Counterfactual Safety Modeling: Estimating the Impact of ATC Staffing and Scheduling Practices on Risky Events

Berk Öztürk*

Air traffic control (ATC) staffing-related factors, such as staffing levels, supervisor presence and controller fatigue, have been cited as contributing factors to recent aviation incidents and accidents. These events have brought into focus the need to develop new tools that can quantitatively assess the impacts of ATC staffing and scheduling practices on rates of risky encounters in the National Airspace System (NAS). This work addresses the question of how staffing impacts risk using a Bayesian causal modeling approach. As a proof of concept, we examine whether the presence of a dedicated supervisor in an ATC tower reduces the likelihood of runway incursions (RIs). Informed by operational, staffing and incident data, the model can identify operational conditions that have a higher likelihood of RIs, and make prescriptions for when and where supervisor presence can be most effective in reducing risk. These kinds of models can help the Federal Aviation Administration prioritize the most effective ATC technologies and staffing changes to increase safety and efficiency in the NAS.

I. Introduction

Aviation accidents or risky encounters often initiate debates about whether factors related to Air Traffic Control (ATC) staffing play a role in the events. Some of these factors include low staffing levels, controller fatigue, the chronic use of overtime, the combination of ATC positions and the presence of a supervisor. This research addresses the question of whether and by how much staffing factors could impact rates of risky events in the National Airspace System (NAS) using a Bayesian causal inference approach, with the goal of guiding policy to mitigate risks.

As a proof of concept, we investigate whether the presence of a dedicated supervisor in an ATC tower may help reduce the rate of runway incursions. A runway incursion (RI) is defined by the Federal Aviation Administration (FAA) as "any occurrence at an aerodrome involving incorrect presence of an aircraft, vehicle or person on the protected area of a surface designated for the landing and take off of aircraft." [1] The FAA closely tracks RIs since they are precursors to potential high-speed collisions between aircraft, or between aircraft and ground vehicles.

RIs are usually due to human error precipitated "by various underlying factors including degraded situational awareness, high workload, unfamiliar settings or procedures, reduced visibility, and communications errors" [2]. The most recent high-profile RI occurred at LaGuardia Airport (LGA) in March 2026, when a regional jet collided with a firefighting vehicle that had received a clearance to cross a runway during an arrival operation [2]. While it is still unclear whether high ATC workload may have contributed to the event, the preliminary National Transportation Safety Board (NTSB) report of the incident notes that the ground controller was in a *concurrent controller / supervisor (CC/S)* role, where they coordinate ground traffic while also performing supervisory duties as a controller-in-charge (CIC). They had also been "coordinating ground operations with an airplane that had performed two rejected takeoffs, followed by a ground emergency at terminal B", which involved the firefighting vehicle from the accident [3]. There is reason to believe that workload was higher than usual during the event, and we know that a controller was in a CC/S role. However, the interplay between ATC workload, supervisor presence and risk is not well-understood and requires further investigation.

In a statement before the US Senate Committee on Commerce, Science, and Transportation on November 9, 2023, Timothy L. Arel, then Chief Operating Officer of the FAA Air Traffic Organization (ATO), described a series of actions taken following a March 2023 safety summit convened in response to an uptick in the most severe (Category A and B) RIs [4]. Some of these included independent safety reviews, additional training, and improvements to signage and surveillance technology. Among these actions was an explicit directive that supervisors "devote their full attention to the operation and airfield during peak traffic hours at each facility" [4], directly indicating the supervisor role as a mechanism for mitigating RI risk. Supervisors have a complex role in the ATC tower where they maintain oversight of

*The MITRE Corporation, Bedford, MA, 01730, USA. Email: bozturk@mitre.org

the overall operations to ensure safety and efficiency, providing guidance to on-position controllers as needed. This involves a combination of managing controllers (e.g. scheduling breaks and assigning positions), coordinating with other facilities, keeping records, supervising trainee controllers, and rarely intervening in the tasks of active controllers to prevent or mitigate unsafe situations. The supervisor role may also be taken on by a licensed controller, called a CIC, who can perform many or most of the tasks that a supervisor would normally perform. A CIC may also be designated as a CC/S, which means that they are controlling traffic while simultaneously performing supervisory duties, as in the LGA accident. We do not have a great understanding of whether multi-tasking CC/Ss are as effective as dedicated supervisors in mitigating safety risk.

