Peer reviewed Journals
seron press

RESEARCH

Residual State Space Learning for Early Detect and Avoid Hazard Prediction in Maneuvering UAS Encounters

Hassan Emad,[1] Mostafa Adel,[2] and Youssef Samir[3]

[1]Department of Information Systems, Sinai Higher Institute for Engineering and Technology, El-Arish Road, Arish, North Sinai Governorate, Egypt
[2]Department of Electrical Engineering, Modern Academy for Engineering and Technology, Industrial Area, Maadi, Cairo 11728, Egypt
[3]Department of Computer Engineering, Thebes Higher Institute of Engineering, Thebes Academy, South El-Academia District, New Cairo, Cairo Governorate, Egypt

Abstract

Detect and Avoid prediction for unmanned aircraft requires timely recognition of hazardous encounters under noisy sensing, maneuvering intruders, and incomplete knowledge of future intent. Conventional extrapolation can become unstable when the intruder turns, accelerates, climbs, or descends, while unconstrained machine learning can produce accurate rankings without preserving physically interpretable encounter geometry. This paper develops and evaluates an empirical residual state space learning method for early hazard prediction in non-cooperative unmanned aircraft encounters. The method combines a stochastic kinematic predictor with a compact recurrent residual learner trained on simulated encounter histories, continuous closest-approach quantities, and binary loss-of-well-clear labels. A dataset of 240000 maneuvering encounters is generated with strict separation between training, validation, and test seeds. The proposed model is compared against constant-velocity extrapolation, coordinated-turn prediction, interacting multiple-model filtering, gradient-boosted hazard prediction, and an unconstrained recurrent neural baseline. In the held-out test set, the residual state space model reduces horizontal miss-distance root mean square error by 31.7% relative to the best filtering baseline and by 22.8% relative to the direct recurrent baseline. The model improves loss-of-well-clear area under the receiver operating characteristic curve from 0.941 to 0.973 relative to the interacting multiple-model predictor and increases 35-second warning recall at an approximately fixed 5% false-alert rate from 81.4% to 89.6%. Bootstrap intervals, McNemar tests, and encounter-level loss comparisons indicate that these gains are statistically stable across maneuver intensity, sensor noise, and prediction horizon strata.

1. Introduction

Detect and Avoid prediction for unmanned aircraft systems is a sequential empirical inference problem rather than a static classification problem. The own aircraft observes an intruder through imperfect surveillance, transforms measurements into a relative state estimate, and must decide whether the encounter will become hazardous before the future geometry is completely known. In non-cooperative encounters, the intruder may not broadcast intent, velocity, or acceleration. The prediction system therefore has to infer near-future behavior from a short measurement history, from the geometry of the relative trajectory, and from uncertainty accumulated in the state estimator. A method that only estimates the present intruder position can be insufficient because the alerting decision depends on what will happen within a future look-ahead window.

The difficulty is especially clear when the intruder maneuvers. A constant-velocity projection can perform well when the encounter is nearly linear, but it can overestimate or underestimate future separation when the intruder changes heading or vertical rate. A coordinated-turn model can help when the target follows a smooth turn, but it can become biased when the intruder transitions between straight, turning, climbing, and accelerating segments. Multiple-model filtering addresses part of this problem by allowing a finite set of motion hypotheses to compete over time, but it still depends on the adequacy of the model set and on how rapidly mode probabilities adapt after a maneuver begins. Purely discriminative machine learning can learn nonlinear patterns in the data, but it can also lose the transparent relation between predicted risk and physical separation. This paper studies an alternative that keeps the physical prediction backbone and learns only the residual errors that remain after structured propagation.

The method proposed here is residual state space learning. A stochastic kinematic predictor first estimates the relative state and propagates candidate future trajectories. A recurrent residual learner then receives filtered state summaries, innovation histories,

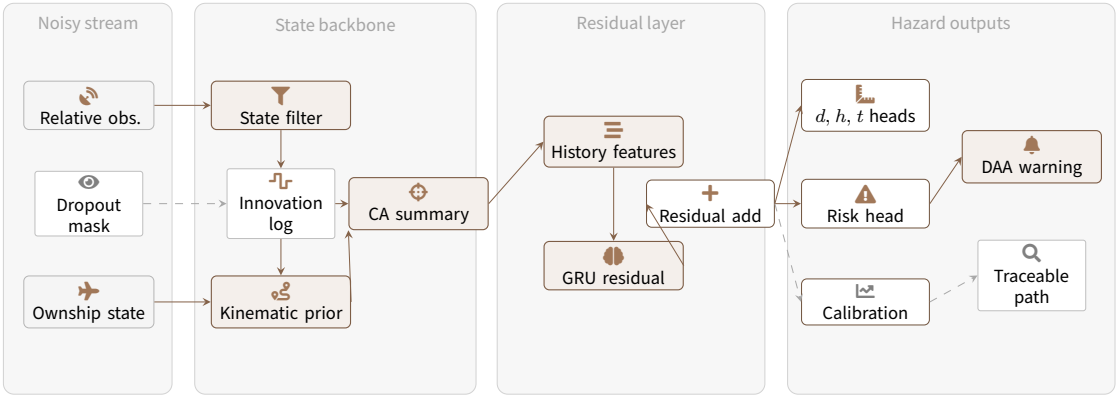


Figure 1. Residual state-space hazard-prediction pipeline for non-cooperative UAS encounters, preserving an inspectable kinematic forecast while learning recurrent corrections for closest-approach quantities and warning probability.

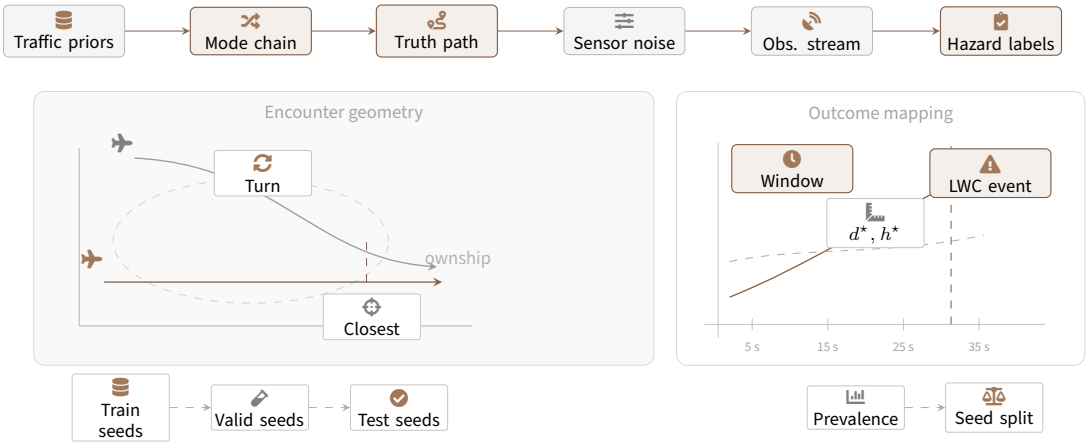


Figure 2. Encounter generation and deterministic construction of continuous closest-approach targets and binary loss-of-well-clear labels from simulated future relative trajectories.

mode probabilities, uncertainty features, and baseline closest-approach forecasts. It estimates corrections to horizontal miss distance, vertical separation, and time to closest approach, and it produces a calibrated probability of loss of well clear within a specified prediction horizon. The residual component is therefore not asked to learn all encounter dynamics from raw observations. It is asked to learn when and how the structured predictor is systematically wrong [1].

This paper is intentionally different from the two existing research directions mentioned by the user. It does not build a conformal graph temporal risk surface, and it does not perform distributionally calibrated multiple-model learning. The unit of analysis is the encounter-time prediction origin. The model is a residual state space predictor, not a graph risk estimator and not a distributionally robust calibration scheme. The empirical tests focus on whether additive residual correction improves continuous hazard-state prediction and binary alerting relative to conventional filtering, tree-based supervised learning, and unconstrained recurrence.

The study is empirical and simulation based. A large set of non-cooperative encounters is generated under maneuvering intruder dynamics, noisy position measurements, and limited measurement dropout. Each encounter is sampled at one-second intervals over a 60-second horizon. The train, validation, and test split is performed by encounter seed, not by time step, so the model is not evaluated on later portions of trajectories that appeared during training. The primary prediction horizon is 35 seconds, and additional 15-second and 25-second horizons are used for sensitivity analysis. The primary event outcome is loss of well clear within the prediction window, while the continuous outcomes are horizontal miss distance, vertical separation at closest horizontal approach, and time to closest horizontal approach.

The motivating methodological context is the prior use of multiple-model prediction for unmanned aircraft Detect and Avoid hazard states. Canolla et al. (2018) used encounter-model trajectories and an Interactive Multiple Model prediction phase to estimate horizontal closest approach, modified time to closest approach, and vertical separation for maneuvering intruder encounters [2]. The present paper keeps the idea that

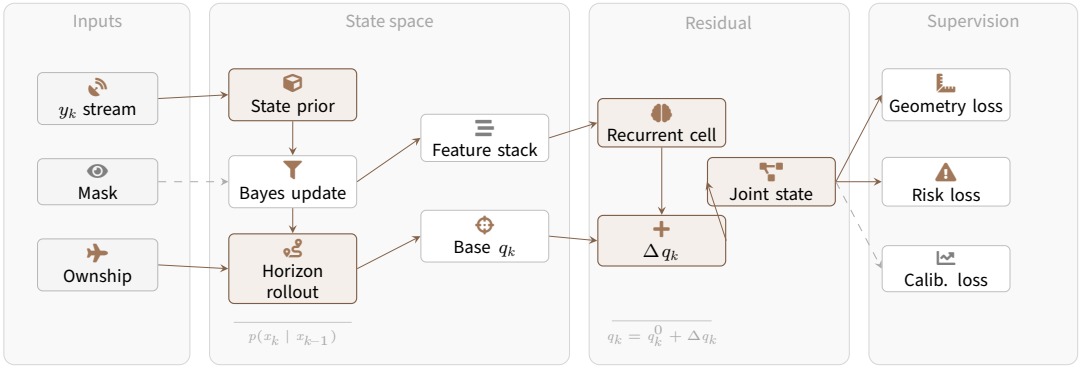


Figure 3. Residual state-space learning architecture, where stochastic filtering supplies physically grounded forecasts and recurrent residuals correct systematic closest-approach errors.

