Credible Uncertainty Quantification under Noise and System Model Mismatch

Penggao Yan, Li-Ta Hsu, *Senior Member, IEEE*, and Rui Sun, *Member, IEEE*

***Abstract*—State estimators often provide self-assessed uncertainty metrics, such as covariance matrices, whose credibility is critical for downstream tasks. However, these self-assessments can be misleading due to underlying modeling violations like noise model mismatch (NMM) or system model misspecification (SMM). This work addresses this problem by developing a unified, multi-metric framework that integrates noncredibility index (NCI), negative log-likelihood (NLL), and energy score (ES) metrics, featuring an empirical location test (ELT) to detect system model bias and a directional probing technique that uses the metrics' asymmetric sensitivities to distinguish NMM from SMM. Monte Carlo simulations reveal that the proposed method achieves excellent diagnosis accuracy $(80-100\%)$ and significantly outperforms single-metric diagnosis methods. The effectiveness of the proposed method is further validated on a real-world ultra-wideband (UWB) positioning dataset. This framework provides a practical tool for turning patterns of credibility indicators into actionable diagnoses of model deficiencies.**

***Index Terms*—State estimation, credibility, calibration, noise model mismatch, system model misspecification**

I. Introduction

STATE estimation problems are widely studied in the tracking, navigation, and control community [1]–[4]. Estimators routinely accompany point estimations with self-assessed uncertainty. For example, Kalman filters (KF) provide covariance matrices, while methods like particle filters produce full predictive distributions. These self-assessments are informative but rest on modeling assumptions that may be violated in practice [5]–[7]. For example, Chauchat et al. [8] highlighted that the optimality of Kalman filters relies on perfect system modeling, which rarely holds in real-world scenarios. Similarly, Ge et al. [9] analyzed the performance degradation when noise covariances are mismatched. Fortunati et al. [10] discussed the fundamental performance bounds of parameter estimation under misspecified models, highlighting the impact of modeling violations such as noise-model mismatch (NMM) (e.g., pessimism or optimism) and system-model misspecification (SMM) [3], [11]. This raises a practical question: can the self-assessment be trusted, to what degree, and in which direction of non-credibility (optimism vs. pessimism)? Following [12], we refer to this as the credibility problem.

This work was supported by the National Natural Science Foundation of China (NSFC)/ Research Grants Council (RGC) of Hong Kong Joint Research Scheme under Grant 42561160140 and N_PolyU502/25. *(Corresponding author: Li-Ta Hsu).*

Penggao Yan and Li-Ta Hsu are with the Department of Aeronautical and Aviation Engineering, Faculty of Engineering, Hong Kong Polytechnic University, Hong Kong (e-mail:lt.hsu@polyu.edu.hk); Rui Sun is with the College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China.

In state estimation, credibility is most often judged by a single statistic, typically the normalized estimation error squared (NEES) and its variants, such as average NEES (ANEES) and noncredibility index (NCI) [1], [2], [11]. By testing the agreement between estimation errors and the nominal covariance, NEES primarily assesses credibility from the calibration perspective, i.e., the consistency between predicted uncertainty and actual outcomes [13]. For example, Blasch et al. [14] investigated the use of NCI and ANEES for nonlinear estimation performance analysis. They used a nonlinear estimation framework to compare filters like UKF and PF by contrasting credibility metrics against absolute RMS errors. Zhang et al. [15] adopted the NCI to quantify the noncredibility of estimation and decision in joint tracking and classification problems. However, prior work has demonstrated that relying on a single metric can be misleading when model assumptions are violated [16]–[18]. Single metrics often suffer from directional asymmetry (responding differently to over- versus under-estimation) and can be overly sensitive to specific modeling choices. These limitations strongly motivate the need for a more comprehensive, multi-metric approach to credibility assessment.

In parallel, the probabilistic-forecasting and ML/DL communities conduct multi-criteria credibility evaluation [19]–[22], most of which are developed based on the theory of proper scoring rules [23], [24]. Notably, the negative log-likelihood (NLL) and energy scores (ES) are widely used to assess calibration and sharpness simultaneously [24]. For example, Ashok et al. [25] utilized NLL to validate the TACTiS-2 model, ensuring the learned distribution matches the true distribution. Additionally, Al-Gabalawy et al. [26] applied NLL minimization to train various deep learning probabilistic models for energy time series, emphasizing the importance of proper scoring rules for calibrated predictions. These practices align with credibility evaluation in state-estimation problems.

To tackle the limitations of single-metric assessment in distinguishing between NMM and SMM, we propose a unified multi-metric credibility evaluation framework that integrates NCI, NLL, and ES. Specifically, we first introduce an empirical location test (ELT) based on energy distance to statistically detect the presence of SMM. If SMM is detected, we mitigate its impact by centering the estimation errors, thereby isolating the potential NMM effects. Subsequently, we employ a directional probing technique that artificially scales the covariance to generate "probes" of NLL and ES. By analyzing the asymmetric responses of these probes, we can robustly distinguish between optimism, pessimism, and SMM. This procedure effectively transforms complex patterns of multiple

metrics into actionable diagnoses of model deficiencies.

The proposed method is evaluated on two types of experiments, including Monte Carlo simulations and a real-world ultra-wideband (UWB) positioning dataset. The simulation results cover six distinct credibility scenarios, revealing that the proposed method achieves a diagnosis accuracy of 80%–100%, significantly outperforming single-metric baselines. Furthermore, the evaluation on the UWB dataset demonstrates the practical applicability of the framework, where it successfully identifies the coexistence of pessimism and SMM in static positioning periods, a nuance that conventional NEES and NCI methods fail to capture. The contributions of this work are threefold:

1) We analytically and experimentally reveal the complementary directional asymmetries of NLL and ES, serving as the theoretical foundation for distinguishing different types of non-credibility.
2) We propose a unified credibility diagnosis framework featuring an ELT for SMM detection and a directional probing mechanism, which provides a robust solution for disentangling NMM and SMM.
3) We experimentally demonstrate the effectiveness and superiority of the proposed framework in both controlled simulations and real-world UWB positioning scenarios.

The rest of this article is organized as follows. Section II analyzes the properties of NEES, NCI, NLL, and ES, pointing out their individual limitations. Section III details the proposed unified credibility diagnosis scheme, including the ELT and directional probing techniques. In Section IV, we examine the performance of the proposed method through Monte Carlo simulations. In Section V, we validate the effectiveness of the framework using the real-world UWB dataset. Section VI gives a summary.

II. Credibility Metrics Analysis

Let the estimatee and its estimate be x and $\hat{x}$. Define the estimation error $e = x - \hat{x}$ with mean μ and covariance Σ. The mean-square error (MSE) is $\mathcal{M} = \mathbb{E}[ee^\top] = \Sigma + \mu\mu^\top$ (equals Σ only when $\mu = 0$). The estimator reports covariance $\hat{\Sigma}$ and (optionally) MSE $\hat{\mathcal{M}}$. For Monte-Carlo (MC) experiments, the k-th run uses $(x_k, \hat{x}_k, e_k, \hat{\Sigma}_k)$ and we run N independent trials. The predictive cumulative distribution function (CDF)/probability density function (PDF) are denoted $\hat{F}(\cdot)$ and $\hat{f}(\cdot)$, respectively.

