2$	0.053	-1.468	0.741	0.963	0.154	0.235
	M3 $\nu 0 = 5$	0.053	-1.464	0.741	0.963	0.155	0.237
	M3 $\nu 0 = 10$	0.053	-1.463	0.741	0.963	0.155	0.239

The observed ranking aligns with the broader probabilistic forecasting literature, where proper scoring rules such as the log score (and its negative form, NLPD) reward calibrated predictive distributions and penalize overconfident or mis-specified uncertainty, often disagreeing with RMSE conclusions when uncertainty is mis-modeled [18]. Likewise, interval diagnostics emphasized that useful uncertainty quantification is not simply wide enough intervals but appropriately sharp intervals that meet nominal reliability targets [11]. The strong degradation observed for M3 $\nu 0 = 0$ (Ra NLPD = 0.692; Rq NLPD = 0.723 with low coverages) is consistent with the known fragility of unregularized variance learning in small, grouped datasets, where variance models can collapse to under-dispersed predictions unless constrained [8].

Under the LOSO test, Ra and Rq behaved near-homoscedastic at the setting level, in the sense that a pooled variance dominates heteroscedastic variants by NLPD, while the unshrunk two-stage variance learner ($\nu 0 = 0$) is clearly unstable. In the case of Rz, the nonlinear mean (M1) provided the best probabilistic evidence, suggesting that for peak-valley sensitive roughness descriptors, mean structure across LP-SS-LT influence more than setting variance modulation. The probabilistic performance was achieved by simple pooled variance surrogates for Ra/ Rq, while Rz benefits from a nonlinear mean but still does not require aggressive heteroscedastic variance learning under NLPD.

3.3 Calibration diagnostics and heteroscedasticity under LOSO

Across the LOSO splits, the calibration diagnostics in Figure 2 indicate that the unregularized heteroscedastic model (M3, $\nu 0 = 0$) exhibits unstable uncertainty behaviour. Several operating points fall materially away from the ideal relationship between predicted dispersion and realized error, consistent with alternating over confidence (points above the line; predicted σ too small for the realized error) and over conservatism (points below the line; predicted σ too large). In contrast, the shrinkage variant (M3, $\nu 0 = 10$) produced visibly tighter agreement with reference line, particularly for Ra and Rq, and this improvement is reflected quantitatively by lower binned calibration discrepancy (Table 4). For Rz, both variants remain comparatively misaligned, which is consistent with the very low realized errors in several settings coupled with a predicted σ floor. These patterns are expected when heteroscedastic noise is learned under group shift and the variance head has enough flexibility to fit idiosyncrasies of the training folds. Probabilistic prediction theory separates calibration from sharpness, and stresses that sharpening without calibration is not a virtue. Scoring rules, such as log score/NLPD, penalize this failure mode directly because overconfident distributions incur large loss on even a small number of misses [19]. Under $\nu 0 = 0$, without stabilization, variance learning can become weakly identified in small-to-moderate samples and under distributional shifts induced by LOSO, yielding

either spurious heteroscedastic structure or an overly optimistic σ for held-out settings. Regularization/ recalibration mechanisms are specifically motivated as a remedy to restore auto-calibration behaviour when the raw predictive uncertainty is either too noisy or systematically biased [21].