While this work was planned and conducted before the LGA accident, the author believes that RIs are a good candidate for further analysis in this and future work, since they are low-probability, high-consequence events that are influenced by staffing-related factors. This limited 8-month effort between August 2024 and April 2025 was supported by MITRE's independent research and development program, and took significant steps towards being able to quantify the impact of ATC staffing and scheduling practices on safety risk. As a proof of concept, we investigated the question of whether dedicated supervisors can help mitigate the risk of RIs compared to CC/Ss. Collaborating with former tower air traffic controllers, we developed a Bayesian causal model of how the presence of a dedicated supervisor may impact rates of RIs. We merged staffing, operational and incident data to create a master dataset informing the models. We trained the models on the data and tested our hypotheses using the posteriors of the estimated variables. To enable policy prescriptions, we developed an exceptional responders model to identify sites and operations where an increase in dedicated supervisors may have the highest chance of reducing the likelihood of RIs. We were then able to quantify the potential reduction in RIs through counterfactual simulation of supervisors at these sites.

II. Literature Review

This work sits at the intersection of three different research fields. The first is quantitative modeling of aviation safety, especially using causal models. The second is research on historical RI trends and how human factors may contribute to RIs. The third is the study of trends in ATC staffing levels and of methods for setting ATC staffing targets. In this section, we highlight works that help guide our investigation. To the best of our knowledge, no prior published work has attempted to quantify the effect of ATC staffing and supervision on safety incident rates using a causal framework. Finally, we provide some works in the field of Bayesian causal inference that inform the mathematical principles and methods used in this work.

Causal safety modeling in aviation: Risk and safety are research areas in aviation that require constant innovation. With a goal of one in a billion fatal incidents per operation, the aviation safety community has developed a number of causal tools for accident risk assessment. Netjasov and Janic [5] survey risk and safety modeling in civil aviation more generally, where one of the four main categories of methods is causal modeling. These take a number of forms including fault-tree analysis, event trees, Bayesian networks, and Monte Carlo simulation. The authors identify some key challenges with these approaches, primarily the complexity of the models, the lack of interpretability of the results, the dependence on expert opinions on determining the causal factors, and thus the lack of generality. They recommend that causal models have predictive capabilities, be able to assess the safety bottlenecks in the existing systems, and be specific enough to address safety questions at individual airports, ATC facilities and airlines. These recommendations are core to the modeling approach in this work, as we will demonstrate in subsequent sections.

One approach that is popular in aviation safety modeling is the combination of event sequence diagrams, fault trees and/or Bayesian belief nets for evaluating risk, linking organizational, human, and technical factors to accident outcomes. These models are developed through a combination of expert elicitation and data from accident investigation records, and have been used to analyze the causal factors in a number of aviation accidents. There are two well-known frameworks that use this approach. The first is Causal model for Air Transportation Safety (CATS), highlighted in works by Ale et al. [6]. The Integrated Safety Assessment Model (ISAM) framework is being developed by the FAA and its partners based on the work of Borener et al. [7]. Their version of the event sequence diagram approach has been used in risk assessments including in the identification of common cause failures [8] and in the study of loss of control events in general aviation [9]. The CATS and ISAM models are useful for understanding the complex interactions between different factors that contribute to aviation accidents, but they are not designed to estimate the effects of specific interventions or policies on safety outcomes. For example, the models can assert that a miscommunication was on the causal path to an incident, but they cannot inform how staffing level or ATC workload may have impacted the likelihood of a miscommunication. While these models are able to determine the most important parameters in safety risk, they fail to provide recommendations on what to do to mitigate specific risks, and how effective those interventions might be.