To provide a concrete foundation for the credibility analysis, we consider three representative scenarios in state estimation:

- **Credible**: $(\hat{\Sigma}_k = \Sigma_k, \mu_k = 0)$
- **NMM**: we characterize the mismatch as an incorrect scaling of the covariance, formalized as $\hat{\Sigma}_k = \rho\,\Sigma_k$ with $\mu_k = 0$, where $\rho > 0$ denotes the scaling factor
- **SMM**: we characterize the mismatch as the presence of a constant estimation bias, expressed as $\mu_k = b_k \neq 0$ with $\hat{\Sigma}_k = \Sigma_k$, where b_k is the bias vector.

These canonical cases serve as the basis for systematically evaluating the behavior and diagnostic power of various credibility metrics in subsequent sections.

A. Normalized Estimation Error Squared (NEES)

$$\epsilon_k \;=\; e_k^\top \hat{\Sigma}_k^{-1} e_k. \tag{1}$$

Consider $e_k \sim \mathcal{N}(\mu_k, \Sigma_k)$, we have

$$\mathbb{E}[\epsilon_k] \;=\; \mathrm{tr}(\hat{\Sigma}_k^{-1}\Sigma_k) \;+\; \mu_k^\top \hat{\Sigma}_k^{-1}\mu_k. \tag{2}$$

Properies.

- **In the credible case:** $\mathbb{E}[\epsilon_k] = d$ (i.e., state dimension).
- **In the NMM case with incorrect scaling covariance:** $\mathbb{E}[\epsilon_k] = d/\rho$. Define the deviation from the expected value as follows:

$$D_k \;=\; \mathbb{E}[\epsilon_k] - d \;=\; d\Big(\tfrac{1}{\rho} - 1\Big)\,. \tag{3}$$

  Notaly, $D_k$is positive for optimism $(\rho < 1)$ and negative for pessimism $(\rho > 1)$. Moreover, $|D_k(\rho)| < |D_k(1/\rho)|$ for $\rho > 1$, i.e., NEES penalizes optimism more severely than pessimism.
- **In the SMM case with constant estimation bias:** $\mathbb{E}[\epsilon_k] = d + \mu_k^T\Sigma_k^{-1}\mu_k$, and $D_k = \mu_k^T\Sigma^{-1}\mu_k$ is always positive and increase with the bias magnitude. Therefore, NEES always penalizes SMM. However, it is difficult to distinguish the SMM and optimism, as in both cases D_k is positive.

B. Noncredibility index (NCI)

$$NCI(\{\hat{x}_k\}) = \frac{10}{N}\sum_{k=1}^{N}\log_{10}(\epsilon_k) - \frac{10}{N}\sum_{k=1}^{N}\log_{10}(\epsilon_k^*)\,, \tag{4}$$

where ϵ_k^* is the NEES of a perfectly credible estimator, calculated as $\epsilon_k^* = e_k^T\mathcal{M}_k^{-1}e_k$. The magnitude of the NCI directly measures the level of non-credibility.

Properties.

- **In the credible case:** NCI is zero.
- **In the NMM case with incorrect scaling covariance:** The NCI is given by

$$NCI(\{\hat{x}_k\}, \rho) = -\frac{10}{N}\sum_{k=1}^{N}\log_{10}\rho. \tag{5}$$

  This value is positive for optimism $(\rho < 1)$ and negative for pessimism $(\rho > 1)$. As $|NCI(\{\hat{x}_k\}, \rho)| = |NCI(\{\hat{x}_k\}, \frac{1}{\rho})|$, NCI penalizes optimism and pessimism equally.
- **In the SMM case with constant estimation bias:** The NCI is always non-negative (see Appendix A). Therefore, NCI always penalizes SMM. However, similar to NEES, it cannot distinguish SMM from optimism, as both result in non-negative NCI values.

C. Negative Log-Likelihood (NLL)

$$\mathrm{NLL}(\hat{F}_k, x_k) = -\log \hat{f}_k(x_k)\,. \tag{6}$$

Since $\hat{f}_k(x_k) \leq 1$, NLL is non-negative. Consider $\hat{F}_k = \mathcal{N}(\hat{x}_k, \hat{\Sigma}_k)$ and $e_k = x_k - \hat{x}_k \sim \mathcal{N}(\mu_k, \Sigma_k)$, we have

$$\mathbb{E}[\mathrm{NLL}] = \tfrac{1}{2}\Big(\mathrm{tr}(\hat{\Sigma}_k^{-1}\Sigma_k) + \mu_k^\top \hat{\Sigma}_k^{-1}\mu_k + \ln|\hat{\Sigma}_k| + d\ln(2\pi)\Big)\,. \tag{7}$$

Properties.

- **In the NMM case with incorrect scaling covariance:**

$$\mathbb{E}[\mathrm{NLL}] = \tfrac{1}{2}\Big(\tfrac{d}{\rho} + d\ln\rho + \ln|\Sigma_k| + d\ln(2\pi)\Big)\,. \tag{8}$$

  Appendix B shows that $|\mathbb{E}[\mathrm{NLL}](\rho)| < |\mathbb{E}[\mathrm{NLL}](1/\rho)|$, indicating that the NLL penalizes optimism more severely than pessimism. This pronounced sensitivity to optimism, which we will later exploit, makes it a powerful tool for diagnosing model misspecification.
- **In the SMM case with constant estimation bias:**

$$\mathbb{E}[\mathrm{NLL}] = \tfrac{1}{2}\big(d + \mu_k^\top \Sigma_k^{-1}\mu_k + \ln|\Sigma_k| + d\ln(2\pi)\big)\,, \tag{9}$$

  revealing that the NLL increases with the bias magnitude. Therefore, the NLL always penalizes SMM. Since the NLL always takes a non-negative value, it is difficult to distinguish NMM and SMM solely by using the NLL.

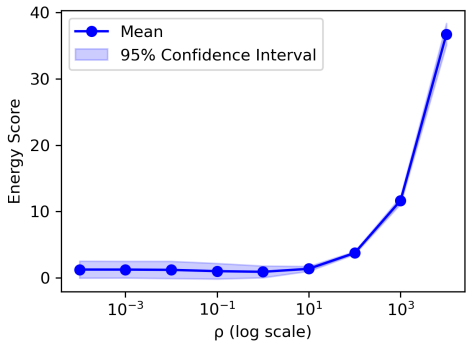
D. Energy Score (ES)

$$\mathrm{ES}(\hat{F}_k, x_k) = \mathbb{E}_{Y\sim\hat{F}_k}\|Y - x_k\|_2 - \tfrac{1}{2}\,\mathbb{E}_{Y,Y'\sim\hat{F}_k}\|Y - Y'\|_2. \tag{10}$$

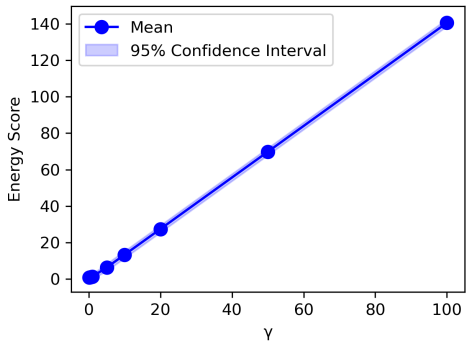
The first term measures calibration (distance to truth), and the second term measures sharpness (concentration). The ES is always non-negative (see Appendix C).

Properties. Due to the complexity of the ES formulation, there are no analytical forms of ES and its expectation. Therefore, we implement Monte-Carlo simulations to study the sensitivity of ES. Specifically, we study the case of a multivariate normal distribution $\hat{F}_k = \mathcal{N}(\hat{x}_k, \hat{\Sigma}_k)$ with $\hat{x}_k \sim \mathcal{N}(x_k + \mu_k, \Sigma_k)$.