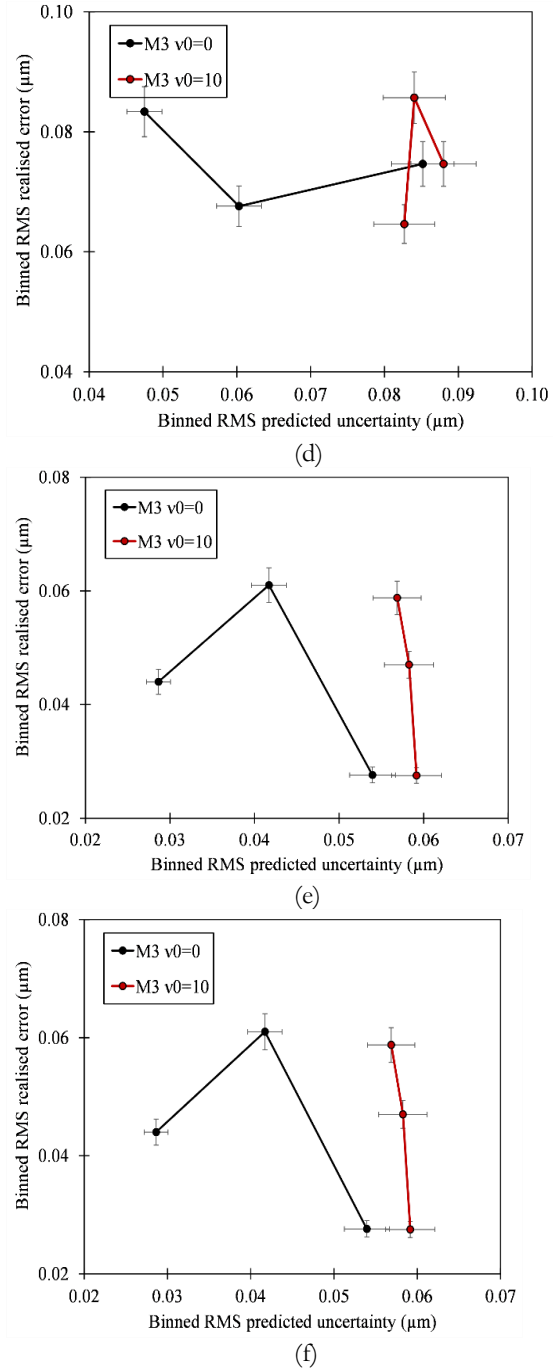
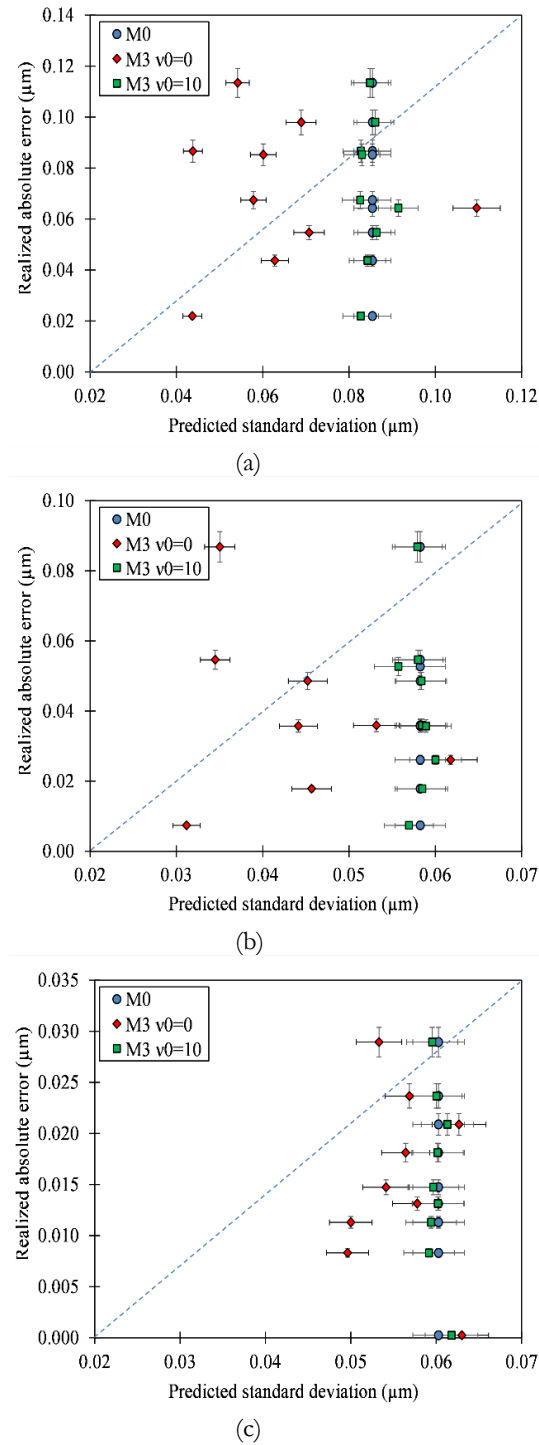


Figure2. LO SO calibration diagnostics for predictive uncertainty (Ra, Rq, Rz): (a-c) pointwise predicted standard deviation vs realized absolute error and (d-f) ENCE binned RMS vs RMS realized error comparing M3 $\nu 0 = 0$ and M3 $\nu 0 = 10$

Figure 3 diagnoses whether the predicted uncertainty of the models track observed within-setting variability. The LO SO uncertainty estimates are consistent with within-setting dispersion (strong Spearman ρ in Table 4), which infers that settings that are intrinsically more variable tend to receive higher predicted σ . However, this ranking success comes with a critical deviation, where $\nu 0 = 10$ largely compresses the σ dynamic range (σ spans only 0.082-0.091 μm for Ra and 0.057-0.061 μm for Rq), so it distinguishes more vs less uncertain settings mainly by small σ shifts rather than by large separations. $\nu 0 = 10$ improved calibration stability but partially sacrifices heteroscedastic resolution, a classic shrinkage trade-off. Meanwhile, $\nu 0 = 0$ spreads σ more widely, but the spread is not reliably aligned with realized error under LO SO.

