

Sim-to-Real Transferability of Spatial and Time-Based Horizontal Probabilistic Conflict Detection

Chen Xia ^{*,1,2} Xuming Lei,³ Bo Peng,³ Yan Xu ³ and Christian Eduardo Verdonk Gallego ²

¹Aeronautical Systems, Air Transport and Airports Department, Universidad Politécnica de Madrid, Madrid, Spain

²CRIDA A.I.E., Madrid, Spain

³State Key Laboratory of CNS/ATM School of Electronic and Information Engineering, Beihang University, Beijing, China

*Corresponding author: cxia@e-crida.enaire.es

Abstract

The increasing availability of open surveillance data has facilitated the development of data-driven trajectory-prediction and conflict-detection models. However, the development of learning-based conflict-resolution agents requires interactive training environments in which resolution actions and their consequences can be explored. Since such training cannot currently be conducted in operational air traffic, simulation environments provide a practical alternative. Nevertheless, discrepancies between simulated and operational aircraft behaviour may affect the transferability of the trajectory-prediction and conflict-detection components on which these agents rely, potentially limiting their performance in operational conditions. This paper investigates the sim-to-real transferability of two probabilistic horizontal conflict-detection formulations: a spatial approach based on predicted aircraft positions and a temporal approach based on arrival times at predefined critical points. Both formulations are trained using trajectories generated with the BlueSky simulator and evaluated on operational scenarios obtained from OpenSky Network data and their corresponding simulated reconstructions to compare in-domain and out-of-domain performance. Detection performance is assessed through missed-alert and false-alert rates across probability thresholds, while prediction errors are examined to interpret cross-domain differences. The results show less favourable overall conflict-detection performance under sim-to-real transfer, associated with increased prediction errors. The spatial formulation exhibits pronounced differences in detection outcomes at corresponding probability thresholds, reaching high missed-alert rates at relatively low thresholds, whereas the temporal formulation retains alerts over a broader threshold range and shows smaller changes in threshold-dependent behaviour between domains. Differences in route geometry and aircraft speed evolution are identified as potential contributors to prediction degradation, with their relative influence depending on the underlying formulation. These findings highlight the importance of representative simulation environments for developing transferable conflict-detection components and subsequent learning-based conflict-resolution systems.

Keywords: Sim-to-real transfer; Conflict detection; BlueSky; OpenSky Network

Abbreviations: JOAS: Journal of Open Aviation Science, ATM: Air Traffic Management

1. Introduction

The development of advanced conflict-resolution support systems relies on trajectory prediction and conflict detection to identify aircraft encounters that may lead to a loss of separation. Recent advances in machine learning and artificial intelligence have encouraged increasingly data-driven approaches to these functions. Large volumes of open Automatic Dependent Surveillance–Broadcast (ADS-B) data, such as those provided by the OpenSky Network, facilitate the development and evaluation of trajectory-prediction and conflict-detection models [1]. However, operational surveillance data contain relatively few encounters approaching a separation infringement, as air traffic controllers generally intervene to prevent losses of separation. Furthermore, developing learning-based

conflict-resolution agents, particularly those relying on reinforcement learning, requires interactive environments in which resolution actions can be applied and their consequences evaluated. Since such training cannot currently be conducted in operational air traffic, simulation provides a practical alternative, enabling controlled scenario generation, systematic variation of traffic, aircraft and environmental conditions, and interactive training of conflict-resolution agents [2].

Nevertheless, the applicability of simulation-trained resolution agents to operational conditions depends, among other factors, on the transferability of their underlying data-driven components. Discrepancies between simulated and operational aircraft behaviour, including differences in route tracking, aircraft-performance characteristics, meteorological conditions and simulator dynamics, may affect trajectory predictions and, consequently, conflict-probability estimates and detection decisions. Assessing sim-to-real transferability is therefore an important step towards understanding the limitations of simulation-based conflict-detection components. In particular, examining formulations that represent predicted aircraft behaviour differently can provide insight into how domain discrepancies influence conflict-detection performance.

This paper investigates the sim-to-real transferability of two complementary probabilistic horizontal conflict-detection formulations. The first is a spatial formulation that estimates conflict probability from predicted aircraft positions and their associated uncertainty, considering a prescribed horizontal-separation threshold [3]. The second is a temporal critical-point formulation that estimates conflict probability from predicted aircraft arrival times at predefined critical points, considering a prescribed temporal-separation threshold [4]. Although both formulations provide conflict-probability estimates, they differ in their prediction variables, uncertainty representations and separation criteria. Their evaluation within a common experimental framework allows these formulation-dependent characteristics to be considered when examining sim-to-real transferability.

The study evaluates both formulations, trained exclusively on simulated trajectories, using operational aircraft encounters and their corresponding simulated reconstructions. Detection performance is evaluated through missed-alert and false-alert rates across probability thresholds, while prediction errors are analysed to investigate potential sources of cross-domain differences. Particular attention is given to differences in overall detection trade-offs and probability-threshold sensitivity, as well as simulation characteristics that may limit transferability.

2. Conflict-Probability Estimation Methods

Two complementary conflict-probability estimation formulations are considered in this study: a spatial formulation based on predicted aircraft positions and a temporal formulation based on predicted arrival times at critical points. Both approaches incorporate prediction uncertainty to estimate conflict probabilities between 0 and 1.

2.0.1 Spatial Formulation

The spatial formulation is adapted from the Gaussian-SE-WTFTP+ trajectory-prediction framework, which incorporates a Squeeze-and-Excitation module and Wavelet Transform-based trajectory prediction [5]. For this study, the approach is adapted to en-route traffic, predicting future horizontal aircraft positions over a 35-minute horizon using the preceding five minutes of observed trajectory data.