- **In the NMM case with incorrect scaling covariance:** We set $x_k = \mathbf{0}$ and $\Sigma_k = \mathrm{I}_2$. We calculate the sample mean of ES by implementing 5,000 Monte Carlo runs for each ρ. Fig. 1a plots the sample mean of ES against ρ. It is evident that the ES penalizes pessimism more severely than optimism, a characteristic that stands in direct contrast to the properties of the NLL. This contrasting sensitivity to pessimism is fundamental to the proposed unified diagnostic approach, as it provides complementary information to the NLL.
- **In the SMM case with constant estimation bias:** We set $\mu_k = \gamma[1,1]^T$, $x_k = \mathbf{0}$, $\Sigma_k = \mathrm{I}_2$. Similarly, we implement the Monte Carlo simulation (5,000 runs for each γ) and plot the relationship between the sample mean of ES and γ in Fig. 1b. Evidently, the ES always penalizes SMM. Since the ES always takes a non-negative value, it is difficult to distinguish NMM and SMM solely by using the ES.



(a)



(b)

Fig. 1: Monte-Carlo simulations show how ES varies against (1) ρ and (2) γ.

Table I summarizes the properties of NEES, NCI, NLL, and ES. While each of the metrics discussed provides a unique perspective on credibility, each metric has its own limitations. NEES and NCI struggle to distinguish optimism from SMM, while NLL and ES show opposite sensitivities to covariance scaling but cannot independently identify the direction of non-credibility. This motivates our development of a unified evaluation scheme that synergistically combines these metrics to provide a more complete and reliable diagnosis.

III. Unified Credibility Diagnosis Scheme

We propose a heuristic procedure to distinguish NMM (optimism or pessimism) and SMM. A heuristic framework is chosen because a purely analytical solution is not straightforward, and the combined effects of SMM and NMM are difficult to formally disentangle. Our step-by-step procedure is therefore designed to navigate this complexity. Fig. 2 gives the flowchart of the proposed algorithm.

A. Identify the impacts of SMM

As discussed in Section II, standard metrics such as NEES, NCI, NLL, and ES are unable to distinguish between NMM and SMM due to their inherent formulations. To overcome this limitation, we propose to use ELT [27] based on energy

TABLE I: Summary of Credibility Metrics and Their Properties

Metric	Credible Case	NMM Case (Scaling ρ)	SMM Case (Bias μ_k)
NEES	$\mathbb{E}[\epsilon_k] = d$	$\mathbb{E}[\epsilon_k] = d/\rho$. Penalizes optimism more severely.	$\mathbb{E}[\epsilon_k]$ increases with the bias magnitude. Difficult to distinguish SMM and optimism.
NCI	$NCI = 0$	Symmetric penalty for optimism ($NCI > 0$) and pessimism ($NCI < 0$).	$NCI \geq 0$. Difficult to distinguish SMM and optimism.
NLL	Minimized	Penalizes optimism more severely than pessimism.	Increases with bias magnitude.
ES	Minimized	Penalizes pessimism more severely than optimism.	Increases with bias magnitude.

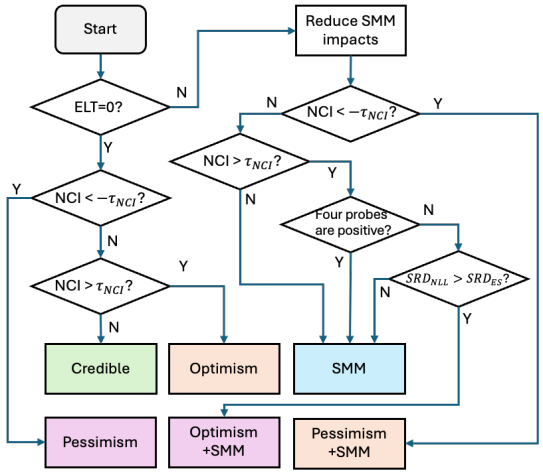


Fig. 2: The flowchart of the proposed unified credibility diagnosis method.

distance to first identify and remove the impacts of SMM, disentangling the combined effects of SMM and NMM.

The energy distance quantifies the dissimilarity between two probability distributions F and G, and is defined as:

$$\begin{aligned}\mathcal{D}_\alpha^2(F,G) = 2\,\mathbb{E}_{X\sim F,\,Y\sim G}\|X-Y\|^\alpha \\ - \mathbb{E}_{X,X'\sim F}\|X-X'\|^\alpha \\ - \mathbb{E}_{Y,Y'\sim G}\|Y-Y'\|^\alpha,\end{aligned} \tag{11}$$

where $\alpha \in (0, 2]$. In this work, we set $\alpha = 1$.

To test for SMM, we examine the distribution of the whitened estimation errors, $s_k = \hat{\Sigma}_k^{-1/2} e_k$. In the absence of SMM, the distribution of $\{s_k\}$ should be symmetric about the origin, i.e., identical to its mirror image $\{-s_k\}$. This leads to the following hypothesis test:

$$\begin{aligned}H_0 &: \mathcal{L}(\{s_k\}) = \mathcal{L}(\{-s_k\}) \\ H_1 &: \mathcal{L}(\{s_k\}) \neq \mathcal{L}(\{-s_k\}),\end{aligned} \tag{12}$$

where $\mathcal{L}(\cdot)$ denotes the empirical distribution. An ELT test is then implemented. Specifically, a test statistic is constructed as the energy distance between the empirical distributions of $\{s_k\}$ and $\{-s_k\}$:

$$\begin{aligned}T_{\text{obs}} &= \mathcal{D}_\alpha^2\big(\mathcal{L}(\{s_k\}),\, \mathcal{L}(\{-s_k\})\big) \\ &= \frac{2}{N(N-1)} \sum_{1\leq i<j\leq N} \left(\|s_i+s_j\|^\alpha - \|s_i-s_j\|^\alpha\right),\end{aligned} \tag{13}$$

where N is the number of samples. When the distribution of estimation errors is centered at the origin, the average pairwise sum and difference distances are balanced, resulting in a small value of T_{obs}. However, if the distribution is systematically shifted away from the origin (indicative of SMM), the average $\|s_i+s_j\|$ increases, leading to a larger value of the test statistic.

To formally assess the statistical significance of the observed energy distance, we employ a sign-flip randomization test. This test is based on the null hypothesis H_0 of central symmetry. Specifically, let $\xi_k \in \{-1, +1\}$ be independent Rademacher random variables. Under H_0, the distribution of the sign-flipped set $\{\xi_1 s_1, \ldots, \xi_N s_N\}$ is identical to that of the original set $\{s_1, \ldots, s_N\}$:

$$\{\xi_1 s_1, \ldots, \xi_N s_N\} \overset{d}{=} \{s_1, \ldots, s_N\}. \tag{14}$$

Therefore, by applying random sign flips to the whitened errors $\{s_k\}$, we can generate samples from the null distribution of the test statistic. The conditional distribution of the energy distance statistic $\mathcal{D}_\alpha^2\big(\mathcal{L}(\{\xi_k s_k\}), \mathcal{L}(\{-\xi_k s_k\})\big)$ serves as the exact reference distribution for our hypothesis test, allowing us to calculate a p-value.