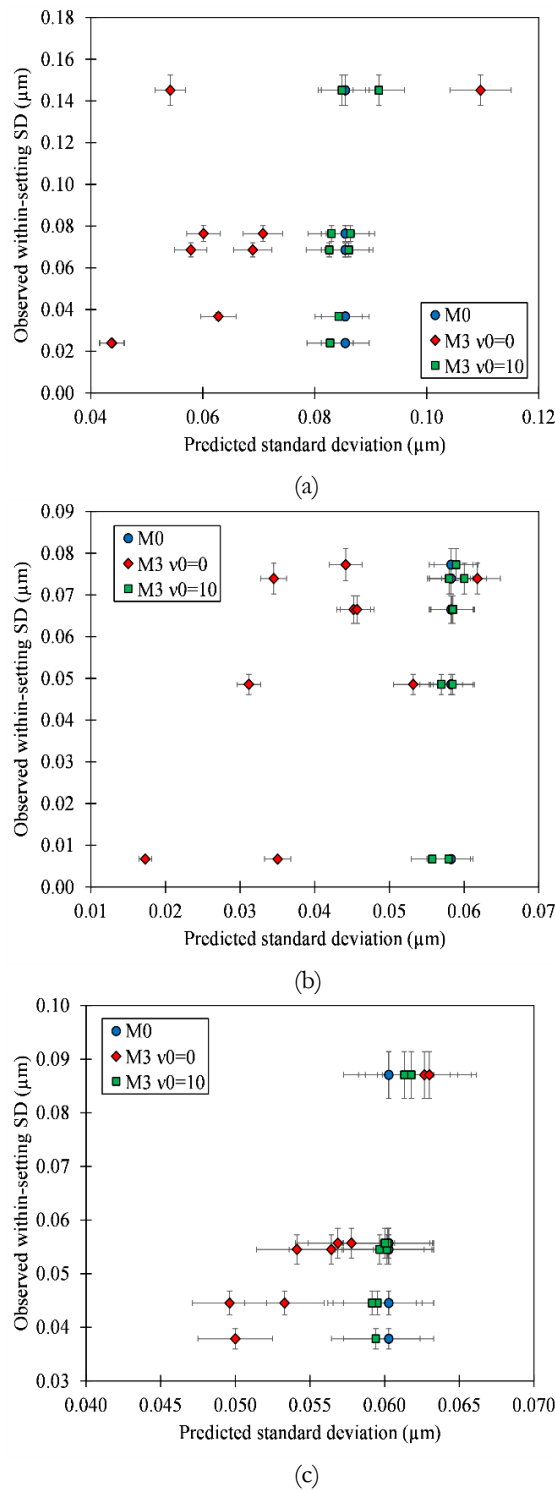


Figure 3. LOSO heteroscedasticity under grouped shift: predicted standard deviation vs observed within-setting standard deviation across settings under M0, M3 $v_0 = 0$ and M3 $v_0 = 10$ for (a) Ra, (b) Rq and (c) Rz

ENCE refers the average binned relative difference between the RMS realized error and the RMS predicted uncertainty (lower is better). M0 is not included in the ENCE and rank correlation columns because the value of sigma is relatively constant under the pooled baseline. The σ values are $0.085 \mu\text{m}$ (Ra), $0.060 \mu\text{m}$ (Rq), and $0.059 \mu\text{m}$ (Rz), so it cannot express heteroscedasticity across settings.

3.4 Decision utility outcomes and reliability frontier

The acceptance/rejection decision that derives from a utility analysis converts probabilistic roughness predictions into a quantile threshold specification with a reliability condition. A setting is accepted operationally only when the probability of meeting spec is predicted to exceed a reliability level r . This was swept to form the reliability frontier. In the LOSO framework, new groups are structurally dissimilar (left out build), so any excess optimism in predicted uncertainty gets immediately registered as over confidence in acceptance. This framing is consistent with the core theoretic decision logic of reject option learning, namely abstain (reject) unless the model can justify acceptance at target confidence. And keep decisions linked to calibrated predictive distributions, not point errors [19]. The acceptance probability and decision flip counts across reliability thresholds are presented in Table 5.

Across Ra and Rq, the pooled baseline (M0) produced a conservative frontier as the reliability threshold tightens from $r = 0.8 - 0.95$. The acceptance set collapses quickly, and at high reliability levels (≥ 0.90) the baseline effectively rejected all settings in this run. The shrinkage variant (M3, $v_0 = 10$) tracks the decision boundary of baseline almost exactly across all r , indicating that the added heteroscedastic machinery is not buying extra accepts once reliability is enforced. Its practical value stabilizing uncertainty enough that it does not create spurious approvals. In contrast, the unshrunk heteroscedastic model (M3, $v_0 = 0$) expands the acceptance region materially, including at strict reliability levels, producing multiple decision flips (accept under M3 but reject under the pooled baseline). For Ra and Rq, M0 collapses to zero acceptance by $r \geq 0.90$; for Rz, M0 retains a small accept set at $r = 0.90$ but collapses to zero by $r = 0.95$. The corresponding decision flips quantify where heteroscedastic calibration changes the accept/reject classification relative to M0, revealing that variance learning mainly affects the boundary cases rather than uniformly shifting all settings. This pattern is a demonstration of miscalibration by under dispersion. If $\hat{\sigma}$ is too small, the model inflates P_{pass} , and the frontier becomes artificially permissive. These findings are in line with study by Yang et al. [22], where probabilistic scoring and calibration diagnostics are emphasized in probabilistic forecasting and downstream decision methods. Sharpness without calibration is fragile, and decision quality degrades when confidence is overstated.

The frontier behavior of the parameters demonstrated in this study is consistent with the reject-option principle i.e., increasing the required confidence r should monotonically reduce acceptance, and models that are better calibrated should exhibit fewer unjustified accepts at the right tail of r . Classical reject-option studies formalized this as a risk coverage trade-off, and probabilistic forecasting theory emphasizes that proper scoring rules reward distributions that are simultaneously sharp and statistically accurate. The setting-level accept/reject outcomes at $r = 0.80$ and 0.85 for Ra, Rq and Rz are presented in Tables 6, 7 and 8 respectively. The setting-level accept/reject outcomes at $r = 0.90$ and 0.95 for Ra, Rq and Rz are presented in Tables S1, S2 and S3, respectively in Supplementary file. The heteroscedastic model, in the present study, exhibited that added flexibility ($v_0 = 0$) does not improve reliable acceptance rather, it increases acceptance by over asserting certainty, creating flips relative to the conservative pooled baseline. This is well aligned with empirical findings across probabilistic modeling domains where naively estimated uncertainties can look good under point metrics while failing calibration checks and leading to overly aggressive decisions once probabilities are included [23].

