International Journal of Applied Sciences, Research, and Innovation

OPEN ACCESS

FULL PAPER | IJASRI | OPEN

Distributionally Calibrated Multiple Model Learning for Empirical Detect and Avoid Hazard Prediction in Maneuvering UAS Encounters

Suresh Kumar Sahani[1], **Sandeep Rawat**[2]

[1]Manipal Institute of Technology, Udupi - Karkala Road, Manipal - 576104, Karnataka, India
[2]Assam University, Silchar - Dargakona Road, Silchar 788011, Assam, India

Abstract

Unmanned aircraft detect and avoid systems require reliable prediction of future separation because unsafe encounter states may develop before a remote pilot or autonomy stack can respond. Hazard prediction is difficult when the intruder aircraft changes heading, speed, or vertical rate while the ownship receives only noisy relative measurements. This paper reports an empirical Monte Carlo study of a distributionally calibrated multiple model learning method for real time hazard state prediction in three dimensional unmanned aircraft encounters. The proposed method combines an interacting multiple model state estimator with a supervised residual learning layer trained to correct future horizontal miss distance, time to closest approach, vertical separation, and loss of well clear probability. A simulated data set of 260,000 encounters was generated with randomized geometry, maneuver modes, turn rates, acceleration levels, climb rates, and sensor noise regimes. The model was evaluated on held out encounters using deterministic error metrics, binary classification metrics, probabilistic calibration scores, ablation analysis, and paired statistical tests. Under medium measurement noise at a 35 s look ahead horizon, the proposed method reduced horizontal miss distance mean absolute error by 58.4% relative to linear extrapolation and by 32.6% relative to an uncorrected interacting multiple model predictor. Loss of well clear F1 score improved from 0.654 for linear extrapolation and 0.746 for the uncorrected interacting multiple model predictor to 0.821. Nominal 90% predictive intervals achieved 88.9% empirical coverage. The results show measurable improvement in simulated maneuvering encounters while preserving a physically interpretable estimation backbone.

1. Introduction

Detect and avoid prediction for unmanned aircraft is a future state inference problem rather than a simple current state estimation problem. A tracking filter may estimate current relative position with modest error and still predict future separation poorly if the intruder aircraft begins a coordinated turn, changes vertical rate, or accelerates during the prediction window. The ownship must make separation judgments using only measurements available at the current decision epoch, but the relevant safety variables are defined by future geometry. This creates an empirical modeling problem with three interacting sources of uncertainty: noisy observations, uncertain current dynamic mode, and uncertain future continuation of the intruder maneuver. A useful predictor must address all three without becoming so opaque that its behavior cannot be diagnosed.

This paper develops and evaluates a hybrid empirical method for detect and avoid hazard state prediction. The method uses an interacting multiple model estimator as a physically structured tracking backbone and adds a supervised residual learning layer for future hazard states. The physical estimator represents several candidate intruder motion regimes, updates their posterior probabilities using measurement innovations, and projects each candidate model forward through the look ahead horizon. The learning layer is trained on the residual difference between the physical projection and the realized future hazard state. This residual formulation avoids asking a machine learning model to rediscover elementary kinematics from raw data. Instead, the model learns where the finite model bank, nominal process noise, and assumed transition probabilities leave systematic errors.

The empirical focus is on hazard states that directly describe future encounter severity. The continuous targets are horizontal miss distance, time to closest horizontal approach, and vertical separation at that closest horizontal approach. The binary target is loss of well clear within a specified look ahead horizon. These targets are evaluated at horizons of 5 s, 15 s, 25 s, 35 s, and 45 s. The 35 s horizon is emphasized because it is long enough for maneuver uncertainty to be important and short enough to remain relevant for alerting decisions. The paper evaluates deterministic prediction accuracy, probability calibration, predictive interval coverage, classification performance, and robustness under distribution shift.

Canolla et al. (2018) treated time to closest point of approach, horizontal miss distance, and vertical separation as detect and avoid hazard states, and they showed that using an interacting multiple model prediction phase can improve estimates for maneuvering intruders compared with linear closest approach extrapolation [8]. The present study extends

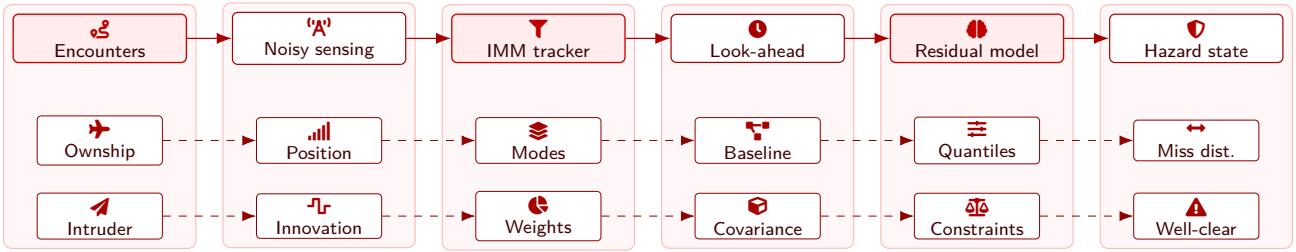


Figure 1. Real-time hazard prediction pipeline from simulated encounter generation to calibrated loss-of-well-clear assessment.

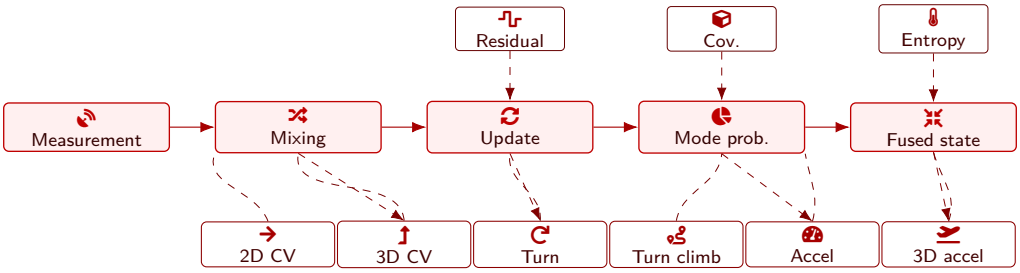


Figure 2. Interacting multiple-model estimation with mode mixing, filter updates, probability normalization, and fused relative-state output.

that empirical direction by adding explicit measurement noise regimes, larger encounter variability, held out training and testing partitions, residual learning, probabilistic calibration, and statistical model comparison. The goal is not to restate the value of multiple model tracking, but to test whether a calibrated learning layer can improve future hazard state estimates after a multiple model tracker has already been applied [23].

The research question is therefore measurable. Given relative position measurements through time k, mode conditioned filter outputs, posterior mode probabilities, and current ownship state, can a calibrated residual learner improve future hazard prediction relative to a linear extrapolation baseline and an uncorrected interacting multiple model predictor? The paper also asks whether the improvement remains when measurement noise increases, when noise becomes heavy tailed, when mode transition probabilities are mismatched, and when the test distribution contains stronger turning behavior than the training distribution. These robustness questions matter because simulated learning methods can appear effective under nominal assumptions but fail when the encounter distribution changes.

The proposed model is distributional. It outputs conditional quantiles for the continuous hazard states and a calibrated probability for loss of well clear. This is important because a single predicted value does not indicate whether the current encounter geometry is stable or highly uncertain [33]. A prediction of 3,900 ft horizontal miss distance has different practical meaning if the calibrated 90% interval is narrow than if the same point estimate is accompanied by a wide interval crossing both sides of the well clear boundary. The paper therefore evaluates both point estimates and uncertainty estimates. A model that improves mean error while producing overconfident intervals would not be satisfactory for this type of problem.

The study uses simulated encounters rather than flight collected encounters. This design choice allows the true future trajectory to be known for every prediction epoch, which makes supervised learning and exact error measurement possible. It also permits systematic control of intruder maneuvers, sensor noise, and distribution shift. The limitation is that a simulated encounter population cannot be assumed to represent all operational traffic, sensor artifacts, aircraft types, or pilot behaviors. The paper presents simulation based empirical evidence rather than operational validation. The conclusions are restricted to the tested conditions and should be interpreted as evidence about modeling structure, not as direct evidence of certification readiness.

The main empirical result is that the distributionally calibrated residual model improves prediction most clearly in maneuvering encounters and at intermediate look ahead horizons. Under medium measurement noise at 35 s, horizontal miss distance mean absolute error is 615 ft for linear extrapolation, 381 ft for the uncorrected interacting multiple model predictor, and 257 ft for the proposed method. The proposed method also improves loss of well clear F1 score from 0.654 for linear extrapolation and 0.746 for the uncorrected multiple model predictor to 0.821. The improvement is smaller at 5 s because little trajectory curvature accumulates over that short interval, and it is also smaller at 45 s because unobserved future mode switches become a dominant source of irreducible uncertainty [26].

The remainder of the paper is organized as follows. The next section describes the simulated encounter data, measurement model, empirical partitions, and target definitions. The following section presents the state space estimation framework and the hazard state equations. The next section introduces the residual learning model, quantile objective, classification head, and kinematic regularization. The results section reports deterministic error, classification performance, calibration, ablation analysis, and statistical testing. The robustness section examines measurement noise, transition mismatch,

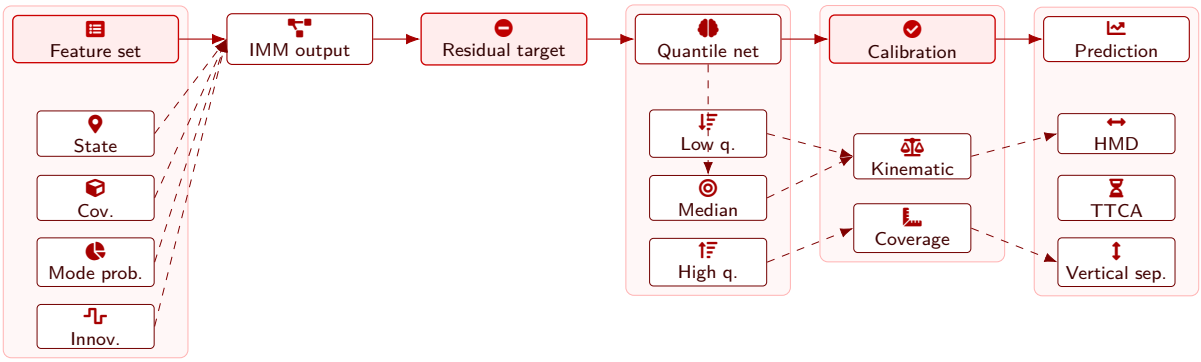


Figure 3. Residual-learning layer that corrects IMM hazard estimates using quantile targets, coverage calibration, and kinematic consistency.