In practice, we perform the following randomization procedure to construct this null distribution: for each iteration $b = 1, \ldots, B$ (usually $B \geq 1000$), we independently sample Rademacher random variables $\xi_k^{(b)} \in \{\pm 1\}$ for each k, apply the sign flips to obtain $\{\xi_k^{(b)} s_k\}$, and then compute the corresponding randomized test statistic:

$$T^{(b)} = \hat{\mathcal{E}}_\alpha(\{\xi_k^{(b)} s_k\}). \tag{15}$$

The (one-sided) randomized p-value is then estimated as

$$p_{\text{ELT}} = \frac{1 + \#\{b : T^{(b)} \geq T_{\text{obs}}\}}{B+1}. \tag{16}$$

Finally, we define the ELT decision as

$$\text{ELT} = \mathbf{1}\{p_{\text{ELT}} < \alpha_{\text{sig}}\}, \tag{17}$$

where α_{sig} is the significance level. An outcome of $\text{ELT} = 1$ indicates statistical evidence for SMM.

B. Evaluation of noncredibility direction without SMM

When $ELT = 0$, we declare the SMM does not exist. The remaining question is whether the estimation is pessimistic, optimistic, or credible. To answer this question, we simply use the NCI metric due to its ability to evaluate the noncredibility direction. Specifically, we define a positive threshold $\tau_{\text{NCI}} > 0$. If NCI is smaller than $-\tau_{\text{NCI}}$, the estimation is said to be pessimistic; If NCI is larger than τ_{NCI}, the estimation is said to be optimistic; Otherwise, the estimation is said to be credible.

C. Evaluation of noncredibility direction with SMM

1) Reduce the impacts of SMM: When $ELT = 1$, we declare the SMM exists. If the SMM and NMM both exist, the effects of SMM will disrupt the assessment of the direction of non-credibility. One intuitive solution is to subtract the sample mean of estimation errors from the estimation, i.e., $\check{x}_k = \hat{x}_k - \frac{1}{N}\sum_{k=1}^{N} e_k$. The corresponding predictive distribution is given by $\check{F}_k = \mathcal{N}(\check{x}_k, \hat{\Sigma}_k)$. The modified estimation $\check{x}_k$ and predictive distribution $\check{F}_k$ will be used to further determine whether the estimation is pessimistic or optimistic.

2) Use NCI to tentatively evaluate pessimism: After mitigating the impact of SMM, we use the NCI to tentatively identify the direction of non-credibility. If $|\mathrm{NCI}(\{\check{x}_k\})| \leq \tau_{\mathrm{NCI}}$, the adjusted estimation $\check{x}_k$ can be considered credible in terms of its noise model. We can therefore conclude that the original estimation is only affected by SMM; If $\mathrm{NCI}(\{\check{x}_k\}) < -\tau_{\mathrm{NCI}}$, this provides strong evidence that the original estimation is affected by both SMM and pessimism; If $\mathrm{NCI}(\{\check{x}_k\}) > \tau_{\mathrm{NCI}}$, the interpretation is more subtle. A positive NCI may result from either residual SMM or genuine optimism. Therefore, the analysis in Section III-C3 is required to distinguish this ambiguity.

3) Directional probes using NLL and ES: In this step, we exploit the asymmetric sensitivities of the NLL and ES to pessimism and optimism. The core idea is to "probe" the credibility of the estimate by artificially scaling its covariance. By observing how NLL and ES react differently to these optimistic and pessimistic probes, we can infer the underlying nature of the uncertainty.

To capture the asymmetric sensitivity of NLL and ES to covariance scaling, we construct the following probes of NLL and ES as follows:

$$\begin{aligned} \Delta^{-}_{\mathrm{NLL}} &= \mathrm{NLL}(\check{F}_k(1/c), x_k) - \mathrm{NLL}(\check{F}_k, x_k) \\ \Delta^{+}_{\mathrm{NLL}} &= \mathrm{NLL}(\check{F}_k(c), x_k) - \mathrm{NLL}(\check{F}_k, x_k) \\ \Delta^{-}_{\mathrm{ES}} &= \mathrm{ES}(\check{F}_k(1/c), x_k) - \mathrm{ES}(\check{F}_k, x_k) \\ \Delta^{+}_{\mathrm{ES}} &= \mathrm{ES}(\check{F}_k(c), x_k) - \mathrm{ES}(\check{F}_k, x_k)\,, \end{aligned} \tag{18}$$

where $\check{F}_k(c) = \mathcal{N}(\check{x}_k, c\hat{\Sigma}_k)$ is the scaled predictive distribution by the scaling factor $c > 1$. Importantly, when $\check{x}_k$ is free from NMM, all four probes are expected to be positive, reflecting the fact that both NLL and ES increase under either pessimistic or optimistic estimation scenarios.

To quantify the asymmetry of NLL and ES's response to optimistic versus pessimistic scaling, we construct the slope relative difference (SDR) as follows:

$$\mathrm{SRD}_{\mathrm{NLL}} = \frac{c|\Delta^{-}_{\mathrm{NLL}}| - |\Delta^{+}_{\mathrm{NLL}}|}{|\Delta^{+}_{\mathrm{NLL}}|}, \mathrm{SRD}_{\mathrm{ES}} = \frac{c|\Delta^{-}_{\mathrm{ES}}| - |\Delta^{+}_{\mathrm{ES}}|}{|\Delta^{+}_{\mathrm{ES}}|}\,, \tag{19}$$

which measures the relative difference between a optimistic probe ($|\Delta^{-}|$) and a pessimistic probe ($|\Delta^{+}|$). The factor c accounts for the unequal step sizes of the probes ($1 \to 1/c$ vs. $1 \to c$), effectively comparing the local slopes of the scoring rule in each direction. Since NLL exhibits stronger sensitivity to optimism than pessimism, it is expected that $\mathrm{SRD}_{\mathrm{NLL}} > \mathrm{SRD}_{\mathrm{ES}}$ when the estimation is optimism. In contrast, $\mathrm{SRD}_{\mathrm{ES}} > \mathrm{SRD}_{\mathrm{NLL}}$ is expected when the estimation is pessimism.

Based on the above findings, we propose the following two-step diagnosis procedure. First, we examine the signs of the probes. If all four probes are positive, it suggests that the bias-corrected estimate $\check{x}_k$ is likely free from NMM, and thus we conclude that the original estimate $\hat{x}_k$ is only affected by SMM. If any of these four metrics are not positive, it indicates that $\check{x}_k$ may still exhibit optimism. To resolve this, we proceed to the second step: comparing the SRD values. If $\mathrm{SRD}_{\mathrm{NLL}} > \mathrm{SRD}_{\mathrm{ES}}$, we confirm $\check{x}_k$ is optimism, and therefore declare that $\hat{x}_k$ is affected by both SMM and optimism. Otherwise, the evidence for optimism is not conclusive, and we revert to concluding that $\hat{x}_k$ is primarily affected by SMM.

D. Pseudocode of the proposed Algorithm

The pseudocode of the proposed algorithm is listed in Algorithm 1.

Algorithm 1 Unified Credibility Evaluation Scheme

```
if ELT = 0 then                                                    ▷ No SMM
    if NCI < −τ_NCI then
        output: pessimism
    else if NCI > τ_NCI then
        output: optimism
    else
        output: credible
    end if
else                                                               ▷ SMM exists
    Subtract mean error from estimation
    if NCI < −τ_NCI then
        output: pessimism + SMM
    else if NCI > τ_NCI then ▷ Maybe also optimism or small pessimism
        if Δ⁻_NLL > 0 and Δ⁺_NLL > 0 and Δ⁻_ES > 0 and Δ⁺_ES > 0 then
    ▷ Both metrics increase when moving away
            output: SMM
        else if SRD_NLL > SRD_ES then
            output: optimism + SMM
        else
            output: SMM
        end if
    else
        output: SMM
    end if
end if
```

IV. Simulation Experiments

A. Experimental Design

We validate the proposed unified evaluation scheme through Monte Carlo simulations with controlled synthetic data covering six distinct credibility scenarios.