There are relatively few published works that use a causal model to assess aviation risk in both the predictive and prescriptive contexts. In the predictive context, Valdés et al.[10] implement hierarchical models to predict the likelihood of safety incidents for air carriers using Mandatory Occurrence Report (MOR) data. More specifically, they are able to predict the number of safety incidents per number of operations or flight hours for each carrier, while partially pooling aircraft fleets within the same carrier to capture variation across the fleets. Haselein et al. [11] use several types of Bayesian networks to evaluate the likelihood of near mid-air collisions (NMACs), and compare their predictive accuracy to traditional machine learning approaches. They are able to determine the factors that are more likely to lead to NMACs, e.g. the type of airspace and instrument versus visual flight rules. Both of these works are successful in the predictive context, but neither evaluate the impact of specific interventions on the predicted risk and thus cannot provide prescriptions to improve safety. Ayra et al. [12] provide the best example of how Bayesian modeling can be effective in prediction and prescription. In this work, the authors implement a Bayesian causal model for evaluating runway overruns at 5 airports in Spain. They are able to evaluate the kinds of conditions and approach behavior under which runway overruns are more likely, and provide runway-level recommendations for mitigating risk. We take a similar approach to evaluate what conditions increase the likelihood of RIs, and make airport-specific recommendations on how supervisor presence can mitigate that risk.

Runway incursion and human factors research: RIs are extensively studied as aviation safety events given their potential for catastrophic consequences. The FAA classifies them by severity from Category A (the most dangerous, involving collision or near-collision) through Category D, and has sustained dedicated runway safety programs aimed at driving their frequency to zero [1, 13]. Despite these programs, Ison [14] shows that RI rates in the United States trended upward from 2001 to 2017 on a per-operation basis, driven by increases in Category C and D RIs, and weak evidence of reductions in Category A and B RIs. Note that this effect was observed despite accounting for the FAA adopting the International Civil Aviation Organization (ICAO) reporting standards for RIs in 2007, effective 2008 [15], which would have caused an increase in RI rates by broadening the definition of RI events to include any unauthorized intrusion into the runway, even if there was no potential conflict with another aircraft. The most up-to-date RI data can be found on the FAA's Runway Safety Statistics website [13], but it is difficult to discern trends since 2017 due to the COVID-19 pandemic and other exogenous factors that have affected air traffic patterns.

The research literature on RIs spans qualitative human factors investigation, content analyses of incident reports, and quantitative panel models. In a literature review of human factors risks related to RIs for the FAA, Knott et al. [16] cite a combination of environmental factors (increase in operations, higher complexity of runway configurations and taxiway connections, and weather conditions) in explaining the increases in RIs between 1993 and 2000. While they accept the perspective that RIs are due to operational errors, they assert that this "serves as a valid basis to *describe* the problem, and not to *solve* the problem". This aligns with the author's point of view that, while such report-based approaches characterize *which* human factors are present, they cannot estimate *how much* different interventions change safety risk. Knott et al. argue that procedures and training targeted to pilots and controllers may mitigate the problems in the short term, but more structural changes that remove the sources of error are required. These include visual aids and runway safety technologies, and while they do not explicitly discuss the high workload problem, they point to technologies or procedures that reduce workload as a way to mitigate errors [16]. They also challenge the assertion that RIs are due to pilot error, citing a MITRE report [17] that "pilots are increasingly being exposed to situations that make them more vulnerable to making errors". In general, while the FAA does determine whether a RI is a pilot deviation or an operational incident linked to actions by ATC, it is often a judgment call as to who bore primary responsibility for the incident. This is an area of debate that the author does not engage with in this paper, choosing to consider all RIs as equally relevant and important to mitigate using ATC staffing recommendations.

Content analyses of NTSB reports can help understand how human errors can manifest by mapping tasks that pilots and controllers were doing to operational errors. Bhargava and Marais [18] examine narratives from 71 NTSB reports that had sufficient detail on human error causes, and find that in 50 out of 71 cases, there was at least one error in ATC instructions, with 173 different hidden reasons for human errors from communication and situational awareness perspectives. In terms of controller errors, they found failure to follow procedures (16), failure to detect conflicts (10), high or increasing workload (11), or miscommunication with other controllers (5) to have occurred before the errors. As the authors themselves point out, there is selection bias in this study since, when a RI is serious enough to warrant an NTSB report, it is likely to have involved controller error [18]. However, they found that there was limited data on the causal factors leading to the errors, and that NTSB reports do not provide much insight into the subtasks that are intended to mitigate RIs. Just like Knott et al. [16], they point out the difficulty in determining whether errors can be attributed to pilots or controllers, instead asserting that it is more important to figure out "how an error in the instruction

propagated through intermediate subtasks, . . . and if it was incorrect, why it was so and why the controller did not catch or correct it" [18].