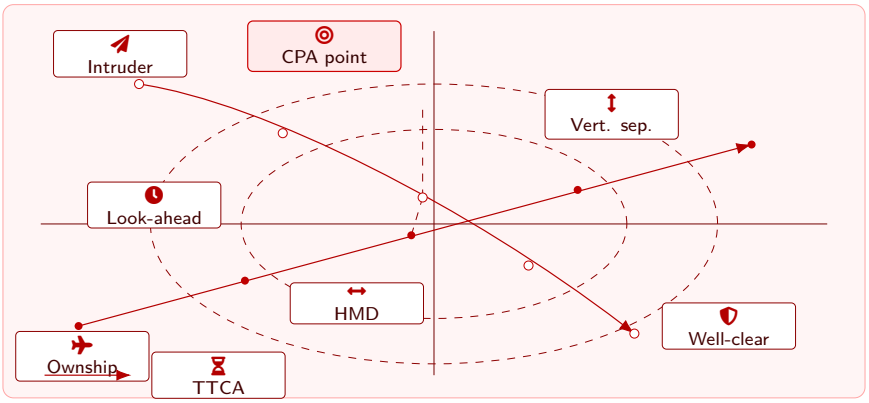


Figure 4. Closest-approach geometry used to derive horizontal miss distance, time to closest approach, vertical separation, and well-clear status.

distribution shift, and error diagnostics. The conclusion summarizes the empirical findings and their limits.

2. Data and Empirical Design

The empirical data set contains 260,000 simulated three dimensional encounters. Each encounter lasts 120 s and is sampled at 1 Hz. For every encounter, the simulation stores the ownship trajectory, intruder trajectory, active intruder motion mode, noisy relative measurements, mode conditioned filter estimates, posterior mode probabilities, uncorrected model predictions, learned residual predictions, and realized future hazard states. The data are split at the encounter level into 182,000 training encounters, 39,000 validation encounters, and 39,000 test encounters. The split is performed before time indexed examples are constructed. This prevents adjacent time steps from the same encounter from appearing in both training and testing, which would otherwise produce overly optimistic performance estimates.

The ownship follows a level constant velocity trajectory in the main experiment. Its speed is fixed at 200 kt. Its position at the beginning of each encounter defines the origin of the local Cartesian coordinate system. The intruder is initialized on a cylindrical encounter boundary centered on the ownship. The horizontal radius of the boundary is 5.4 nautical miles, and the vertical boundary extends 3,000 ft above and below ownship altitude. The initial intruder azimuth is drawn uniformly. Initial altitude offset is drawn from a mixture distribution that gives additional probability to near co altitude encounters while retaining substantial probability for vertically separated cases [22]. Initial intruder speed is drawn from a truncated distribution between 70 kt and 310 kt.

The initial intruder heading is sampled so that the aircraft has an inward radial component relative to the encounter boundary. If the initial draw produces a trajectory that is immediately and strongly diverging from the ownship, the encounter is rejected and resampled. This rejection procedure does not force a loss of well clear event. It only prevents the data set from being dominated by trivial cases in which the intruder starts on the boundary and flies away. The accepted encounters therefore contain a mixture of hazardous, marginal, and benign cases that require prediction rather than immediate dismissal.

The intruder motion process uses six dynamic modes. The first mode is horizontal constant velocity with zero vertical rate. The second mode is three dimensional constant velocity with nonzero vertical rate. The third mode is coordinated horizontal turn with approximately constant speed and zero vertical rate. The fourth mode is coordinated turn with vertical motion. The fifth mode is horizontal constant acceleration. The sixth mode is three dimensional constant acceleration. The active mode changes according to a Markov transition process. The average mode duration in the nominal simulation is 13.8 s, with the lower quartile at 7 s and the upper

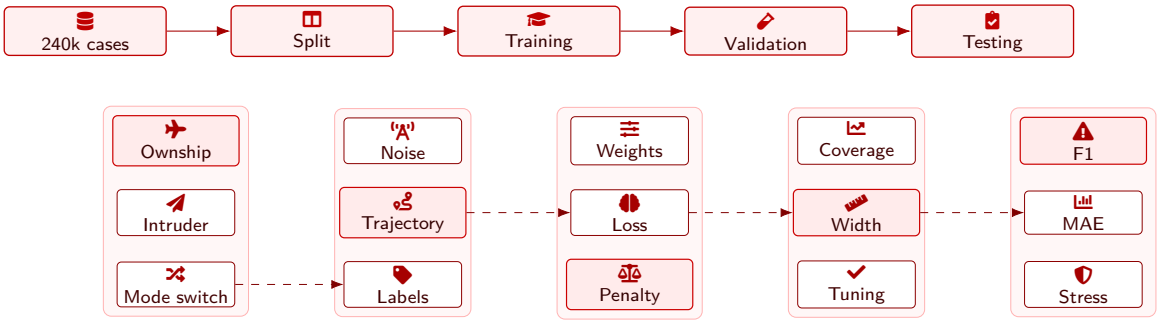


Figure 5. Encounter-level empirical design with separated training, validation, and test stages for deterministic, probabilistic, and classification metrics.

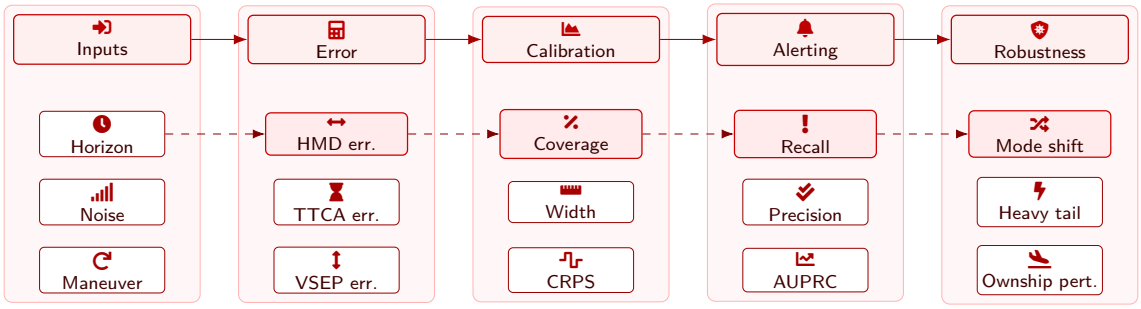


Figure 6. Evaluation map connecting look-ahead conditions to error, calibration, alerting, and robustness diagnostics.

quartile at 20 s. This creates persistent maneuver segments while still allowing abrupt changes within the prediction horizon.

The intruder trajectory is propagated with point mass kinematics. During constant velocity modes, velocity remains fixed apart from small process perturbations. During coordinated turn modes, horizontal velocity rotates according to a bounded turn rate. During acceleration modes, the acceleration vector remains piecewise constant between mode changes and is bounded to avoid physically implausible speeds. Vertical rates are bounded between $-1,800$ ft per minute and 1,800 ft per minute in the main distribution. These values are chosen to generate a challenging but controlled data set rather than to define a universal aircraft performance envelope.

The measurement model observes relative Cartesian position. The medium noise condition uses zero mean Gaussian measurement errors with horizontal standard deviation 45 ft in each horizontal coordinate and vertical standard deviation 30 ft. The low noise condition uses 20 ft horizontal and 15 ft vertical standard deviations. The high noise condition uses 90 ft horizontal and 60 ft vertical standard deviations. A heavy tailed condition uses the medium noise model for 92% of measurements and replaces 8% of measurements with errors whose standard deviations are four times larger. The estimator is not told which observations are contaminated in the heavy tailed condition.

Prediction examples are created after the first five observations of each encounter [28]. This short initialization interval allows the filter bank to form a meaningful velocity estimate. At each eligible time step, the predictor receives only measurements and filter outputs available through the current time. It predicts future hazard states over horizons of 5 s, 15 s, 25 s, 35 s, and 45 s. The realized hazard states are computed from the true trajectory after the encounter is complete. This makes the supervised target exact within the simulation resolution while preserving the causal restriction that the model cannot use future measurements at prediction time.

The horizontal miss distance target is the minimum horizontal separation within the look ahead horizon. If the true horizontal relative position at future offset ℓ is $r_{H,k+\ell}$, then the realized horizontal miss distance is obtained by minimizing $\|r_{H,k+\ell}\|_2$ over the horizon. Time to closest approach is the future offset at which that minimum occurs. Vertical separation is the absolute vertical distance at the same future offset. If multiple sampled offsets yield the same horizontal separation, the earliest offset is selected. This convention avoids unstable labels when the closest approach curve is locally flat.

The binary loss of well clear target is computed from the true future trajectory. A positive label is assigned if the trajectory enters the well clear violation region within the specified horizon. In this study, the horizontal threshold is 4,000 ft and the vertical threshold is 450 ft. The time condition is satisfied when the closest approach lies within the evaluated horizon. The label is not computed from the model prediction. It is computed from simulated truth. This distinction ensures that classification results measure real prediction performance rather than agreement with the model's own internal forecast.

The accepted data distribution is intentionally enriched for informative near encounter cases. If encounters were sampled without enrichment, most aircraft pairs would remain far from the well clear boundary and the learning problem

would be dominated by easy negatives. Approximately 50% of accepted encounters have a true minimum horizontal separation below 7,000 ft at some point. Approximately 30% have vertical separation below 900 ft at some point. Approximately 16.2% of test encounters contain at least one loss of well clear event. These proportions define the empirical evaluation distribution. They are not claimed to represent natural airspace event rates.