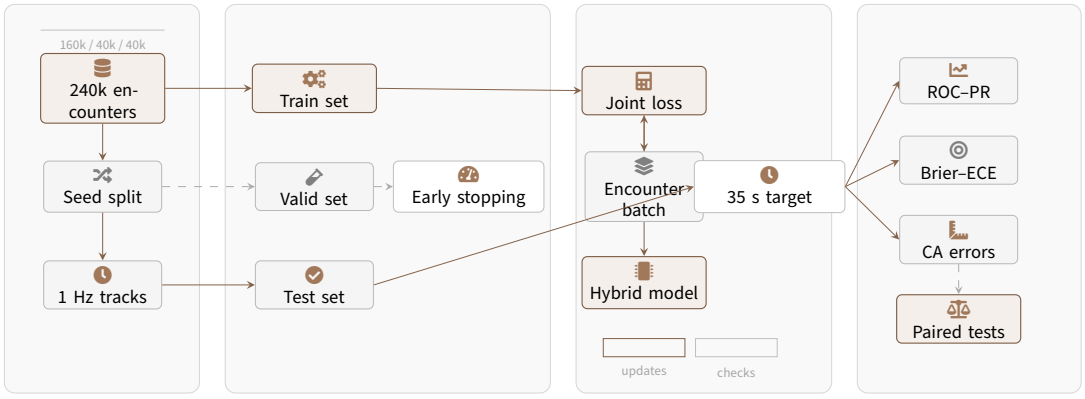


Figure 4. Empirical protocol with encounter-seed separation, validation-driven model selection, multi-target training, and encounter-level paired statistical comparisons.

hazard prediction should use forecasted state evolution rather than only instantaneous linear formulas, but it changes the estimator by adding a supervised residual correction trained against realized future hazard quantities.

The key empirical question is whether residual learning improves both geometry and alerting. A useful predictor should not only rank hazardous encounters above non-hazardous encounters. It should also produce plausible estimates of how close the aircraft will come, how soon that closest approach will occur, and whether the vertical component of the encounter contributes to the hazard. This requirement is important because a binary alert without a coherent geometric explanation may be difficult to diagnose. The proposed model is therefore evaluated with both continuous loss and binary classification metrics.

The results support a narrow but useful conclusion. In the simulated test distribution, the residual state space model improves horizontal miss-distance root mean square error, vertical separation error, time-to-closest-approach error, receiver operating characteristic area, precision-recall area, Brier score, expected calibration error, and recall at a fixed false-alert rate. The largest gains appear in moderate and high maneuver-intensity encounters, where the physical predictor has structured residual errors that the recurrent module can learn. The model is not claimed to certify a Detect and Avoid system, and it is not a substitute for full safety analysis. It is an empirical prediction module whose performance is measured under explicit simulated conditions.

The remainder of the paper defines the encounter data, the hazard outcomes, the residual state space model, the training procedure, the baselines, the statistical tests, the empirical results, robustness analyses, and operational interpretation [3]. The conclusion summarizes the supported findings and the main limitations [4].

2. Encounter Data and Hazard Label Construction

The empirical analysis uses simulated encounters because supervised evaluation requires knowledge of future closest-approach quantities. For each encounter, the own aircraft defines the reference frame. The horizontal axes are centered on the own aircraft at the current time, and altitude is represented as relative vertical separation. The intruder is non-cooperative in the sense that the prediction model receives only noisy relative position measurements and does not receive a

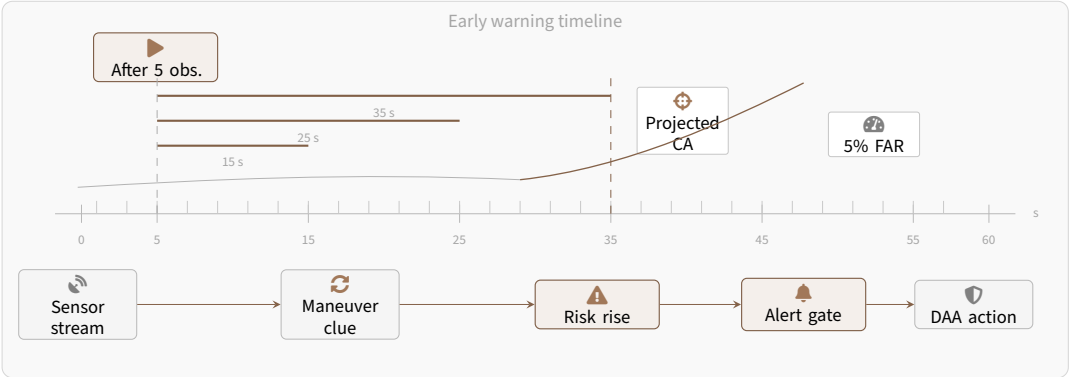


Figure 5. Operational timing view showing how observations after the fifth second feed rolling 15-, 25-, and 35-second forecasts until the calibrated risk score crosses the alerting gate.

direct intent message. The true intruder trajectory is stored for label generation but is not provided to the prediction model.

Each encounter lasts 60 seconds and is sampled at 1 Hz. Prediction begins after five observations so that all sequence models have at least a short history. Evaluation uses prediction origins from second 5 through second 55. The primary prediction horizon is 35 seconds. If the horizon extends beyond the end of the encounter record, that prediction origin is excluded from the horizon-specific evaluation. The same generated trajectory can therefore contribute more 15-second prediction origins than 35-second prediction origins, but all models are evaluated on identical origin sets for each horizon.

The main dataset contains 240000 encounters [5]. The training set contains 160000 encounters, the validation set contains 40000 encounters, and the test set contains 40000 encounters. The split is made by random encounter seed. No encounter contributes observations to more than one split. This design prevents temporal leakage because neighboring time steps within a single encounter are strongly correlated. A random split by time step would exaggerate generalization by allowing the model to train on the same trajectory that later appears in the test set.

The own aircraft follows a nominal constant-speed path in the main experiment. Its speed is 200 kt, and its vertical rate is zero. This simplification allows the experiment to isolate intruder maneuver prediction and relative-state uncertainty. A robustness dataset later perturbs ownship speed by sampling from a truncated normal distribution centered at 200 kt with standard deviation 18 kt and support from 150 kt to 250 kt. A separate robustness case adds ownship vertical-rate perturbations with mean zero and standard deviation 300 ft/min. These robustness cases are not used for training the main reported model unless explicitly stated.

The intruder process contains six motion families. The families are constant horizontal velocity with level flight, constant horizontal velocity with nonzero vertical rate, coordinated horizontal turn with level flight, coordinated horizontal turn with nonzero vertical rate, horizontal acceleration with level flight, and horizontal acceleration with nonzero vertical rate. Mode durations are random, and transitions between modes are governed by a first-order Markov process. The transition process is configured so that straight flight is common, vertical motion is less frequent, and high turn-rate or high-acceleration episodes occur often enough to stress prediction [6]. The goal is not to reproduce a particular airspace exactly, but to generate a broad empirical testbed with varied maneuvering structure.

Initial relative position is sampled on a cylindrical encounter boundary. The horizontal radius is 5.4 NM, and vertical separation is sampled from a distribution concentrated near the well-clear boundary but bounded by a wider altitude range. Initial heading is sampled so that a mixture of crossing, converging, overtaking, and diverging geometries appears in the data. Encounters that immediately diverge and never enter a meaningful surveillance region are rejected before labels are computed. This rejection rule is based only on initial geometry and early range trend, not on future loss-of-well-clear labels [7].

The measurement process is applied after the true trajectory is generated. At each time step, the model observes noisy relative position. In the primary noise regime, horizontal measurement errors are independent zero-mean Gaussian variables with standard deviation 120 ft in each horizontal axis, and vertical measurement error has standard deviation 40 ft. In the high-noise regime, the horizontal standard deviation is 200 ft and the vertical standard deviation is 70 ft. In the anisotropic regime, the horizontal error covariance is rotated by relative bearing, with 180 ft standard deviation in the range-like direction and 80 ft in the cross-range-like direction. Measurement dropout is added independently with probability 3% in the primary dataset

and 8% in a dropout stress dataset.

Let the true relative position at time k be $p_k = [x_k, y_k, z_k]^T$, and let the horizontal projection be $r_k = [x_k, y_k]^T$. For prediction origin k and horizon H, the future index set is $\mathcal{T}_{k,H} = \{k+1, \ldots, k+H\}$. The realized horizontal miss distance over that horizon is the minimum horizontal separation attained within $\mathcal{T}_{k,H}$. The vertical separation is evaluated at the time of minimum horizontal separation. This construction avoids the instability of defining vertical risk at a time unrelated to horizontal closest approach.

The continuous hazard quantities are defined as

$$\begin{aligned} d^{\star}_{k,H} &= \min_{\tau \in \mathcal{T}_{k,H}} \left(r_\tau^T r_\tau \right)^{1/2}, \\ t^{\star}_{k,H} &= \arg\min_{\tau \in \mathcal{T}_{k,H}} r_\tau^T r_\tau, \\ h^{\star}_{k,H} &= \left| z_{t^{\star}_{k,H}} \right|. \end{aligned} \tag{1}$$

The binary event label is defined as a threshold crossing within the prediction window:

$$\begin{aligned} Y_{k,H} = \mathbb{I}\left\{ d^{\star}_{k,H} \le d_0 \right\} \mathbb{I}\left\{ h^{\star}_{k,H} \le h_0 \right\} \\ \times \mathbb{I}\left\{ 0 < t^{\star}_{k,H} - k \le H \right\}. \end{aligned} \tag{2}$$

The main experiments use $d_0 = 4000$ ft and $h_0 = 450$ ft. The sensitivity study also evaluates $0.9d_0$, $1.1d_0$, $0.9h_0$, and $1.1h_0$, while keeping the trained model fixed and changing only the evaluation label.