Table 4. Summary of LOSO calibration and heteroscedasticity diagnostics

Target	Model	Binned ENCE	Spearman ρ (standard deviation vs realized absolute error)	Spearman ρ (standard deviation vs observed within-setting standard deviation)	$\sigma_{\min}$	$\sigma_{\max}$
Ra	M3 $\nu 0 = 0$	0.333	0.000	0.809	0.043	0.111
Ra	M3 $\nu 0 = 10$	0.130	0.365	0.994	0.082	0.091
Rq	M3 $\nu 0 = 0$	0.352	-0.050	0.843	0.018	0.059
Rq	M3 $\nu 0 = 10$	0.185	-0.150	0.990	0.057	0.061
Rz	M3 $\nu 0 = 0$	0.976	-0.258	0.945	0.033	0.063
Rz	M3 $\nu 0 = 10$	0.876	-0.179	0.977	0.058	0.062

Table 5. Acceptance probability and decision flip counts across reliability thresholds

Metric	Reliability level r	Accept probability (M0)	Accept probability (M3 $\nu 0 = 0$)	Accept probability (M3 $\nu 0 = 10$)	Decision flips vs M0 ($\nu 0 = 0$)	Decision flips vs M0 ($\nu 0 = 10$)
Ra	0.80	0.556	1	0.556	4	0
Ra	0.85	0.333	0.556	0.333	4	0
Ra	0.90	0	0.444	0	4	0
Ra	0.95	0	0.111	0	1	0
Rq	0.80	0.111	0.667	0.111	5	0
Rq	0.85	0	0.333	0	3	0
Rq	0.90	0	0	0	0	0
Rq	0.95	0	0	0	0	0
Rz	0.80	1	1	1	0	0
Rz	0.85	0.667	0.889	0.667	0	0
Rz	0.90	0.222	0.333	0.222	0	0
Rz	0.95	0	0	0	0	0

*Acceptance probability = fraction of settings accepted at reliability r ; flips are vs pooled baseline M0.

Table 6. Setting level accept/reject outcomes at $r = 0.80$ and 0.85 (Ra)

LP (W)	SS (mm/s)	LT (μm)	P_{pass} M0	P_{pass} M3($\nu 0 = 0$)	P_{pass} M3($\nu 0 = 10$)	M0@0.80	$\nu 0 = 0@0.80$	$\nu 0 = 10@0.80$	M0@0.85	$\nu 0 = 0@0.85$	$\nu 0 = 10@0.85$
300	800	20	0.811	0.980	0.811	A	A	A	R	A	R
300	900	40	0.801	0.978	0.801	A	A	A	R	A	R
300	1000	80	0.702	0.946	0.702	R	A	R	R	A	R
330	800	40	0.801	0.978	0.801	A	A	A	R	A	R
330	900	80	0.876	0.992	0.876	A	A	A	A	A	A
330	1000	20	0.815	0.981	0.815	A	A	A	R	A	R
360	800	80	0.722	0.953	0.722	R	A	R	R	A	R
360	900	20	0.717	0.951	0.717	R	A	R	R	A	R
360	1000	40	0.865	0.991	0.865	A	A	A	A	A	A

*A=Accept, R=Reject

Table 7. Setting level accept/reject outcomes at $r = 0.80$ and 0.85 (Rq)

LP (W)	SS (mm/s)	LT (μm)	P_{pass} M0	P_{pass} M3($\nu 0 = 0$)	P_{pass} M3($\nu 0 = 10$)	M0@0.80	$\nu 0 = 0@0.80$	$\nu 0 = 10@0.80$	M0@0.85	$\nu 0 = 0@0.85$	$\nu 0 = 10@0.85$
300	800	20	0.768	0.965	0.768	R	A	R	R	A	R
300	900	40	0.797	0.975	0.797	R	A	R	R	A	R
300	1000	80	0.843	0.985	0.843	A	A	A	R	A	R
330	800	40	0.793	0.974	0.793	R	A	R	R	A	R

330	900	80	0.836	0.983	0.836	A	A	A	R	A	R
330	1000	20	0.795	0.975	0.795	R	A	R	R	A	R
360	800	80	0.793	0.974	0.793	R	A	R	R	R	R
360	900	20	0.836	0.983	0.836	A	A	A	R	R	R
360	1000	40	0.806	0.978	0.806	R	A	R	R	R	R

*A=Accept, R=Reject

Table 8. Setting level accept/reject outcomes at $r = 0.80$ and 0.85 (R_z)

LP (W)	SS (mm/s)	LT (μm)	P_{pass} M0	P_{pass} M3($\nu 0 = 0$)	P_{pass} M3($\nu 0 = 10$)	M0@0.80	$\nu 0 =$ 0@0.80	$\nu 0 =$ 10@0.80	M0@0.85	$\nu 0 =$ 0@0.85	$\nu 0 =$ 10@0.85
300	800	20	0.848	0.864	0.849	A	A	A	R	A	R
300	900	40	0.864	0.855	0.86	A	A	A	A	A	A
300	1000	80	0.898	0.908	0.899	A	A	A	A	A	A
330	800	40	0.867	0.896	0.87	A	A	A	A	A	A
330	900	80	0.900	0.924	0.903	A	A	A	A	A	A
330	1000	20	0.831	0.82	0.825	A	A	A	R	R	R
360	800	80	0.903	0.915	0.904	A	A	A	A	A	A
360	900	20	0.834	0.881	0.838	A	A	A	R	A	R
360	1000	40	0.851	0.895	0.855	A	A	A	A	A	A