Quantitative panel models provide a complementary perspective to the content analyses. Omosebi et al. [19] find, across 30 large-hub US airports, that runway intersection density is the strongest structural predictor of incursion rates, showing the importance of controlling for runway configuration when evaluating risk. They also show that increased operational rates tend to increase RI rates; having more operations on legacy runway and taxiway infrastructure causes more congestion and increases the likelihood of a RI. In terms of policy recommendations, they focus on the impacts of Airport Surface Detection Equipment, Model X (ASDE-X) and Runway Status Light (RWSL) technology, where they show that RWSLs can reduce RI risk by more than half, and that ASDE-X can reduce risk by one-third. However, they do not provide any recommendations from an ATC staffing or workload perspective.

ATC staffing research: Air traffic controllers are critical to the safety and efficiency of the NAS. They perform complex tasks that require high levels of situational awareness, decision-making, and communication skills, and they do so with very high accuracy, bolstering the safety record of air travel. According to USA Facts, which is a curator of government data, air travel is the nation's "safest form of transit"; from 2003 to 2023, passengers in cars and trucks were injured at a rate of 42.2 per 100 million miles traveled. In comparison, rail transit passengers were injured at 6.9 per 100 million, and aircraft passengers were injured at a rate of only 0.004 per 100 million miles [20].

However, there are structural factors that can potentially challenge this safety record. There is chronic understaffing in ATC facilities, due to a combination of retirements, attrition, recruiting challenges, the time it takes to train new controllers and the COVID-19 pandemic. As of September 2024, over 40% of facilities are understaffed, based on the FAA's own targets at 290 facilities [21]. At the same time, the United States had 3.9% fewer controllers in 2024 than in 2013, while experiencing 6.5% more traffic. Faced with these shortfalls, the FAA has had to rely on a combination of short-term solutions including higher use of overtime, and combination of controller and supervisor roles, which can lead to fatigue and higher workload for controllers. The FAA has an ambitious 2025-2028 workforce plan [22] to bring staffing in line with expected air traffic levels. In this plan, they also point to a National Academies of Sciences, Engineering, and Medicine (NASEM) Transportation Research Board (TRB) study from June 2025 that concluded that the "legacy FAA staffing standard models are sound, but incorporating additional considerations and input would enhance the models." [22]

The TRB study describes, compares and reviews the FAA's two primary staffing target modeling processes, which are led by the Office of Financial and Labor Analysis (AFN) and Collaborative Resource Working Group (CRWG) [23]. The AFN model has been continuously updated since the 1960s to estimate the number of controllers needed to meet staffing demands on the 90th percentile day at each facility to safely manage traffic. It has three methods for estimating the required staffing, which are the regression method, the 5-hour Time on Position method, and the Minimum Watch Standard method. Each method produces different staffing targets, which are resolved by taking the maximum across the three approaches, with some adjustments on a per-facility basis. A key variable of the AFN model is the availability factor, which is the inverse proportion of time that controllers are expected to be on-position. In comparison, the CRWG model was developed in 2016 and was geared to specifically account for the "non-operational needs of the controller workforce", better estimating availability across different facilities. The CRWG model has 4 steps, where Steps 1 and 2 estimate a staffing target for a busy day, similar to AFN's 90th percentile day. Then Step 3 estimates time required for non-operational needs, and Step 4 adjusts based on availability. It is generally agreed upon that the CRWG model usually produces higher staffing targets than the AFN model. It should also be noted that the CRWG approach relies on historical staffing levels for its validation and calibration, which were initially influenced by the outputs of the AFN model [23], making it somewhat dependent on that model.