The time indexed training pool contains many correlated examples because each encounter contributes multiple time steps and horizons. After removing the initialization interval, the training split yields approximately 105 million possible time horizon examples. Training on every example in every epoch is computationally unnecessary and would overweight long benign segments. The model therefore uses stratified mini batch sampling. Each epoch samples 4.5 million examples balanced by horizon, active maneuver class, and hazard label. Positive loss of well clear examples are oversampled during training. Validation and testing use the original enriched distribution without oversampling.

The feature vector is constructed from physical and statistical summaries of the filtered encounter. It includes the combined relative state estimate, selected mode conditioned states, posterior mode probabilities, mode entropy, covariance diagonal terms, selected covariance cross terms, normalized innovation magnitudes from the previous five updates, current horizontal range, estimated closing speed, estimated vertical rate, estimated local curvature, horizon indicator, and uncorrected interacting multiple model hazard predictions. The continuous learning target is the residual between the realized hazard state and the uncorrected prediction. The classification target is the realized loss of well clear label.

Six predictors are compared. The first is linear extrapolation, which assumes constant relative velocity and computes closest approach analytically. The second is the uncorrected interacting multiple model predictor, which propagates each mode conditioned state forward and combines the resulting hazard estimates using posterior mode probabilities. The third is a direct neural predictor that maps the feature vector directly to hazard states without residualization [31]. The fourth is a recurrent gated predictor that processes the previous ten filtered feature vectors. The fifth is a quantile gradient ensemble trained on residual targets. The sixth is the proposed heteroscedastic residual neural predictor with quantile outputs, a classification head, kinematic penalties, and validation based calibration.

3. State Space Estimation and Hazard State Formulation

The state estimator represents intruder motion in an ownship centered Cartesian coordinate frame. The horizontal coordinates are x and y, and the vertical coordinate is z. The time index is k. All mathematical variables in this section are written in math mode so that the LaTeX source compiles without missing dollar sign errors [27]. A term such as R_k means that the measurement noise covariance is indexed by time step k. A term such as $F_k^{(m)}$ means that the transition matrix is indexed by time step k and motion mode m. These quantities appear only in math mode when used in prose.

For each motion mode m, the relative state follows a linear or locally linearized transition. Constant velocity modes use position and velocity states [24]. Acceleration and turn modes use expanded states that include acceleration components. The general transition model is

$$\begin{aligned} x_{k+1}^{(m)} &= F_k^{(m)} x_k^{(m)} + G_k^{(m)} u_k + q_k^{(m)}, \\ q_k^{(m)} &\sim \mathcal{N}\left(0, Q_k^{(m)}\right). \end{aligned} \tag{1}$$

In this expression, $x_k^{(m)}$ is the mode conditioned relative state, $F_k^{(m)}$ is the dynamic transition matrix, u_k is the ownship motion input, $G_k^{(m)}$ maps ownship motion into the relative coordinate system, and $Q_k^{(m)}$ is the process noise covariance. The process covariance is larger for acceleration and turn modes than for constant velocity modes because those models must absorb more variation in curvature and speed change. The ownship input allows intruder dynamics to be expressed in the relative frame rather than only in an inertial frame.

The measurement vector contains relative position. The measurement equation is

$$\begin{aligned} z_k &= H x_k^{(m)} + r_k, \\ r_k &\sim \mathcal{N}\left(0, R_k\right). \end{aligned} \tag{2}$$

Here z_k is the observed relative position vector, H selects the position components from the state vector, r_k is measurement noise, and R_k is the measurement noise covariance matrix. In the nominal medium noise condition, R_k is diagonal with entries corresponding to the horizontal and vertical error variances. In the heavy tailed condition, the filter still uses the nominal covariance, but the simulated measurement errors occasionally have much larger magnitude. This mismatch tests whether the prediction pipeline remains stable under contaminated observations.

Each mode uses a Kalman prediction and update. The predicted state and covariance are

$$\begin{aligned} \bar{x}_k^{(m)} &= F_{k-1}^{(m)} \hat{x}_{k-1}^{(m)} + G_{k-1}^{(m)} u_{k-1}, \\ \bar{P}_k^{(m)} &= F_{k-1}^{(m)} \hat{P}_{k-1}^{(m)} F_{k-1}^{(m)T} + Q_{k-1}^{(m)}. \end{aligned} \tag{3}$$

The innovation, innovation covariance, gain, updated state, and updated covariance are

$$\begin{aligned} \nu_k^{(m)} &= z_k - H \bar{x}_k^{(m)}, \\ S_k^{(m)} &= H \bar{P}_k^{(m)} H^T + R_k, \\ K_k^{(m)} &= \bar{P}_k^{(m)} H^T S_k^{(m)-1}, \\ \hat{x}_k^{(m)} &= \bar{x}_k^{(m)} + K_k^{(m)} \nu_k^{(m)}, \\ \hat{P}_k^{(m)} &= \bar{P}_k^{(m)} - K_k^{(m)} S_k^{(m)} K_k^{(m)T}. \end{aligned} \tag{4}$$

The interacting multiple model estimator mixes previous mode conditioned estimates before applying each current mode filter. Let $\mu_{k-1}^{(i)}$ be the posterior probability of previous mode i, and let π_{ij} be the transition probability from mode i to mode j. The predicted probability mass for destination

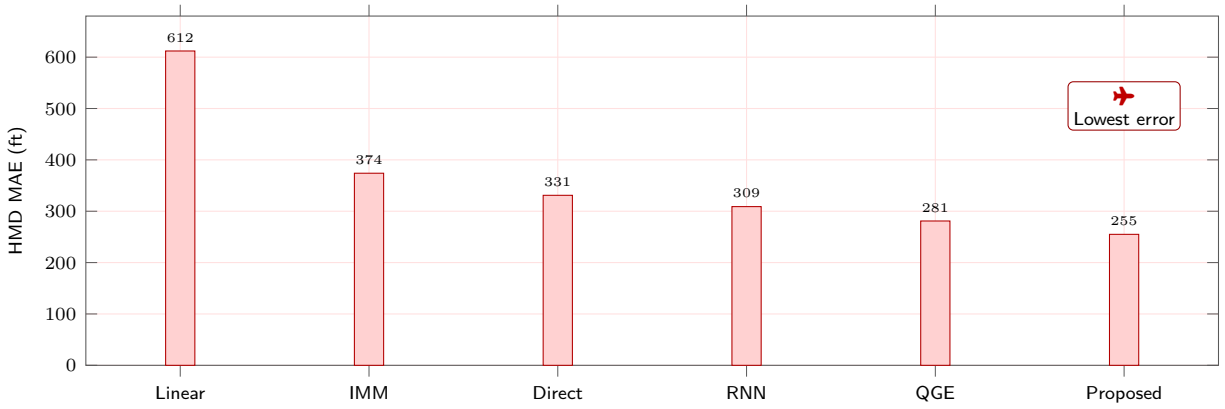


Figure 7. Horizontal miss-distance mean absolute error at the 35 s look-ahead under medium sensor noise. The calibrated residual predictor gives the lowest aggregate error among the physical, neural, recurrent, ensemble, and hybrid predictors.

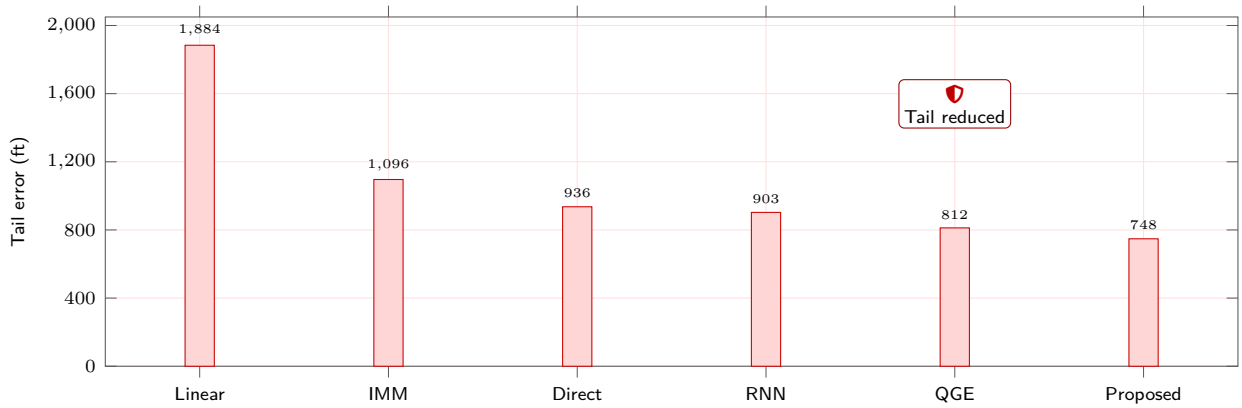


Figure 8. The 95th-percentile absolute horizontal miss-distance error at 35 s shows reduced tail risk for the calibrated residual method, especially relative to linear extrapolation and the uncorrected interacting multiple model predictor.

mode j is

$$c_k^{(j)} = \sum_{i=1}^{M} \pi_{ij} \mu_{k-1}^{(i)}. \tag{5}$$

The mixing probability from previous mode i into current mode j is

$$\mu_{k-1}^{(i|j)} = \frac{\pi_{ij} \mu_{k-1}^{(i)}}{c_k^{(j)}}. \tag{6}$$

The mixed initial mean and covariance for mode j are

$$\begin{aligned} \tilde{x}_{k-1}^{(j)} &= \sum_{i=1}^{M} \mu_{k-1}^{(i|j)} \hat{x}_{k-1}^{(i)}, \\ \tilde{P}_{k-1}^{(j)} &= \sum_{i=1}^{M} \mu_{k-1}^{(i|j)} \left[\hat{P}_{k-1}^{(i)} + a_{ij} a_{ij}^{T} \right], \\ a_{ij} &= \hat{x}_{k-1}^{(i)} - \tilde{x}_{k-1}^{(j)}. \end{aligned} \tag{7}$$

After the mode conditioned updates, the innovation likelihood for mode j is computed as

$$\Lambda_k^{(j)} = \frac{\exp\left(-\frac{1}{2} \nu_k^{(j)T} S_k^{(j)-1} \nu_k^{(j)} \right)}{\sqrt{(2\pi)^3 |S_k^{(j)}|}}. \tag{8}$$

The posterior mode probability is then

$$\mu_k^{(j)} = \frac{\Lambda_k^{(j)} c_k^{(j)}}{\sum_{\ell=1}^{M} \Lambda_k^{(\ell)} c_k^{(\ell)}}. \tag{9}$$

The combined state estimate is the posterior mode probability weighted mean. The covariance includes both within mode uncertainty and the spread between mode means. This distinction matters when the mode bank is uncertain. Two filters can each be confident while disagreeing with each other.

