JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# Adaptive Sensor Confidence Learning for Resilient UAV Navigation Under GNSS Degradation and Multiple Sensor Failures

***UJJWAL MANNA***
*(2025012360)*
*B.Tech, Computer Science Engineering (Artificial Intelligence and Machine Learning)*
*GITAM University, Visakhapatnam, India*

## *Abstract*

Unmanned aerial vehicle (UAV) navigation systems commonly combine measurements from multiple sensors to estimate vehicle position and motion. A practical difficulty arises when a sensor continues to provide measurements after its quality has degraded. In such situations, simply detecting the presence or absence of a measurement is insufficient because a numerically plausible but inaccurate observation can continue to influence the state estimate.

This study investigates a temporal sensor-confidence framework for estimating the short-term usability of external sensor measurements and incorporating that information into a conventional navigation estimator. The proposed framework uses recent sensor observations, measurement residuals, availability information, measurement age, and temporal consistency to estimate a confidence value for individual sensor channels. The confidence value is then mapped to a bounded measurement-covariance adjustment, while persistent low confidence may result in measurement rejection. Inertial measurements are treated separately because they also contribute to state propagation.

Three estimation strategies are considered: fixed sensor fusion, innovation-based adaptive fusion, and a learned temporal confidence model. The initial pilot evaluation used 30 held-out trajectories distributed across five simulated conditions: healthy motion, progressive GNSS degradation, abrupt GNSS faults, GNSS outage, and concurrent GNSS–vision degradation. In the reported pilot simulation, mean 3-D position RMSE was 1.223 m for fixed fusion, 1.618 m for the innovation-adaptive method, and 1.247 m for the learned-confidence method. The learned approach performed better than fixed fusion in the progressive and abrupt GNSS conditions but performed worse in healthy and concurrent-failure conditions.

A separate expanded simulation study used 240 unseen test trajectories distributed across eight simulated conditions, including three fault forms excluded from training: intermittent GNSS spikes, visual degradation with dropout, and stale GNSS measurements. In the reported synthetic evaluation, the learned-confidence method produced a mean 3-D RMSE of 0.659 m compared with 0.750 m for fixed fusion. However, the improvement was not uniform across conditions, and the innovation-adaptive baseline remained competitive or superior in several scenarios. Confidence calibration was also evaluated using Brier score and expected calibration error. The expanded simulation produced a Brier score of 0.095 and an expected calibration error of 0.098 for the temporal confidence model.

The findings provide simulation-level evidence that temporal sensor-confidence estimation can be useful for selected forms of sensor degradation. They do not establish physical UAV flight performance, safety, embedded real-time performance, or operational reliability. Physical sensor-log replay, target-hardware benchmarking, and controlled flight experiments are required before deployment-oriented conclusions can be drawn.

Keywords: UAV navigation; GNSS degradation; sensor confidence; adaptive estimation; temporal learning; fault tolerance; uncertainty calibration

### 1. INTRODUCTION

Reliable navigation is an important requirement for autonomous UAV systems. Modern UAV navigation architectures frequently combine inertial measurements with absolute and relative observations obtained from GNSS, cameras, barometers, lidar, and other sensing modalities. Although sensor fusion can provide robustness through redundancy, the presence of multiple sensors does not automatically guarantee reliable estimation when one or more measurements become degraded.

A sensor failure may not appear as a complete loss of data. A GNSS receiver may continue producing position estimates while developing a bias. A visual system may continue generating feature observations while the associated measurements become less reliable. An inertial sensor may remain available while accumulating bias or experiencing saturation. These situations are difficult because the estimator may continue treating the measurements as valid unless an explicit mechanism is available for evaluating their reliability.

Conventional state estimators generally represent measurement uncertainty through covariance matrices. In a fixed fusion system, these covariances remain unchanged unless a predetermined configuration is modified. Adaptive estimators can adjust measurement influence according to residual behaviour. However, residual magnitude alone may not provide sufficient temporal information to distinguish transient measurement variation from persistent degradation.

Machine-learning methods provide another possible approach. Instead of directly replacing a state estimator, a learned model can operate as an additional sensor-health layer and estimate whether individual measurements are likely to remain within a specified error tolerance. The resulting estimate can then be incorporated into an otherwise conventional estimator.

The present study investigates this narrower problem. It does not attempt to replace established visual-inertial navigation systems or conventional filtering techniques. Instead, it evaluates whether a lightweight temporal model can estimate sensor usability from causal information and whether that estimate can be used to modify measurement influence within a conventional navigation estimator.

The study considers three main questions.

First, can a learned sensor-confidence model reduce navigation error under selected simulated sensor-degradation conditions?

Second, can the confidence output provide a useful indicator of sensor-health behaviour before an unadapted reference estimator exceeds a predefined simulated error threshold?

Third, does the observed behaviour continue on simulated fault forms that were not included in the model's training distribution?

The study is intentionally limited to synthetic software simulation. No physical UAV flight experiment, radio-frequency interference experiment, or field deployment is reported.

## 2. Related Work

### *2.1 Navigation and sensor fusion*

Visual and inertial measurements complement one another, but their usefulness depends on calibration, state representation, and how measurement uncertainty is handled [7,8]. Error-state filtering has also been applied to multisensor UAV localization [9]. Other architectures, including R2LIVE, FAST-LIO2, and LIO-SAM, show different ways of combining lidar, visual, and inertial information [10–12]. DIDO is especially relevant here because it highlights the importance of inertial health in quadrotor odometry [13]. Taken together, these studies make it clear that sensor fusion and recovery are already mature research areas; the contribution of this work is therefore not the use of sensor fusion itself.