The NASEM TRB committee's task was to inform "the FAA's interest in developing objective, science-based approaches for setting future ATC staffing targets to ensure the safe and efficient operation of the NAS." [23] To that end, they developed 6 criteria as foundational for tools that can set targets, which were traceability, relevance, adaptability, validity, data quality, and ongoing monitoring and periodic adjustment, and rated both the AFN and CRWG models on these criteria. Among their recommendations was for the FAA continue to examine its workload models, establish a system to periodically evaluate the need for future updates and provide structured means to account for local factors in setting staffing targets. The author believes that safety-risk-driven staffing adjustments can be one such local factor, and could mitigate the overreliance of the aforementioned approaches on historical data and surveys which may not have captured the entire safety risk landscape. The TRB committee makes it clear in the study that they did not have safety indicators available to them to evaluate the safety impacts of staffing levels or overtime use, but they do recommend additional studies on this front, e.g. comparing the frequency of close calls between facilities that are 10% above or

below their staffing targets, as a way to evaluate the safety impacts of staffing levels.

As for supervisor staffing, the institutional structure of the supervisory role in ATC towers has itself been subject to debate. The FAA's CIC program arose from a 1998 collective bargaining agreement that allowed controllers to perform supervisory duties when a dedicated supervisor is not present, in order to avoid compromising safety in the face of rising supervisor attrition [24]. As noted in [4], it is expected that supervisors mitigate safety incidents as part of their role, but to the best of the author's knowledge there is no evidence on the comparative effectiveness of CICs, CC/Ss or professional supervisors at enabling safe operations. There is some evidence from the literature to support the idea that dedicated supervisors (professional managers or CICs) mitigate safety risk. In support of the FAA's ATO, Schroeder et al. [25] analyzed the operational error (OE) literature. They point to several studies that the author could not find in the literature, one of which found a "statistically significant relationship between the decline in supervisory staffing and an overall increase in OEs" [25], where the change explained 14% of the overall variance. However, the relationship between supervisor staffing and safety risk can be complex due to the potential hierarchical interplay between managers and on-position controllers. Schopf et al. [26] find that high trust in supervisors can actually be associated with low safety citizenship behavior, looking at survey data from 49 employees of a European air navigation service provider. As a possible mechanism, they argue that controllers may believe that trustworthy supervisors will address potential safety concerns, and thus be personally less vigilant.

Causal inference methodology: The analysis in this paper is grounded in causal modeling and inference. Pearl [27] establishes the mathematical basis for answering causal questions from observational data via directed acyclic graph (DAG)-based causal models and do-calculus. McElreath [28] provides the statistical and modeling foundation for the Bayesian multilevel models used in this paper. DAGitty [29] was used to visualize and verify our causal graphs, and to derive the correct adjustment sets for hypothesis testing.

III. Overview of Hypothesis and Testing Approach

The question we address is whether the presence of a dedicated supervisor has an impact on the RI rate at an airport. We posit that a supervisor may impact the likelihood of a RI through two mechanisms. The first we call the *base rate effect*, where they may lower the base rate of incursions, in terms of likelihood per arrival or departure operation, regardless of the arrival and departure rates, or ATC staffing level. This effect could be a direct effect, or may be mediated by operations during *rare configurations*. Rare configurations are defined as those that are seen less often by controllers, as a proportion of yearly operations. These may occur due to unusual weather conditions, or as transitions between more common runway configurations, and could result in higher RI rates through a combination of more complex traffic patterns and less operational experience. The supervisor may mitigate risk during these configurations by providing guidance in complex situations and through their expertise in managing controller positions.

Supervisors may also have *rate-dependent effects* on the RI rate, where they may enable controllers to handle higher arrival rates, departure rates, or workloads (measured as number of operations per staff). There may be risks inherent in higher arrival and departure rates independently of workload, when aircraft are operating in more congested air and ground traffic, and thus may make a collision more likely. It is also possible that, when workload increases (i.e. less staff per operation) independent of operational rates, supervisors may intervene to ensure that safety levels are maintained, or their presence may reduce the likelihood of human error. We separate these rate-dependent effects into the arrival rate effect, the departure rate effect and the workload effect, to be able to determine whether supervisors have a stronger impact in any one of these dimensions.

We relied on a Bayesian causal modeling approach for the analysis, which has the following advantages:

Ability to generate counterfactuals: Bayesian causal models allow for the evaluation of *counterfactuals*. In other words, they allow us to simulate the effect of an intervention, all other things equal, without having to do an experiment. This is key for asking questions such as, what is the safety impact of adding a dedicated supervisor at the airport, with constant staffing, configuration and weather conditions?