Conflict probability estimation is performed from the predicted horizontal positions and their associated uncertainty, considering the relative-position distribution between aircraft and a horizontal separation threshold of 5 NM. Conflict probability is evaluated at 1 s intervals throughout the available prediction horizon, up to a maximum of 35 minutes. The maximum instantaneous conflict probability across these target prediction steps is retained as the conflict probability used for

detection.

2.0.2 Temporal Formulation

The temporal formulation estimates conflict probability from the temporal separation between flights at predefined critical points, corresponding to locations where the planned trajectories interact. This approach is consistent with the use of critical points in air traffic controllers' conflict reasoning [6] and follows the temporal conflict-detection methodology proposed in [4].

The formulation employs a multilayer perceptron model to predict the time required for an aircraft to reach a predefined critical point along its reference trajectory, whose position is specified by the arc-length variable γ [4, 7]. The predicted arrival-time uncertainty is then used to estimate the probability that the temporal separation between aircraft at a predefined critical point falls below 60 s, following the conflict criterion adopted in [4]. The same predefined critical points are used for the corresponding aircraft encounters in Sim-SCN and Real-SCN to ensure consistency in the temporal conflict-probability estimation. Although the temporal formulation uses a 60 s separation threshold, the subsequent detection evaluation uses observed horizontal separation infringements below 5 NM as reference events for both formulations.

Unlike the spatial formulation, which uses predicted horizontal positions and their uncertainty, the temporal formulation represents predictions and their associated uncertainty in terms of arrival times and requires predefined critical points for conflict-probability estimation.

3. Experiment Plan

This section describes the experimental framework used to assess the sim-to-real transferability of the two conflict-probability estimation formulations. Both models are trained exclusively on simulated trajectories and evaluated on operational traffic scenarios and their corresponding simulated reconstructions.

3.1 Airspace and Operational Dataset

The area of interest (AoI) is the LECMBDP collapsed sector in Spanish upper airspace, comprising three elementary sectors: LECMBLU, LECMDGU and LECMPAU. Two intersecting cruise traffic flows are selected for the analysis: a West–East flow (WE) and a South–North flow (SN). Their representative planned routes intersect within LECMPAU.

Historical aircraft trajectories are obtained from the OpenSky Network and processed to identify the selected flows using a clustering procedure based on principal component analysis and Recursive Density-Based Clustering Algorithm with Noise [8, 9]. The extracted trajectories are used to characterise the operational traffic and construct the real-domain evaluation scenarios, hereafter denoted as Real-SCN.

The operational trajectories also provide the traffic characteristics used to construct the corresponding simulated evaluation scenarios (Sim-SCN). Both evaluation datasets are derived from the same underlying traffic demand, with comparable initial flight conditions and entry times, but differ in their subsequent trajectory evolution. Real-SCN therefore constitutes the out-of-domain evaluation dataset, whereas Sim-SCN constitutes the in-domain evaluation dataset.

3.2 Simulated Training Dataset

The training dataset is generated using the BlueSky air traffic simulator [10]. For each selected traffic flow, 3,000 flights are simulated for model training and native testing.

Aircraft follow representative waypoint sequences identified from recurrent route structures in the operational data. For each flow, the aircraft type is randomly assigned and the Mach number is randomly sampled between 0.70 and 0.88. Flight levels and initial conditions are specified in the simulation scenarios, with aircraft performance modelled using the Base of Aircraft Data (BADA). Lateral and vertical navigation (LNAV and VNAV) are enabled.

To ensure complete trajectories within the AoI, aircraft are introduced outside its boundary when necessary. A simulation boundary extending 50 NM beyond the AoI is used to retain the relevant trajectory segments before spatial filtering.

Meteorological conditions are obtained from ERA5 reanalysis data [11]. Wind fields are derived from the zonal and meridional wind components at pressure levels of 300, 250, 225 and 200 hPa, and vertically interpolated to FL350, FL360, FL370 and FL380. The resulting wind fields have an hourly temporal resolution and a horizontal grid resolution of 0.5 degrees.

For each simulation scenario, a wind field is sampled and applied consistently to all simulated flights throughout the scenario. In the current implementation, wind affects aircraft speed along the assigned route without changing its geometry. Meteorological variability therefore primarily affects the temporal evolution of the generated trajectories.

The resulting synthetic dataset is used exclusively for model development, including training and native testing, and remains separate from the scenarios used for the transferability evaluation.

3.3 Matched Simulated Evaluation Dataset

To assess transferability under comparable traffic conditions, the simulated evaluation dataset (Sim-SCN) is constructed by reproducing selected operational traffic scenarios in BlueSky. Unlike the simulated training dataset, the evaluation scenarios preserve the temporal overlap between traffic flows, allowing horizontal infringements to be analysed.

Each simulated aircraft is initialised using information from its corresponding operational trajectory. The reconstruction preserves the reference route, requested flight level and entry time into the AoI. As aircraft-type information is unavailable, the A320 is assigned by default to all simulated flights. Initial speed is approximated from the observed ground speed, flight level, heading and wind conditions.

Simulations are performed with a time step of 1 s. Meteorological conditions are selected from ERA5 wind fields corresponding to the operational scenario dates. For each scenario, the wind field nearest to the median aircraft entry time is selected and held fixed throughout the simulation.

After aircraft injection, trajectories evolve according to BlueSky's aircraft-performance and navigation models. Consequently, differences in initial-condition approximations, route representation and aircraft-performance assumptions may produce different trajectory evolution and encounter geometries across domains.