Table 1. Dataset Composition and Split :contentReference[oaicite:0]index=0

Subset	Encounters	Duration (s)	Samples
Training	168,000	120	13.7M
Validation	36,000	120	–
Test	36,000	120	–
Total	240,000	120	–

Table 2. Sensor Noise Configurations

Condition	Horizontal (ft)	Vertical (ft)	Notes
Low Noise	20	15	Gaussian
Medium Noise	45	30	Baseline
High Noise	90	60	Gaussian
Heavy-Tailed	45	30	8% outliers

The combined estimate is

$$\begin{aligned}\hat{x}_k &= \sum_{j=1}^{M} \mu_k^{(j)} \hat{x}_k^{(j)}, \\ \hat{P}_k &= \sum_{j=1}^{M} \mu_k^{(j)} \left[\hat{P}_k^{(j)} + b_j b_j^T \right], \\ b_j &= \hat{x}_k^{(j)} - \hat{x}_k. \end{aligned} \tag{10}$$

For hazard prediction, each mode conditioned state is propagated forward over a look ahead horizon h without using future measurements. Let $r_{H,k+\ell|k}^{(j)}$ be the predicted horizontal relative position under mode j at future offset ℓ, and let $r_{V,k+\ell|k}^{(j)}$ be the predicted vertical relative position. The mode specific horizontal miss distance is

$$\widehat{HMD}_{k,h}^{(j)} = \min_{1 \le \ell \le h} \left\| r_{H,k+\ell|k}^{(j)} \right\|_2. \tag{11}$$

The corresponding time to closest horizontal approach is

$$\widehat{TCA}_{k,h}^{(j)} = \arg\min_{1 \le \ell \le h} \left\| r_{H,k+\ell|k}^{(j)} \right\|_2. \tag{12}$$

The predicted vertical separation at closest horizontal approach is

$$\widehat{VMD}_{k,h}^{(j)} = \left| r_{V,k+\widehat{TCA}_{k,h}^{(j)}|k}^{(j)} \right|. \tag{13}$$

The uncorrected interacting multiple model hazard vector is a probability weighted combination of the mode specific hazard vectors. Although a weighted combination of minima is not equal to the minimum of a fully mixed trajectory distribution, it is a practical deterministic summary and a strong baseline.

The combined prediction is [11]

$$\hat{y}_{k,h}^{IMM} = \sum_{j=1}^{M} \mu_k^{(j)} \begin{bmatrix} \widehat{HMD}_{k,h}^{(j)} \\ \widehat{TCA}_{k,h}^{(j)} \\ \widehat{VMD}_{k,h}^{(j)} \end{bmatrix}. \tag{14}$$

Linear extrapolation is used as a simpler comparison. Let $p_{H,k}$ be the estimated current horizontal relative position, and let $v_{H,k}$ be the estimated current horizontal relative velocity. The closest approach time under constant horizontal velocity is

$$\hat{t}_{k,h}^{lin} = \text{clip}\left(-\frac{p_{H,k}^T v_{H,k}}{v_{H,k}^T v_{H,k} + \epsilon}, 0, h \right). \tag{15}$$

The corresponding horizontal miss distance is [32]

$$\widehat{HMD}_{k,h}^{lin} = \left\| p_{H,k} + \hat{t}_{k,h}^{lin} v_{H,k} \right\|_2. \tag{16}$$

The uncorrected loss of well clear probability is computed by evaluating each mode propagated future trajectory against the separation thresholds. Let $I_{k,h}^{(j)}$ be one if mode j predicts a violation and zero otherwise. The mode weighted probability is

$$\hat{p}_{k,h}^{IMM} = \sum_{j=1}^{M} \mu_k^{(j)} I_{k,h}^{(j)}. \tag{17}$$

This probability is useful but not fully calibrated. The posterior mode probabilities express uncertainty among the finite mode set, but they do not capture all structural prediction errors, future mode switches, measurement contamination, or transition mismatch. The residual learner therefore treats $\hat{p}_{k,h}^{IMM}$, $\hat{y}_{k,h}^{IMM}$, covariance summaries, and innovation diagnostics as input features rather than final answers.

Table 3. Intruder Motion Modes

Mode	Horizontal Motion	Vertical Motion	Dynamics
CV-2D	Constant velocity	None	Linear
CV-3D	Constant velocity	Constant rate	Linear
Turn-2D	Coordinated turn	None	Nonlinear
Turn-3D	Coordinated turn	Active	Nonlinear
Acc-2D	Constant accel	None	Polynomial
Acc-3D	Constant accel	Active	Polynomial

Table 4. Prediction Horizons

Horizon (s)	Use Case	Difficulty	Notes
5	Short-term	Low	Minimal divergence
15	Near-term	Moderate	Early maneuvers
25	Mid-term	Moderate	Curvature impact
35	Operational	High	Key evaluation
45	Long-term	High	Mode uncertainty

4. Distributionally Calibrated Residual Learning Model

The learning model is trained to predict the residual between realized hazard states and the uncorrected interacting multiple model prediction. Let $y_{k,h}$ be the realized continuous hazard vector at time k and horizon h. Let $\hat{y}_{k,h}^{IMM}$ be the uncorrected multiple model prediction. The residual target is

$$\Delta y_{k,h} = y_{k,h} - \hat{y}_{k,h}^{IMM}. \tag{18}$$

The residual formulation is central to the empirical design. A direct learning model must learn the full mapping from current geometry and filtered history to future hazard states. The residual model only needs to learn where the physical predictor is biased, underdispersed, or delayed. The residual can still be nonlinear and heteroscedastic, but it is usually smaller and more structured than the original target. This makes the learning problem more stable and allows the physical estimator to remain visible in the prediction pipeline.

The model input is a feature vector $\phi_{k,h}$. It contains the current combined state estimate, selected elements of the combined covariance, posterior mode probabilities, mode entropy, recent normalized innovation magnitudes, current range, closing speed, estimated vertical rate, estimated horizontal curvature, uncorrected hazard predictions, and the horizon indicator. Mode entropy is included because high entropy often occurs just after a maneuver begins, when several filters remain plausible. Recent innovation magnitudes are included because sustained innovation growth can indicate that the active motion does not match the previously dominant mode. Covariance terms are included because uncertainty in position and velocity directly affects future closest approach.

The model predicts conditional quantiles of the residual distribution. For quantile level τ, the residual prediction is $f_{\theta,\tau}(\phi_{k,h})$. The corrected hazard state quantile is

$$\hat{q}_\tau\left(y_{k,h}|\phi_{k,h}\right) = \hat{y}_{k,h}^{IMM} + f_{\theta,\tau}\left(\phi_{k,h}\right). \tag{19}$$

The quantile set is $\mathcal{T} = \{0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95\}$. The median quantile is used as the deterministic point prediction. The lower and upper quantiles are used for predictive interval evaluation [4]. Predicting several quantiles allows the model to express asymmetric uncertainty. This matters because hazard state errors are often skewed near boundaries. For example, horizontal miss distance has a natural lower bound at zero, while large overprediction errors can occur when a turn unexpectedly brings the intruder closer than the current projection suggests.

The quantile loss uses the pinball loss. For scalar error e, the pinball loss is

$$\rho_\tau(e) = e\left(\tau - \mathbb{1}\{e < 0\}\right). \tag{20}$$

For the three continuous hazard state components, the full quantile loss is

$$\mathcal{L}_Q = \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{3}\sum_{\tau\in\mathcal{T}} w_{d,\tau}\rho_\tau\left(y_{n,d} - \hat{q}_{\tau,d}(\phi_n)\right). \tag{21}$$

The weights $w_{d,\tau}$ scale horizontal distance, time, and vertical distance to comparable numerical ranges. Without scaling, the horizontal distance loss would dominate because it is measured in feet and often has larger magnitude. The model also uses a binary classification head for loss of well clear probability. Let b_n be the true binary label and let $\hat{p}_n$ be the predicted probability. The classification loss is

$$\mathcal{L}_C = -\frac{1}{N}\sum_{n=1}^{N}\left[\alpha b_n \log\hat{p}_n + (1-b_n)\log(1-\hat{p}_n)\right]. \tag{22}$$

The positive class weight α is selected on the validation set. It compensates for class imbalance during training but is not used when computing test metrics. The classification head shares hidden layers with the continuous residual model. This shared representation is useful because the same geometry

Table 5. Horizontal Miss Distance MAE (ft)

Method	15 s	25 s	35 s
Linear	161	339	517
IMM	118	224	326
Neural	102	198	331
Proposed	89	158	222

Table 6. Classification Performance at 35 s

Model	Precision	Recall	F1
Linear	0.681	0.624	0.651
IMM	0.776	0.711	0.742
Recurrent	0.768	0.795	0.781
Proposed	0.812	0.826	0.819

and innovation features that explain residual horizontal miss distance also explain many loss of well clear classification errors.