Ability to work with sparse data: Risky encounters are rare events, and Bayesian causal models can work with limited data to gain insights and provide reliable estimates despite the low sample counts through the use of prior distributions. In this case, we had ~14.5 million operations across 40 airports, but only 317 RIs.

Ability to quantify uncertainty rigorously: Bayesian causal models allow for the quantification of probability distributions over the effects of variables, rather than just point estimates, through the use of posterior distributions. This allows us to gauge both the strength of each effect and the confidence of the result.

Within the Bayesian causal modeling paradigm, the hypotheses may be expressed as a causal DAG. Figure 1

describes the aforementioned hypotheses, where each edge corresponds to a causal relationship between each random variable (RV) and/or data. More explicitly, we describe the relationships in the hypothesis through three effects.

- The base-rate direct effect of a supervisor is represented by the green edge from supervisor availability to the runway incursion rate. This is the magnitude of the risk reduction by a supervisor that is independent of the operational rates, and the use of a rare configuration.
- The base rate effect of a supervisor is also mediated through rare configurations. The rare configuration direct effect is the green edge from the rare configuration to the RI rate. Given that the rare configuration effect is dependent on the presence of a supervisor, the supervisor effect that is mediated by rare runway configurations is simply the difference between the rare configuration direct effect with and without supervisors.
- The rate-dependent effects of supervisors on risk go from arrival rate, departure rate and ATC workload to RI rate. These are stratified by supervisor availability, so that the rate-dependent supervisor effect is the difference between each rate effect with and without supervisors. Thus we can figure out if operational rates impact risk, and whether the presence of a supervisor impacts that risk.

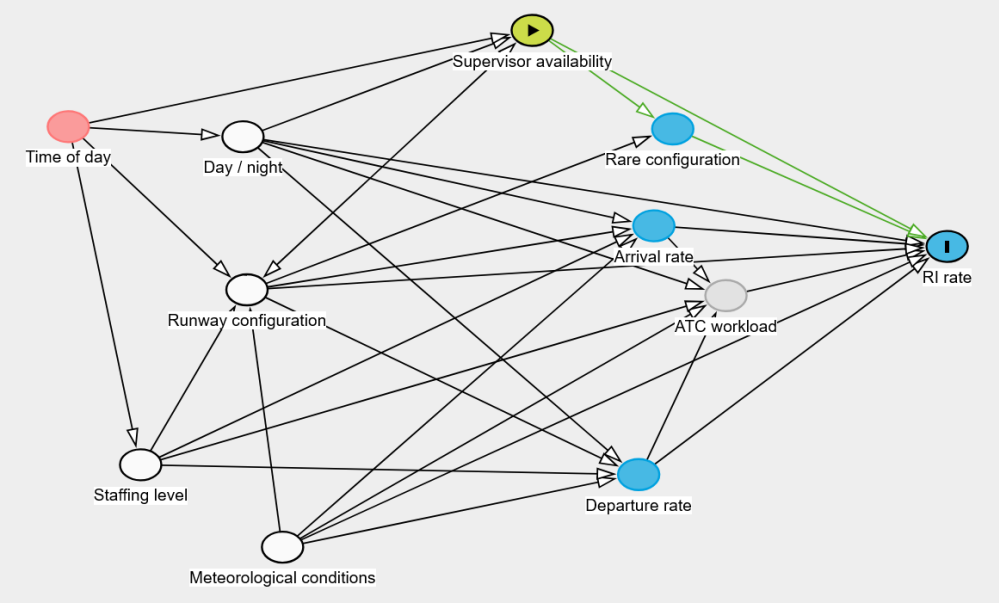


Fig. 1 Directed acyclic graph describing how supervisors may impact the runway incursion likelihood.

In addition to the supervisor availability, rare configuration, arrival rate, ATC workload, departure rate and RI rate nodes, Figure 1 includes the potential confounding variables for the analysis, which are time of day, day/night, runway configuration, staffing level and meteorological conditions. These are primarily *fork* confounders [28], which affect both the independent and dependent variables of interest, in our case the supervisor availability and the likelihood of RIs. The fork confounders can be eliminated through simple stratification, i.e., by training separate or partially pooled models for different levels of the independent variable, as we will demonstrate in Section V.