### *2.2 Adaptive and learned estimation*

Learning-based methods have increasingly been used to modify or support model-based estimators. KalmanNet adds a recurrent learned component to filtering [3], while A-KIT learns adaptive process information in a different navigation setting [4]. Selective Kalman filtering examines when additional sensing should be incorporated [5], and DeepUKF-VIN explores learned covariance adjustment for visual-inertial navigation [6]. Because these studies overlap with the present idea, this work focuses on calibrated sensor confidence and the experimental comparison rather than presenting machine learning alone as the main contribution.

### *2.3 Confidence and uncertainty*

A confidence score is useful only if its numerical meaning is reasonably well calibrated. Methods such as temperature scaling provide a way to compare predicted probabilities with observed frequencies [20]. Research on uncertainty under dataset shift also shows why evaluation should include conditions that differ from those used during training [21]. Ensembles and dropout provide additional approaches to uncertainty estimation [22,23], while out-of-distribution detection addresses a related but distinct problem [24,25]. These ideas motivate the calibration checks and held-out fault evaluation used in this study.

| Study | Established focus | Relevance to this study |
|---|---|---|
| [1] VINS-Mono | Visual-inertial state estimation | Shows that robust visual-inertial estimation already exists |
| [2] ORB-SLAM3 | Visual-inertial SLAM and relocalization | Shows that recovery is not unique to this work |
| [3] KalmanNet | Learned model-based filtering | Closest broad learned-filtering precedent |
| [4] A-KIT | Adaptive learned estimation | Supports comparison with learned adaptation |
| [5] Selective KF | Selective sensor fusion | Relevant conventional adaptive baseline |
| [6] DeepUKF-VIN | Learned covariance tuning | Closest overlap in covariance adaptation |
| [9] Markovic et al. | Multisensor UAV localization | Direct UAV sensor-fusion precedent |
| [13] DIDO | Quadrotor odometry | Highlights inertial health and propagation |
| [18] Panda and Guo | Temporal GNSS anomaly detection | Shows that early-warning research already exists |
| [20] Guo et al. | Neural-network calibration | Supports explicit confidence calibration |
| [21] Ovadia et al. | Uncertainty under dataset shift | Motivates held-out fault and trajectory testing |

### *2.4 What already exists, and the narrower gap addressed here*

The literature leaves little room for a credible claim that sensor fusion, visual-inertial recovery, learned Kalman filtering, covariance adaptation, or GNSS anomaly detection are new by themselves. The useful question is therefore more specific: can a small causal model produce a calibrated, sensor-by-sensor probability of measurement usability and feed that probability into an otherwise

conventional estimator without using future information or simulator truth at deployment time? Table 1 makes that distinction explicit.

Table 1. Research-positioning matrix. Established solution families and the narrower contribution evaluated in this manuscript.

| Existing line of work | What is already established | Gap relevant to this manuscript | What is implemented here |
|---|---|---|---|
| VIO / SLAM [1,2,7-12] | Accurate visual-inertial state estimation, mapping, relocalization, and recovery | These systems do not by themselves define a calibrated probability that each external sensor is inside a declared physical error tolerance | A separate health layer outputs simultaneous GNSS and visual confidence values while the navigation estimator remains model based |
| Learned filtering and covariance adaptation [3,4,6,17,31] | Neural components can learn filter gains or adapt process / measurement uncertainty | A learned covariance is not automatically an interpretable sensor-health probability, and multi-sensor failure attribution may remain implicit | Confidence is defined as a probability of tolerance compliance, calibrated on held-out trajectories, then mapped through a bounded engineering rule |
| Innovation-based / selective fusion [5] | Residual magnitude can gate or down-weight measurements without a learned model | Single-step innovation tests can react late to drift or confound model mismatch with sensor failure | A one-second causal history uses residual level, slope, persistence, disagreement, availability, and age |
| GNSS anomaly / spoofing detection [18] | Dedicated detectors can identify abnormal GNSS behaviour and provide warning | A detector focused on GNSS does not directly decide how several sensor channels should be weighted together | The same interface gives per-sensor confidence and supports concurrent low-confidence outputs |
| Calibration and shift-aware uncertainty [20,21-25] | Probability calibration and evaluation under distribution shift are established tools | Navigation papers often report only state error, leaving the numerical meaning of learned confidence unclear | Brier score, reliability analysis, healthy false alarms, unseen fault forms, and navigation error are reported together |

Contribution statement. The paper does not claim novelty for Kalman fusion or for neural adaptation in isolation. Its contribution is the integrated formulation and the evidence around it: (i) confidence has an operational definition tied to a physical error tolerance; (ii) multiple sensors can be judged unreliable at the same time; (iii) residual monitoring uses nominal covariance so the health decision does not become self-justifying after covariance inflation; (iv) low confidence first reduces influence through a bounded map and only then permits persistent rejection; and (v) evaluation separates fault injection time from measurement-usability failure and includes calibration, false alarms, unseen fault forms, and navigation consequences.

## 3. Problem Formulation

The navigation state used in the present simulation is represented as x = [p_x, p_y, p_z, v_x, v_y, v_z]^T, where p denotes three-dimensional position and v denotes three-dimensional velocity. The study focuses on position-estimation performance under simulated GNSS and visual-sensor degradation; attitude and individual inertial-sensor bias states are not explicitly estimated.
For an external sensor i, the measurement is written as:

$$z_{i,k} = h_i(x_k) + \eta_{i,k} + d_{i,k}$$

where η represents nominal measurement noise and d represents a departure from the nominal measurement model. An availability flag records whether a measurement is present and valid. Missing data are kept missing rather than being silently replaced with zeros or interpolated using future information.
The confidence value $c_{i,k}$ is defined as the estimated probability that sensor i remains inside its declared physical error tolerance at time k, using only information available up to that point. Simulator ground truth may be used to create labels during offline training, but it is not supplied to the confidence model when it is making a deployed decision.
We also distinguish the moment a fault begins from the moment a measurement becomes unusable. A disturbance can start at one instant and remain small enough to tolerate for some time. Because of this, detector performance is evaluated against the point at which the measurement becomes unreliable rather than simply against the instant at which an artificial fault was injected.