Setup. We consider a 2-D state estimation problem ($d = 2$) across six scenarios with 50 trials each. Each trial comprises 100 Monte Carlo runs. True states follow $x_k \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathrm{true}})$ where Σ_{true} is randomized per trial via QR decomposition with eigenvalues uniformly distributed in $[0.5, 2.0]$ to mitigate conditioning effects. State estimates are generated as $\hat{x}_k \sim \mathcal{N}(x_k + \mu_k, \Sigma_{\mathrm{true}})$, and claimed covariances are $\hat{\Sigma}_k = \rho \cdot \Sigma_{\mathrm{true}}$.

Scenarios. Six scenarios are considered. Except for the credible scenario, each scenario uses a wide parameter range to examine the algorithm's robustness:

1) **Credible:** $\mu_k = \mathbf{0}$, $\rho = 1$
2) **Optimism:** $\mu_k = \mathbf{0}$, $\rho \sim \mathrm{Uniform}[0.1, 0.8]$
3) **Pessimism:** $\mu_k = \mathbf{0}$, $\rho \sim \mathrm{Uniform}[1.25, 10]$

4) **SMM:** μ_k with randomized direction and $\|\mu_k\| \sim \text{Uniform}[1.6, 2.4]$, $\rho = 1$
5) **Optimism + SMM:** μ_k with randomized direction and $\|\mu_k\| \sim \text{Uniform}[1.6, 2.4]$, $\rho \sim \text{Uniform}[0.1, 0.8]$
6) **Pessimism + SMM:** μ_k with randomized direction and $\|\mu_k\| \sim \text{Uniform}[1.6, 2.4]$, $\rho \sim \text{Uniform}[1.25, 10]$

Algorithm Parameters. $\tau_{\text{NCI}} = 0.5$ dB, $\alpha_{\text{sig}} = 0.05$, and directional probe scaling $c = 2$.

B. Results and Analysis

Table II compares the classification accuracy of the proposed algorithm with baseline algorithms, including NCI and NEES. Since NLL and ES cannot independently assess the direction of non-incredibility, they are not considered in benchmarking. Single-metric methods exhibit severe limitations: The NEES-based method fails in most scenarios except detecting some "Credible" scenarios (6.0%). Similarly, the NCI-based method only shows 26.0% accuracy in the "Optimism" scenarios. This validates the necessity of the multi-metric approach for comprehensive credibility assessment. The proposed method consistently outperforms all single-metric approaches across scenarios, achieving 80.0%-100.0% accuracy.

TABLE II: Diagnosis accuracy of the proposed method and single-metric baselines

	Proposed	NEES	NCI
Credible	**94.0%**	6.0%	0%
Optimism	**84.0%**	0.0%	26.0%
Pessimism	**90.0%**	0.0%	0.0%
SMM	**80.0%**	0.0%	0.0%
Optimism+SMM	**82.0%**	0.0%	0.0%
Pessimism+SMM	**100.0%**	0.0%	0.0%

Table III summarizes the confusion patterns of the diagnosis results of the proposed method. For Optimism+SMM scenarios, 9 cases are classified as "SMM"—not incorrect since Optimism+SMM inherently contains SMM, but incomplete as it misses the optimism component. Similarly, 10 pure "SMM" cases are classified as "Pessimism + SMM," indicating the algorithm correctly identifies the bias but erroneously detects pessimism in the bias-corrected estimates. In "Optimism" scenarios, 6 cases are misclassified as "Credible," suggesting the algorithm conservatively requires stronger evidence for optimism detection, while 2 cases trigger "Optimism + SMM," indicating spurious bias detection in purely optimistic conditions. The confusion patterns demonstrate that "misclassifications" often represent partial but meaningful detections rather than complete algorithmic failures, supporting the framework's utility for practical credibility assessment.

V. UWB Positioning Experiment

This section evaluates the proposed framework on the STAR-loc dataset [28], a real-world dataset for stereo and range-based localization. Specifically, we use the UWB measurements collected under configuration s3 with landmark set v2 and grid trajectory (starloc_data_grid_s3_uwb.csv), which comprises range measurements between a mobile UWB tag and eight fixed anchors with known coordinates. The surveyed range from anchor to tag is also provided in the dataset, which enables the calculation of authentic UWB range measurement errors.

TABLE III: Confusion matrix of the diagnosis results of the proposed method

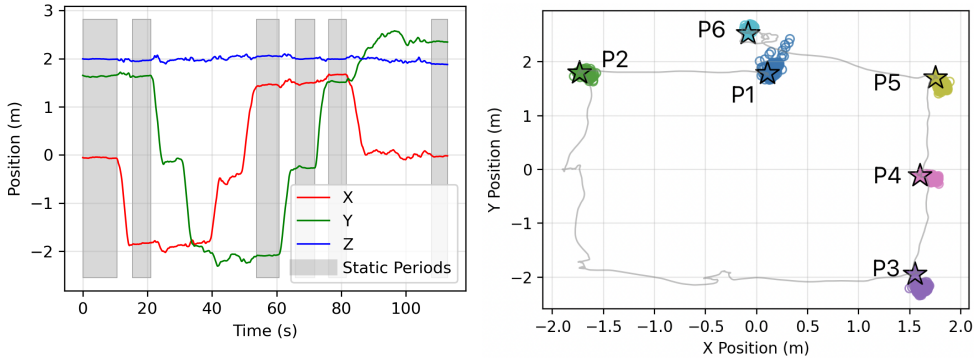
True Scenario	Diagnosis Results					
	Credible	O	P	SMM	O+SMM	P+SMM
Credible	47	0	0	3	0	0
O[1]	6	42	0	0	2	0
P[1]	1	0	45	0	0	4
SMM	0	0	0	40	0	10
O+SMM	0	0	0	9	41	0
P+SMM	0	0	0	0	0	50

[1] O: Optimism; P: Pessimism.