For the conflict-detection evaluation, encounters are selected according to the horizontal proximity of aircraft belonging to the two flows. The selected sample includes aircraft pairs with a minimum horizontal separation below 10 NM, including encounters in which the horizontal separation falls below 5 NM while adequate vertical separation is maintained. Such encounters allow horizontal conflict-detection performance to be examined without requiring an operational loss of separation.

3.4 Transferability Evaluation Protocol

The transferability analysis compares the conflict-detection performance of both formulations in Sim-SCN and Real-SCN without retraining the models. Predictions are initiated at the earliest eval-

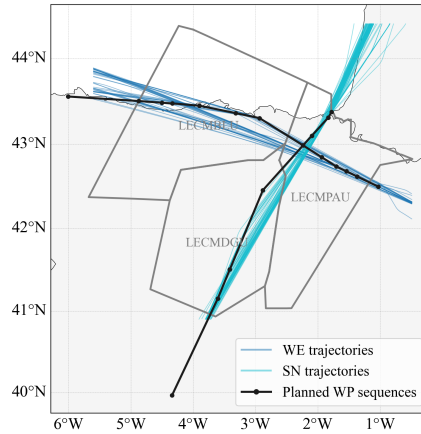


Figure 1. Operational trajectories of the selected evaluation flights in the two intersecting traffic flows.

uation point for which the corresponding model inputs are available, generally near aircraft entry into the AoI. For the spatial formulation, prediction requires five minutes of historical trajectory observations, which may delay its evaluation point relative to the temporal formulation.

3.4.1 Conflict-detection evaluation

For each encounter, an estimated conflict probability is converted into a binary detection decision using a probability threshold p_{thr} . An alert is issued when the estimated probability exceeds the threshold; otherwise, the alert is suppressed.

Reference infringement labels are determined using the complete realised trajectories of the corresponding flight pairs. An encounter is labelled positive when its horizontal separation falls below the prescribed 5 NM threshold and negative otherwise.

Detection performance is evaluated through correct alerts (CA), false alerts (FA), correct suppressions (CS) and missed alerts (MA). A MA corresponds to an observed horizontal separation infringement for which no alert is issued, whereas a FA corresponds to an alert issued for an encounter without an observed horizontal separation infringement. The missed-alert and false-alert rates are defined as $MA/(CA+MA)$ and $FA/(FA+CS)$, respectively. These rates are used instead of absolute counts to facilitate comparisons between datasets containing different numbers of positive and negative detections.

The probability threshold is varied from 0.01 to 0.99 to examine the trade-off between missed alerts and false alerts.

The analysis is conducted separately for the spatial and temporal formulations and for the two evaluation domains. Differences in detection performance between Sim-SCN and Real-SCN are used to assess sim-to-real transferability.

3.4.2 Supporting trajectory-prediction evaluation

To support the interpretation of conflict-detection results, trajectory-prediction errors are also evaluated in both domains.

For the spatial formulation, prediction errors are expressed as horizontal-position deviations between predicted and observed positions in nautical miles, whereas the temporal formulation uses signed arrival-time prediction errors in seconds. The global μ and σ are calculated as the arithmetic

averages of the corresponding residual means and standard deviations across the evaluated target points. Additionally, the global root mean squared error (RMSE) is computed by equally weighting the mean squared prediction errors across the evaluated target points.

For the spatial formulation, errors are additionally reported at prediction horizons of 45 and 900 s, while temporal errors are examined at reference-trajectory distances of 5 and 100 NM.

Because the formulations predict different variables, their prediction errors are not compared directly. Instead, their changes between simulated and operational domains are examined to help explain the corresponding differences in conflict-detection performance.

4. Results and Analysis

4.1 Conflict-Detection Transferability

Conflict-detection transferability is assessed by comparing the MA rates and FA rates obtained with both formulations in the simulated (Sim-SCN) and operational (Real-SCN) evaluation datasets. Detection outcomes are evaluated against horizontal separation infringement labels determined separately from the realised trajectories in each domain. The same 52 flight pairs are considered in Sim-SCN and Real-SCN, ensuring consistent encounter selection.

Figure 2 illustrates the trade-off between MA and FA rates across the considered probability thresholds for both formulations and evaluation domains.

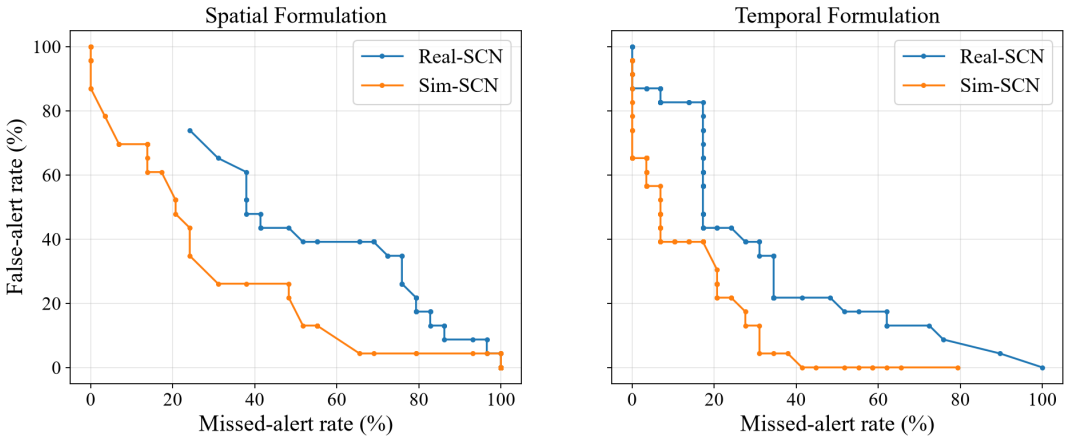


Figure 2. MA/FA trade-offs by probability threshold for the spatial and temporal formulations in Sim-SCN and Real-SCN.