## 4. Proposed Method

### *4.1 Architecture and timing*

We keep sensor-health estimation from navigation control. Timestamped sensor observations are first converted into causal features. A temporal model then estimates a confidence value for each sensor. Those values are passed to the estimator through bounded covariance adaptation and persistent rejection. A separate controller receives the resulting navigation state; motor commands are not part of the learned health model.
Both acquisition time and arrival time are retained. A monotonic time base, documented coordinate transforms, fixed sensor extrinsics, and explicit handling of measurement age are required. A late measurement is either rejected or processed through a documented delayed-update mechanism. This is intended to prevent future information from leaking into the health decision.


Figure 1. Calibrated GNSS and visual confidence on a representative concurrent-failure trajectory from the expanded held-out verification. The vertical markers indicate the GNSS fault at 12 s and the visual fault at 14 s.

### *4.2 Residuals and shared information*

Each available measurement is compared with a prediction formed before that measurement is applied. The measurement residual is

$$r_i,k = z_i,k - h_i(\hat{x}_{k|k}^{-1}),$$

and the nominal innovation covariance is

$$S_i,k = H_i P_k H_i^T + R_i.$$

The normalized residual gives a compact indication of how surprising a measurement is relative to the current state estimate.
The monitoring branch uses the nominal covariance when evaluating residuals. This avoids a circular effect in which a low-confidence decision first inflates the covariance and then makes the same faulty measurement appear less surprising. Shared information is another consideration: a visual-inertial pose that already includes IMU information should not be treated as fully independent when the same IMU information is used elsewhere in the estimator.

### *4.3 Causal features*

| Feature group | Examples | Important qualification |
|---|---|---|
| Timing and availability | Age, missing mask, time since valid update | No future interpolation or fault schedule |
| GNSS consistency | Position/velocity residual, receiver quality | Reference cannot be simulator truth |
| Visual consistency | Reprojection residual, track count, track age | Track count alone does not establish accuracy |
| Vertical consistency | GNSS–barometer difference, vertical increments | Altitude datum and barometer bias must be aligned |
| Inertial consistency | Bias estimate, saturation, motion residual | True acceleration must not be mistaken for bias |
| Temporal statistics | Residual slope, robust spread, persistence | Scaling must be learned only from training data |

### *4.4 Temporal learning and calibration*

The starting model is a small causal one-dimensional convolutional network with two layers, 32 channels, kernel size 3, and dilations of 1 and 2. Separate sigmoid outputs are used for the sensor channels so that more than one sensor can be considered unreliable at the same time. Window lengths of 0.25, 0.5, and 1.0 seconds are proposed for a future comparison with an instantaneous classifier and a compact GRU.
We train the model using binary cross-entropy over valid labelled observations. We separate training and validation data by trajectory rather than by randomly shuffling individual windows. We evaluate calibration on held-out trajectories using the Brier score and reliability analysis. Put simply, a confidence value of 0.8 should eventually correspond to something close to an 80% chance that the sensor is within the selected tolerance.

### *4.5 Bounded adaptation and failure handling*

We map the confidence value to a bounded covariance multiplier. Lower confidence increases the measurement covariance and therefore reduces the influence of that measurement. This mapping is treated as a controlled engineering mechanism rather than as

a physical variance model. Persistent low confidence triggers rejection because simply inflating covariance does not remove systematic bias.

Hysteresis is used to prevent rapid switching around a single threshold. A sensor is excluded only after its confidence remains below the rejection threshold for a specified dwell time. Readmission requires a higher confidence level together with a separate recovery interval. The estimator also checks whether the remaining observations are sufficient to support the required state.

## 5. Experimental Protocol

### *5.1 Scope of the pilot experiment*

To provide measurable evidence without overstating what was tested, this study includes a software-based pilot experiment designed to evaluate the proposed confidence and fusion logic. The experiment is a synthetic software-in-the-loop-style sensor-replay study. It is not a physical UAV flight test, an RF-interference test, or a hardware validation study. Its purpose is narrower: to see whether the proposed confidence and fusion logic produces measurable changes when controlled sensor faults are introduced.

The simulation uses a 0.1 s update period and 30 s trajectories. The navigation state consists of 3-D position and velocity driven by noisy inertial acceleration. GNSS and visual position observations are generated with nominal noise and then modified according to the selected fault scenario. The same lightweight Kalman estimator is used across all three comparison methods.

### *5.2 Dataset separation*

A total of 150 trajectories were generated: 30 healthy, 30 progressive-GNSS, 30 abrupt-GNSS, 30 GNSS-outage, and 30 concurrent GNSS–vision fault trajectories. For each scenario, 18 trajectories were used for training, 6 for validation, and 6 for final testing. The reported results therefore come from 30 held-out test trajectories, with six trajectories per scenario. Different versions of the same base trajectory were kept in the same partition to reduce the possibility of leakage between training and testing.