A. Data preprocessing in UWB Positioning Experiments

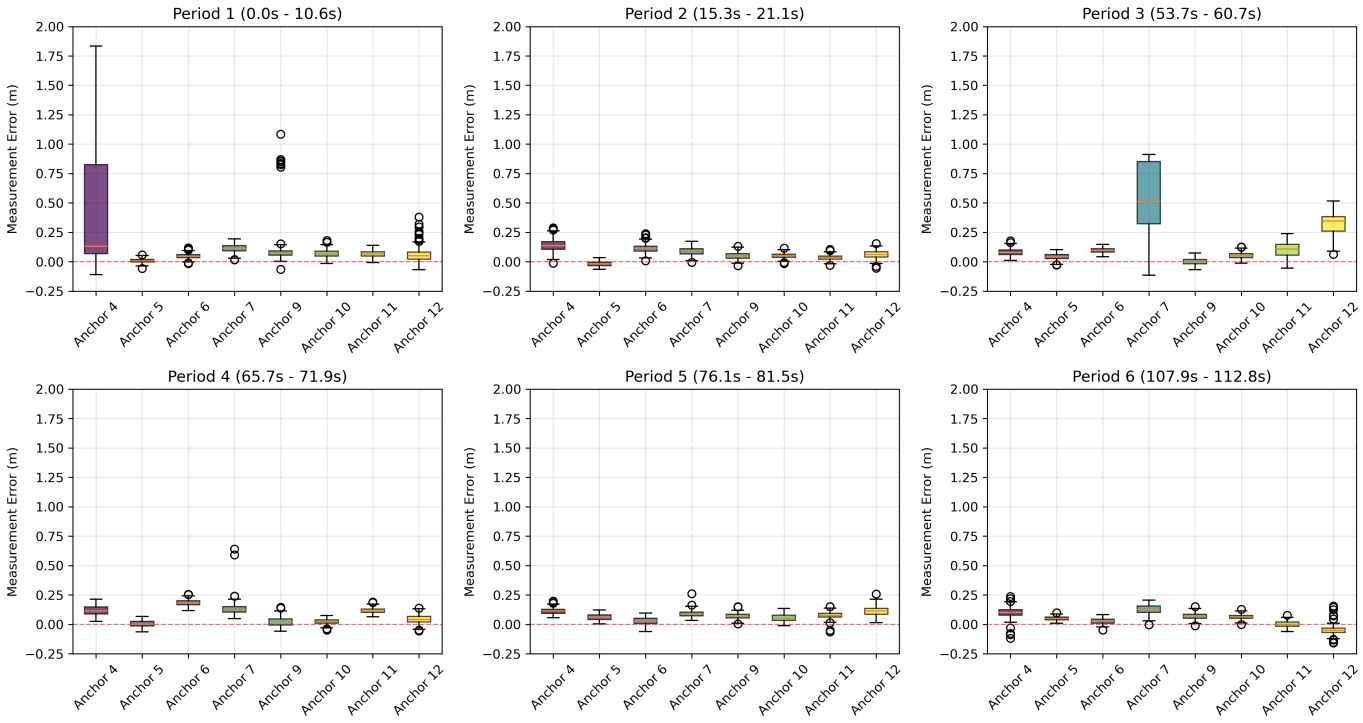
The data processing pipeline encompasses two primary stages. Initially, we implement a velocity-based segmentation strategy to identify static periods. We select periods where the tag's three-dimensional velocity remains below 0.1 m/s for a minimum duration of 4 seconds as static periods, ensuring stable measurement conditions for subsequent analysis. Fig. 3 shows the segmentation of the dataset in terms of the xyz-coordinate series and 2D-trajectory. Fig. 4 plots the distribution of UWB range measurement errors with respect to each anchor, revealing significant systematic biases in measurements across all static periods. Fig. 5 shows boxplots of reported uncertainties (standard deviations) of UWB range measurements with respect to each anchor. The red stars indicate the empirical standard deviation, which is calculated using the UWB range measurement errors. The comparison clearly demonstrates that most reported measurement uncertainties are pessimistic.

In the second stage, we apply a temporal aggregation strategy in each static period to address the challenge of asynchronous UWB measurements. Specifically, UWB range measurements collected within a 0.03-second window from distinct anchors are grouped into packets. For each packet, we apply the WLS algorithm to get the two-dimensional position estimation, where each measurement is weighted by the inverse of its reported covariance, as indicated in the dataset. This approach ensures each positioning solution is computed using a consistent set of near-simultaneous measurements. Fig. 3b shows the positioning results in each static period. As can be seen, the positioning results do not surround the ground truth, indicating the presence of SMM. This is consistent with the findings in Fig. 4. Given the observations in Figs. 4, 5, and 3b, an ideal credibility assessment method should identify the positioning estimation of such a dataset as both SMM and pessimism.



(a) (b)

Fig. 3: (a) The coordinates of the UWB tag against time, where the static period is marked as the shaded area; (b) The 2-D trajectory of the UWB tag and the positioning solutions at six static locations. The 'star' stands for ground-truth location, and the 'circle' represents the positioning estimation.



(a)

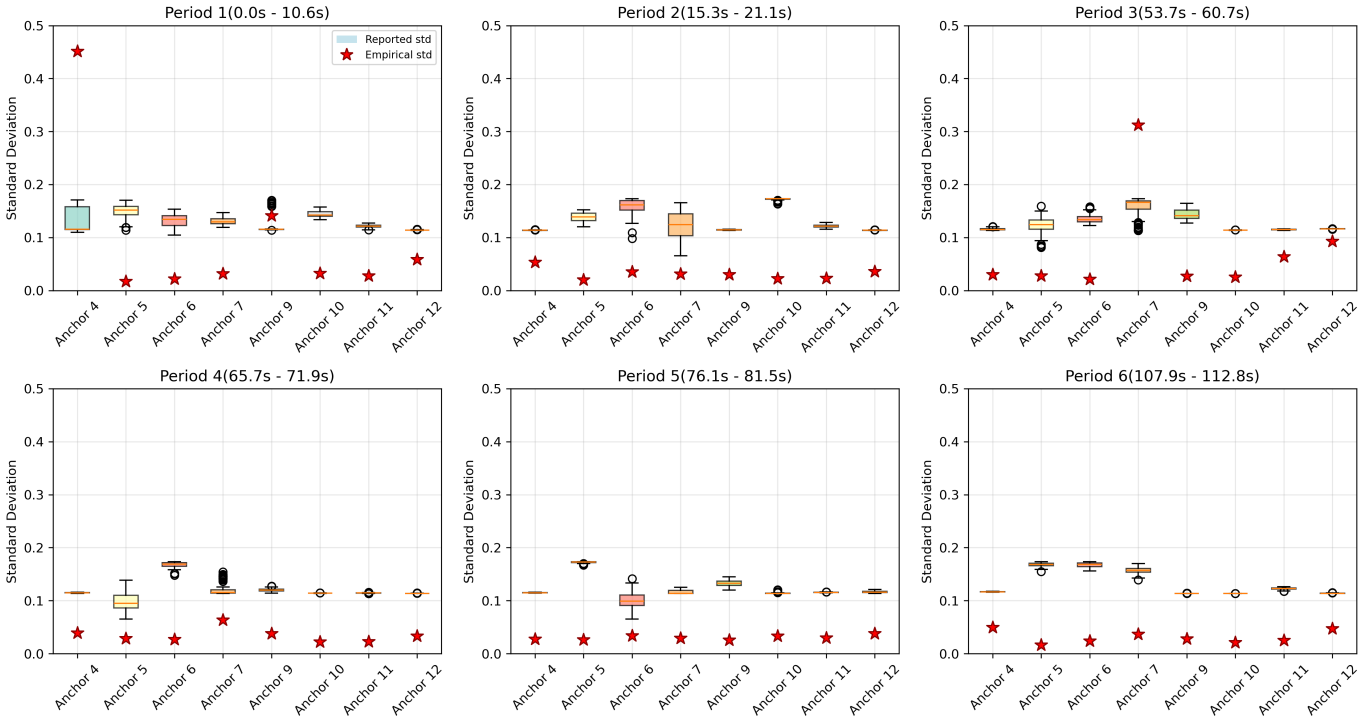
Fig. 4: The distribution of UWB range measurement errors (related to each anchor) at each static period.

B. Diagnosis Result Analysis

Table IV compares the credibility diagnosis results between our proposed method and the two baseline approaches. As predicted by our initial data analysis, the proposed method correctly identifies the combination of pessimism and SMM in periods 2-6. In contrast, baseline methods failed to achieve this complete diagnosis. Moreover, both the NCI and NEES based methods fail to identify the SMM in estimation. These results demonstrate that the proposed method provides a more nuanced and complete credibility assessment compared to baseline methods, successfully capturing both the SMM and NMM revealed in the dataset.

VI. Conclusions

In this work, we introduced a unified framework for evaluating the credibility of state estimators by comprehensively integrating a suite of metrics, including NCI, NLL, and ES. Experimental results from Monte Carlo simulations and real-world UWB positioning experiments confirmed the superiority of our multi-metric approach. The proposed method achieved high classification accuracy across a range of challenging scenarios, whereas traditional single-metric methods show unreliable results. This work offers practitioners a powerful tool to not only validate estimator performance but also to diagnose specific modeling failures.