Regarding sim-to-real transferability, both formulations exhibit less favourable conflict-detection performance in the out-of-domain operational scenarios, Real-SCN, as reflected in the observed MA/FA trade-offs. The differences between domains are more pronounced for the spatial formulation. In particular, the spatial formulation exhibits an MA rate of approximately 20% in Real-SCN even at the lowest evaluated probability threshold of 0.01, indicating that a proportion of observed horizontal separation infringements are assigned very low estimated conflict probabilities and remain undetected within the evaluated threshold range.

Beyond the overall MA/FA trade-offs, Figure 3 and Table 1 reveal substantial differences in probability-threshold sensitivity between the two formulations that are not directly observable from the trade-off curves. For the spatial formulation in particular, both the MA and FA rates change markedly at relatively low probability thresholds. The MA rate reaches 100% at thresholds of approximately 0.4–0.5,

while the FA rate has already decreased to zero. This behaviour indicates that the spatial formulation generally produces lower estimated conflict probabilities for the evaluated samples, with relatively modest probability thresholds resulting in the suppression of most alerts. The transition towards complete alert suppression also occurs at lower thresholds in Real-SCN than in Sim-SCN.

Table 1 provides a quantitative comparison of detection outcomes at a probability threshold of 0.35. At this threshold, both formulations exhibit higher MA and FA rates in Real-SCN, illustrating a change in the detection outcomes. These results complement the overall trade-off curves and highlight the importance of considering both detection metrics when assessing domain-dependent performance.

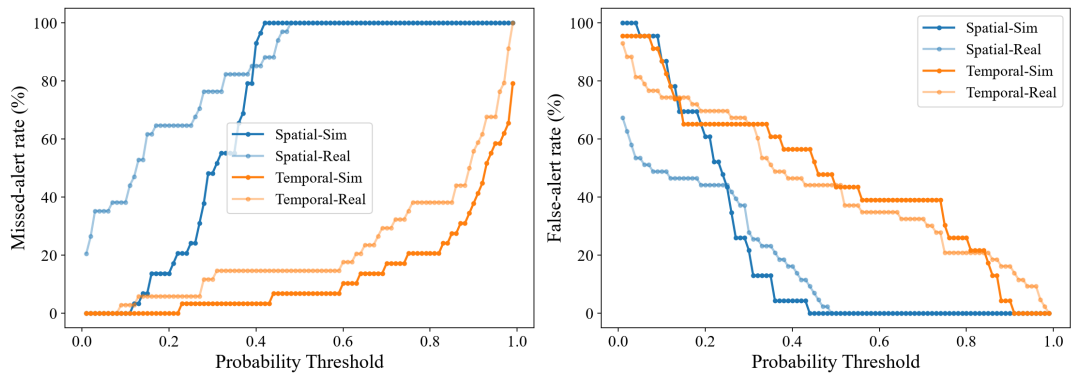


Figure 3. Missed-alert (left) and false-alert (right) rates by probability threshold for the spatial and temporal formulations in Sim-SCN and Real-SCN.

Table 1. Conflict-detection performance at a probability threshold of 0.35.

Formulation	Scenario	CA	FA	CS	MA	MA rate (%)	FA rate (%)
Spatial	Sim-SCN	13	3	20	16	55.17	13.04
Spatial	Real-SCN	6	5	18	23	79.31	21.74
Temporal	Sim-SCN	28	14	9	1	3.45	60.87
Temporal	Real-SCN	24	16	7	5	17.24	69.57

In comparison, the temporal formulation maintains relatively low MA rates over a broader range of probability thresholds, approaching 100% MA only near the upper end of the evaluated threshold range. Its FA rate also decreases progressively with increasing thresholds, reaching zero at slightly different thresholds in Sim-SCN and Real-SCN. Consequently, the spatial formulation reaches the limiting combination of 100% MA and 0% FA at substantially lower thresholds than the temporal formulation. However, this combination does not represent improved detection performance, as it indicates the suppression of alerts for both encounters with and without observed horizontal separation infringements.

The distinct threshold-dependent behaviour of the two formulations and the deterioration observed under sim-to-real transfer may be partly associated with differences in their prediction representations and uncertainty characteristics (see Section 4.2). Larger prediction errors may affect the reliability of estimated conflict probabilities and reduce the ability to distinguish encounters with and without observed horizontal separation infringements. Nevertheless, differences in encounter composition and reference infringement labels between Sim-SCN and Real-SCN may also contribute to the observed performance differences between domains.

Overall, both formulations exhibit degraded conflict-detection trade-offs when transferred from simulated to operational scenarios, although they differ substantially in their probability-threshold sensitivity. The spatial formulation generally produces lower estimated conflict probabilities and reaches complete alert suppression at comparatively low thresholds, whereas the temporal formulation retains alerts over a wider threshold range. These findings highlight the importance of considering both the overall MA/FA trade-off and threshold-dependent detection behaviour when assessing the sim-to-real transferability of probabilistic conflict-detection formulations.

4.2 Trajectory-Prediction Transferability

To support the analysis of conflict-detection transferability, the prediction accuracy of both formulations is evaluated in Sim-SCN and Real-SCN (see Tables 2 and 3). For the spatial formulation, prediction errors are measured as horizontal-position deviations in nautical miles, whereas the temporal formulation uses arrival-time errors in seconds. The μ and σ of these errors are reported globally and at two representative prediction points, corresponding to shorter and longer prediction horizons for the spatial formulation and reference-trajectory distances for the temporal formulation. These statistics illustrate how prediction errors vary with the prediction target.