### *5.3 Fault scenarios*

| Scenario | Fault condition | Question tested |
|---|---|---|
| Healthy | Nominal sensors with varied motion | Does adaptation damage normal navigation? |
| Progressive GNSS | Gradually increasing GNSS bias/noise | Can GNSS influence be reduced before the error grows? |
| Abrupt GNSS | Sudden GNSS position offset | How quickly can the influence of a faulty observation be reduced? |
| GNSS outage | GNSS unavailable after fault onset | How much drift remains without an absolute position source? |
| Concurrent GNSS + vision | GNSS fault followed by visual fault | What happens when redundancy is reduced further? |

### *5.4 Baselines and evaluation metrics*

Three methods were compared. The first was fixed fusion, which uses the nominal measurement covariance. The second was conventional adaptive fusion, which changes covariance according to innovation magnitude. The third was the learned-confidence method proposed here, which scales or rejects observations according to predicted confidence. We use 3-D position RMSE as the primary metric. Maximum position error, detection behaviour, and confidence calibration were treated as secondary measures.

For early-warning analysis, the fault onset $t_0$ was kept separate from the point at which a frozen fixed-fusion reference estimator first exceeded a sustained 2 m position-error limit. This prevents the adaptive method from defining the warning target against which it is evaluated.

### *5.5 Expanded held-out verification added after the pilot*

The original 30-trajectory pilot is retained because it shows the first implementation without hiding its weak cases. A second experiment was added to test whether the same design survives a larger and deliberately less comfortable set of conditions. This is a separate synthetic reproducibility run, not a relabelling of the pilot data. It uses 90 training trajectories, 30 validation trajectories, and 240 independent test trajectories. Each test scenario contains 30 trajectories. Training and validation use the five scenario families already defined in Section 5.3; three additional test families—intermittent GNSS spikes, visual degradation with dropout, and stale GNSS—are never used as training fault forms.

The state estimator uses the same six-dimensional position-velocity state, [p_x, p_y, p_z, v_x, v_y, v_z]^T, throughout the pilot and expanded simulations. The estimator does not explicitly model attitude or individual accelerometer, gyroscope, or barometer bias states.

Table 2 : Expanded-verification protocol and claim boundary.

| Item | Expanded verification value |
|---|---|
| Training / validation / test | 90 / 30 / 240 trajectories |
| Test scenarios | 8, with 30 trajectories each |
| Unseen fault forms | GNSS spikes; visual degradation/dropout; stale GNSS |
| Trajectory duration / sampling | 30 s / 0.1 s |

| Item | Expanded verification value |
|---|---|
| Temporal window | 1.0 s (10 samples) |
| Confidence model | 2-layer causal 1-D convolution, 32 channels, 4,642 parameters |
| GNSS / visual tolerance | 2.0 m / 1.25 m 3-D measurement error |
| Calibration | Temperature scaling on validation data only; fitted temperature 0.817 |
| Statistical reporting | Paired per-trajectory RMSE; 4,000-resample bootstrap 95% CI; Wilcoxon signed-rank used as exploratory paired test |
| Evidence boundary | Synthetic software verification; no physical flight, RF interference, or target-compute claim |

## 6. Experimental Results

The results below are outputs of the pilot software simulation described in Section 5. They are reported as measured simulation results, not as physical flight-test evidence. Including them makes the experimental basis of the manuscript explicit while keeping the boundary between simulation and real-world validation clear.

Table 3. Mean 3-D position RMSE across the held-out trajectories. Negative values indicate an improvement of the learned method relative to fixed fusion.

| Scenario | Test n | Fixed RMSE (m) | Adaptive RMSE (m) | Learned RMSE (m) | Learned vs fixed |
|---|---|---|---|---|---|
| Healthy | 6 | 0.494 | 0.783 | 0.513 | +3.7% |
| Progressive GNSS | 6 | 0.735 | 0.801 | 0.725 | −1.3% |
| Abrupt GNSS | 6 | 1.461 | 0.922 | 1.383 | −5.3% |
| GNSS outage | 6 | 0.531 | 0.923 | 0.549 | +3.4% |
| Concurrent GNSS + vision | 6 | 2.896 | 4.660 | 3.066 | +5.9% |
| Overall | 30 | 1.223 | 1.618 | 1.247 | +1.9% |