*A=Accept, R=Reject

4. Discussion

4.1 Stable texture metrics vs. extreme-defect metrics: mechanistic separation of R_a/R_q and R_z

Across the investigated LP-SS-LT window, the descriptive statistics suggest a stable corridor region for average amplitude roughness. R_a and R_q shift only modestly across settings, and their within-setting scatter, while nontrivial does not explode in particular corners of the design space (Table 2). In that region, R_a/R_q are dominated by repeatable meso-scale texture contributions that persist across settings but do not change their stochastic intensity when contour/strategy/ orientation was held constant. This is consistent with LPBF surface formation studies on stainless steels, where vertical wall roughness is often governed by a mixture of relatively stationary phenomena (track overlap, staircase geometry, scan strategy) and a weaker parameter-to-roughness transfer function once a stable processing corridor is entered [13]. Within such a corridor, a pooled (setting-invariant) noise model can be hard to beat under grouped generalization, because the residual variance does not exhibit strong, repeatable input dependence once the model is forced to extrapolate to an unseen setting under LOSO [14]. R_z , however, is not just a larger R_a but an extreme-sensitive descriptor. R_z is disproportionately driven by rare peak attachments and deep valleys caused by sporadic events spatter redeposition, agglomerate attachment, balling traces, local lack-of-fusion pockets, and intermittent melt-pool instability. The occurrence rate can be weakly linked to energy input but strongly modulated by local powder scale randomness and transient thermal/ fluid instabilities [15]. That mechanistic difference explains why R_z can look qualitatively different even when its setting-wise means vary in a narrow band (Table 2). In the present study, this manifests as R_z having within-setting scatter comparable to its between-setting mean spread, which makes the inference problem intrinsically variance-dominated (Table 2) and therefore especially sensitive to how uncertainty is learned and regularized under LOSO.

4.2 Small-group pathology under LOSO: shrinkage as a calibration safeguard.

The calibration diagnostics and benchmarking collectively point to a small group pathology. With three replicates per setting, setting-wise

variance is weakly identified, and an unconstrained variance learner can easily become under-dispersed (too-small σ) or unstable fold when confronted with grouped shift [6]. By pooling noisy group variances toward a global residual variance with strength controlled by $\nu 0$, the variance targets become less idiosyncratic to whichever settings happen to be in the training split, thereby improving fold-to-fold stability and reducing overconfident tail behavior [5]. In the present study, this mechanism is visible in two coupled ways: the unregularized heteroscedastic model (M3, $\nu 0 = 0$) degrades NLPD and coverage for R_a and R_q relative to pooled baselines, while shrinkage (e.g., $\nu 0 = 10$) restores calibration behavior and makes the heteroscedastic model track the conservative decision boundary of M0 (Tables 3-5). Shrinkage compresses the σ dynamic range effectively admitting that the data do not support strong, stable heteroscedastic structure at the setting level under LOSO. So, the model becomes less tempted to errors by learning spurious variance patterns [8].

For R_z , despite the mechanistic possibility of extreme variability, the LOSO evidence suggests that elaborate variance learning is not reliably rewarded by proper scoring, and the best NLPD is achieved by a nonlinear mean with pooled variance (M1), while M3 variants remain close but do not dominate (Table 3). With nine settings and three replicates each, the dataset can support nonlinearity in the conditional mean for R_z , but it cannot robustly identify a fine-grained input variance surface that generalizes to an unseen setting. Shrinkage therefore functions as a reality check that prevents overfitting variance structure that is not transportable across settings [5].

4.3 Benchmarking interpretation

The benchmarking distinguishes point accuracy from distributional adequacy. RMSE verifies whether $\hat{\mu}$ is close to y on average, whereas NLPD evaluates the entire predictive distribution and penalizes both mis-centered means and mis-calibrated variances, especially overconfidence, because even a small number of outcomes incurs large log loss [24]. A model can look competitive (or even best) under RMSE while performing poorly under NLPD when its uncertainty is too narrow or structurally mis-specified. In the present study, related to R_z , the quantile baseline (M2) attains low RMSE but

catastrophically poor NLPD (Table 3), implying that its implied uncertainty representation is not aligned with the empirical error behavior under LOSO. Conversely, the pooled variance baselines can win NLPD for Ra and Rq precisely as a constant σ is closer to the truth than a noisy learned σ surface that alternates between under and over dispersion across folds (Tables 3, 4). This divergence is the methodological contribution. In small, grouped DMLS datasets, distributional scoring should be treated as the primary arbiter of uncertainty model quality, and RMSE should be interpreted as a necessary but not sufficient condition for decision grade usefulness [25].

4.4 Decision framing

The decision utility operationalizes uncertainty as it converts predictions into acceptance probabilities under a specific τ and then enforces a reliability requirement r , which creates a reliability frontier (Table 5). In this method, miscalibration has a direct, mechanical consequence. If $\hat{\sigma}$ is underestimated (under-dispersion), then $\Phi((\tau - \hat{\mu}(x))/\hat{\sigma}(x))$ is inflated, and settings cross the τ -threshold too easily. The model becomes artificially permissive, generating acceptance that is not supported by out-of-sample evidence. The results show this failure mode sharply for Ra and Rq: M3 with $\nu 0 = 0$ increases acceptance probability relative to M0 across r and triggers multiple decision flips, while the shrinkage variant ($\nu 0 = 10$) collapses those flips to zero and essentially reproduces the conservative frontier of the pooled baseline. Bjerregård et al. [11] proposed that the results of the quantification of decision grade uncertainty is not enough. Uncertainties must be calibrated in the tails relevant to threshold decision; otherwise, the downstream accept/reject policy becomes systematically biased toward false approvals. Recalibration and shrinkage stabilization, therefore, are not cosmetic, but rather theoretical decision safeguards against overconfident approvals under distribution shift [8].