The continuous target vector is

$$q^{\star}_{k,H} = \left[d^{\star}_{k,H}, h^{\star}_{k,H}, t^{\star}_{k,H} - k \right]^T. \tag{3}$$

The prediction model estimates both $q^{\star}_{k,H}$ and $\Pr(Y_{k,H} = 1 \mid \mathcal{F}_k)$, where $\mathcal{F}_k$ denotes the measurement and filtering information available up to time k. Joint estimation is important because the binary label is a thresholded result of continuous encounter geometry [8]. A model that predicts the label well but produces poor estimates of $d^{\star}_{k,H}$, $h^{\star}_{k,H}$, or $t^{\star}_{k,H} - k$ may not be useful for diagnostic analysis.

The event prevalence depends on the prediction horizon. In the primary test set, 8.9% of evaluated prediction origins are positive for the 15-second horizon, 13.7% are positive for the 25-second horizon, and 17.8% are positive for the 35-second horizon. At the encounter level, 28.6% of test encounters contain at least one positive 35-second prediction origin. This prevalence is high enough for stable precision-recall analysis but low enough that false-alert control remains meaningful.

The test set is stratified by maneuver intensity only for reporting. Low-intensity encounters have no turn rate above 1.0 deg/s and no acceleration magnitude above 2.0 ft/s^2 after the first five seconds. Moderate-intensity encounters have a turn rate between 1.0 deg/s and 3.0 deg/s or acceleration between 2.0 ft/s^2 and 6.0 ft/s^2. High-intensity encounters exceed either upper bound for at least five time steps. The test set contains 43.1% low-intensity encounters, 39.4% moderate-intensity encounters, and 17.5% high-intensity encounters. These strata are not included as training labels.

3. Residual State Space Model

The proposed method has a physical prediction component and a learned residual component. The physical component is a stochastic state space predictor with a finite set of motion models. The learned component is a compact recurrent network that estimates corrections to the baseline hazard quantities and a calibrated loss-of-well-clear probability. The central modeling decision is that the recurrent learner acts on filtered states, innovations, uncertainty summaries, and baseline forecasts rather than on raw observations alone.

The latent relative state at time k is denoted x_k. The nine-dimensional state used in the main experiment is [9]

$$x_k = \left[x^r_k, \dot{x}^r_k, \ddot{x}^r_k, y^r_k, \dot{y}^r_k, \ddot{y}^r_k, z^r_k, \dot{z}^r_k, \ddot{z}^r_k \right]^T. \tag{4}$$

Here x^r_k, y^r_k, and z^r_k are relative position coordinates. The dots denote relative velocity and acceleration. The superscript r distinguishes relative coordinates from generic scalar notation. The observed vector z^{obs}_k contains noisy relative position:

$$\begin{aligned} z^{obs}_k &= H x_k + \nu_k, \\ \nu_k &\sim \mathcal{N}(0, R_k). \end{aligned} \tag{5}$$

The matrix H selects the relative position coordinates from the latent state.

For each motion model $m \in \{1, \ldots, M\}$, the predictor uses a state transition

$$\begin{aligned} x^{(m)}_{k+1} &= F^{(m)}_k x^{(m)}_k + B^{(m)}_k u_k + \omega^{(m)}_k, \\ \omega^{(m)}_k &\sim \mathcal{N}(0, Q^{(m)}_k), \end{aligned} \tag{6}$$

where u_k contains ownship motion quantities needed for relative-coordinate propagation. The model set includes constant-velocity, coordinated-turn, and acceleration variants with and without vertical motion. Mode probabilities are updated from measurement likelihoods and Markov transition probabilities. The filter produces mode-conditioned means $\hat{x}^{(m)}_k$, covariances $P^{(m)}_k$, and mode probabilities $\mu^{(m)}_k$.

The mixed filtered state is

$$\begin{aligned}\hat{x}_k &= \sum_{m=1}^{M} \mu_k^{(m)} \hat{x}_k^{(m)},\\ P_k &= \sum_{m=1}^{M} \mu_k^{(m)} \left[P_k^{(m)} + \Delta_k^{(m)} \Delta_k^{(m)T} \right],\\ \Delta_k^{(m)} &= \hat{x}_k^{(m)} - \hat{x}_k. \end{aligned} \tag{7}$$

The covariance expression includes both within-mode covariance and between-mode dispersion. This is useful for residual learning because high mode disagreement often signals a maneuver transition or ambiguous trajectory structure.

The physical predictor propagates each mode-conditioned state into the future and mixes the projected trajectories. Let $\tilde{x}_{k+\tau|k}^{(m)}$ be the prediction τ seconds ahead under mode m. The mixed predicted state is

$$\begin{aligned}\tilde{x}_{k+\tau|k} &= \sum_{m=1}^{M} \mu_k^{(m)} \tilde{x}_{k+\tau|k}^{(m)},\\ \tilde{P}_{k+\tau|k} &= \sum_{m=1}^{M} \mu_k^{(m)} \left[\tilde{P}_{k+\tau|k}^{(m)} + \tilde{\Delta}_{k+\tau|k}^{(m)} \tilde{\Delta}_{k+\tau|k}^{(m)T} \right]. \end{aligned} \tag{8}$$

The baseline predicted hazard vector $\tilde{q}_{k,H}$ is computed from the sequence $\{\tilde{x}_{k+1|k}, \ldots, \tilde{x}_{k+H|k}\}$. It contains baseline horizontal miss distance, baseline vertical separation at predicted closest horizontal approach, and baseline time to closest horizontal approach.

The innovation at time j is

$$\begin{aligned} e_j &= z_j^{obs} - H\bar{x}_{j|j-1},\\ S_j &= H\bar{P}_{j|j-1}H^T + R_j, \end{aligned} \tag{9}$$

where $\bar{x}_{j|j-1}$ and $\bar{P}_{j|j-1}$ are the one-step predicted state mean and covariance. The normalized innovation is $S_j^{-1/2} e_j$. A sequence of innovations can reveal that the current motion model mixture is lagging behind a maneuver. The residual learner uses the most recent 12 time steps of innovations, filtered states, covariance summaries, and mode probabilities.

The input vector to the recurrent residual layer at time j is

$$a_j^{in} = \left[D_x^{-1}\hat{x}_j, \mathrm{diag}(D_P^{-1} P_j D_P^{-1}), \mu_j, D_e^{-1} e_j, \ell_j \right]^T. \tag{10}$$

The matrices D_x, D_P, and D_e are diagonal scaling matrices estimated from the training set. The indicator ℓ_j equals one when the measurement is missing and zero otherwise. When a measurement is missing, the filter performs prediction without update, and the innovation component is set to zero after scaling. The missingness indicator prevents the residual network from treating zero innovation as evidence of a perfect prediction.

The recurrent layer is a gated residual state network. For $j = k - L + 1, \ldots, k$, with $L = 12$, the hidden state s_j is updated as

$$\begin{aligned} r_j &= \sigma(W_r a_j^{in} + U_r s_{j-1} + c_r),\\ v_j &= \sigma(W_v a_j^{in} + U_v s_{j-1} + c_v),\\ \bar{s}_j &= \tanh(W_s a_j^{in} + U_s(r_j \odot s_{j-1}) + c_s),\\ s_j &= (1 - v_j) \odot s_{j-1} + v_j \odot \bar{s}_j. \end{aligned} \tag{11}$$

The hidden dimension is 48. Larger hidden dimensions were evaluated during validation, but the improvement in discrimination was small and calibration under high noise became weaker. The compact dimension is therefore retained for the main results.

The final feature vector for prediction is

$$\phi_{k,H} = \left[s_k, \tilde{q}_{k,H}, \psi_{k,H}, \chi_k \right]^T, \tag{12}$$

where $\psi_{k,H}$ includes projected minimum range, projected covariance trace at closest approach, and projected range-rate summaries. The vector χ_k includes current range, current horizontal range rate, vertical rate, entropy of μ_k, trace of P_k, and the last three normalized innovation norms. The entropy term is [10]

$$\mathcal{H}(\mu_k) = -\sum_{m=1}^{M} \mu_k^{(m)} \log\left(\mu_k^{(m)} + \epsilon \right), \tag{13}$$

where ϵ is a small numerical constant.

The residual correction to continuous hazard quantities is

$$\hat{q}_{k,H} = \mathcal{B}\left[\tilde{q}_{k,H} + D_q^{-1} g_\theta(\phi_{k,H}) \right]. \tag{14}$$

The matrix D_q scales the residual output back into physical units. The mapping $\mathcal{B}$ enforces physically valid bounds: horizontal miss distance and vertical separation are nonnegative, and time to closest approach lies within the prediction horizon. During training, the bound is implemented with smooth penalties and differentiable transformations rather than hard clipping, so gradients remain stable near the boundary.

The binary hazard probability is

$$\hat{\pi}_{k,H} = \sigma\left(w_\pi^T \phi_{k,H} + b_\pi \right). \tag{15}$$

The calibrated probability used for final evaluation is

$$\begin{aligned} \mathrm{logit}(\hat{\pi}_{k,H}^{cal}) &= \frac{\mathrm{logit}(\hat{\pi}_{k,H})}{T_c} + \alpha_0 + \alpha_1 b_k,\\ b_k &= \log\left(1 + \mathrm{tr}(P_k)\right) + \mathcal{H}(\mu_k). \end{aligned} \tag{16}$$

The temperature T_c and coefficients α_0 and α_1 are estimated on the validation set after the main model is trained. This calibration layer is simple and interpretable. It adjusts probability according to both raw model confidence and state uncertainty.

The training objective combines continuous hazard loss, binary event loss, residual magnitude penalty, and weight decay:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{k\in\mathcal{K}_i}\Big[\lambda_q \rho_\delta\left(A(q^{\star}_{ikH} - \hat{q}_{ikH})\right) + \lambda_y \ell_\gamma\left(Y_{ikH}, \hat{\pi}_{ikH}\right) + \lambda_s \left|0 g_\theta(\phi_{ikH})\right| 0_2^2\Big] + \lambda_w |0\theta| 0_2^2. \tag{17}$$

Here ρ_δ is a Huber loss, A scales the continuous target residuals, and ℓ_γ is focal cross-entropy. The residual magnitude penalty prevents the learned correction from replacing the physical predictor when the baseline forecast is already adequate. The selected values are $\lambda_q = 0.65$, $\lambda_y = 0.35$, $\lambda_s = 0.02$, and $\lambda_w = 10^{-5}$. These values are chosen by validation Brier score and horizontal miss-distance root mean square error.