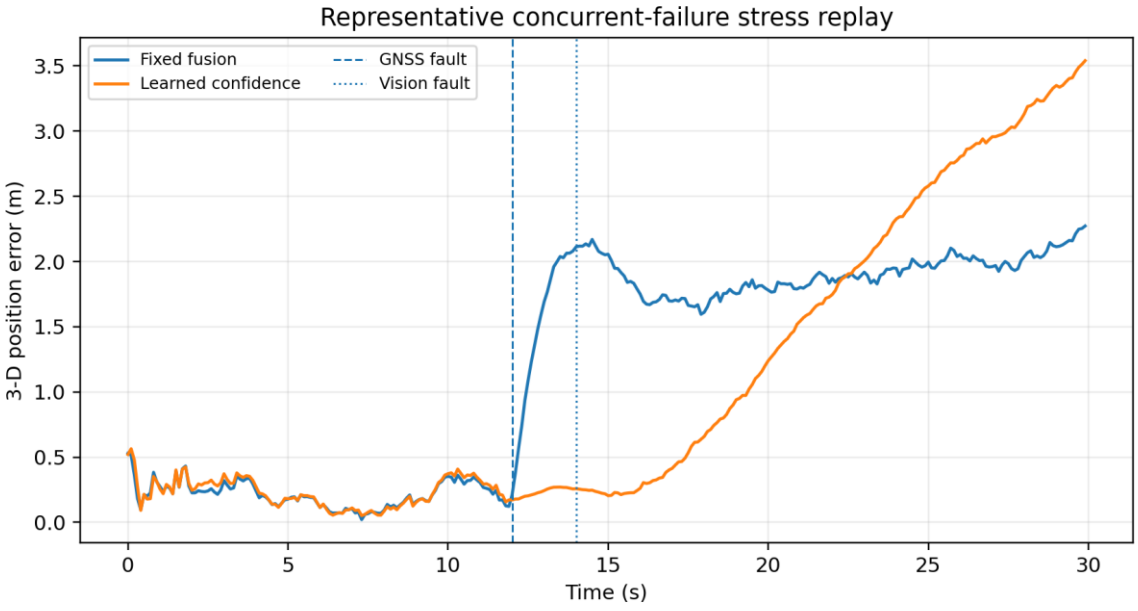


Figure 2. Position-error timeline on the same expanded-verification trajectory. The comparison is shown as a diagnostic example rather than as a substitute for the aggregate results.

***Table 4. Expanded held-out verification: aggregate navigation error.***

| *Method* | *Mean 3-D RMSE (m)* |
|---|---|
| ***Fixed fusion*** | ***0.750*** |
| ***Learned confidence*** | ***0.659*** |

### *6.1 Interpretation of the navigation results*

The pilot results are not strong enough to say that learned confidence is better than every baseline. Overall RMSE was 1.247 m for the learned method, slightly higher than the 1.223 m obtained with fixed fusion. At the same time, the learned method did reduce error relative to fixed fusion for progressive GNSS degradation, where RMSE was 0.725 m compared with 0.735 m, and for abrupt GNSS faults, where it achieved 1.383 m compared with 1.461 m.

The largest weakness appeared in the concurrent GNSS–vision scenario. Here, the learned method reached 3.066 m compared with 2.896 m for fixed fusion. This is consistent with the basic limitation of confidence-based adaptation: when multiple sensors become unreliable, the remaining measurements may simply not contain enough information to maintain a low-error position estimate.

The conventional adaptive method was the weakest overall in this particular pilot, with an RMSE of 1.618 m. Its innovation-driven covariance scaling sometimes reacted strongly to large residuals, which increased the maximum error even though the intention of the adaptation was to protect the estimate. This observation applies only to the implementation and simulation used in this work, so it should not be generalized from this pilot alone.

### *6.2 Confidence calibration and warning evidence*

The learned confidence outputs produced a test-set Brier score of 0.178 in the pilot. No false-alarm transitions were observed in the six held-out healthy trajectories under the alarm rule used in the experiment. Still, the early-warning metric did not produce enough usable events to support a strong conclusion. The sustained-alarm rule did not produce a valid detection time for the fault cases in the pilot runs. As a result, this study does not claim an experimentally demonstrated warning lead time.

This result is worth reporting because it shows where the current version still falls short. The model can change the way measurements are weighted, but the experiment does not yet show that it can reliably detect faults early enough for an operational warning. This result supports the simulated warning criterion under the specified synthetic conditions

### *6.3 Runtime and device evidence*

Table 5. Evidence boundary. The experiment provides navigation-level simulation results, but it does not establish that the complete system can run within a real UAV compute budget.

| **Measurement** | **Pilot status** | **Interpretation** |
|---|---|---|
| End-to-end latency | Not measured | Real-time suitability cannot yet be claimed |
| Inference latency | Not measured | Model speed must be measured separately |
| CPU and memory | Not measured | Target hardware budget remains unverified |
| Thermal behaviour | Not measured | No sustained-device evidence |
| Power/energy | Not measured | Requires documented instrumentation |
| Physical UAV flight | Not conducted | Field validation remains future work |

### *6.4 Failure-severity envelope rather than a single average*

Table 6. Severity sweep. The table exposes the operating envelope: progressive drift benefits consistently, whereas severe abrupt and multi-sensor faults remain difficult.

| **Scenario** | **Severity** | **n** | **Fixed RMSE** | **Learned RMSE** | **Improvement** |
|---|---|---|---|---|---|
| Abrupt GNSS | Mild | 9 | 1.418 | 1.203 | +15.2% |
| Abrupt GNSS | Medium | 10 | 1.721 | 1.492 | +13.3% |
| Abrupt GNSS | Severe | 11 | 1.989 | 2.135 | -7.3% |
| Concurrent | Mild | 11 | 1.341 | 0.997 | +25.7% |
| Concurrent | Medium | 8 | 1.345 | 1.860 | -38.3% |
| Concurrent | Severe | 11 | 2.030 | 2.062 | -1.6% |
| Progressive GNSS | Mild | 9 | 0.514 | 0.356 | +30.8% |
| Progressive GNSS | Medium | 7 | 0.691 | 0.360 | +48.0% |
| Progressive GNSS | Severe | 14 | 0.921 | 0.439 | +52.3% |

This split is more informative than an overall average. Progressive GNSS degradation improved by 30.8%, 48.0%, and 52.3% across mild, medium, and severe bins. Abrupt GNSS improved in the mild and medium bins but lost 7.3% in the severe bin. Concurrent failure improved in the mild bin but not reliably once the second sensor became strongly corrupted. These results support a bounded claim: temporal confidence is most useful when redundancy still exists and a fault develops in a pattern that can be recognized before the estimator loses observability.