For Rz, the frontier behaves more smoothly, with high acceptance at $r = 0.80$ and a monotone contraction as r tightens (Table 5), which is consistent with the reject option principle. Increasing r should reduce acceptance unless the probability mass of the model below τ is genuinely large [19]. In the present study, more flexible uncertainty ($\nu 0 = 0$) does not create more reliable acceptance but it creates more acceptance by over asserting certainty, and the only heteroscedastic variant that remains defensible under LOSO is the one whose regularization prevents spurious probability inflation (Tables 4, 5).

5. Practical implications

This study has direct practical implications for process qualification and parameter selection in LPBF/DMLS when replicate scatter competes with setting-to-setting differences. The best mean predictor is not necessarily the safest basis for approving a parameter setting, because accept/reject decisions are driven by the tails of the predictive distribution, not by RMSE alone. In the Taguchi L9 array (LP 300-360 W, SS 800-1000 mm·s⁻¹, LT 20-80 μ m; $n = 27$ with three replicates/setting), the between-setting movement in roughness is narrow and overlaps substantially, especially for Rz where within-setting SD is comparable to the range of setting means. For Ra and Rq, the LOSO evidence indicates that a pooled (homoscedastic) variance model paired with a low-capacity mean provides the most reliable probabilistic behaviour (best NLPD: Ra - 0.392; Rq -1.006). Operationally, this means that in replicated datasets, treating the residual scatter as setting-invariant is often the safest default for average amplitude roughness metrics, because any setting-wise heteroscedastic pattern does not transport cleanly to an unseen setting under grouped shift [24]. For Rz, performance improves primarily through mean function flexibility (best NLPD

from nonlinear mean with pooled variance: -1.490). Hence, if the function is Rz, the practical lever is often to improve the mean model (feature/ structure) rather than to overfit a variance surface from sparse replicates.

When a two-stage heteroscedastic learner is used with only three replicates per setting, unregularized variance learning becomes a known failure mode. It can produce under-dispersed, damaging coverage and inflating decision probabilities, even if RMSE appears acceptable. The unshrunk heteroscedastic model degraded NLPD for Ra/Rq (Ra NLPD 0.692; Rq NLPD 0.723), while empirical Bayes shrinkage restored calibration (Ra ENCE 0.333 to 0.130; Rq ENCE 0.352 to 0.185). If the objective is learning a variance model from sparse replicate groups, treat shrinkage (or an equivalent regularization/ recalibration step) as a process control knob, not a modeling preference otherwise, risk converting sampling noise into confidence amplifies [25].

Under percentile specifications (q90 scenario) and a reliability frontier sweep, the pooled baseline becomes highly conservative: for Ra and Rq, acceptance collapses to zero at $r \geq 0.90$ in this dataset. This is not a model defect but an operational signal that, within the studied corridor and replicate variability, the data do not justify high confidence approvals for Ra/Rq without either loosening the reliability requirement, widening the process window / adding new controllable factors, or adding more replicates to reduce uncertainty. The unshrunk heteroscedastic model inflated acceptance probabilities and introduced multiple decision flips relative to the pooled baseline; for example, at $r = 0.90$ for Ra, acceptance moved from 0 (pooled) to 0.444 (unshrunk heteroscedastic). In contrast, the stabilized shrinkage variant eliminated these flips and essentially reproduced the conservative frontier. The heteroscedastic modelling should never be deployed in qualification workflows unless calibration is verified under grouped generalization, because miscalibration does not merely reduce accuracy, it creates false approvals.

6. Limitations and boundary conditions of the study

The following are the limitations and boundary conditions:

- Taguchi L9 window with nine LP-SS-LT settings and three replicates per setting was adopted, which limits the identifiability of fine-grained variance structure and reduces the capacity to distinguish true heteroscedasticity from sampling noise.
- Hatch spacing, scan strategy contour offset, contour passes, preheat temperature conditions were held constant. Therefore, the reported calibration/decision behavior should be interpreted as conditional on this fixed build.
- The study does not quantify uncertainty contributions from orientation, support strategy, part position, local gas flow, or edge thermal conditions.

The probabilistic scoring and decision probabilities assume a Gaussian predictive distribution with learned mean and variance. If the true conditional error distribution is skewed, then Gaussian prediction can under-represent tail risk.