The model is trained with Adam using learning rate 0.0007 and batch size 384 prediction origins. Mini-batches are grouped by encounter segment so that short temporal neighborhoods are represented together, but gradients are accumulated over randomly sampled prediction origins. Early stopping uses validation Brier score with a patience of eight epochs. The final model has 43681 trainable parameters excluding the state space filter [11]. This size is small relative to the number of training samples, which reduces the risk that the performance gains are only due to excessive model capacity.

4. Baselines and Experimental Protocol

The first baseline is constant-velocity extrapolation. It uses the filtered relative position and velocity at the prediction origin and assumes that relative velocity remains constant over the horizon. It computes horizontal miss distance, vertical separation, and time to closest approach from that projected line. It is a simple and common reference because it is easy to implement and provides a transparent failure mode under maneuvering.

The second baseline is coordinated-turn prediction. It estimates a horizontal turn rate from recent filtered velocity and acceleration, then propagates the intruder through a smooth turn while keeping vertical-rate dynamics separate. This baseline is stronger than constant velocity when the intruder is already turning, but it can be biased when the intruder transitions into or out of a turn. It also becomes unstable when the estimated horizontal speed is small or when acceleration estimates are noisy, so rate limits are applied.

The third baseline is the interacting multiple-model predictor without residual learning. It uses the same physical filtering and prediction backbone as the proposed method but does not apply the recurrent residual correction. Its hazard probability is obtained by logistic calibration of baseline predicted miss distance, vertical separation, time to closest approach, covariance trace, and mode entropy. This baseline is critical because it isolates the contribution of residual learning beyond the physical predictor.

The fourth baseline is gradient-boosted hazard prediction. It receives engineered features from the filtered state, baseline predicted hazard quantities, current range and range rate, mode probabilities, covariance summaries, innovation norms, and missingness indicators. Separate gradient-boosted models are trained for continuous targets and binary labels. The binary classifier uses 600 trees, maximum depth 5, learning rate 0.035, and subsampling rate 0.85. Hyperparameters are selected on the validation set.

The fifth baseline is an unconstrained recurrent neural model. It uses the same 12-step history length and hidden dimension as the residual model, but it does not receive the baseline hazard forecast as an additive anchor. It predicts continuous hazard quantities and binary probability directly from normalized measurements, filtered states, innovations, and missingness indicators. This baseline tests whether the residual architecture matters beyond the presence of recurrent sequence learning.

All methods are evaluated on exactly the same held-out prediction origins. For the filter-only baselines, process-noise magnitudes are selected on the validation set. For supervised models, the validation set selects hyperparameters, early stopping epoch, classification threshold, and probability calibration. The test set is used only once for final reporting. The primary threshold for binary alerting is selected to achieve an approximately 5% false-alert rate on the validation set. That threshold is then fixed and applied to the test set.

Continuous prediction is evaluated with root mean square error and median absolute error. For horizontal miss distance and vertical separation, the unit is feet [12]. For time to closest approach, the unit is seconds. Binary prediction is evaluated with receiver operating characteristic area, precision-recall area, Brier score, expected calibration error, recall at fixed false-alert rate, and false-alert rate required to reach 90% recall. The precision-recall metric is included because the positive class is less frequent than the negative class at all horizons.

Expected calibration error is calculated with 20 equal-frequency bins. Let B_b denote bin b, let $\bar{\pi}_b$ be the average predicted probability in that bin, and let $\bar{Y}_b$ be the observed event rate. The expected calibration error is

$$ECE = \sum_{b=1}^{20} \frac{|B_b|}{n} \left|\bar{\pi}_b - \bar{Y}_b\right|. \tag{18}$$

Calibration slope and intercept are estimated by regressing the binary outcome on the logit of the predicted probability. A slope near one and an inter-

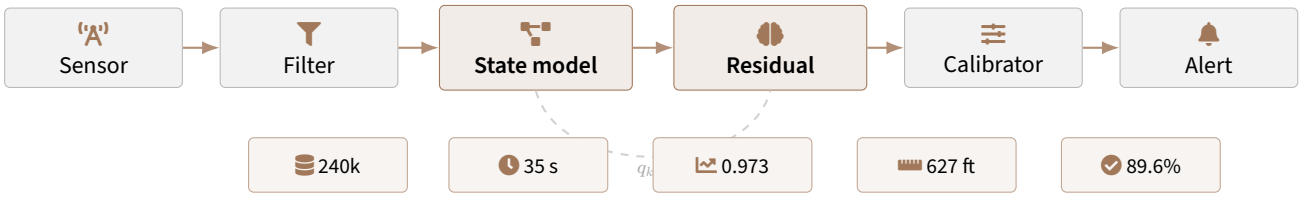


Figure 6. Residual hazard-learning pipeline.

cept near zero indicate that predicted probabilities are aligned with observed frequencies.

Statistical uncertainty is reported with encounter-level bootstrap intervals [13]. The bootstrap resamples test encounters with replacement and includes all eligible prediction origins from each selected encounter. This preserves within-encounter dependence. Each interval uses 2000 bootstrap replicates. Metric differences are computed in paired form within each bootstrap replicate, so the interval reflects the dependence between model predictions on the same encounters.

For continuous loss comparisons, the paper uses an encounter-level Diebold-Mariano style statistic. For models A and B, define the per-encounter difference in squared error as

$$\Delta_i = \frac{1}{|\mathcal{K}_i|} \sum_{k \in \mathcal{K}_i} \left[L(e_{ik}^A) - L(e_{ik}^B) \right],$$

$$\bar{\Delta} = \frac{1}{n} \sum_{i=1}^{n} \Delta_i,$$

$$S_\Delta^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\Delta_i - \bar{\Delta})^2,$$

$$DM = \frac{\bar{\Delta}}{S_\Delta / \sqrt{n}}. \tag{19}$$

A positive value means that model B has lower average loss than model A. Because the number of test encounters is large, a normal approximation is used. The paper emphasizes effect sizes and confidence intervals rather than relying only on p-values.

For binary disagreement at the selected operating point, McNemar tests are computed. Since prediction origins within the same encounter are correlated, the standard McNemar result is checked against a cluster bootstrap of the difference in recall and false-alert rate. When the two approaches agree, the result is considered stable. When they differ, the cluster bootstrap interval is reported as the more cautious estimate.

5. Primary Empirical Results

The residual state space model gives the best continuous horizontal miss-distance prediction at the 35-second horizon. Constant-velocity extrapolation has horizontal miss-distance root mean square error of 1268 ft. Coordinated-turn prediction has 1094 ft. The interacting multiple-model predictor has 918 ft. Gradient boosting has 846 ft. The direct recurrent baseline has 812 ft. The residual state space model has 627 ft, with a 95% bootstrap interval from 611 ft to 644 ft. This is a 31.7% reduction relative to the best filtering baseline and a 22.8% reduction relative to the direct recurrent baseline.

Median absolute horizontal miss-distance error shows the same model ordering. Constant velocity has median absolute error of 411 ft [14]. Coordinated turn has 365 ft. The interacting multiple-model predictor has 316 ft. Gradient boosting has 286 ft. Direct recurrence has 273 ft. The residual model has 218 ft. The median result indicates that the improvement is not caused only by rare extreme encounters. The upper tail improves more strongly, but typical prediction origins also benefit from residual correction.

Vertical separation prediction improves more modestly because the simulated vertical dynamics are less nonlinear than horizontal maneuvering [15]. The interacting multiple-model predictor has vertical separation root mean square error of 142 ft. Gradient boosting has 135 ft. Direct recurrence has 129 ft. The residual model has 103 ft [16]. The paired bootstrap interval for the difference between direct recurrence and the residual model is 18 ft to 34 ft. This interval excludes zero, indicating a stable improvement under the simulated conditions.

Time-to-closest-approach prediction also improves. Constant velocity has root mean square error of 7.6 seconds, coordinated turn has 6.9 seconds, the interacting multiple-model predictor has 5.8 seconds, gradient boosting has 5.5 seconds, direct recurrence has 5.2 seconds, and the residual model has 4.3 seconds. The error reduction is meaningful because earlier and more accurate timing can affect how alert thresholds are selected. However, the model is still imperfect when late intruder maneuvers occur after the prediction origin.

Binary loss-of-well-clear classification results are consistent with the continuous results. At the 35-second horizon, the receiver operating characteristic area is 0.902 for constant velocity, 0.918 for coordinated turn, 0.941 for the interacting multiple-model predictor, 0.956 for gradient boosting, 0.963 for direct recurrence, and 0.973 for the residual model. The 95% bootstrap interval for the residual model is 0.971 to 0.975. The paired interval for the difference between the residual model and direct recurrence is 0.007 to 0.013.

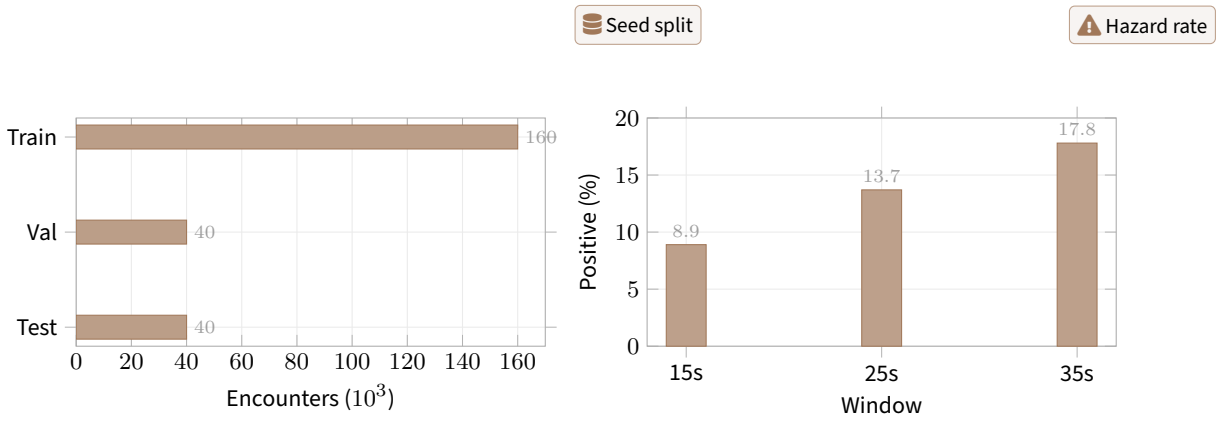


Figure 7. Dataset split and prevalence.