For the spatial formulation, subscripts 45 and 900 denote prediction horizons of 45 and 900 s, respectively. For the temporal formulation, subscripts 5 and 100 denote distances of 5 and 100 NM along the reference trajectory.

Table 2. Spatial formulation: horizontal-position-error statistics (NM).

Domain	Flow	RMSE	μ_t	σ_t	μ_{45}	σ_{45}	μ_{900}	σ_{900}
Real-SCN	South-North	25.87	17.16	17.97	9.86	27.39	19.78	14.36
Sim-SCN	South-North	9.60	6.98	5.17	4.11	2.13	4.51	6.13
Real-SCN	West-East	30.57	24.89	16.03	15.48	17.83	29.64	16.49
Sim-SCN	West-East	9.82	7.75	2.66	1.85	0.31	9.43	2.88

The spatial formulation exhibits substantially larger prediction errors in Real-SCN than in Sim-SCN for both traffic flows, as shown in Table 2. For the South–North flow, the global mean horizontal-position error increases from 6.98 to 17.16 NM, while the RMSE increases from 9.60 to 25.87 NM. Similarly, for the West–East flow, the global mean error increases from 7.75 to 24.89 NM, accompanied by an RMSE increase from 9.82 to 30.57 NM. The average error standard deviations also increase, indicating greater variability in the prediction residuals at the evaluated target points.

Table 3. Temporal formulation: signed arrival-time-error statistics (s).

Domain	Flow	RMSE	μ_γ	σ_γ	μ_5	σ_5	μ_{100}	σ_{100}
Real-SCN	South-North	46.68	-34.13	6.04	1.50	2.51	-37.44	6.76
Real-SCN	West-East	17.46	-1.50	1.22	0.99	0.40	-0.92	1.25
Sim-SCN	South-North	27.52	-15.90	5.20	-3.13	0.62	-16.65	6.27
Sim-SCN	West-East	16.53	0.02	0.82	1.01	0.18	0.76	0.73

The temporal formulation also exhibits differences between domains, although their magnitude varies between flows (Table 3). For the South–North flow, the global mean arrival-time error changes from -15.90 s in Sim-SCN to -34.13 s in Real-SCN, accompanied by an RMSE increase from 27.52 to 46.68 s. This indicates both a larger negative prediction bias and greater overall error magnitude in the operational domain. In contrast, the West–East flow exhibits smaller changes in the global mean error, from 0.02 to -1.50 s, and in RMSE, from 16.53 to 17.46 s. These results suggest that temporal

prediction transferability is flow-dependent, with substantially greater degradation observed for the South–North flow.

Overall, both formulations exhibit reduced prediction accuracy when applied to operational trajectories, although the magnitude and characteristics of the differences vary across formulations and traffic flows. These results support the interpretation of the conflict-detection results and motivate a closer examination of the discrepancies between simulated and operational aircraft behaviour.

4.3 Analysis of Domain-Shift Effects

The differences in conflict-detection performance observed in Section 4.1 can be examined in relation to the prediction uncertainty and the discrepancies between simulated and operational flight behaviour. In particular, discrepancies in route geometry and speed profile may contribute to the observed transferability differences of both formulations, although their relative influence may depend on the underlying prediction representation. Route-geometry discrepancies are expected to have a more direct influence on the spatial formulation, which predicts future aircraft positions based on past positions, whereas speed-evolution discrepancies are particularly relevant to the temporal formulation, which predicts arrival times along reference trajectories. Nevertheless, these effects are not independent, as aircraft speed also influences spatial trajectory evolution, while route geometry can affect predicted arrival times. The following analysis examines these potential mechanisms through representative trajectory and prediction examples.

Prediction uncertainty plays an important role in both conflict-probability formulations, particularly at longer prediction horizons. For the spatial formulation, the predicted positional uncertainty can extend considerably beyond the 5 NM horizontal separation threshold. Consequently, even when the predicted mean positions indicate a horizontal separation below this threshold, the estimated conflict probability may remain relatively low depending on the dispersion of the predicted relative-position distribution. This may result in missed alerts when the estimated probability falls below the alert threshold despite an observed horizontal separation infringement. The temporal formulation instead considers uncertainty in arrival times at a predefined critical point, using a temporal separation threshold of 60 s. Differences in uncertainty representation may therefore contribute to the distinct conflict-probability distributions and detection behaviours observed in Section 4.1.

As both formulations are based on data-driven algorithms, another relevant factor is the discrepancy between the trajectory patterns represented in the simulated training data and those observed operationally in the evaluation data. In the simulated dataset, aircraft follow predefined waypoint sequences under LNAV and VNAV guidance, resulting in consistent route geometries. Operational trajectories, however, may deviate from these representative routes due to differences in flight execution, air traffic control instructions or meteorological conditions. Consequently, the spatial predictor may encounter trajectory patterns in Real-SCN that are not adequately represented during training.

Figure 4 illustrates this behaviour for a selected flight pair. In Sim-SCN, the predicted and simulated trajectories remain relatively close during the initial segments, with deviations becoming more noticeable around subsequent heading changes. In Real-SCN, larger discrepancies between predicted and observed trajectories are apparent. These differences are consistent with the larger spatial prediction errors reported in Section 4.2. The green markers indicate the prediction evaluation points, which occur after entry into the AoI when the five-minute observation history required by the spatial formulation becomes available.