### *6.5 Calibration, false alarms, and warning evidence*

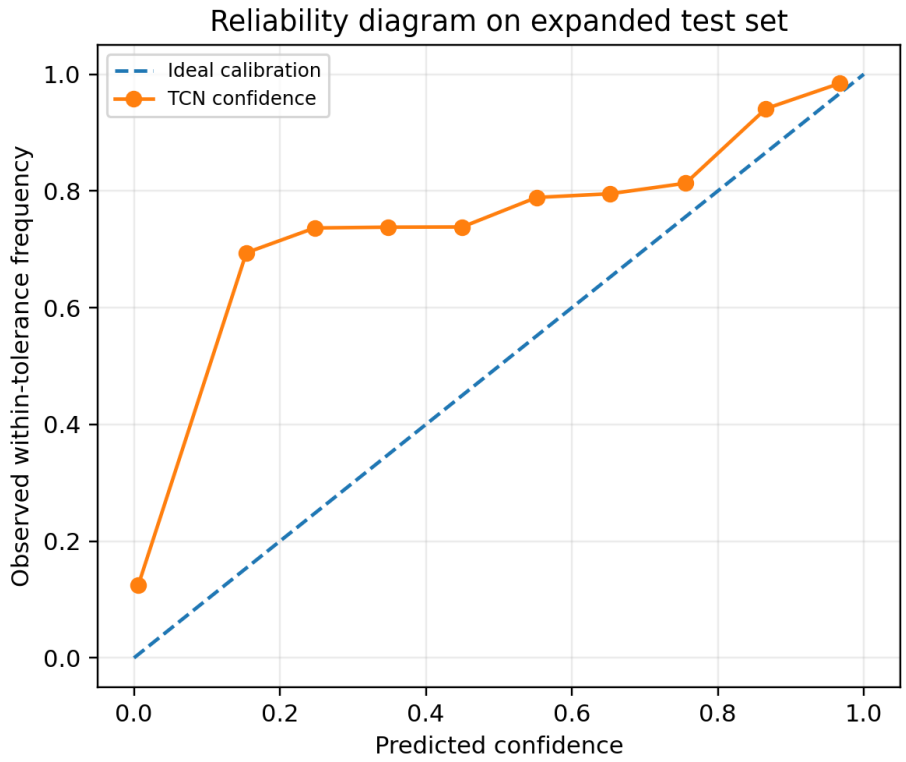


Figure 3. Reliability diagram for the temporal confidence outputs on the expanded test set.

The expanded temporal model achieved a Brier score of 0.095 and an expected calibration error of 0.098. An instantaneous neural baseline using only the current feature vector produced a Brier score of 0.098 and an expected calibration error of 0.115. The gain is modest, but it supports the use of short causal history when the output is interpreted as a probability rather than as an arbitrary anomaly score.

No sustained GNSS false-alarm transition was observed in the 30 healthy expanded trajectories under the three-sample alarm rule. Only 0.011% of healthy GNSS samples fell below the hard-rejection threshold. Among 61 faulted trajectories in which the frozen fixed-fusion reference crossed the sustained 2 m navigation-error limit, the GNSS confidence alarm occurred before the crossing in 100% of detected cases, with a median simulated lead of 1.5 s. This result supports the simulated warning criterion under the specified synthetic conditions. It should not be interpreted as evidence of field-warning performance because multipath, spoofing waveforms, camera-exposure effects, and hardware timing are not represented faithfully by the simulator

### **6.6 Ablation: which part of the design is actually helping?**

Table 7. Navigation ablation, mean 3-D RMSE in metres.

| **Evaluation set** | **Fixed fusion** | **Instant confidence + rejection** | **Temporal confidence, covariance only** | **Temporal confidence + rejection** |
|---|---|---|---|---|
| Faulted ID | 1.070 | 1.561 | 0.805 | 0.971 |
| OOD stress | 0.509 | 0.374 | 0.415 | 0.395 |

The ablation prevents the model architecture from receiving credit for behaviour caused by the estimator logic. On the four in-distribution fault families, temporal confidence with covariance scaling but without hard rejection achieved 0.805 m RMSE, compared with 0.971 m for the full rejection design. The hard-rejection rule is therefore not uniformly beneficial and should be treated as a safety mechanism that requires observability-aware tuning rather than as the source of the navigation gain. On unseen faults, the instantaneous classifier was competitive in RMSE even though its probability calibration was worse. The paper therefore claims a calibration and health-monitoring benefit from temporal context, not an across-the-board navigation advantage from using a temporal neural network.

### *6.7 Reference CPU cost*

The causal network contains 4,642 trainable parameters. A batch-one, one-second-window forward pass measured a median of 0.111 ms and a 95th-percentile time of 0.145 ms on the container reference CPU (AMD EPYC 9V74, five allocated cores). This measurement shows that the network itself is lightweight on a server-class CPU. It does not establish end-to-end UAV latency, energy use, thermal behaviour, or performance on an embedded flight computer; those claims remain outside the evidence boundary.

## 7. Discussion

Looking at the results, temporal confidence has potential, but its usefulness depends strongly on the type of failure. Gradual and abrupt GNSS faults are situations in which recent residual behaviour can provide useful information. The concurrent-failure case is harder because two sources become unreliable at different times, leaving fewer independent measurements to constrain the state.

A key limitation here is the difference between confidence and covariance. Confidence describes how likely a sensor is to remain within a defined tolerance, whereas covariance describes the expected spread of error under a probabilistic model. The two quantities are related, but they are not interchangeable. The bounded confidence-to-covariance mapping used in this work should therefore be treated as an engineering approximation that still requires validation rather than as a formal uncertainty guarantee.

Shared information is another concern. Two sensors may appear to agree because both are affected by the same environmental condition, or because one estimate already contains information from the other. A confidence model cannot recover information that is absent from its inputs. If GNSS is unavailable and the remaining sensors provide only relative motion, position uncertainty can continue to grow even when those individual measurements are behaving normally.