(a)

Fig. 5: The distribution of reported standard deviations at each static period. The red stars indicate the empirical standard deviation, which is calculated using the UWB range measurement errors.

TABLE IV: Diagnosis results for UWB positioning on STAR-loc dataset

Period ID	Labeled	Proposed	NEES	NCI
1	Pessimism + SMM	SMM	Pessimism	Optimism
2	Pessimism + SMM	Pessimism + SMM	Pessimism	Optimism
3	Pessimism + SMM	Pessimism + SMM	Pessimism	Pessimism
4	Pessimism + SMM	Pessimism + SMM	Pessimism	Pessimism
5	Pessimism + SMM	Pessimism + SMM	Pessimism	Pessimism
6	Pessimism + SMM	Pessimism + SMM	Credible	Pessimism

APPENDIX A
NON-NEGATIVITY OF NCI

Step 1: Simplify the Matrix Inverse

Let $\hat{\Sigma}_k = \Sigma_k$ and $\mu_k \neq 0$, we have

$$NCI(\{\hat{x}_k\}) = \frac{10}{N}\sum_{k=1}^{N} \log_{10} \frac{e_k^T \Sigma_k^{-1} e_k}{e_k^T (\Sigma_k + \mu_k \mu_k^T)^{-1} e_k}.$$

The denominator contains the inverse of a sum of a matrix and a rank-1 matrix, which can be simplified using the Woodbury matrix identity:

$$(\Sigma_k + \mu_k \mu_k^T)^{-1} = \Sigma_k^{-1} - \Sigma_k^{-1}\mu_k(1 + \mu_k^T \Sigma_k^{-1}\mu_k)^{-1}\mu_k^T \Sigma_k^{-1}. \tag{20}$$

Since $(1 + \mu_k^T \Sigma_k^{-1}\mu_k)$ is a scalar, we can write the inverse as:

$$(\Sigma_k + \mu_k \mu_k^T)^{-1} = \Sigma_k^{-1} - \frac{\Sigma_k^{-1}\mu_k \mu_k^T \Sigma_k^{-1}}{1 + \mu_k^T \Sigma_k^{-1}\mu_k}. \tag{21}$$

Step 2: Simplify the Denominator

Now, substitute this simplified inverse into the denominator of the original expression:

$$e_k^T (\Sigma_k + \mu_k \mu_k^T)^{-1} e_k = e_k^T \left(\Sigma_k^{-1} - \frac{\Sigma_k^{-1}\mu_k \mu_k^T \Sigma_k^{-1}}{1 + \mu_k^T \Sigma_k^{-1}\mu_k}\right) e_k \tag{22}$$

$$= e_k^T \Sigma_k^{-1} e_k - \frac{e_k^T \Sigma_k^{-1}\mu_k \mu_k^T \Sigma_k^{-1} e_k}{1 + \mu_k^T \Sigma_k^{-1}\mu_k}. \tag{23}$$

Since $e_k^T \Sigma_k^{-1}\mu_k$ and $\mu_k^T \Sigma_k^{-1} e_k$ are scalars and are transposes of each other, they are equal. The denominator can be simplified to:

$$\frac{(e_k^T \Sigma_k^{-1} e_k)(1 + \mu_k^T \Sigma_k^{-1}\mu_k) - (e_k^T \Sigma_k^{-1}\mu_k)^2}{(1 + \mu_k^T \Sigma_k^{-1}\mu_k)}. \tag{24}$$

Step 3: Simplify the Overall Expression

Substituting the simplified denominator into the NCI expression, the finalized NCI is given by:

$$\log_{10}\left(\frac{(e_k^T \Sigma_k^{-1} e_k)(1 + \mu_k^T \Sigma_k^{-1}\mu_k)}{(e_k^T \Sigma_k^{-1} e_k)(1 + \mu_k^T \Sigma_k^{-1}\mu_k) - (e_k^T \Sigma_k^{-1}\mu_k)^2}\right). \tag{25}$$

Given that $\mu_k \neq 0$, it follows that $\mu_k^T \Sigma_k^{-1} \mu_k$ is always positive, and $e_k^T \Sigma_k^{-1} e_k$ is non-negative. We can use the Cauchy-Schwarz inequality for the inner product defined by Σ_k^{-1}:

$$(e_k^T \Sigma_k^{-1} \mu_k)^2 \leq (e_k^T \Sigma_k^{-1} e_k)(\mu_k^T \Sigma_k^{-1} \mu_k)\,. \tag{26}$$

Therefore, the denominator is always positive. It is obvious that the numerator is always greater than or equal to the denominator, because $(e_k^T \Sigma_k^{-1} \mu_k)^2$ is non-negative. Therefore, the term in the bracket is always equal to or larger than 1, and thus NCI is always non-negative.

Appendix B
Asymmetric Properties of NLL

Eq. (8) contains the term $1/\rho + \ln(\rho) > 0$. For $\rho > 1$, we have $1/\rho + \ln(\rho) > 0$ and $\rho - \ln(\rho) > 0$. The difference between these two expressions is:

$$\Delta(\rho) = (\rho - \ln(\rho)) - \left(\frac{1}{\rho} + \ln(\rho)\right) = \rho - \frac{1}{\rho} - 2\ln(\rho)\,. \tag{27}$$

To determine the sign of $\Delta(\rho)$ for $\rho > 1$, we calculate the derivative of $\Delta(\rho)$ with respect to ρ:

$$\frac{d\Delta}{d\rho} = 1 + \frac{1}{\rho^2} - \frac{2}{\rho} = \left(1 - \frac{1}{\rho}\right)^2\,. \tag{28}$$

For $\rho > 1$, the term $(1 - 1/\rho)$ is positive, which implies that $(1 - 1/\rho)^2 > 0$. Thus, $d\Delta/d\rho > 0$, and $\Delta(\rho)$ is a strictly increasing function for $\rho > 1$.

At the boundary point $\rho = 1$, we have:

$$\Delta(1) = 1 - \frac{1}{1} - 2\ln(1) = 0\,. \tag{29}$$

Since $\Delta(\rho)$ is increasing for $\rho > 1$ and $\Delta(1) = 0$, it follows that $\Delta(\rho) > 0$ for all $\rho > 1$. This proves that for any $\rho > 1$:

$$\rho - \ln(\rho) > \frac{1}{\rho} + \ln(\rho)\,. \tag{30}$$

Therefore,

$$|\mathbb{E}[\mathrm{NLL}](\rho)| < |\mathbb{E}[\mathrm{NLL}](1/\rho)|\,. \tag{31}$$

Appendix C
Non-negativity of ES

The energy distance is a metric for measuring the distance between two probability distributions. For two independent random variables X and Y, the energy distance is defined as:

$$\mathcal{E}(X,Y) = 2\mathbb{E}\|X - Y\|_2 - \mathbb{E}\|X - X'\|_2 - \mathbb{E}\|Y - Y'\|_2 \tag{32}$$

where X, X' are independent and identically distributed (i.i.d.) and Y, Y' are i.i.d. A key property of the energy distance is that it is always non-negative. $\mathcal{E}(X,Y) \geq 0$, and $\mathcal{E}(X,Y) = 0$ if and only if the distributions of X and Y are identical.

The energy score is a special case of the energy distance. We can interpret the true observed outcome x_k as a degenerate probability distribution—a Dirac measure, which is a distribution that places all its probability mass at a single point, x_k. Let's call this distribution G_k.