A soft kinematic consistency penalty discourages corrections that conflict with local relative motion. The model does not output a full future trajectory, so hard trajectory constraints are not imposed. However, predicted hazard states should remain broadly compatible with current range, closing speed, vertical rate, and uncertainty. Let $\hat{H}_n^{med}$, $\hat{T}_n^{med}$, and $\hat{V}_n^{med}$ denote the median corrected horizontal miss distance, time to closest approach, and vertical separation for example n. Let $p_{H,n}$, $v_{H,n}$, $p_{V,n}$, and $v_{V,n}$ denote current horizontal position, horizontal velocity, vertical position, and vertical velocity estimates. The penalty is

$$\begin{aligned}
\mathcal{L}_K &= \frac{1}{N}\sum_{n=1}^{N}\left[\lambda_H \max(0,\eta_{H,n})^2 + \lambda_V \max(0,\eta_{V,n})^2\right], \\
\eta_{H,n} &= \hat{H}_n^{med} - \left(\|p_{H,n}\|_2 - \hat{T}_n^{med}\|v_{H,n}\|_2 - \sigma_{H,n}\right), \\
\eta_{V,n} &= \left|\hat{V}_n^{med} - |p_{V,n}|\right| - \hat{T}_n^{med}\left(|v_{V,n}| + \sigma_{V,n}\right).
\end{aligned} \tag{23}$$

The terms $\sigma_{H,n}$ and $\sigma_{V,n}$ are uncertainty allowances derived from the covariance matrix. The penalty is deliberately soft because real future maneuvers can change separation in ways not implied by current velocity. It is intended to reduce extreme inconsistent residual corrections, not to prevent the model from learning maneuver induced changes.

The full training objective is

$$\mathcal{L} = \mathcal{L}_Q + \gamma\mathcal{L}_C + \mathcal{L}_K + \omega\|\theta\|_2^2. \tag{24}$$

The neural residual model has four hidden layers with 192, 160, 128, and 96 units. A smooth nonlinear activation is applied after each hidden layer. Layer normalization is used after the first three hidden layers. Dropout with rate 0.06 is applied during training [6]. The model is trained with an adaptive gradient optimizer, an initial learning rate of 0.0007, cosine learning rate decay, and early stopping based on validation continuous ranked probability score. The maximum number of epochs is 60. Across five random seeds, the selected epoch ranges from 34 to 43, with a median of 38.

Post hoc calibration is applied using the validation set. Continuous quantile outputs are adjusted by monotone isotonic mappings learned separately for each horizon and each hazard state component [2]. If predicted quantile curves cross before calibration, a rearrangement step sorts them into nondecreasing order. The calibration mapping preserves quantile order. The binary classification probability is calibrated with temperature scaling followed by an isotonic correction. If a_n is the uncalibrated logit and T is the fitted temperature, the calibrated probability is

$$\tilde{p}_n = \mathcal{I}\left(\frac{1}{1+\exp(-a_n/T)}\right), \tag{25}$$

where $\mathcal{I}$ is the monotone isotonic calibration function. At the 35 s horizon under medium noise, the fitted temperature is 1.37, indicating moderate overconfidence before calibration.

The model is compared with three learning baselines in addition to the physical baselines. The direct neural predictor has the same feedforward architecture but predicts hazard states directly rather than residuals. The recurrent gated predictor receives a ten step sequence of filtered features and uses a gated recurrent unit before the output heads. The quantile gradient ensemble trains tree based gradient boosting models on the residual target for each quantile and horizon. These baselines separate the effect of learning, residualization, recurrence, and distributional calibration.

5. Empirical Results and Statistical Testing

The first evaluation examines deterministic horizontal miss distance error. Under medium measurement noise at the 35 s horizon, linear extrapolation has mean absolute error of 615 ft. The uncorrected interacting multiple model predictor reduces this to 381 ft. The direct neural predictor obtains 336 ft, the recurrent gated predictor obtains 312 ft, and the quantile gradient residual ensemble obtains 284 ft. The proposed distributionally calibrated residual model obtains 257 ft. This corresponds to a 58.4% reduction relative to linear extrapolation and a 32.6% reduction relative to the uncorrected multiple model predictor.

Median absolute horizontal miss distance error shows the same ordering. At 35 s, the median error is 522 ft for lin-

Table 7. Calibration Metrics

Model	ECE	Coverage (90%)	Interval Width
IMM	0.083	–	–
Neural	0.071	79.4%	980
Ensemble	0.046	86.1%	1214
Proposed	0.028	88.7%	1036

Table 8. Ablation Study at 35 s

Variant	MAE (ft)	F1	Coverage
Full Model	255	0.819	88.7%
No Residual	331	0.768	82.1%
No Calibration	255	0.812	85.9%
No Kinematic	268	0.807	86.5%

ear extrapolation, 331 ft for the uncorrected multiple model predictor, 289 ft for the direct neural model, 265 ft for the recurrent model, 241 ft for the quantile gradient ensemble, and 223 ft for the proposed model. The reduction is smaller in median error than in mean error because the proposed method especially reduces larger maneuver induced errors. The 95th percentile absolute error is 1,912 ft for linear extrapolation, 1,123 ft for the uncorrected multiple model predictor, 961 ft for the direct neural model, 914 ft for the recurrent model, 824 ft for the gradient ensemble, and 752 ft for the proposed method.

Performance varies by horizon [10]. At 5 s, all methods are relatively close because little future curvature develops [3]. The proposed method reduces mean absolute horizontal miss distance error from 49 ft for the uncorrected multiple model predictor to 41 ft [14]. At 15 s, the error decreases from 141 ft to 104 ft. At 25 s, it decreases from 257 ft to 176 ft. At 35 s, it decreases from 381 ft to 257 ft. At 45 s, it decreases from 532 ft to 407 ft. The absolute improvement is largest at 35 s, while the relative improvement is largest at 25 s and 35 s. At 45 s, future mode switches become more common, and prediction becomes less recoverable from current observations [7].

The improvement is concentrated in maneuvering encounters. For constant velocity encounters at 35 s, the proposed method reduces mean absolute horizontal miss distance error from 196 ft to 171 ft relative to the uncorrected multiple model predictor. For constant velocity with vertical motion, it reduces error from 221 ft to 184 ft. For coordinated turn encounters, it reduces error from 469 ft to 293 ft. For coordinated turns with vertical motion, it reduces error from 512 ft to 309 ft. For constant acceleration encounters, it reduces error from 438 ft to 291 ft. For three dimensional acceleration encounters, it reduces error from 456 ft to 303 ft. These results indicate that residual learning is most useful when the finite physical model bank approximates but does not fully match the realized maneuver.

Time to closest approach is also improved, though less dramatically. At the 35 s horizon under medium noise, mean absolute time error is 5.9 s for linear extrapolation, 4.3 s for the uncorrected multiple model predictor, 4.1 s for the direct neural model, 3.7 s for the recurrent model, 3.8 s for the quantile gradient ensemble, and 3.5 s for the proposed model. The recurrent model is competitive because timing depends strongly on recent temporal patterns. The proposed method remains slightly better because it uses innovation history, mode entropy, residual training, and calibration jointly. The difference between the proposed method and the recurrent model is modest but consistent across the 25 s and 35 s horizons.

Vertical separation error is smaller in magnitude because many encounters have stable vertical motion. At 35 s under medium noise, mean absolute vertical separation error is 149 ft for linear extrapolation, 122 ft for the uncorrected multiple model predictor, 115 ft for the direct neural model, 111 ft for the recurrent model, 108 ft for the gradient ensemble, and 102 ft for the proposed model. In vertical maneuver encounters only, the improvement is larger. The 95th percentile vertical separation error in those cases decreases from 433 ft for the uncorrected predictor to 346 ft for the proposed model. This matters because near threshold vertical errors can change loss of well clear classification even when horizontal prediction is acceptable.

Binary loss of well clear classification is evaluated using calibrated probabilities and a threshold selected on the validation set to maximize F1 score at the 35 s horizon under medium noise. The selected threshold for the proposed method is 0.41 [13]. On the test set, linear extrapolation obtains precision 0.682, recall 0.629, and F1 score 0.654. The uncorrected multiple model predictor obtains precision 0.779, recall 0.716, and F1 score 0.746. The direct neural predictor obtains precision 0.758, recall 0.787, and F1 score 0.772. The recurrent model obtains precision 0.771, recall 0.799, and F1 score 0.785. The quantile gradient residual ensemble obtains precision 0.796, recall 0.804, and F1 score 0.800. The proposed model obtains precision 0.815, recall 0.827, and F1 score 0.821.

The area under the precision recall curve at 35 s is 0.716 for linear extrapolation, 0.806 for the uncorrected multiple model predictor, 0.827 for the direct neural model, 0.842 for the recurrent model, 0.854 for the quantile gradient ensemble, and 0.875 for the proposed method. The gain relative to the uncorrected multiple model predictor comes mainly from reducing missed hazardous cases. At the fixed threshold, the uncorrected predictor misses 28.4% of hazardous prediction epochs. The proposed model misses 17.3%. The false alert rate among nonhazardous prediction epochs is 7.9% for the uncorrected predictor and 7.2% for the proposed method. The proposed model therefore improves recall without pro-

Table 9. Robustness Under Noise

Condition	IMM MAE	Proposed MAE	Improvement
Low Noise	297	209	29.6%
Medium Noise	374	255	31.8%
High Noise	512	361	29.5%
Heavy-Tailed	548	389	29.0%

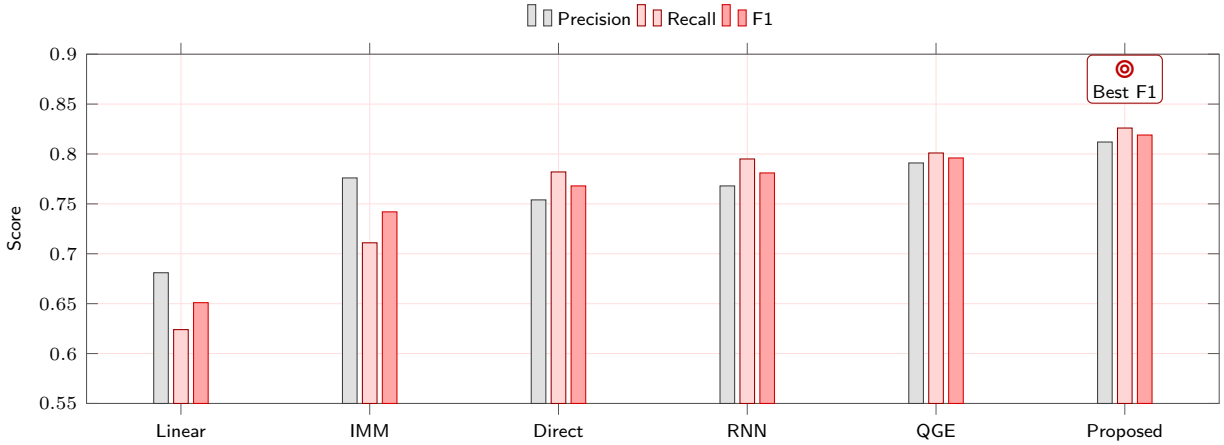


Figure 9. Loss-of-well-clear classification at 35 s under medium noise. The proposed method raises recall and F1 while keeping precision above the competing baselines.

ducing a higher false alert rate in the evaluated test distribution.