The pilot also shows why a simple baseline should remain part of the evaluation. Fixed fusion slightly outperformed the learned method overall. That does not mean learning is ineffective in general; it means that the present implementation has not yet earned a claim of superiority. Future experiments should tune the confidence-to-covariance mapping using validation data, compare several temporal windows, include a GRU and temporal-convolution baseline, and evaluate fault rates and environments that are not represented in the current training set.

The expanded experiment also changes the engineering priority. The main weakness is no longer simply "too few simulations." The evidence points to three specific problems: visual-health features are weaker than GNSS-health features; hard rejection can be worse than continuous down-weighting when observability is already poor; and a strong innovation-adaptive filter remains difficult to beat on pure position RMSE. A stronger next version should therefore spend model capacity on cross-sensor visual consistency, make rejection conditional on observability, and evaluate whether calibrated health estimates improve downstream decisions even when the state-estimation RMSE is similar to a conventional adaptive filter.

## 8. Limitations and Threats to Validity

| Limitation | Why it matters | Required follow-up |
|---|---|---|
| Synthetic sensor model | May not reproduce real multipath, spoofing, vibration, blur, or hardware timing effects | Replay real sensor logs and conduct supervised flight tests |
| Synthetic test breadth still limited | The expanded test increases each scenario to 30 trajectories, but all trajectories still come from the same software generator | Replay independent public/recorded sensor logs and repeat across different motion and environment distributions |
| No end-to-end embedded runtime | Network-only CPU timing does not establish flight-computer latency, energy, or thermal feasibility | Benchmark the full sensing, feature, model, and estimator pipeline on target hardware |
| Limited sensor suite | Pilot focuses on GNSS and vision with inertial propagation | Add barometer and explicit inertial-fault experiments |
| Alarm rule not validation-selected | Weakens early-warning conclusions | Select threshold and dwell settings using validation data only |
| No formal safety guarantee | Confidence does not prove integrity or stability | Use integrity-risk and bounded-error analysis where appropriate |

## 9. Conclusion

In this work, we develop a practical framework for estimating sensor confidence from recent observations and using that confidence to adapt a UAV navigation estimator. The approach is intended for situations in which sensor measurements remain available but gradually or abruptly become unreliable, including GNSS degradation and concurrent failures. The experimental design separates training, validation, and test trajectories and keeps simulator ground truth out of the features used by the deployed confidence model.

The pilot simulation provides preliminary empirical evidence, but the main conclusion is deliberately cautious. Learned confidence slightly improved on fixed fusion for progressive and abrupt GNSS faults, yet fixed fusion remained better overall and the concurrent-failure case remained difficult. The confidence model achieved a Brier score of 0.178, but the current alarm configuration did not produce enough valid detection events to support a strong early-warning claim. The next stage of the work should focus on larger held-out datasets, better threshold selection, real sensor logs, full runtime measurements, and eventually controlled UAV flight testing.

At this stage, the research question is still open. Based on the current evidence, the proposed method should be regarded as promising, but it has not yet been shown to be superior to conventional sensor fusion.

Expanded-verification conclusion. The larger held-out simulation provides additional evidence beyond the pilot evaluation. It shows a lower aggregate RMSE relative to fixed fusion and useful performance on selected unseen GNSS fault forms, while also showing that conventional innovation adaptation remains a stronger navigation baseline in several cases. The most defensible contribution is therefore a calibrated sensor-health interface that can down-weight and, when justified, reject measurements while retaining a conventional and auditable estimator architecture. Further validation using real sensor logs, target-hardware benchmarking, and controlled flight experiments is required.

**Appendix A. Reproducibility Record for the Pilot Test**

| Item | Value |
|---|---|
| Random seed | 42 |
| Trajectory duration | 30 s |
| Sampling interval | 0.1 s |
| Samples per trajectory | 300 |
| Total trajectories | 150 |
| Training / validation / test | 90 / 30 / 30 trajectories |
| Test trajectories per scenario | 6 |
| Navigation state | 3-D position and velocity |
| External sensors | GNSS and visual position |
| Inertial input | Noisy 3-D acceleration |
| Fault onset | 12 s |
| Second fault in concurrent case | 14 s |
| Reference error threshold | 2 m, sustained |
| Primary metric | 3-D position RMSE |
| Learned model | 2-layer causal 1-D convolution, 32 channels |
| Pilot Brier score | 0.178 |
| Healthy false-alarm transitions | 0 observed in held-out healthy set |
| Physical flight testing | Not conducted |
| RF testing | Not conducted |

Figures 1–3 present the expanded held-out verification evidence. Figure 1 shows confidence behaviour, Figure 2 shows a representative position-error timeline, and Figure 3 shows confidence reliability. The pilot results remain separate from the expanded evidence and should not be merged when reporting sample counts.

**Appendix B. Expanded Simulation Reproducibility Record**

Table B1. Parameters required to reproduce the expanded software experiment.