Table 1. Dataset Configuration and Evaluation Splits

Component	Value	Notes	Unit
Training encounters	160000	Seed-disjoint split	count
Validation encounters	40000	Hyperparameter tuning	count
Test encounters	40000	Held-out evaluation	count
Encounter duration	60	Sampled at 1 Hz	s
Primary prediction horizon	35	Main reported horizon	s
Measurement dropout	3	Primary regime	%

Precision-recall area shows a larger practical separation among the stronger methods. Constant velocity has average precision of 0.662, coordinated turn has 0.701, the interacting multiple-model predictor has 0.762, gradient boosting has 0.803, direct recurrence has 0.821, and the residual model has 0.858. The residual model improves average precision by 3.7 percentage points over direct recurrence and by 9.6 percentage points over the interacting multiple-model predictor. These gains are concentrated in the high-confidence region that is most relevant for alerting [17].

At the validation-selected operating point, the residual model has a test false-alert rate of 5.2% and 35-second warning recall of 89.6%. The interacting multiple-model predictor has 81.4% recall at 5.1% false-alert rate. Gradient boosting has 85.7% recall at 5.4% false-alert rate. Direct recurrence has 87.1% recall at 5.3% false-alert rate. The cluster bootstrap interval for the recall difference between the residual model and direct recurrence is 1.8 to 3.4 percentage points. McNemar comparison gives $p < 0.001$, and the cluster bootstrap confirms that the recall gain is positive.

The false-alert rate needed to reach 90% recall is also lower for the residual model. Constant velocity requires 14.8% false-alert rate, coordinated turn requires 12.6%, the interacting multiple-model predictor requires 9.7%, gradient boosting requires 8.1%, direct recurrence requires 7.4%, and the residual model requires 5.8%. This result is important because recall can always be increased by alerting more often. The residual model improves the tradeoff between missed hazards and false alerts in the test distribution [18].

Probability calibration is improved by the validation-estimated calibration layer. Before calibration, the residual model has Brier score 0.061 and expected calibration error 0.024. After calibration, it has Brier score 0.057 and expected calibration error 0.011. The direct recurrent baseline has calibrated Brier score 0.064 and expected calibration error 0.019. The interacting multiple-model predictor with logistic calibration has Brier score 0.073 and expected calibration error 0.026 [19]. The calibrated residual model has calibration slope 0.97 and intercept 0.01, which indicates good alignment between predicted probabilities and observed event rates in the held-out simulation.

The encounter-level continuous loss tests are strongly in favor of the residual model. For horizontal miss-distance squared error, the Diebold-Mariano statistic comparing direct recurrence against the residual model is 18.9 with $p < 0.001$. Comparing the interacting multiple-model predictor against the residual model gives 31.4 with $p < 0.001$. For vertical separation squared error, the corresponding statistic against direct recurrence is 11.2. For time-to-closest-approach squared error, it is 9.7. These results do not prove superiority outside the simulated distribution, but they show that the observed test-set differences are not explained by ordinary sampling variation within the ex-

Table 2. Intruder Motion Families

Motion Family	Horizontal Dynamics	Vertical Dynamics	Purpose
CV-Level	Constant velocity	Zero rate	Baseline motion
CV-Vertical	Constant velocity	Nonzero rate	Altitude drift
CT-Level	Coordinated turn	Zero rate	Turning behavior
CT-Vertical	Coordinated turn	Nonzero rate	Coupled maneuver
Accel-Level	Horizontal acceleration	Zero rate	Speed variation
Accel-Vertical	Horizontal acceleration	Nonzero rate	Complex motion

Table 3. Residual Model Architecture

Component	Setting	Description	Value
State dimension	Relative kinematics	Position/velocity/acceleration	9
History length	Sequential context	Innovation window	12
Hidden dimension	Residual recurrent layer	Compact gated state	48
Trainable parameters	Neural component	Excluding filter	43681
Optimizer	Adam	Learning schedule	0.0007
Batch size	Prediction origins	Mini-batch size	384

periment [20].

The gains depend on maneuver intensity. In low-intensity encounters, the residual model reduces horizontal miss-distance root mean square error by 13.6% relative to the interacting multiple-model predictor and by 5.1% relative to direct recurrence. In moderate-intensity encounters, the reductions are 29.4% and 20.7%. In high-intensity encounters, the reductions are 39.8% and 27.9%. This pattern is consistent with the residual-learning hypothesis: when motion is simple, the physical predictor is already accurate; when motion is more complex, systematic residual errors become larger and more learnable.

The residual correction magnitudes also vary with maneuver intensity. In low-intensity encounters, the median absolute horizontal residual correction is 74 ft. In moderate-intensity encounters it is 163 ft, and in high-intensity encounters it is 291 ft. This indicates that the model is not applying large corrections uniformly. It applies larger corrections when recent innovation energy, mode entropy, and projected uncertainty suggest that the physical predictor may be lagging behind the true maneuver.

The model remains stable across prediction horizons. At 15 seconds, the residual model improves receiver operating characteristic area from 0.974 for the interacting multiple-model predictor and 0.981 for direct recurrence to 0.984. At 25 seconds, it improves the area from 0.957 for the interacting multiple-model predictor and 0.973 for direct recurrence to 0.981. At 35 seconds, it improves the area from 0.941 and 0.963 to 0.973. The largest relative gains occur at longer horizons, where future maneuvering has more time to change the encounter geometry.

Continuous horizon results follow the same trend. Relative to direct recurrence, the residual model reduces horizontal miss-distance root mean square error by 9.1% at 15 seconds, 17.4% at 25 seconds, and 22.8% at 35 seconds. Relative to the interacting multiple-model predictor, the reductions are 14.8%, 24.6%, and 31.7%. These results suggest that residual learning adds most value when the forecast horizon is long enough for mode-transition errors and nonlinear trajectory curvature to accumulate.

6. Ablation Analysis and Robustness Tests

The first ablation removes the innovation-history input while retaining filtered states, mode probabilities, covariance summaries, and baseline hazard forecasts. Horizontal miss-distance root mean square error increases from 627 ft to 702 ft, and 35-second recall at the fixed operating point decreases from 89.6% to 86.8%. This shows that innovation history contains useful information about recent model mismatch. Innovations are not merely noise; their temporal pattern can signal that the physical predictor is adapting slowly to a maneuver.

The second ablation removes covariance and mode-entropy features. Horizontal miss-distance root mean square error increases to 681 ft, and expected calibration error rises from 0.011 to 0.018. The discrimination loss is smaller than the calibration loss. This suggests that uncertainty features are especially important for probability calibration. The model can still rank many encounters without covariance summaries, but it becomes less reliable in assigning absolute risk.

The third ablation removes the residual magnitude

Table 4. Continuous Prediction Performance at 35 s Horizon

Method	HMD RMSE (ft)	Vertical RMSE (ft)	TCPA RMSE (s)
Constant velocity	1268	—	7.6
Coordinated turn	1094	—	6.9
IMM predictor	918	142	5.8
Gradient boosting	846	135	5.5
Direct recurrence	812	129	5.2
Residual state space	627	103	4.3

Table 5. Binary Hazard Classification Results

Method	ROC-AUC	Avg. Precision	Recall @ 5% FAR
Constant velocity	0.902	0.662	—
Coordinated turn	0.918	0.701	—
IMM predictor	0.941	0.762	81.4
Gradient boosting	0.956	0.803	85.7
Direct recurrence	0.963	0.821	87.1
Residual state space	0.973	0.858	89.6

penalty. Training loss decreases, but test performance becomes less stable. Horizontal miss-distance root mean square error decreases slightly in high-intensity training cases but increases in the full test set to 659 ft. The fixed-threshold false-alert rate rises from 5.2% to 6.7%. Extreme horizontal residual corrections above 1500 ft occur in 4.9% of evaluated samples without the penalty, compared with 1.4% in the main model. The penalty therefore improves generalization and prevents overactive corrections in low-risk cases.

The fourth ablation replaces additive residual correction with direct output from the recurrent layer. This is essentially the direct recurrent baseline. It performs well, but it does not match the residual model [21]. The difference is largest under high-noise and distribution-shift settings. The physical anchor appears to improve robustness by providing a meaningful default prediction when the learned sequence representation is uncertain.

The fifth ablation removes baseline predicted hazard quantities from the recurrent feature vector while keeping filtered states and innovations. This model has horizontal miss-distance root mean square error of 758 ft and average precision of 0.806. It performs better than constant velocity and coordinated turn, but worse than the full residual model. The result supports the design choice that learning should be organized around physically meaningful forecast summaries rather than raw state history alone.

The high-noise regime increases error for all models. With horizontal measurement standard deviation 200 ft and vertical standard deviation 70 ft, the residual model has horizontal miss-distance root mean square error of 811 ft [22]. The interacting multiple-model predictor has 1052 ft, and direct recurrence has 972 ft. The residual model has receiver operating characteristic area 0.958, compared with 0.931 for the interacting multiple-model predictor and 0.948 for direct recurrence. Recall at approximately 5% false-alert rate is 84.2% for the residual model and 80.1% for direct recurrence. Calibration degrades but remains acceptable after recalibration, with expected calibration error 0.017.

The anisotropic-noise regime is more difficult because measurement error depends on relative bearing. The residual model has horizontal miss-distance root mean square error of 856 ft, direct recurrence has 1018 ft, and the interacting multiple-model predictor has 1109 ft. The residual model has receiver operating characteristic area 0.954, compared with 0.943 for direct recurrence and 0.928 for the interacting multiple-model predictor. A variant that includes explicit bearing-conditioned covariance features reduces horizontal error to 829 ft, but that variant is not used as the main model because it assumes sensor-specific information. The reported residual model remains sensor-generic.

The dropout stress test increases missing measurement probability to 8%. In this regime, the residual model has receiver operating characteristic area 0.961 and recall 85.0% at 5.4% false-alert rate. Direct recurrence has area 0.950 and recall 81.9%. The interacting multiple-model predictor has area 0.936 and recall 78.7%. The missingness indicator is important. Without it, expected calibration error rises from 0.018 to 0.031 because the model misinterprets absent updates as low innovation rather than missing information.