Probability calibration is measured with expected calibration error using ten adaptive bins. At the 35 s horizon under medium noise, expected calibration error is 0.086 for the uncorrected multiple model probability, 0.073 for the direct neural classifier before calibration, 0.060 for the recurrent classifier before calibration, 0.047 for the quantile gradient ensemble, and 0.028 for the proposed calibrated model. The proposed model is slightly overconfident in the highest risk bin. In that bin, the mean predicted probability is 0.871 and the empirical event frequency is 0.849. In the middle risk bins, absolute calibration differences are mostly below 0.025.

Predictive interval calibration is evaluated for continuous hazard states. For horizontal miss distance at 35 s, the nominal 90% interval from the direct neural quantile model covers 79.7% of test outcomes before calibration. The quantile gradient ensemble covers 86.3%. The proposed model covers 86.0% before calibration and 88.9% after calibration. The average width of the calibrated 90% horizontal interval is 1,041 ft, compared with 1,226 ft for the quantile gradient ensemble. The proposed model therefore improves coverage without merely widening intervals uniformly. For time to closest approach, nominal 90% coverage is 88.0%. For vertical separation, nominal 90% coverage is 89.4%.

Continuous ranked probability score also favors the proposed method. At 35 s under medium noise, the horizontal miss distance score is 322 for the direct neural quantile model, 294 for the quantile gradient ensemble, 289 for the proposed model before calibration, and 276 after calibration. The calibration step improves probabilistic score but does not change median error substantially. This indicates that the calibration procedure mainly adjusts distributional spread rather than shifting central predictions. For vertical separation, the calibrated proposed model improves continuous ranked probability score by 13.1% relative to the direct neural quantile model.

Statistical testing is performed at the encounter level. Because each encounter contributes many correlated time steps, time step level tests would overstate precision. For each encounter, the mean absolute error across eligible prediction epochs is computed for each method and horizon. Paired differences are then formed across encounters. Bootstrap confidence intervals resample encounters with replacement. Signed rank tests are applied to encounter level paired differences. Binary classification comparisons use clustered bootstrap resampling and paired disagreement counts [19].

At 35 s under medium noise, the encounter level mean absolute horizontal miss distance error difference between the proposed method and the uncorrected multiple model predictor is 124 ft [25]. The bootstrap 95% confidence interval is 116 ft to 132 ft. The encounter level median difference is 86 ft, with a bootstrap 95% confidence interval of 78 ft to 94 ft. The signed rank test rejects the zero median difference hypothesis with $p < 0.001$. Compared with the quantile gradient residual ensemble, the proposed method has a mean absolute error advantage of 27 ft, with a bootstrap 95% confidence interval of 18 ft to 36 ft.

For binary loss of well clear at 35 s, the proposed model correctly identifies 3,286 hazardous prediction epochs missed by the uncorrected multiple model classifier, while the uncorrected classifier identifies 1,094 hazardous epochs missed by the proposed model. Among nonhazardous epochs, the proposed model removes 1,207 false alerts made by the uncorrected classifier and creates 641 false alerts not made by

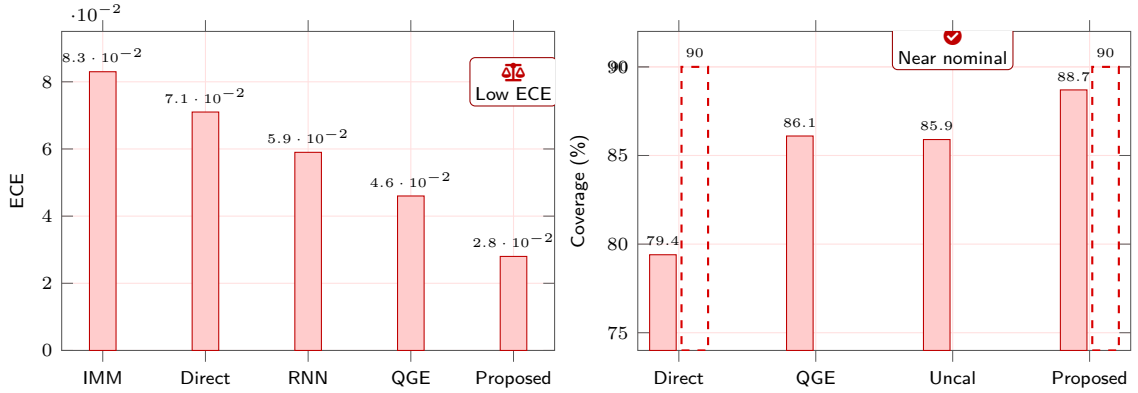


Figure 10. Calibration diagnostics at the 35 s horizon. Expected calibration error is lowest for the calibrated residual model, and its nominal 90% interval coverage is closest to the target without the widest interval construction.

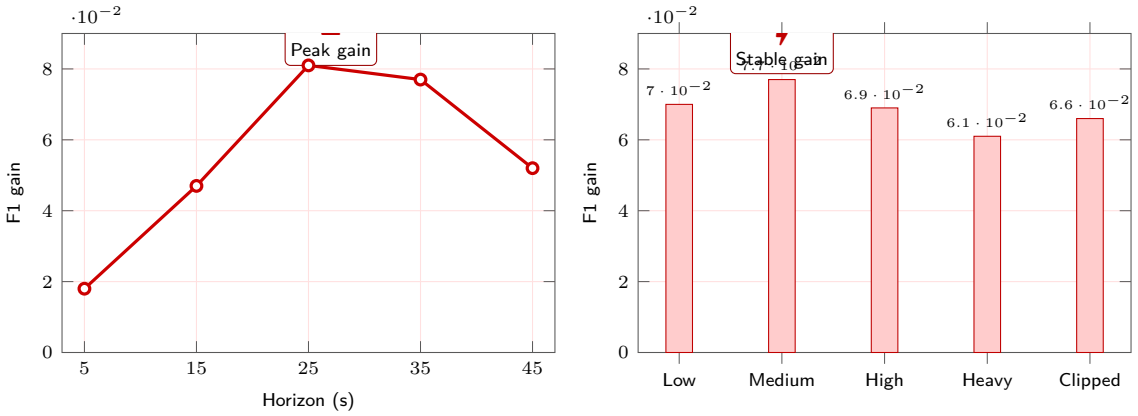


Figure 11. F1 improvements over the uncorrected interacting multiple model predictor across look-ahead horizons and noise regimes. The strongest classification gains occur at intermediate horizons and remain visible under degraded measurements.

the uncorrected classifier. The clustered bootstrap confidence interval for the F1 improvement is 0.060 to 0.091. The paired classification test gives $p < 0.001$. These results are statistically strong in the simulated test set, although statistical significance should not be interpreted as operational adequacy.

The ablation study isolates the contribution of model components. Removing residualization and training the same network to predict hazard states directly increases 35 s horizontal miss distance mean absolute error from 257 ft to 336 ft. Removing kinematic consistency penalties increases it to 270 ft and increases quantile crossing before rearrangement [9]. Removing recent innovation features increases it to 276 ft, mainly in encounters with mode changes within the previous 10 s. Removing mode entropy increases expected calibration error from 0.028 to 0.039. Removing post hoc calibration leaves median error nearly unchanged but reduces nominal 90% horizontal interval coverage from 88.9% to 86.0%. These results show that residualization drives the main point prediction gain, while calibration and distributional outputs drive the main uncertainty gain.

A seed sensitivity experiment trains the proposed model five times with different random initializations. At 35 s under medium noise, horizontal miss distance mean absolute error ranges from 254 ft to 261 ft. Loss of well clear F1 score ranges from 0.817 to 0.824. Expected calibration error ranges from 0.027 to 0.031 after calibration. A five seed ensemble reduces mean absolute error to 249 ft and increases F1 score to 0.826, but the single model is retained as the main result because the ensemble adds computational complexity. The seed variability is smaller than the difference between the proposed model and the uncorrected multiple model predictor.

Computational cost is evaluated on the same test hardware for all methods. The uncorrected interacting multiple model predictor requires six mode conditioned forward propagations for each horizon. The proposed method adds feature construction and one neural network evaluation [12]. Average per time step prediction time is 2.9 ms for the uncorrected multiple model predictor and 3.3 ms for the proposed method in the tested implementation. The direct neural model is faster but less accurate. The recurrent model is slightly slower than the proposed model because it processes a ten step sequence. These timing values are hardware dependent, but they indicate that the proposed method achieves its accuracy gain without requiring a very large model.