Differences between domains are also observed in aircraft speed evolution. Figure 5(a) compares the ground-speed profiles of the selected flight pair. Although speed variations are present in both domains, partly associated with heading changes, the simulated trajectories exhibit more abrupt speed changes than their operational counterparts. These discrepancies are particularly noticeable

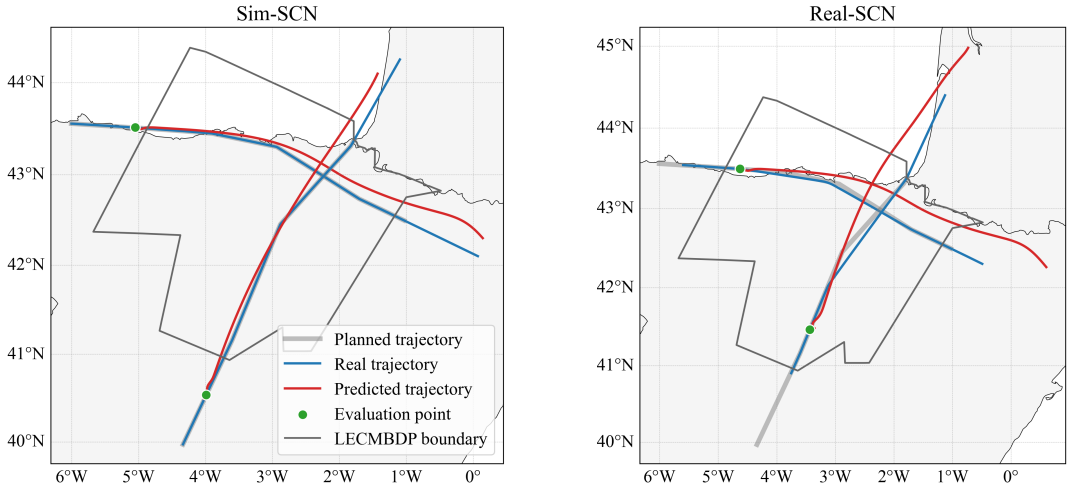


Figure 4. Comparison of spatial trajectory predictions for a selected aircraft pair in simulated (left) and operational (right) scenarios.

for the South–North flight, whereas the West–East flight shows comparatively smaller differences.

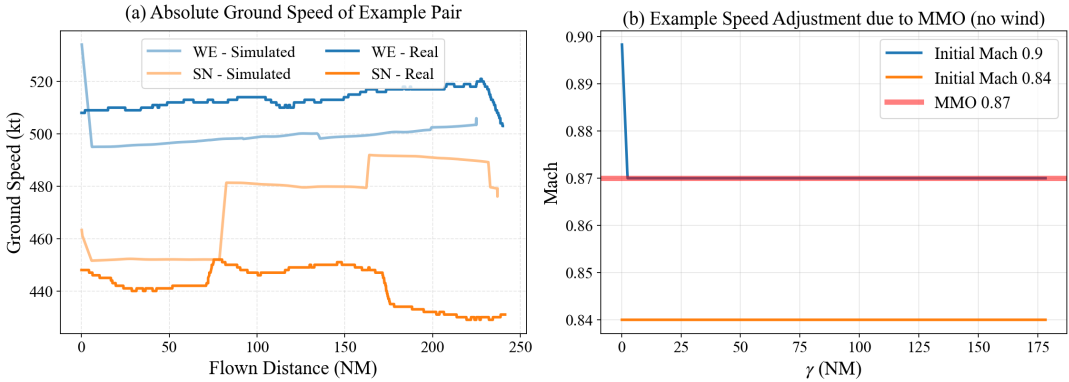


Figure 5. Aircraft speed evolution: (a) ground-speed profiles of the selected aircraft pair in simulated and operational scenarios; (b) Mach-number adjustment in BlueSky for flights initialised below and above the assigned maximum operating Mach number, without wind.

An additional source of speed variation in Sim-SCN arises from BlueSky’s aircraft-performance constraints. When an aircraft is initialised at a speed exceeding the maximum operating Mach number (MMO) or maximum operating speed of its assigned aircraft model, the simulator adjusts its speed automatically to satisfy the corresponding limit. This behaviour is illustrated in Figure 5(b), which compares two additional flights simulated without wind. The flight initialised below the Mach limit maintains its assigned speed across the whole trajectory, whereas the flight initialised above the limit undergoes a speed reduction until the maximum operating Mach number is reached.

Since the A320 is assigned by default in Sim-SCN, some aircraft may be initialised at speeds exceeding the performance limits of the assigned model. The resulting speed adjustments may not occur in their operational counterparts, contributing to cross-domain speed-profile discrepancies.

These speed-profile discrepancies are particularly relevant to the temporal formulation, which estimates arrival times along reference trajectories. Figure 6 presents the temporal prediction residuals for the same aircraft pair. For the West–East flight, the residuals remain comparatively similar between Sim-SCN and Real-SCN, reaching approximately 20 s towards the end of the evaluated trajectory. In contrast, the South–North flight exhibits a considerably larger difference between domains. While the simulated residual remains relatively small, reaching approximately -20 s, the operational residual decreases progressively to approximately -120 s. This behaviour is consistent with the larger differences in speed evolution observed for the South–North flight and the flow-dependent temporal prediction errors reported in Section 4.2.

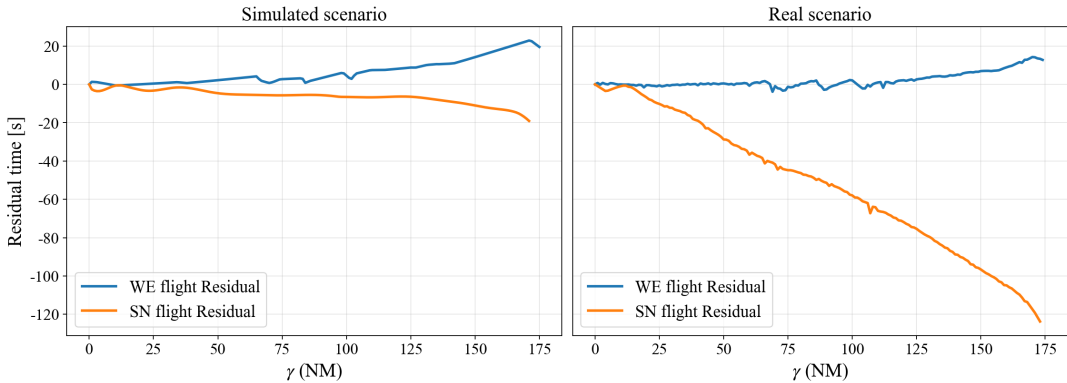


Figure 6. Comparison of temporal trajectory predictions for a selected aircraft pair in simulated (left) and operational (right) scenarios.