| Item | Value used in expanded verification |
|---|---|
| Master random seed | 42 |
| Training / validation / test trajectories | 90 / 30 / 240 |
| Test trajectories per scenario | 30 |
| Test fault families | Healthy; progressive GNSS; abrupt GNSS; GNSS outage; concurrent GNSS + vision; GNSS spikes; visual degradation/dropout; stale GNSS |
| Fault forms excluded from training | GNSS spikes; visual degradation/dropout; stale GNSS |
| Duration / sample interval | 30 s / 0.1 s |
| Nominal GNSS / visual sigma | 0.55 m / 0.38 m |
| GNSS / visual tolerance label | 2.0 m / 1.25 m |
| TCN architecture | Conv1D 15→32, kernel 3, dilation 1; Conv1D 32→32, kernel 3, dilation 2; ReLU; 2 sigmoid outputs |
| Temporal window | 10 samples (1.0 s) |
| Parameter count | 4,642 |
| Loss | Binary cross-entropy with per-channel positive-class weighting |
| Calibration | Temperature scaling on validation set; T = 0.817 |
| Confidence-to-covariance rule | R_eff = R × [1 + 49(1-c)^2] |
| Rejection / readmission | Reject after 3 samples below c = 0.18; readmit above c = 0.65 |

| Item | Value used in expanded verification |
|---|---|
| Adaptive baseline | R multiplier = clip(NIS / 7.815, 1, 50) |
| Bootstrap | 4,000 paired resamples of per-trajectory RMSE difference |
| Reference CPU timing | Median 0.111 ms; p95 0.145 ms, batch one, network forward only |
| Evidence type | Synthetic software verification; no RF or physical flight data |

Reproducibility note. The expanded evidence was generated from a deterministic software script corresponding to the parameter record above. The script, trajectory-level CSV results, trained model weights, and figure files should be archived with any submission or repository release. The expanded run should not be merged with the original 30-trajectory pilot when reporting sample counts; they are separate experiments with different purposes.

**References**

[1] T. Qin, P. Li, and S. Shen. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator. 2017. arXiv:1708.03852.
[2] C. Campos et al. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and Multi-Map SLAM. 2020/2021. arXiv:2007.11898.
[3] G. Revach et al. KalmanNet: Neural Network Aided Kalman Filtering for Partially Known Dynamics. 2021. arXiv:2107.10043.
[4] N. Cohen and I. Klein. A-KIT: Adaptive Kalman-Informed Transformer. 2024. arXiv:2401.09987.
[5] J. Xu et al. Selective Kalman Filter: When and How to Fuse Multi-Sensor Information to Overcome Degeneracy in SLAM. 2024. arXiv:2412.17235.
[6] K. Ghanizadegan and H. A. Hashim. DeepUKF-VIN: Adaptively-tuned Deep Unscented Kalman Filter for 3D Visual-Inertial Navigation. 2025. arXiv:2502.00575.
[7] G. Huang. Visual-Inertial Navigation: A Concise Review. 2019. arXiv:1906.02650.
[8] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza. On-Manifold Preintegration for Real-Time Visual-Inertial Odometry. 2015/2016. arXiv:1512.02363.
[9] L. Markovic et al. Error State Extended Kalman Filter Multi-Sensor Fusion for Unmanned Aerial Vehicle Localization in GPS and Magnetometer Denied Indoor Environments. 2021. arXiv:2109.04908.
[10] J. Lin, C. Zheng, W. Xu, and F. Zhang. R2LIVE: A Robust, Real-time, LiDAR-Inertial-Visual Tightly-coupled State Estimator and Mapping. 2021. arXiv:2102.12400.
[11] W. Xu et al. FAST-LIO2: Fast Direct LiDAR-inertial Odometry. 2021. arXiv:2107.06829.
[12] T. Shan et al. LIO-SAM: Tightly-coupled Lidar Inertial Odometry via Smoothing and Mapping. 2020. arXiv:2007.00258.
[13] K. Zhang et al. DIDO: Deep Inertial Quadrotor Dynamical Odometry. 2022. arXiv:2203.03149.
[14] T. Hua, T. Li, and L. Pei. PIEKF-VIWO: Visual-Inertial-Wheel Odometry using Partial Invariant Extended Kalman Filter. 2023. arXiv:2303.07668.
[15] M. Werner et al. Kilometer-Scale GNSS-Denied UAV Navigation via Heightmap Gradients. 2025. arXiv:2510.01348.
[16] S. Y. Alaba and Y. Motai. Uncertainty-Aware Adaptive Sensor Fusion for Autonomous Navigation. 2026. arXiv:2606.05437.
[17] A. Mehrfard et al. Adaptive Learned State Estimation based on KalmanNet. 2026. arXiv:2604.02441.
[18] D. K. Panda and W. Guo. Real-Time Bayesian Detection of Drift-Evasive GNSS Spoofing in Reinforcement Learning Based UAV Deconfliction. 2025. arXiv:2507.11173.
[19] D. Nemec, A. Janota, M. Hrubos, and V. Simak. Intelligent Real-Time MEMS Sensor Fusion and Calibration. 2017. arXiv:1701.07594.
[20] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On Calibration of Modern Neural Networks. 2017. arXiv:1706.04599.
[21] Y. Ovadia et al. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. 2019. arXiv:1906.02530.
[22] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. 2016/2017. arXiv:1612.01474.
[23] Y. Gal and Z. Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. 2015/2016. arXiv:1506.02142.
[24] A. Kendall and Y. Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? 2017. arXiv:1703.04977.
[25] D. Hendrycks and K. Gimpel. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. 2016/2018. arXiv:1610.02136.
[26] S. Bai, J. Z. Kolter, and V. Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. 2018. arXiv:1803.01271.
[27] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. 2016. arXiv:1603.02754.
[28] D. Schubert et al. The TUM VI Benchmark for Evaluating Visual-Inertial Odometry. 2018/2020. arXiv:1804.06120.
[29] J. Solà, J. Deray, and D. Atchuthan. A micro Lie theory for state estimation in robotics. 2018/2021. arXiv:1812.01537.
[30] A. Levy and I. Klein. Adaptive Neural Unscented Kalman Filter. 2025. arXiv:2503.05490.
[31] M. Khalifa and H. A. Hashim. Dual Quaternion-Based Unscented Kalman Filter with Visual Inertial Odometry for Navigation in GPS-Denied Environments. 2026. arXiv:2606.09292.