7. Conclusions

This study developed a decision-oriented uncertainty quantification framework for as-built surface roughness in LPBF/DMLS-fabricated SS316L produced on an EOS M 290 using a Taguchi L9 design (laser power 300-360 W, scan speed 800-1000 mm·s⁻¹, layer thickness 20-80 μ m) with replicated fabrication ($n=27$; three specimens per setting) and standardized profilometry (Ra, Rq, Rz). To test transfer to unseen parameter settings, models were benchmarked using LOSO grouped cross-validation with

probabilistic scoring via NLPD alongside interval coverage/ width diagnostics. The following are the key findings:

- Across the nine settings, roughness means occupied narrow bands ($R_a=5.748\text{--}5.952\text{ }\mu\text{m}$; $R_q = 6.673\text{--}6.811\text{ }\mu\text{m}$; $R_z = 28.828\text{--}28.892\text{ }\mu\text{m}$), while within-setting scatter was non-negligible, particularly for R_z where the mean range ($0.064\text{ }\mu\text{m}$) was comparable to typical within-setting SD ($0.038\text{--}0.087\text{ }\mu\text{m}$).
- A low-capacity linear mean with pooled (homoscedastic) variance achieved the best NLPD for R_a (-0.392) and R_q (-1.006), indicating weak, non-transportable setting-level heteroscedasticity under LOSO. Whereas R_z favored a nonlinear mean with pooled variance (best NLPD= -1.490).
- Unregularized two stage variance learning was unstable (e.g., NLPD = 0.692 for R_a ; 0.723 for R_q), but empirical Bayes shrinkage restored calibration (R_a ENCE 0.333 to 0.130 ; R_q ENCE 0.352 to 0.185).
- Decision analysis under percentile specifications showed the effect of calibration operationally. At strict reliability ($r \geq 0.90$), pooled models rejected all settings for R_a and R_q , while unshrunk heteroscedasticity inflated acceptance (R_a accept probability 0 to 0.444 at $r = 0.90$).

Through reliability-based decision analysis, it has been shown that uncertainty calibration affects the acceptance of the process. By reducing the number of settings accepted, stricter reliability thresholds and shrinkage-stabilized uncertainty led to a conservative and stable decision frontier. The proposed methodology links grouped validation, probabilistic scoring, calibration diagnostics and reliability-driven accept/reject decisions providing a more defensible basis for qualification of the surface roughness in LPBF/DMLS of SS316L.

7. References

- [1] V. V. Phani Babu, V. K. GB, and P. Barmavatu, "An experimental investigation on surface quality of 3D metal printed SS316L by direct metal laser sintering technique," *International Journal on Interactive Design and Manufacturing (IJIDE)*, vol. 19, no. 1, pp. 393-403, 2025. <https://doi.org/10.1007/s12008-024-01801-6>
- [2] B. Pappula, J. K. Jatavallabhula, V. P. B. Vemuri, R. Chintalapudi, G. N. Kannaiyan, and S. Makgato, "Machine Learning-Enhanced Reverse Engineering for Dimensional Accuracy Optimization in DMLS-Fabricated SS316L Components," *Materials Today Communications*, p. 114293, 2025. <https://doi.org/10.1016/j.mtcomm.2025.114293>
- [3] A. Adin, E. T. Krainski, A. Lenzi, Z. Liu, J. Martínez-Minaya, and H. Rue, "Automatic cross-validation in structured models: Is it time to leave out leave-one-out?," *Spatial Statistics*, vol. 62, p. 100843, 2024. <https://doi.org/10.1016/j.spasta.2024.100843>
- [4] Q. He, C. Hong, L. Xu, and L. Li, "Robust portfolio management: A novel multi-task learning model fusing predicted returns and residual data under the framework of Mean-VaR," *Journal of the Operational Research Society*, vol. 76, no. 7, pp. 1466-1480, 2025. <https://doi.org/10.1080/01605682.2024.2438333>
- [5] Y. Kwon and Z. Zhao, "On F-modelling-based empirical Bayes estimation of variances," *Biometrika*, vol. 110, no. 1, pp. 69-81, 2023. <https://doi.org/10.1093/biomet/asac019>
- [6] V. Dheur and S. B. Taieb, "A large-scale study of probabilistic calibration in neural network regression," in *International Conference on Machine Learning*, 2023: PMLR, pp. 7813-7836.
- [7] N. I. Bosse, S. Abbott, A. Cori, E. Van Leeuwen, J. Bracher, and S. Funk, "Scoring epidemiological forecasts on transformed scales," *PLoS computational biology*, vol. 19, no. 8, p. e1011393, 2023. <https://doi.org/10.1371/journal.pcbi.1011393>
- [8] P. Pernot, "Calibration in machine learning uncertainty quantification: beyond consistency to target adaptivity," *APL Machine Learning*, vol. 1, no. 4, 2023. <https://doi.org/10.1063/5.0174943>
- [9] T. Yin, G. Panapitiya, E. D. Coda, and E. G. Saldanha, "Evaluating uncertainty-based active learning for accelerating the generalization of molecular property prediction," *Journal of Cheminformatics*, vol. 15, no. 1, p. 105, 2023. <https://doi.org/10.1186/s13321-023-00753-5>
- [10] T. Löfström, H. Löfström, U. Johansson, C. Sönströd, and R. Matela, "Calibrated explanations for regression," *Machine Learning*, vol. 114, no. 4, pp. 1-34, 2025. <https://doi.org/10.1007/s10994-024-06642-8>
- [11] M. B. Bjerregård, J. K. Möller, and H. Madsen, "An introduction to multivariate probabilistic forecast evaluation," *Energy and AI*, vol. 4, p. 100058, 2021. <https://doi.org/10.1016/j.egyai.2021.100058>
- [12] S. Allen, J. Koh, J. Segers, and J. Ziegel, "Tail calibration of probabilistic forecasts," *Journal of the American Statistical Association*, no. just-accepted, pp. 1-20, 2025. <https://doi.org/10.5194/egusphere-egu24-6281>
- [13] L. Hao, W. Wang, J. Zeng, M. Song, S. Chang, and C. Zhu, "Effect of scanning speed and laser power on formability, microstructure, and quality of 316L stainless steel prepared by selective laser melting," *Journal of Materials Research and Technology*, vol. 25, pp. 3189-3199, 2023. <https://doi.org/10.1016/j.jmrt.2023.06.144>
- [14] T. Zhang and L. Yuan, "Understanding surface roughness on vertical surfaces of 316 L stainless steel in laser powder bed fusion additive manufacturing," *Powder Technology*, vol. 411, p. 117957, 2022. <https://doi.org/10.1016/j.powtec.2022.117957>
- [15] S. Chowdhury et al., "Laser powder bed fusion: a state-of-the-art review of the technology, materials, properties & defects, and numerical modelling," *Journal of Materials Research and Technology*, vol. 20, pp. 2109-2172, 2022. <https://doi.org/10.1016/j.jmrt.2022.07.121>
- [16] A. Leicht, M. Fischer, U. Klement, L. Nyborg, and E. Hryha, "Increasing the productivity of laser powder bed fusion for stainless steel 316L through increased layer thickness," *Journal of Materials Engineering and Performance*, vol. 30, no. 1, pp. 575-584, 2021. <https://doi.org/10.1007/s11665-020-05334-3>
- [17] H. Lei and A. Bellotti, "Reliable prediction intervals with directly optimized inductive conformal regression for deep learning," *Neural Networks*, vol. 168, pp. 194-205, 2023. <https://doi.org/10.1016/j.neunet.2023.09.008>
- [18] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *Journal of the American statistical Association*, vol. 102, no. 477, pp. 359-378, 2007. <https://doi.org/10.1198/016214506000001437>
- [19] T. Gneiting and M. Katzfuss, "Probabilistic forecasting," *Annual Review of Statistics and Its Application*, vol. 1, no. 1, pp. 125-151, 2014. <https://doi.org/10.1146/annurev-statistics-062713-085831>
- [20] B. S. Reddy et al., "From metrics to decisions: a robust evidential methodology for dimensional accuracy and manufacturability in DMLS SS316L," *The International*