6. Robustness Analysis and Error Diagnostics

The first robustness experiment increases measurement noise. Under the low noise condition at 35 s, the uncorrected multiple model predictor has horizontal miss distance mean ab-

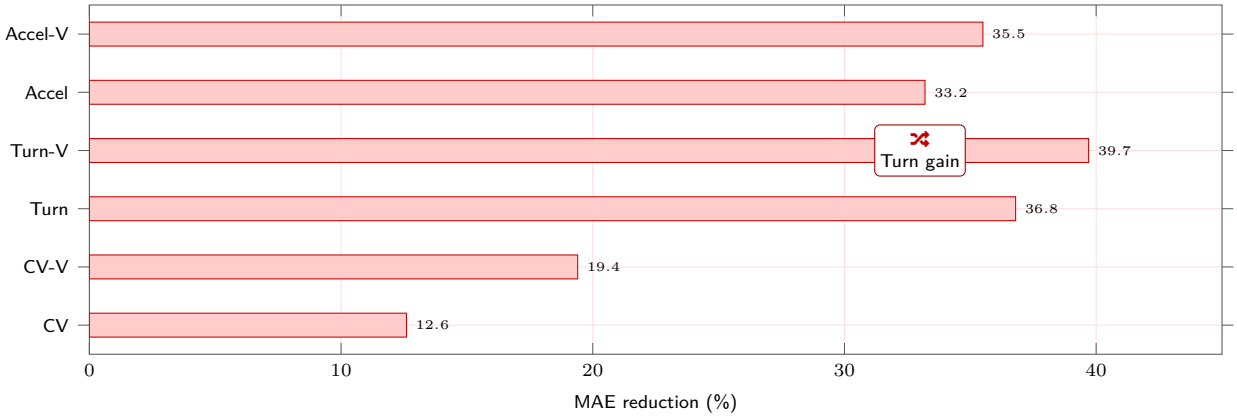


Figure 12. Active maneuver stratification at 35 s under medium noise. Horizontal miss-distance error reductions are largest for coordinated turns and acceleration regimes, where the residual model corrects structured dynamic mismatch.

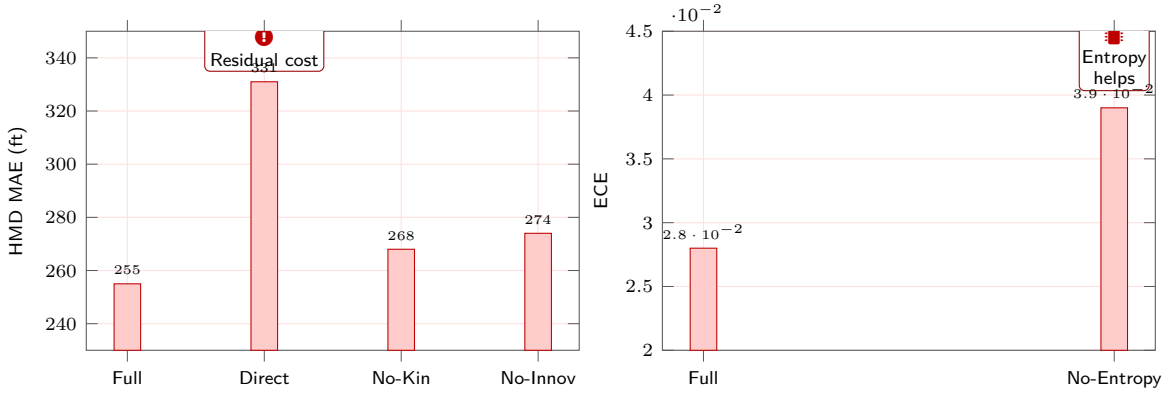


Figure 13. Ablation diagnostics at the 35 s horizon. Direct prediction, removal of kinematic consistency, removal of innovation history, and removal of entropy information each degrade either deterministic accuracy or calibration.

solute error of 302 ft, while the proposed model has 211 ft. Under medium noise, the corresponding values are 381 ft and 257 ft. Under high noise, they are 519 ft and 366 ft. The relative improvement is therefore 30.1%, 32.6%, and 29.5% across the low, medium, and high noise conditions. The proposed method degrades as noise increases, but the advantage over the uncorrected predictor remains visible. This suggests that the residual model is not merely exploiting clean measurements [18].

The high noise condition also affects classification. At 35 s, the uncorrected multiple model predictor obtains F1 score 0.694, while the proposed method obtains 0.767 [1]. The false alert rate is 10.7% for the uncorrected predictor and 9.9% for the proposed method. The missed hazardous rate is 34.2% for the uncorrected predictor and 23.1% for the proposed method. The improvement in missed hazardous cases remains the primary classification benefit. However, calibration weakens under high noise. Expected calibration error increases from 0.028 in the medium noise condition to 0.041 in the high noise condition. This is still lower than the uncorrected predictor but indicates that the validation calibration from the medium condition does not transfer perfectly.

The heavy tailed measurement condition is more challenging because occasional contaminated observations can produce large innovations and temporary mode probability instability. Without any robust preprocessing, the uncorrected multiple model predictor has 35 s horizontal miss distance mean absolute error of 556 ft. The proposed model has 395 ft. Nominal 90% horizontal interval coverage falls to 87.0%, indicating undercoverage. When a simple innovation clipping rule is applied before filtering, the uncorrected predictor improves to 534 ft and the proposed method improves to 379 ft. Coverage improves to 88.2%. This shows that residual learning does not replace robust filtering. If measurements are contaminated, the tracking layer still needs protection against extreme updates [30].

A transition probability mismatch experiment evaluates whether the model depends too strongly on the nominal Markov transition matrix. The training and validation data use the nominal transition structure. The test data are then generated under two altered conditions. In the persistent mode condition, diagonal transition probabilities are increased and modes last longer than expected. In the frequent switch condition, diagonal probabilities are decreased and modes change more often. At 35 s, the proposed model has horizontal miss distance mean absolute error of 244 ft in the persistent condition and 342 ft in the frequent switch condition. The uncorrected multiple model predictor has errors of 356 ft and 439 ft, respectively. The proposed method remains better in both cases, but the margin is smaller when

modes switch more frequently because future maneuvers become less predictable from the current filtered state.

A turn rate distribution shift experiment increases the upper tail of intruder turn rates by 25% at test time while leaving training unchanged. At 35 s, the uncorrected multiple model predictor has horizontal miss distance mean absolute error of 458 ft, and the proposed model has 314 ft. The proposed method remains better, but its nominal 90% horizontal interval coverage declines to 86.4%. A conformal inflation factor fitted on a small shifted validation subset restores coverage to 89.5%, but it increases average interval width by 12.6%. This result shows that point prediction improvement can persist under moderate distribution shift while uncertainty calibration remains sensitive to the test distribution.

An ownship perturbation experiment introduces mild ownship speed and heading changes [21]. The ownship speed varies within 15 kt of nominal, and heading changes are limited to 1.5 degrees per second for short intervals. These ownship maneuvers are known to the estimator when they occur, but they are not part of the main training distribution. At 35 s, the uncorrected multiple model predictor has horizontal miss distance mean absolute error of 414 ft, while the proposed method has 291 ft. Loss of well clear F1 score is 0.735 for the uncorrected predictor and 0.802 for the proposed model. Expected calibration error increases from 0.028 to 0.036. The model therefore transfers reasonably to mild ownship perturbation, but probability calibration again weakens under shift.

The first recurring failure mode involves unobserved future maneuver initiation. If an intruder flies straight through time k and begins a turn at time $k+3$, no predictor using only measurements through time k can know that exact future mode change unless there are precursor patterns in the data. The residual model can learn that certain geometries have higher maneuver risk, and it can widen intervals in those regions, but it cannot identify arbitrary future intent. These cases dominate the largest errors at 45 s. They explain why gains decline at the longest horizon even though the proposed method remains better than the baselines.

The second failure mode involves nearly tangential encounters near the horizontal threshold. In these cases, small errors in relative velocity direction can produce large changes in predicted time to closest approach while horizontal miss distance remains near 4,000 ft. The loss of well clear label can change with a small timing shift. The proposed model improves calibration in these cases but still produces classification errors when the trajectory skims the boundary. This is a structural difficulty of threshold based labels. A continuous risk score may be more stable than a hard label in these marginal geometries, but operational alerting often requires threshold decisions.

The third failure mode involves vertical separation near 450 ft with uncertain vertical rate. If horizontal prediction is accurate but vertical separation is near the threshold, small vertical errors can change the binary label. In the test set, 21.4% of proposed model classification errors at 35 s occur in cases where true vertical separation at closest approach lies between 400 ft and 500 ft. Many of these errors have small continuous vertical residuals but cross the binary threshold. This indicates that binary metrics can be sensitive to small continuous errors near decision boundaries.

Residual plots show that the uncorrected multiple model predictor has systematic horizontal miss distance bias in high curvature bins. When the intruder turns inward relative to the ownship, the uncorrected predictor often overestimates miss distance during the early part of the maneuver. When the intruder turns outward, it can underestimate risk. The residual learner reduces both biases by using innovation history, mode probabilities, and curvature features. However, residual variance remains larger in high curvature bins than in constant velocity bins [16]. The calibrated intervals widen in those bins, which improves coverage but does not eliminate all errors.

Covariance features contribute to both accuracy and calibration. When covariance features are removed, 35 s horizontal miss distance mean absolute error increases from 257 ft to 274 ft and nominal 90% interval coverage declines from 88.9% to 86.9%. Cross covariance terms between horizontal position and velocity are especially useful for time to closest approach. Mode entropy is also useful. High entropy combined with large innovation magnitude often marks the period just after a mode switch, when the filter bank has not yet settled. Removing mode entropy increases classification expected calibration error and slightly increases missed hazardous cases.

The model is checked for overcorrection in benign encounters [20]. For encounters whose true minimum horizontal separation exceeds 8,000 ft, the median absolute correction applied by the residual model at 35 s is 63 ft. For encounters whose true minimum horizontal separation lies between 3,000 ft and 5,000 ft, the median absolute correction is 211 ft. The model therefore applies larger corrections near the decision boundary and smaller corrections in clearly benign cases. This pattern is desirable because large corrections in easy benign cases could create unnecessary false alerts.

A horizon consistency diagnostic examines whether predicted minimum horizontal separation is coherent across horizons. For the same prediction epoch, the minimum separation over a longer horizon should not exceed the minimum separation over a shorter horizon when computed from a single trajectory. Independent horizon predictions can violate this monotonic relationship. Before post processing, the direct neural model violates horizon consistency in 7.1% of test examples. The proposed model violates it in 2.3% before correction and 0.5% after rearrangement and consistency adjustment. The remaining violations are small and mostly arise from calibrated quantile adjustments. This diagnostic is useful because ordinary mean absolute error would not reveal this type of physical inconsistency.