Although the presented examples do not isolate the individual contribution of each factor to conflict-detection performance, they illustrate mechanisms that may help explain the domain-dependent detection behaviour reported in Section 4.1.

5. Discussion

The results indicate that the spatial and temporal formulations both exhibit less-favourable conflict detection performance in operational scenarios than in simulated scenarios, associated with reduced trajectory-prediction accuracy in operational scenarios. In addition, the threshold-response curves show that spatial conflict probabilities are generally lower than those produced by the temporal formulation, especially for the operational sample. Consequently, the spatial MA rate rises to 100% and its FA rate falls to zero at comparatively low probability thresholds. The temporal formulation retains nonzero detection rates over a much broader threshold range, even though its operational MA/FA trade-off also deteriorates. Thus, both discrimination and threshold sensitivity must be considered when assessing transferability; a reduction in false alerts that arises solely from suppressing nearly all alerts does not indicate improved conflict detection. These findings suggest that sim-to-real transferability depends not only on the prediction model but also on how its outputs are translated into conflict probabilities and detection decisions.

Nevertheless, as the two formulations employ different conflict representations and separation thresholds, their comparison is intended to illustrate formulation-dependent transferability rather than establish the general superiority of either approach.

The differences between domains may be partly related to the characteristics of the simulated training data. Aircraft in the training scenarios follow representative waypoint sequences, whereas oper-

ational trajectories exhibit greater variability and deviations from these routes. This discrepancy is particularly relevant to the spatial formulation, which relies on learned trajectory patterns and may therefore struggle to predict manoeuvres that are insufficiently represented during training. Furthermore, the spatial formulation was originally developed for shorter prediction horizons during climb and descent, and its direct adaptation to longer-horizon en-route prediction may contribute to the observed prediction errors.

For the temporal formulation, route deviations have a less direct influence because predictions are expressed in terms of distance along a reference trajectory rather than geographical position. Nevertheless, the formulation remains dependent on the assumed route geometry and the location of the critical point. Differences in aircraft speed evolution are particularly relevant because the model predicts arrival times along the reference trajectory. The simulated evaluation scenarios exhibit speed-profile discrepancies associated with aircraft-performance constraints and assumptions used during scenario reconstruction. In particular, assigning a default aircraft model when the actual type is unavailable may introduce speed adjustments that differ from operational behaviour. Such discrepancies can affect arrival-time predictions and, consequently, the estimated temporal separation between aircraft.

Several limitations should be considered when interpreting these findings. First, the simulated training trajectories follow predefined routes without conflict-resolution interventions, limiting the representation of manoeuvring behaviour observed in operational traffic. Introducing a wider variety of route patterns and simulated manoeuvres could improve the representativeness of the training data. Second, aircraft-performance assumptions and simplified meteorological effects may contribute to discrepancies between domains. Third, predictions are evaluated at relatively early encounter stages, when long prediction horizons can produce substantial uncertainty. Evaluating predictions closer to the potential conflict could provide additional insight into detection performance at shorter horizons.

The evaluation design also introduces limitations. The analysis is restricted to two en-route traffic flows, limiting the generalisability of the findings. Another consideration is that the temporal formulation adopts a 60 s temporal-separation criterion from the reference formulation, whereas detection performance is evaluated against observed horizontal separation infringements below 5 NM. As these criteria are not strictly equivalent, their differences may also contribute to the resulting detection outcomes. Finally, the analysis considers only sim-to-real transferability. Evaluating the reverse transfer direction, from operationally trained models to simulated scenarios, could provide complementary insights into the suitability of simulation environments for training conflict-resolution agents, with the ultimate goal of improving their transferability to operational conditions.

Overall, the results highlight the importance of representing operational trajectory variability and aircraft-performance characteristics when developing simulation-based training environments for learning-based conflict-resolution systems that rely on probabilistic conflict-detection components. Improving the representativeness of simulated trajectories, refining aircraft-model assignment, and examining the sensitivity of the conflict-probability formulations to prediction uncertainty and critical-point selection are therefore relevant directions for improving sim-to-real transferability.

6. Conclusions

This study evaluated the sim-to-real transferability of two probabilistic aircraft conflict-detection formulations trained exclusively on simulated trajectories and evaluated on simulated and operational scenarios. The results show less favourable conflict-detection performance in operational scenarios, accompanied by increased trajectory-prediction errors. The spatial formulation additionally produces generally lower estimated conflict probabilities for the evaluated encounters, reaching complete alert suppression at comparatively low probability thresholds. This behaviour is especially

pronounced in operational scenarios and demonstrates the importance of examining probability-threshold responses as well as overall detection trade-offs. The temporal formulation retains alerts over a broader threshold range but also exhibits degradation in its operational trade-off.