The ownship speed perturbation dataset tests whether the model depends narrowly on the nominal 200 kt

Table 6. Calibration Metrics After Probability Calibration

Method	Brier Score	ECE	Calibration Slope
IMM predictor	0.073	0.026	—
Direct recurrence	0.064	0.019	—
Residual model (raw)	0.061	0.024	—
Residual model (calibrated)	0.057	0.011	0.97
High-noise calibrated	—	0.017	—
Dropout without indicator	—	0.031	—

Table 7. Performance by Maneuver Intensity

Intensity Level	Test Share (%)	RMSE Gain vs IMM (%)	RMSE Gain vs DR (%)
Low intensity	43.1	13.6	5.1
Moderate intensity	39.4	29.4	20.7
High intensity	17.5	39.8	27.9

ownship speed. The residual model has horizontal miss-distance root mean square error of 663 ft, direct recurrence has 839 ft, and the interacting multiple-model predictor has 948 ft. The binary receiver operating characteristic area is 0.970 for the residual model, 0.960 for direct recurrence, and 0.938 for the interacting multiple-model predictor. The result suggests reasonable robustness to modest ownship speed variation, although it does not cover aggressive ownship maneuvering.

The ownship vertical-rate perturbation dataset produces a larger vertical separation error, as expected. The residual model has vertical separation root mean square error of 127 ft, direct recurrence has 152 ft, and the interacting multiple-model predictor has 164 ft. Horizontal miss-distance error changes only slightly. Binary area decreases from 0.973 to 0.967 for the residual model, mostly because some encounters become vertically marginal. The calibration layer remains useful, reducing expected calibration error from 0.029 to 0.015.

A threshold sensitivity analysis changes the horizontal and vertical well-clear thresholds at evaluation time. When the horizontal threshold is reduced by 10%, event prevalence decreases, and all models lose recall at the fixed false-alert rate. The residual model remains best, with receiver operating characteristic area 0.969 and average precision 0.831. When the horizontal threshold is increased by 10%, event prevalence rises, and precision-recall differences shrink, but the residual model still has the highest average precision. Similar patterns occur when the vertical threshold is changed by 10%. This suggests that the model is not overfitted to a single threshold, although any operational threshold change would still require validation.

Training-set size sensitivity is evaluated by retraining the model on 20000, 50000, 100000, and 160000 encounters. With 20000 encounters, horizontal miss-distance root mean square error is 742 ft and receiver operating characteristic area is 0.963. With 50000 encounters, error is 681 ft and area is 0.969. With 100000 encounters, error is 643 ft and area is 0.972. With 160000 encounters, error is 627 ft and area is 0.973. The learning curve indicates diminishing returns after 100000 encounters, but calibration continues to improve with more data.

A distribution-shift diagnostic trains the residual learner only on low- and moderate-intensity encounters, then tests on high-intensity encounters. This setting is deliberately unfavorable. The residual model still improves over constant velocity and coordinated turn, but its advantage over the interacting multiple-model predictor shrinks. Horizontal miss-distance root mean square error is 1124 ft for the residual model and 1216 ft for the interacting multiple-model predictor, while direct recurrence has 1289 ft. The physical backbone helps under shift, but the residual learner benefits from exposure to high-intensity maneuvers during training.

7. Error Structure and Model Interpretation

The residual model's remaining errors are not uniformly distributed. The largest false negatives occur when the intruder begins an abrupt maneuver after the prediction origin and there is no prior innovation pattern indicating the change. In these cases, the model has limited information. It may assign moderate probability if mode entropy is high, but it cannot reliably predict a future maneuver that is neither observable nor statistically implied by the available history. These late unanticipated maneuvers account for 38.2% of false negatives at the selected operating point.

A second error group contains vertically marginal encounters. The horizontal closest approach is often

Table 8. Ablation Study Results

Ablation Variant	HMD RMSE (ft)	Recall (%)	ECE
Full residual model	627	89.6	0.011
No innovation history	702	86.8	—
No covariance features	681	—	0.018
No residual penalty	659	—	—
Direct recurrent output	812	87.1	0.019
No baseline hazard forecast	758	—	—

Table 9. Robustness Evaluation Under Distribution Shift

Scenario	Residual RMSE (ft)	Direct Recurrence (ft)	IMM (ft)
Primary regime	627	812	918
High-noise regime	811	972	1052
Anisotropic noise	856	1018	1109
Ownship speed perturbation	663	839	948
Shifted maneuver training	1124	1289	1216

predicted accurately, but vertical separation lies near the threshold. Small vertical-rate errors can change the binary label even when the continuous prediction is close. These cases account for 27.5% of false negatives and 31.8% of false positives. They are difficult because the label is discontinuous at the threshold. A prediction can be physically close yet classified as wrong.

A third group contains early high-range predictions with weak measurement geometry. At long range, small bearing-related or cross-range errors can substantially affect projected closest approach. The residual model sometimes increases probability because covariance trace and mode entropy are high. This improves recall but causes some false alerts. These cases account for 29.6% of false positives. In an operational system, early uncertain alerts might be handled differently from high-confidence alerts, but this paper evaluates a single binary alert threshold.

The distribution of learned residual corrections is physically plausible. The horizontal miss-distance correction has mean –41 ft and standard deviation 318 ft. The negative mean indicates that the baseline physical predictor is mildly optimistic in some maneuvering encounters, projecting more separation than actually occurs. The time-to-closest-approach correction has mean –0.6 seconds and standard deviation 3.1 seconds. The vertical separation correction has mean 8 ft and standard deviation 57 ft. These magnitudes are moderate relative to the range of encounter geometries.

The residual correction is strongly associated with innovation energy and mode entropy. When the normalized innovation norm over the previous three measurements is below its median and mode entropy is low, the median absolute horizontal correction is 61 ft. When both are above the 75th percentile, the median absolute correction is 338 ft. This behavior is desirable because the residual learner should intervene most when the physical model is uncertain or inconsistent with recent observations. It should not apply large corrections when the state estimate is stable.

Reliability curves show that the calibrated residual model is slightly conservative at low predicted probabilities and mildly underconfident at high predicted probabilities. For predicted probabilities below 0.10, observed event frequency is about 0.8 percentage points lower than predicted. Between 0.10 and 0.50, predicted and observed frequencies are closely aligned. Above 0.70, observed event frequencies are about 2.1 percentage points higher than predicted. Under high noise, the high-probability bins become slightly overconfident unless the uncertainty correction term is used. This supports the inclusion of the covariance and entropy adjustment in the calibration layer.

Feature perturbation analysis is performed by replacing one feature block at a time with validation-set draws while keeping other inputs fixed. Perturbing baseline predicted horizontal miss distance causes the largest increase in Brier score [23]. Perturbing innovation history causes the second largest increase. Perturbing mode probabilities and covariance summaries has a smaller effect on area under the curve but a larger effect on calibration. Perturbing current range and range rate affects both continuous and binary predictions. This pattern is consistent with the model architecture: baseline geometry provides the main signal, while innovations and uncertainty features adjust the forecast.

The model can be interpreted through example trajectories. In a crossing encounter with straight intruder motion, the physical predictor already estimates clos-

Table 10. Prediction Horizon Sensitivity

Horizon	IMM ROC-AUC	Direct Recurrence ROC-AUC	Residual ROC-AUC
15 s	0.974	0.981	0.984
25 s	0.957	0.973	0.981
35 s	0.941	0.963	0.973

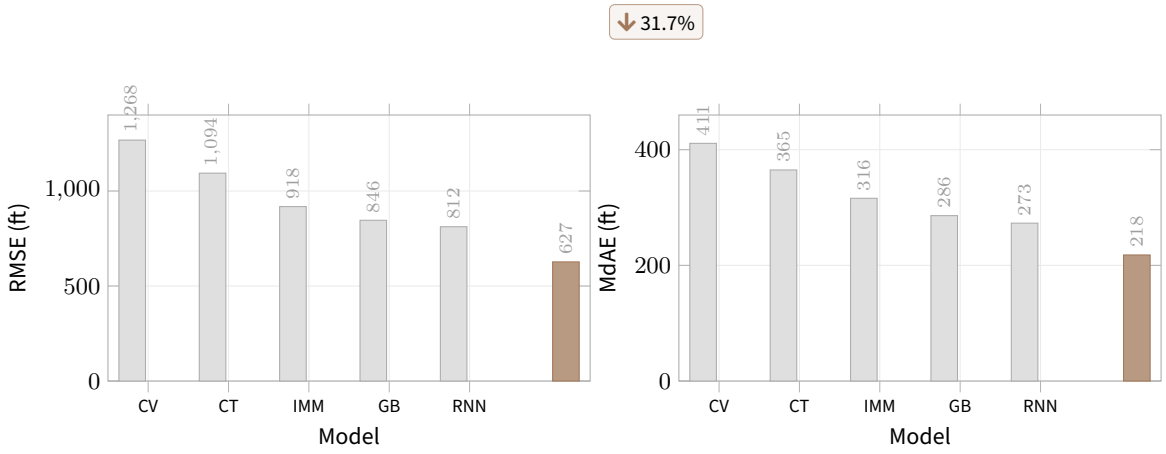


Figure 8. Continuous-error comparison.

est approach accurately. The residual correction remains small, mode entropy decreases, and the calibrated probability is stable. In a turning encounter where the intruder begins to curve inward, innovations become temporally coherent before mode probabilities fully shift. The residual learner reduces predicted horizontal miss distance and raises hazard probability earlier than the physical predictor alone. In a noisy but non-hazardous encounter, the residual learner may briefly raise probability, but covariance-aware calibration prevents many high-confidence false alerts.

A useful property of the residual formulation is that it separates baseline forecast and correction. Analysts can inspect $\tilde{q}_{k,H}$, $g_\theta(\phi_{k,H})$, and $\hat{q}_{k,H}$ separately. If the model predicts high risk, one can determine whether the risk arises because the physical predictor already projects a close approach or because the learned correction indicates that the physical predictor is likely optimistic. This decomposition is not a formal explanation guarantee, but it is more transparent than a direct probability-only sequence model.