Let the predictive distribution be F_k and the true distribution be the Dirac measure G_k. We can now express the energy distance between F_k and G_k:

$$\begin{aligned}\mathcal{E}(F_k, G_k) =& 2\mathop{\mathbb{E}}_{\substack{Y\sim F_k \\ X\sim G_k}}\|Y - X\|_2 - \mathop{\mathbb{E}}_{\text{i.i.d.}}{}_{Y,Y'\sim F_k}\|Y - Y'\|_2 \\ &- \mathop{\mathbb{E}}_{\text{i.i.d.}}{}_{X,X'\sim G_k}\|X - X'\|_2\,.\end{aligned} \tag{33}$$

The first term is $\mathbb{E}\|Y - X\|_2$. Since X is always equal to x_k, this simplifies to $\mathbb{E}_{Y\sim F_k}\|Y - x_k\|_2$. The second term, $\mathbb{E}\|Y - Y'\|_2$, is identical to the second term of the ES definition. The third term is $\mathbb{E}\|X - X'\|_2$. Since both X and X' are always equal to x_k, their distance is always zero, so this term is 0. Therefore, the energy distance can be simplified as:

$$\mathcal{E}(F_k, G_k) = 2\mathbb{E}_{Y\sim F_k}\|Y - x_k\|_2 - \mathbb{E}_{Y,Y'\sim F_k}\|Y - Y'\|_2\,. \tag{34}$$

By comparing this to the ES definition, we can see that:

$$ES(F_k, x_k) = \tfrac{1}{2}\mathcal{E}(F_k, G_k)\,. \tag{35}$$

It is known that the energy distance $\mathcal{E}(F_k, G_k)$ is always non-negative; and therefore, the energy score, which is half of the energy distance, must also be non-negative. The minimum value of the energy score is 0, and this occurs when the predictive distribution F_k perfectly matches the true distribution G_k (i.e., when the model predicts the exact true outcome with 100% certainty).

References

[1] V. P. Dubey, J. Saha, S. Bhaumik, and A. Dey, "Tracking an underwater target in a large surveillance region with sensor location uncertainty," *IEEE Access*, vol. 11, pp. 140 007–140 021, 2023.

[2] Z. Chen, H. Biggie, N. Ahmed, S. Julier, and C. Heckman, "Kalman filter auto-tuning with consistent and robust bayesian optimization," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 60, no. 2, pp. 2236–2250, 2024.

[3] Y. Gao, Z. Lv, and L. Zhang, "Two-step trajectory spoofing algorithm for loosely coupled gnss/imu and nis sequence detection," *IEEE access*, vol. 7, pp. 96 359–96 371, 2019.

[4] J. Wang, X. Chen, C. Shi, and J. Liu, "Robust m-estimation-based ickf for gnss outlier mitigation in gnss/sins navigation applications," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–17, 2023.

[5] K. Haggag, S. Lange, T. Pfeifer, and P. Protzel, "A credible and robust approach to ego-motion estimation using an automotive radar," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6020–6027, 2022.

[6] J. Dunik, O. Straka, M. Simandl, and E. Blasch, "Random-point-based filters: Analysis and comparison in target tracking," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 51, no. 2, pp. 1403–1421, 2015.

[7] M. Xia, T. Zhang, J. Wang, L. Zhang, Y. Zhu, and L. Guo, "The fine calibration of the ultra-short baseline system with inaccurate measurement noise covariance matrix," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–8, 2021.

[8] P. Chauchat, J. Vilà-Valls, and P. Closas, "Robust information filtering under model mismatch for large-scale dynamic systems," *IEEE Control Systems Letters*, vol. 6, pp. 158–163, 2021.

[9] Q. Ge, T. Shao, Z. Duan, and C. Wen, "Performance analysis of the kalman filter with mismatched noise covariances," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 4014–4019, 2016.

[10] S. Fortunati, F. Gini, M. S. Greco *et al.*, "Performance bounds for parameter estimation under misspecified models: Fundamental findings and applications," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 142–157, 2017.

[11] K. Li, X. Chen, and G. Zhou, "Maneuvering target tracking in constraint coordinates with radar measurements," in *2016 IEEE Radar Conference (RadarConf)*. IEEE, 2016, pp. 1–6.

[12] X. R. Li, Z. Zhao, and X.-B. Li, "Evaluation of estimation algorithms: Credibility tests," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 42, no. 1, pp. 147–163, 2011.
[13] X. R. Li and Z. Zhao, "Measuring estimator's credibility: Noncredibility index," in *2006 9th International Conference on Information Fusion*. IEEE, 2006, pp. 1–8.
[14] E. P. Blasch, O. Straka, J. Dunik, and M. Simandl, "Multitarget tracking performance analysis using the non-credibility index in the nonlinear estimation framework (nef) toolbox," in *Proceedings of the IEEE 2010 National Aerospace & Electronics Conference*. IEEE, 2010, pp. 17–24.
[15] Y. Zhang and X. Li, "Joint tracking and classification noncredibility index," in *2023 26th International Conference on Information Fusion (FUSION)*. IEEE, 2023, pp. 1–8.
[16] S. Chen, Q. Zhang, D. Lin, and S. Wang, "Generalized loss based geometric unscented kalman filter for robust power system forecasting-aided state estimation," *IEEE Signal Processing Letters*, vol. 29, pp. 2353–2357, 2022.
[17] P. Denti, P. Vicini, A. Bertoldo, and C. Cobelli, "Glucose minimal model population analysis: likelihood function profiling via monte carlo sampling," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2008, pp. 4932–4935.
[18] M. Walker, M. Reith-Braun, and U. Hanebeck, "Weaknesses of the anees and new calibration measures for multivariate predictions," in *2025 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, 2025.
[19] D. Bhatt, K. Mani, D. Bansal, K. Murthy, H. Lee, and L. Paull, "f-cal: Aleatoric uncertainty quantification for robot perception via calibrated neural regression," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6533–6539.
[20] S. Chai, Z. Xu, and Y. Jia, "Conditional density forecast of electricity price based on ensemble elm and logistic emos," *IEEE transactions on smart grid*, vol. 10, no. 3, pp. 3031–3043, 2018.
[21] L. Dang, J. Yang, M. Liu, and B. Chen, "Differential equation-informed neural networks for state-of-charge estimation," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–15, 2023.
[22] X. Chen, H. Zhang, F. Zhao, Y. Cai, H. Wang, and Q. Ye, "Vehicle trajectory prediction based on intention-aware non-autoregressive transformer with multi-attention learning for internet of vehicles," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.
[23] 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.
[24] T. Gneiting and M. Katzfuss, "Probabilistic forecasting," *Annual Review of Statistics and Its Application*, vol. 1, pp. 125–151, 2014.
[25] A. Ashok, É. Marcotte, V. Zantedeschi, N. Chapados, and A. Drouin, "Tactis-2: Better, faster, simpler attentional copulas for multivariate time series," *arXiv preprint arXiv:2310.01327*, 2023.
[26] M. Al-Gabalawy, N. S. Hosny, and A. R. Adly, "Probabilistic forecasting for energy time series considering uncertainties based on deep learning algorithms," *Electric Power Systems Research*, vol. 196, p. 107216, 2021.
[27] G. J. Székely and M. L. Rizzo, "Energy statistics: A class of statistics based on distances," *Journal of statistical planning and inference*, vol. 143, no. 8, pp. 1249–1272, 2013.
[28] F. Dümbgen, M. A. Shalaby, C. Holmes, C. C. Cossette, J. R. Forbes, J. L. Ny, and T. D. Barfoot, "Star-loc: Dataset for stereo and range-based localization," *arXiv preprint arXiv:2309.05518*, 2023.