Both formulations exhibit reduced prediction accuracy in operational scenarios, although the characteristics of the degradation differ. Route-geometry and speed-evolution discrepancies may affect both formulations, with their relative influence depending on the underlying prediction representation. The observed results suggest that route-pattern discrepancies may be particularly relevant to the spatial formulation, whereas speed-evolution discrepancies may be especially relevant to the temporal formulation. However, the present analysis does not isolate or quantify the individual contributions of these factors to the observed transferability differences. The findings highlight the importance of improving the representativeness of simulated trajectories and aircraft-performance characteristics to enhance the transferability of simulation-trained conflict-detection components intended for learning-based conflict-resolution systems.

7. References

Acknowledgement

This study was conducted using data from the Copernicus Climate Change Service, Climate Data Store.

Author contributions

- Chen Xia: Conceptualization, Data Curation, Formal Analysis, Methodology, Resources, Software, Supervision, Validation, Writing (Original Draft), Writing (Review and Editing)
- Xuming Lei: Conceptualization, Methodology, Software, Writing (Review and Editing)
- Bo Peng: Conceptualization, Methodology, Software, Writing (Review and Editing)
- Yan Xu: Conceptualization, Supervision, Writing (Review and Editing)
- Christian Eduardo Verdonk Gallego: Conceptualization, Supervision, Writing (Review and Editing)

Open data statement

The aircraft surveillance and meteorological data used in this study are available from the OpenSky Network and the Copernicus Climate Change Service, respectively. Additional simulation data were generated using the BlueSky air traffic simulator.

Reproducibility statement

The methodology, experimental setup, and evaluation procedures are described in the paper to facilitate reproducibility. The source code and supporting scripts are currently upon reasonable request. A public repository is planned for a future release.

References

- [1] Matthias Schäfer, Martin Strohmeier, Vincent Lenders, Ivan Martinovic, and Matthias Wilhelm. "Bringing up OpenSky: A large-scale ADS-B sensor network for research". In: *IPSN-14 proceedings of the 13th international symposium on information processing in sensor networks*. IEEE. 2014, pp. 83–94.

- [2] Marc Brittain and Peng Wei. “Autonomous Air Traffic Controller: A Deep Multi-Agent Reinforcement Learning Approach”. In: *CoRR* abs/1905.01303 (2019). arXiv: 1905.01303 [cs.MA]. URL: <https://arxiv.org/abs/1905.01303>. 431 432 433
- [3] Xuming Lei, Yan Xu, Bo Peng, and Kaiquan Cai. “A Probability Collision Risk Assessment and Decision Support Framework for UAV Intrusion in Airport Terminal Airspace”. In: June 2026, pp. 1371–1379. DOI: 10.1109/ICUAS69441.2026.11598669. 434 435 436
- [4] Chen Xia, Christian Eduardo Verdonk Gallego, Adrián Fabio Bracero, Víctor Fernando Gómez Comendador, and Rosa María Arnaldo Valdés. “Trajectory Predictor and Conflict Detection Figures of Merit for a Performance-Based Adaptive Air Traffic Monitoring System”. In: *Aerospace* 11.2 (2024), p. 155. DOI: 10.3390/aerospace11020155. URL: <https://doi.org/10.3390/aerospace11020155>. 437 438 439
- [5] Bo Peng, Yan Xu, Xuming Lei, and Kaiquan Cai. “Probabilistic Trajectory Prediction with Safety Envelope Generation in Terminal Airspace: A Gaussian-SE-WTFTP+ Framework”. Unpublished manuscript. 440 441 442 443
- [6] Jonathan Histon and R. John Hansman. “Air Traffic Controller Operating Modes and Cognitive Complexity Regulation”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 55.1 (2011), pp. 345–349. DOI: 10.1177/1071181311551071. eprint: <https://doi.org/10.1177/1071181311551071>. URL: <https://doi.org/10.1177/1071181311551071>. 444 445 446 447
- [7] Christian Eduardo Verdonk Gallego, Víctor Fernando Gómez Comendador, Manuel Angel Amaro Carmona, Rosa María Arnaldo Valdés, Francisco Javier Sáez Nieto, and Miguel García Martínez. “A machine learning approach to air traffic interdependency modelling and its application to trajectory prediction”. In: *Transportation Research Part C: Emerging Technologies* 107 (2019), pp. 356–386. ISSN: 0968-090X. DOI: <https://doi.org/10.1016/j.trc.2019.08.015>. URL: <https://www.sciencedirect.com/science/article/pii/S0968090X18316024>. 448 449 450 451 452 453
- [8] Maxime Gariel, Ashok N. Srivastava, and Eric Feron. “Trajectory Clustering and an Application to Airspace Monitoring”. In: *IEEE Transactions on Intelligent Transportation Systems* 12.4 (2011), pp. 1511–1524. DOI: 10.1109/TITS.2011.2160628. 454 455 456
- [9] Christian Eduardo Verdonk Gallego, Victor Fernando Gómez Comendador, Francisco Javier Saez Nieto, and Miguel Garcia Martinez. “Discussion On Density-Based Clustering Methods Applied for Automated Identification of Airspace Flows”. In: *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*. 2018, pp. 1–10. DOI: 10.1109/DASC.2018.8569219. 457 458 459 460
- [10] Jacco M. Hoekstra and Joost Ellerbroek. “BlueSky ATC Simulator Project: An Open Data and Open Source Approach”. In: *Proceedings of the 7th International Conference on Research in Air Transportation*. Philadelphia, Pennsylvania, USA, June 2016. URL: <https://resolver.tudelft.nl/uuid:d1131a90-f0ea-4489-a217-ad29987689a1>. 461 462 463 464
- [11] Copernicus Climate Change Service (C3S). *Complete ERA5 Global Atmospheric Reanalysis*. Climate Data Store, European Centre for Medium-Range Weather Forecasts, 2023. DOI: 10.24381/cds.143582cf. URL: <https://doi.org/10.24381/cds.143582cf>. 465 466 467