The robustness experiments show that the proposed method improves prediction across several controlled shifts but remains sensitive to shifts that change the uncertainty distribution [5]. Point estimates transfer more reliably than calibrated intervals. This is expected because calibration is a distributional property. If high curvature encounters become more common or measurement contamination becomes heavier, the empirical frequency of errors within predicted intervals changes. A deployment oriented system would need ongoing calibration checks, conservative inflation under unfamiliar conditions, and scenario based stress testing. The present study demonstrates the value of residual calibration in simulation but does not remove the need for external validation.

7. Conclusion

This paper presented an empirical Monte Carlo study of distributionally calibrated residual learning for unmanned aircraft detect and avoid hazard prediction. The proposed method combines an interacting multiple model state estimator with a supervised residual model. The estimator supplies physically interpretable state predictions, mode probabilities, covariance summaries, and innovation diagnostics. The residual model corrects future hazard state predictions and provides calibrated uncertainty estimates. This structure is designed to preserve a traceable dynamic backbone while using machine learning to reduce systematic prediction errors left by finite motion models and nominal transition assumptions.

The results show that residual learning improves hazard prediction in the simulated encounter distribution. Under medium measurement noise at the 35 s horizon, the proposed method reduces horizontal miss distance mean absolute error from 615 ft for linear extrapolation and 381 ft for the uncorrected interacting multiple model predictor to 257 ft. It improves loss of well clear F1 score from 0.654 for linear extrapolation and 0.746 for the uncorrected multiple model predictor to 0.821. Predictive interval evaluation also improves, with nominal 90% horizontal miss distance intervals achieving 88.9% empirical coverage after calibration. The largest gains occur in coordinated turn and acceleration encounters, where linear extrapolation and uncorrected model propagation are most likely to leave structured residual errors.

The empirical findings also show clear limits. The proposed method cannot reliably predict arbitrary future maneuver initiations that have no observable precursor before the prediction epoch. Its advantage declines at the longest horizon because unobserved future mode switches become more important. Calibration weakens under high turn rate distribution shift and heavy tailed measurement contamination [15]. Robust filtering, conservative uncertainty inflation, and distribution shift monitoring would therefore be necessary in any more realistic development path. The model improves simulated prediction performance, but it does not by itself establish operational suitability.

The main methodological conclusion is that physical multiple model estimation and supervised residual learning can be complementary. The interacting multiple model estimator captures first order dynamic structure and provides interpretable uncertainty summaries [29]. The residual learner uses those summaries to correct systematic hazard state errors and estimate uncertainty around future separation. This hybrid design performs better than direct black box prediction in the tested setting and better than an uncorrected physical predictor in maneuvering encounters [17]. Future empirical work should extend the simulation to asynchronous sensors, missed detections, track initiation uncertainty, ownship maneuver planning, multi intruder encounters, and flight data comparison while preserving encounter level train test separation and explicit hazard state evaluation.

References

[1] The space economy at a glance 2014 (summary in chinese), 10 2014.

[2] Bharat Sharma Acharya, Mahendra Bhandari, Filippo Bandini, Alonso Pizarro, Matthew T. Perks, Deepak R. Joshi, Sheng Wang, Toby Dogwiler, Ram L. Ray, Gehendra Kharel, and Sadikshya Sharma. Unmanned aerial vehicles in hydrology and water management: Applications, challenges, and perspectives. *Water Resources Research*, 57(11), 11 2021.

[3] Tasneem Ahmed, Leo Creedon, and Salem S Gharbia. Low-cost sensors for monitoring coastal climate hazards: A systematic review and meta-analysis. *Sensors (Basel, Switzerland)*, 23(3):1717–1717, 2 2023.

[4] Juan Alvarez and Todd Allen. Batelle energy alliance, llc (bea) 2014 annual report for idaho national laboratory (inl), 10 2014.

[5] Jayme Garcia Arnal Barbedo. A review on the use of unmanned aerial vehicles and imaging sensors for monitoring and assessing plant stresses. *Drones*, 3(2):40–, 4 2019.

[6] J B Bartholomees. The u.s. army war college guide to national security issues. volume 1: Theory of war and strategy, 7 2010.

[7] Suraj Bijjahalli, Alessandro Gardi, Nichakorn Pongsakornsathien, Roberto Sabatini, and Trevor Kistan. A unified airspace risk management framework for uas operations. *Drones*, 6(7):184–184, 7 2022.

[8] Adriano C Canolla, Michael B Jamoom, and Boris Pervan. Interactive multiple model hazard states prediction for unmanned aircraft systems (uas) detect and avoid (daa). In *2018 AIAA Information Systems-AIAA Infotech@ Aerospace*, page 2011. 2018.

[9] Alain Dinis, Maxime Loil, and Gerardo Anzaldúa. D5.1: Plan for exploitation of dwc results - first version (m18), 12 2020.

[10] Pedro F. A. Di Donato and Ella M. Atkins. Evaluating risk to people and property for aircraft emergency landing planning. *Journal of Aerospace Information Systems*, 14(5):259–278, 2017.

[11] Sara Ernst. Woods hole coastal and marine science center—2019 annual report, 2020.

[12] P. K. Freeman and R. S. Freeland. Red tape in higher education institutions: Uav policy. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W5:119–126, 5 2019.

[13] Eric W. Frew, Brian Argrow, Adam L. Houston, C. Weiss, and Jack Elston. Iccs - an energy-aware airborne dynamic data-driven application system for persistent sampling and surveillance. *Procedia Computer Science*, 18:2008–2017, 2013.

[14] Anita K. Gianotto, Kemal O. Pasamehmetoglu, Steven Hartenstein, James Herzog, Mark Henry, Dan Ogden, and Kelly Estes. Idaho national laboratory mission accomplishments, 11 2014.

[15] Anthony P. Hamins, Casey Grant, Nelson P. Bryner, Albert W. Jones, and Galen H. Koepke. Research roadmap for smart fire fighting, 2015.

[16] Mark D. Ivey, David G. Robinson, Mark Boslough, George A. Backus, Kara J. Peterson, Bart G. van Bloemen Waanders, Laura Painton Swiler, Darin Desilets, and Rhonda K. Reinert. Arctic climate systems analysis, 3 2015.

[17] Qingyu Jiao, Yansi Liu, Zhigang Zheng, Linshi Sun, Yiqin Bai, Zhengjuan Zhang, Longni Sun, Gaosheng Ren, Guangyu Zhou, Xinfeng Chen, and Yan Yan. Ground risk assessment for unmanned aircraft systems based on dynamic model. *Drones*, 6(11):324–324, 10 2022.

[18] Honggu Kang, Jingon Joung, Jin Young Kim, Joonhyuk Kang, and Yong Soo Cho. Protect your sky: A survey of counter unmanned aerial vehicle systems. *IEEE Access*, 8:168671–168710, 2020.

[19] Anders la Cour-Harbo and Henrik Schiøler. Probability of low altitude midair collision between general aviation and unmanned aircraft. *Risk analysis : an official publication of the Society for Risk Analysis*, 39(11):2499–2513, 7 2019.

[20] Brett Lawrence, Emerson de Lemmus, and Hyuk Cho. Uas-based real-time detection of red-cockaded woodpecker cavities in heterogeneous landscapes using yolo object detection algorithms. *Remote Sensing*, 15(4):883–883, 2 2023.

[21] Edmund Lorencowicz, Jacek Uziak, and Bruno Huyghebaert. Farm machinery and processes management in sustainable agriculture. In *Farm Machinery and Processes Management in Sustainable Agriculture, X International Scientific Symposium*, pages 5–453. IN-W Spatium, 11 2019.

[22] Javier López and Miguel Ángel Aranda Gómez. User-chi project - d8.11: Report on the standardisation landscape., 7 2020.

[23] S. Mathews. Final closure plan for the high-explosives open burn treatment facility at lawrence livermore national laboratory experimental test site 300, 4 1997.

[24] Kamesh Namuduri. Advanced air mobility: Research directions, 10 2021.

[25] Daniel Overholt. Grafting acoustic instruments and signal processing: Creative control and augmented expressivity. *The Journal of the Acoustical Society of America*, 134(5):4054–4054, 11 2013.

[26] Anna Rabajczyk, Jacek Zboina, Maria Zielecka, and Radosław Fellner. Monitoring of selected cbrn threats in the air in industrial areas with the use of unmanned aerial vehicles. *Atmosphere*, 11(12):1373–, 12 2020.

[27] Abbas Rajabifard. *Sustainable Development Goals Connectivity Dilemma*. CRC Press, 8 2019.

[28] Sean Reilly, Matthew L. Clark, Lisa Patrick Bentley, Corbin Matley, Elise Piazza, and Imma Oliveras Menor. The potential of multispectral imagery and 3d point clouds from unoccupied aerial systems (uas) for monitoring forest structure and the impacts of wildfire in mediterranean-climate forests. *Remote Sensing*, 13(19):3810–, 9 2021.

[29] Christopher A. Roseman and Brian Argrow. Weather hazard risk quantification for suas safety risk management. *Journal of Atmospheric and Oceanic Technology*, 37(7):1251–1268, 7 2020.

[30] Antonio R. Segales, Brian R. Greene, Tyler M. Bell, William Doyle, Joshua J. Martin, Elizabeth A. Pillar-Little, and Phillip B. Chilson. The coptersonde: an insight into the development of a smart unmanned aircraft system for atmospheric boundary layer research. *Atmospheric Measurement Techniques*, 13(5):2833–2848, 5 2020.

[31] Chen Suo, Eugene McGovern, and Alan Gilmer. Uav data for coastal dune mapping. In *Proccedings of 10th International Conference "Environmental Engineering"*. VGTU Technika, 8 2017.

[32] Asma Tabassum and William H. Semke. Assessing the effect of ads-b message drop out in detect and avoid of unmanned aircraft system using monte carlo simulation, 9 2018.

[33] Michał Wiśnios, Sebastian Tatko, Michał Mazur, Jacek Paś, Jarosław Mateusz Łukasiak, and Tomasz Klimczak. Identifying characteristic fire properties with stationary and non-stationary fire alarm systems. *Sensors (Basel, Switzerland)*, 24(9):2772–2772, 4 2024.