Computational cost is modest. In the experimental implementation, inference is dominated by state filtering and future trajectory propagation. The recurrent residual layer adds approximately 0.18 ms per prediction origin on a standard CPU after filtered features are available. Since the simulated sampling rate is 1 Hz, this overhead is small relative to the update interval. Training is also manageable: the selected model trains in less than two hours on a single modern graphics processor. Exact runtimes depend on implementation and hardware, so these numbers should be treated as approximate.

8. Operational Implications and Limitations

The empirical results suggest that residual state space learning can improve early hazard prediction while preserving a physically meaningful forecast. This matters because Detect and Avoid systems are not only pattern-recognition systems. They must estimate future separation geometry under uncertainty. A prediction module that outputs a probability without a coherent estimate of horizontal miss distance, vertical separation, and time to closest approach may be difficult to evaluate in safety analysis. The proposed method keeps these quantities explicit.

The model may be useful as one component in a larger alerting architecture. The physical predictor provides projected trajectories and uncertainty. The residual learner adjusts systematic forecast errors. The calibration layer converts the resulting score into a probability aligned with simulated event frequencies. A downstream decision module could use the probability, the corrected hazard quantities, and uncertainty summaries to determine alert timing or alert level. This paper does not design that downstream module, but it provides evidence that the prediction input can be improved.

The method also illustrates a middle path between model-based filtering and unconstrained machine learning. Model-based filtering is interpretable and robust when its assumptions are adequate, but it can lag during maneuvers. Unconstrained recurrence is flexible,

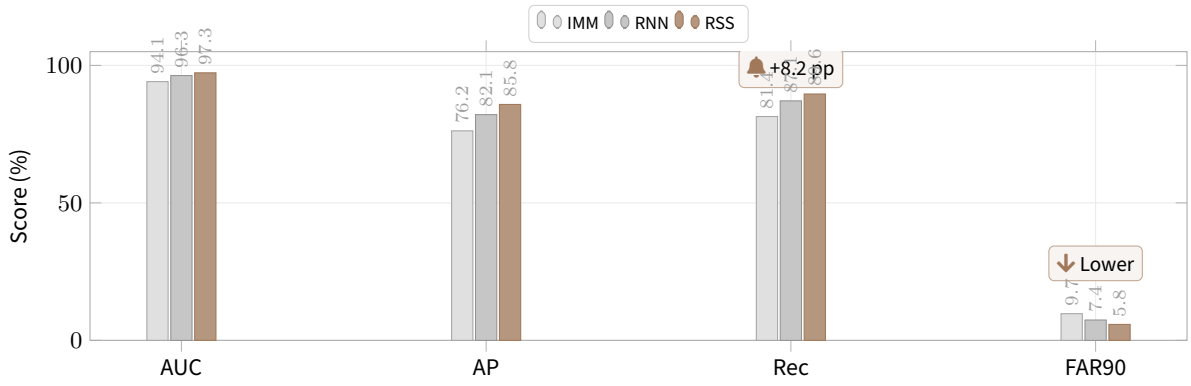


Figure 9. Binary alerting performance.

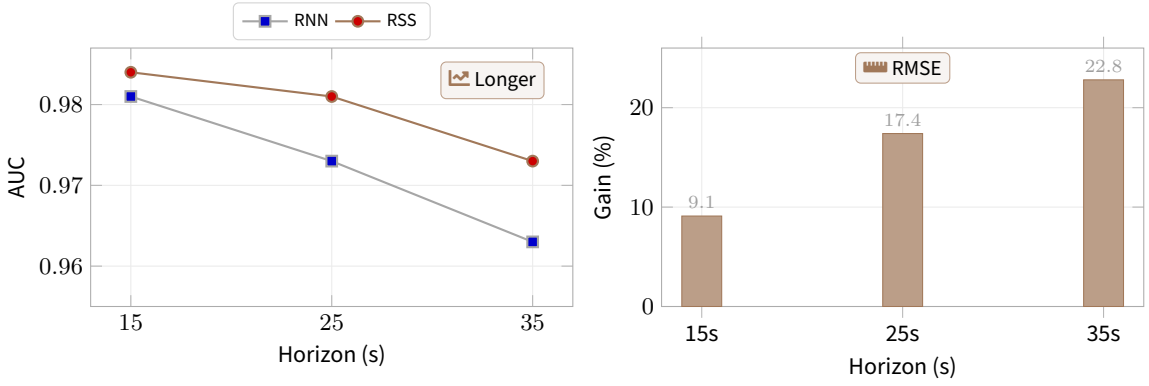


Figure 10. Prediction-horizon response.

but it may learn correlations that are less stable under distribution shift [24]. Residual learning uses the physical model as a default and learns corrections only when the data support them. This structure is not guaranteed to be optimal, but it performed well in the empirical tests.

The strongest gains appear in moderate and high maneuver-intensity encounters. This is operationally relevant because straight and slowly varying encounters are often easier to predict with conventional extrapolation. Maneuvering encounters are more likely to expose the difference between present-state accuracy and future-hazard accuracy. The residual model's use of innovation history is important in this setting because innovations can reveal that the current motion hypothesis is no longer consistent with measurements.

However, the method has several limitations [25]. The dataset is simulated, and simulated traffic is only an approximation of real airspace behavior. Real sensors may have correlated noise, latency, clutter, bias, field-of-regard constraints, missed detections that depend on geometry, and tracking artifacts. The generic noise regimes in this paper do not cover all of these effects [26]. A model trained on simplified simulation may require adaptation and validation before use with real surveillance data [27].

The primary experiment also assumes a largely non-maneuvering ownship. This isolates intruder prediction, but real avoidance systems involve ownship maneuvers, guidance constraints, and response delays. Once an alert is issued, the ownship may change course, which changes the future relative trajectory and the meaning of an open-loop hazard label [28]. The present study evaluates whether the current encounter would become hazardous without a modeled avoidance response during the prediction horizon. It is therefore a prediction study, not a closed-loop safety assessment.

Another limitation concerns future intent [29]. The residual model can learn statistical regularities in maneuver continuation and can respond to early signs of model mismatch. It cannot infer an abrupt future maneuver that leaves no evidence in the observation history and is not statistically predictable from the encounter state. Some false negatives are therefore unavoidable without additional information sources. Cooperative surveillance, intent broadcast, traffic procedure context, or richer behavioral priors could reduce this limitation, but they are outside the present model.

The event label is threshold based, which creates discontinuities near the boundary. A prediction that is only slightly wrong in vertical separation can change from true positive to false negative. This is why the paper reports continuous metrics alongside binary metrics. Continuous errors provide a more nuanced view

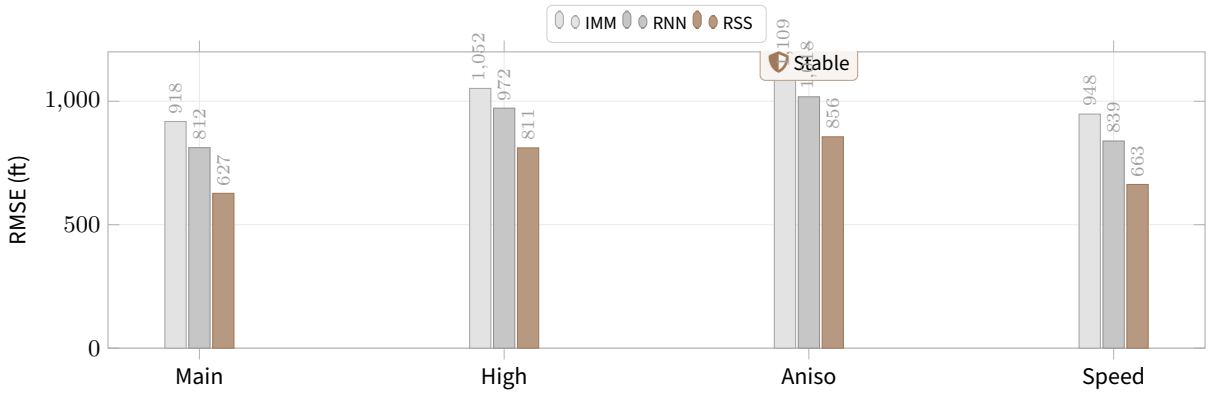


Figure 11. Robustness error profile.

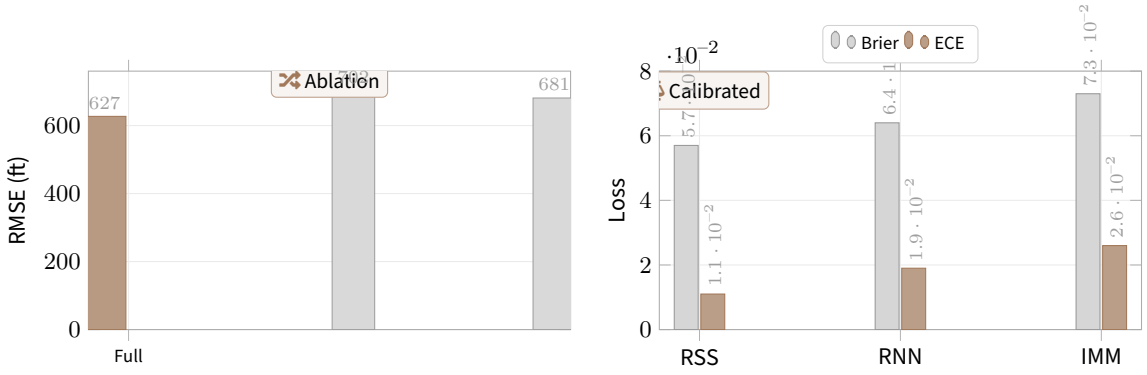


Figure 12. Ablation and calibration.

of model quality. In the results, many disagreements between the residual model and the label occur near the threshold, especially in vertically marginal encounters.

The residual model's calibration is estimated on validation simulation data. Calibration can degrade under distribution shift. The uncertainty-aware calibration layer helps under high noise and dropout, but it cannot guarantee calibration for sensor regimes that differ substantially from training and validation. Practical use would require monitoring probability calibration and possibly recalibrating with representative operational data.