For computational efficiency, it is possible to separate the causal model visualized in Figure 1 into two separate models. The first is an *airport throughput model*, which contains the relationship between staffing and airport throughput. This is a critical component of the causal model since supervisor availability, staffing level and operational rates are confounded by time of day, and clearly affect RI likelihood through both the direct effect of operational rates but also ATC workload. However, instead of stratifying by time of day, we choose to stratify the throughput model by day/night, Instrument Meteorological Conditions (IMC)/Visual Meteorological Conditions (VMC) conditions and runway configuration, since these are more useful and parsimonious operational parameters to consider. The output of the throughput model is a probabilistic forecast of how many arrival and departure operations an airport can support during a specific runway configuration, with a specific staffing level and operational parameters. This can then be used to estimate the relative ATC workload, an unobserved or *latent* variable, by mapping the observed throughput to the forecasted throughput as a quantile.

The second model is a *RI likelihood model*, which contains the relationship between operational rates, workload (as

output from the airport throughput model), and supervisor availability on RI likelihood. In this model RIs are treated as binomial random events whose probabilities are informed by the arrival rate, departure rate, ATC workload, the baseline rates of RIs at the airport and the use of a rare configuration, conditioned on the presence of a supervisor. More details on the models are included in Section V.

IV. Data Processing and Fusion

A substantial component of the work was cleaning, validating and fusing various data sources to be able to test the hypotheses. The data used in the study are outlined below:

- **Operational data**: From MITRE's Transportation Data Platform (TDP) [30] for operational rates, weather and derived configurations,
- **Staffing data**: From FAA's workforce management system (CRU-X/ART), for both controller on-position data and supervisor data, with caveats explained below,
- **Incident data**: From Aviation Safety Information Analysis and Sharing (ASIAS) / Office of Accident Investigation & Prevention (AVP), for RI data.

The scope of the analysis was constrained by the amount of staffing data available; we were able to obtain CRU-X/ART data from across the NAS between 09/30/2023 and 11/06/2024. We focused on tower operations and RI events at 40 ASDE-X and ADS-B Airport Surface Surveillance Capability (ASSC) airports, to ensure high data quality. This period corresponds to ~14.5 million operations and 317 observed RIs.

The data was binned on an hourly basis for both the airport throughput model and RI likelihood model, in order to facilitate the evaluation of runway configurations. Staffing was modeled as on-position controller minutes per hour by aggregating over the CRU-X/ART data. The CRU-X/ART data does provide to-the-minute level resolution with respect to on-position times, so this aggregation does lose relevant details such as which positions were active and whether position changes occurred. However, these factors were outside the scope of this proof-of-concept work.

RIs were represented as binomial random events occurring during each hour. Figure 2 shows the number of hours of operations in the dataset as well as the number of hours with observed RIs. While rare, there were a handful of hours where more than one RI was documented. These cases were checked manually to confirm that the RIs were part of the same event and then pruned down to one event.

Airport	Hours	RIs
ANC	8931	7
ATL	9607	10
BDL	7905	0
BOS	9081	24
BWI	8754	2
CLE	8695	2
CLT	9134	3
CVG	9310	7
DCA	8179	14
DEN	8838	15
DFW	9258	23
DTW	8586	7
EWR	9161	6
FLL	8810	4

Airport	Hours	RIs
HOU	8440	4
IAD	8862	5
JFK	9067	5
LAS	8832	10
LAX	9663	17
LGA	8236	8
MCI	8626	2
MCO	8739	5
MDW	8272	17
MEM	9395	4
MIA	9515	11
MKE	8378	0
MSP	8682	7
MSY	8612	1

Airport	Hours	RIs
ORD	9654	20
PHL	9162	6
PHX	8765	4
PIT	8667	5
PVD	7434	1
SAN	8103	9
SDF	9553	4
SEA	8791	3
SFO	8815	13
SLC	8315	14
SNA	6667	13
STL	8519	5

Fig. 2 Number of operational hours evaluated at each airport, as well as the total runway incursions observed over the time period considered.