- Journal of Advanced Manufacturing Technology, pp. 1-27, 2026. <https://doi.org/10.1007/s00170-026-17524-9>
- [21] M. V. Wüthrich and J. Ziegel, "Isotonic recalibration under a low signal-to-noise ratio," *Scandinavian Actuarial Journal*, vol. 2024, no. 3, pp. 279-299, 2024. <https://doi.org/10.1080/03461238.2023.2246743>
- [22] C.-L. Yang, T. Y. Meles, A. A. Yilma, and M. M. Teshome, "Uncertainty aware predictive maintenance using a hybrid Transformer with Monte Carlo Dropout and conformal prediction," *Ain Shams Engineering Journal*, vol. 17, no. 2, p. 103992, 2026. <https://doi.org/10.1016/j.asej.2026.103992>
- [23] C. Qian, D. Ries, F. Liang, and J. Adams, "Regression Recalibration by Learning PIT Map Values," *Technometrics*, pp. 1-12, 2025. <https://doi.org/10.1080/00401706.2025.2464004>
- [24] X. Wang and W. Xiong, "Uncertainty quantification and composition optimization for alloy additive manufacturing through a CALPHAD-based ICME framework," *npj Computational Materials*, vol. 6, no. 1, p. 188, 2020. <https://doi.org/10.1038/s41524-020-00454-9>
- [25] S. Mahadevan, P. Nath, and Z. Hu, "Uncertainty quantification for additive manufacturing process improvement: Recent advances," *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part B: Mechanical Engineering*, vol. 8, no. 1, p. 010801, 2022. <https://doi.org/10.1115/1.4053184>

Nomenclature

C1	Circle 1
CDF	Cumulative Distribution Function
DMLS	Direct metal laser sintering
E1, E2	Ellipses 1 and 2
ENCE	Expected Normalized Calibration Error
EOS	Electro optical systems
IF1, IF2, IF3, IF4	Inner fillets for surfaces of the geometry
LOSO	Leave-One-Setting-Out
LP	Laser power
LPBF	Laser powder bed fusion
LT	Layer thickness
M0, M1, M2, M3	Models for capacity control
MAE	Mean absolute error

MPI	Mean prediction interval width
NLPD	Negative log predictive density
OF1, OF2, OF3, OF4	Outer fillets for surfaces of the geometry
PI	The prediction interval
Ra	The arithmetic mean roughness
Rq	The root mean square roughness
Rz	The maximum height of the profile
RMSE	Root mean square error
SD	Standard deviation
SS	Scan speed
$y_{g,r}^{(m)}$	The observed roughness for metric
m	The metric
n	The number of replicates
g	The group
r	The replicate index
X_g	The predictor vector
$\mathcal{T}_k$	The test set
$\mu^{(m)}(X)$	The low-capacity mean function
$y^{(m)}$	The conditional expectation of roughness
$Var(y^{(m)} X)$	The variance
$\epsilon(m)$	Zero-mean noise term
$h^{(m)}(X)$	The variance model
$\tilde{s}_g^{2(m)}$	The sample variance of residuals
ν_0	The sensitivity/robustness parameter
$p_{acc}(X)$	The probability of roughness
$\Phi(\cdot)$	The standard normal CDF
τ	Roughness limit/specification threshold
ϱ	Spearman coefficient
k	LOSO fold index
$\hat{\mu}^{(m)}(X)$	Predicted mean roughness
$\sigma^{2(m)}(X)$	Conditional variance
α	Miscoverage level for prediction intervals