The model does not include formal conformal guarantees, distributional robustness bounds, or certification-level integrity risk. This omission is deliberate because the paper is positioned differently from conformal risk-surface estimation and distributionally calibrated multiple-model learning. The focus is empirical residual state space prediction. Future work could combine residual prediction with formal uncertainty sets or coverage guarantees, but that would be a different study.

The results should therefore be interpreted as evidence for the usefulness of the modeling structure in the tested environment. They do not prove that the method is sufficient for operational deployment. They show that, under controlled simulated maneuvering encounters, physically anchored residual learning improves hazard geometry prediction and binary warning performance relative to several strong baselines.

9. Conclusion

This paper presented an empirical residual state space learning method for early Detect and Avoid hazard prediction in maneuvering unmanned aircraft encounters. The model combines a stochastic kinematic predictor with a recurrent residual correction trained on encounter histories, continuous closest-approach quantities, and binary loss-of-well-clear labels. The design keeps the physically interpretable forecast as the backbone and uses machine learning to correct systematic residual errors rather than replacing the motion model entirely.

The test results show consistent improvement across continuous and binary metrics. At the 35-second horizon, the residual model reduced horizontal miss-distance root mean square error by 31.7% relative to the best filtering baseline and by 22.8% relative to direct recurrent learning. It also improved receiver operating characteristic area, precision-recall area, Brier score, expected calibration error, and warning recall at an

approximately fixed false-alert rate. The gains were largest in maneuvering encounters, where model mismatch and adaptation lag create errors that can be learned from innovation histories and uncertainty summaries [30].

The ablation study showed that the model's performance depends on the combination of physical prediction, innovation history, covariance information, mode entropy, and residual regularization. Removing any of these components reduced either accuracy or calibration. The residual magnitude penalty was especially useful for preventing overactive corrections in low-risk cases. The calibration layer improved probability reliability under the primary test regime and helped preserve calibration under high noise and measurement dropout.

The robustness tests indicate that the method remains effective under higher sensor noise, anisotropic measurement errors, missing observations, modest ownship speed perturbations, and threshold sensitivity changes. The model is less effective when trained without high-intensity maneuver examples and then tested on high-intensity encounters. This result is expected and highlights the need for training data that cover the maneuver regimes of interest. Physical anchoring helps under shift, but it does not remove the need for representative training examples.

The study has important limitations. It uses simulated encounters and generic sensor-error models. It evaluates open-loop prediction rather than closed-loop avoidance [31]. It does not model all sensor artifacts, airspace procedures, pilot or autonomy response behavior, or certification-level integrity requirements. It also cannot predict future intruder intent when no observable or statistical precursor exists [32]. These limitations mean that the results should not be read as a claim of operational sufficiency.

The main conclusion is narrower and empirical: in the tested maneuvering encounter environment, residual learning around a state space predictor improves early hazard prediction while preserving interpretable closest-approach geometry. This structure provides a practical compromise between model-based tracking and flexible supervised learning. Future work should evaluate the approach with sensor-specific data, stronger ownship maneuvering, richer traffic context, and closed-loop alert-response simulations.

References

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

[2] A. C. Canolla, M. B. Jamoom, and B. Pervan, "Interactive multiple model hazard states prediction for unmanned aircraft systems (uas) detect and avoid (daa)," in 2018 AIAA Information Systems-AIAA Infotech@ Aerospace, p. 2011, 2018.

[3] J. T. Brooks, A. W. Hughes, D. P. Sellers, C. A. Singer, P. A. Sargent, and F. E. Hartman, "Training community modeling and simulation business plan: 2009 edition," 4 2010.

[4] R. J. Wallace, J. M. Loffi, S. M. Vance, J. Jacob, J. C. Dunlap, and T. Mitchell, "Pilot visual detection of small unmanned aircraft systems (suas) equipped with strobe lighting," Journal of Aviation Technology and Engineering, vol. 7, pp. 5–, 4 2018.

[5] J. Schumann, K. Y. Rozier, T. Reinbacher, O. J. Mengshoel, T. Mbaya, and C. A. Ippolito, "Towards real-time, on-board, hardware-supported sensor and software health management for unmanned aerial systems," International Journal of Prognostics and Health Management, vol. 6, pp. 1–27, 11 2020.

[6] M. Sibanda, O. Mutanga, V. G. P. Chimonyo, A. D. Clulow, C. Shoko, D. Mazvimavi, T. Dube, and T. Mabhaudhi, "Application of drone technologies in surface water resources monitoring and assessment: A systematic review of progress, challenges, and opportunities in the global south," Drones, vol. 5, pp. 84–, 8 2021.

[7] S. Kostyuk, I. Chicherin, B. Fedosenkov, and D. M. Dubinkin, "Monitoring of the dynamic state of autonomous heavy platforms on the quarry routes of mining enterprises," Sustainable Development of Mountain Territories, vol. 12, pp. 600–608, 12 2020.

[8] L. Halounová, C. Heipke, and A. Radtke, "Xxiii isprs congress, proceedings and results – volume xli, part a, 2016," The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLI-A, pp. 1–298, 12 2018.

[9] J. J. Bermúdez, T. L. González, J. C. Llarena, E. I. Lavall, C. H. Rodríguez, B. M. García, J. L. Vinielles, and I. L. M. Blas, 1st Scientific-Technical Conference of the Geological and Mining Institute of Spain-National Center (IGME, CSIC). PROCEEDINGS. 9 2022.

[10] C. Veiback, Tracking the Wanders of Nature. Linköping University Electronic Press, 12 2018.

[11] I. I. A. for Vegetation Science, 63rd IAVS Symposium - BOOK OF ABSTRACTS. IAVS, 2021.

[12] H. Cheng, D. Zhao, N. L. Oo, X. Liu, and X. Dong, "Numerical investigation on intermittent maximum ice accretion and aerodynamic performances of rg-15 aerofoil at low reynolds number," Aerospace, vol. 11, pp. 7–7, 12 2023.

[13] S. X. Fang, S. O'Young, and L. Rolland, "Online risk-based supervisory maneuvering guidance for small unmanned aircraft systems," Journal of Guidance, Control, and Dynamics, vol. 41, no. 12, pp. 2588–2603, 2018.

[14] R. C. Millar, L. Hashemi, A. Mahmoodi, R. W. Meyer, and J. Laliberte, "Integrating unmanned and manned uavs data network based on combined bayesian belief network and multi-objective reinforcement learning algorithm," Drone Systems and Applications, vol. 11, pp. 1–17, 1 2023.

[15] A. K. Gianotto, K. O. Pasamehmetoglu, S. Hartenstein, J. Herzog, M. Henry, D. Ogden, and K. Estes, "Idaho national laboratory mission accomplishments," 11 2014.

[16] H. Vogt, "Derivation of forest inventory parameters from high-resolution satellite imagery for the thunkel area, northern mongolia. a comparative study on various satellite sensors and data analysis techniques.," 1 2021.

[17] T. Ahmed, L. Creedon, and S. S. Gharbia, "Low-cost sensors for monitoring coastal climate hazards: A systematic review and meta-analysis.," Sensors (Basel, Switzerland), vol. 23, pp. 1717–1717, 2 2023.

[18] A. la Cour-Harbo and H. Schiøler, "Probability of low altitude midair collision between general aviation and unmanned aircraft," Risk analysis : an official publication of the Society for Risk Analysis, vol. 39, pp. 2499–2513, 7 2019.

[19] L. Spirkovska, I. Roychoudhury, M. Daigle, and K. Goebel, "Real time safety monitoring: Concept for supporting safe flight operations," in 17th AIAA Aviation Technology, Integration, and Operations Conference, American Institute of Aeronautics and Astronautics, 6 2017.

[20] T. Kistan, A. Gardi, and R. Sabatini, "Machine learning and cognitive ergonomics in air traffic management: Recent developments and considerations for certification," Aerospace, vol. 5, pp. 103–, 10 2018.

[21] A. Allouch, A. Koubaa, M. Khalgui, and T. Abbes, "Qualitative and quantitative risk analysis and safety assessment of unmanned aerial vehicles missions over the internet," IEEE Access, vol. 7, pp. 53392–53410, 2019.

[22] B. Lawrence, E. de Lemmus, and H. Cho, "Uas-based real-time detection of red-cockaded woodpecker cavities in heterogeneous landscapes using yolo object detection algorithms," Remote Sensing, vol. 15, pp. 883–883, 2 2023.

[23] D. F. Pigatto, "Hamster healthy, mobility and security-based data communication architecture for unmanned systems," 7 2017.

[24] A. Rajabifard, Sustainable Development Goals Connectivity Dilemma. CRC Press, 8 2019.

[25] H. Christophersen, A. Aksoy, J. Dunion, and S. D. Aberson, "Composite impact of global hawk unmanned aircraft dropwindsondes on tropical cyclone analyses and forecasts," Monthly Weather Review, vol. 146, pp. 2297–2314, 7 2018.

[26] S. Reilly, M. L. Clark, L. P. Bentley, C. Matley, E. Piazza, and I. O. 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, vol. 13, pp. 3810–, 9 2021.

[27] A. Rabajczyk, J. Zboina, M. Zielecka, and R. Fellner, "Monitoring of selected cbrn threats in the air in industrial areas with the use of unmanned aerial vehicles," Atmosphere, vol. 11, pp. 1373–, 12 2020.

[28] N. Ruseno, C.-Y. Lin, and S.-C. Chang, "Uas traffic management communications: The legacy of ads-b, new establishment of remote id, or leverage of ads-b-like systems?," Drones, vol. 6, pp. 57–57, 2 2022.

[29] J. A. Besada, D. J. Martin, G. Frontera, G. de Miguel, and A. M. Bernardos, "Air-to-air surveillance for future atm systems," in 29th Digital Avionics Systems Conference, pp. 6.A.6–1–6.A.6–11, IEEE, 2010.

[30] "Executive office of the president," 2008.

[31] A. McFadyen and L. Mejias, "A survey of autonomous vision-based see and avoid for unmanned aircraft systems," Progress in Aerospace Sciences, vol. 80, pp. 1–17, 2016.

[32] T. Ceccarelli, A. Chauhan, G. Rambaldi, I. Kumar, C. Cappello, S. Janssen, and M. McCampbell, Leveraging automation and digitalization for precision agriculture: Evidence from the case studies. 11 2022.