An important component of data processing was the determination of when a dedicated supervisor was available. We made the reasonable assumption that there must be a supervisor available at all times, but that this may take two

forms. We define a *dedicated supervisor* as either a CIC or a professional manager whose sole role is to perform the supervisor's duties. The other is a *CC/S*, where a controller is simultaneously directing traffic and performing the supervisor's role. The CRU-X/ART data specifies when a CIC or CC/S is on position, but not when a professional manager is on position. We assume that a CC/S would only be on-position when a dedicated supervisor is not available.

In this study, we are primarily interested in seeing the difference in RI likelihood under a dedicated supervisor versus a concurrent supervisor. If a dedicated supervisor was available more than 80% of the time during a given hour, we made the assumption that that particular hour had a dedicated supervisor (a binary variable). It was generally the case that the supervisor availability was obvious; there were a minority of time periods where there was $< 80\%$ availability but more than 0%, which we classified as having a CC/S.

V. Modeling

A. Airport throughput model

The throughput model captures how staffing at each facility is related to arrival and departure rates during each hour of operations. The throughput model is a necessary component of the analysis, due to the need to adjust for staffing level and operational rates when considering the RI likelihood. Otherwise, the safety impact of a supervisor may be confounded by the simultaneous presence or absence of sufficient staff; since the two would affect RI likelihoods simultaneously, it would be impossible to differentiate between the effect of the supervisor versus the on-position staff.

The throughput model allows us to control for those effects and more accurately evaluate the safety impact of a supervisor. Modeling throughput is challenging in practice since airports differ in fleet mix, weather exposure, controller experience, and infrastructure. We therefore adopted a parsimonious model that takes on-position controller hours and the observed arrival load ($\beta_i \in [0, 1]$, representing arrivals as a proportion of total operations) as inputs, and returns an ensemble forecast of hourly arrival and departure throughput through its posterior predictive output. This throughput distribution is then used to derive ATC workload in each hour by expressing the historically observed throughput as a percentile of the predicted throughput distribution. For example, if 15 arrivals are observed at Boston Logan Airport (BOS) in a historical hour, and under the same operational conditions the model predicts the median throughput to be 15 arrivals, the arrival workload would be estimated to be 50%.

The model is built on the following assumptions:

- The primary variables affecting airport throughput are the level of staffing, runway configuration, IMC/VMC, and day/night conditions.
- The level of staffing is quantified in on-position controller hours in a given hour. More complex approaches could attempt to control for the presence of different on-position roles, but that was outside the scope of this study.
- We control for airport-level confounds by training a separate model for each site.
- The variation in throughput based on the runway configuration, day/night and IMC/VMC is considered through stratification within each model, i.e., training separate parameters for each condition.

Below, we formally describe the throughput model in the language of causal models and RVs. For each airport, a different Bayesian causal model is trained, so the airport indices are omitted in this section. Let $i = 1, \ldots, n$ index hourly observations at a given airport. For each hour i, let

$$A_i = \text{number of arrivals}, \quad D_i = \text{number of departures}, \quad S_i = \text{on-position controller hours}, \quad \beta_i = \frac{A_i}{A_i + D_i},$$

occurring in that particular hour. Additionally, each observation has an associated runway configuration with some numerical index $c_i \in \{1, 2, \ldots, k\}$, day/night status $d_i \in \{1, 2\}$, and meteorological condition $v_i \in \{1, 2\}$ distinguishing between IMC and VMC.

In practice, deciding on how many runway configurations there are at each airport (and thus the value of k) is difficult, given the likelihood of unusual configurations with low operational counts. To resolve this issue, we aggregated runway configurations derived from TDP across the 14 months of data. We rank-ordered them by how common they were at each airport, and what proportion of arrival and departure operations they played a role in. For both arrivals and departures, we considered either the top 15 configurations or the number of configurations involved in 90% or more of operations separately, whichever was the lower number. Then we took the union of the runway configurations, and labeled the others as miscellaneous.

Given the above data, arrival and departure throughput are each modeled separately as the following truncated

Gaussian RVs,

$$A_i \sim \mathcal{N}_{[0,\infty)}\left(\mu_i^{\text{arr}}, \sigma^{\text{arr}}_{c_i,d_i,v_i}\right), \qquad \mu_i^{\text{arr}} = \alpha^{\text{arr}}_{c_i,d_i,v_i}\,\beta_i\,S_i,$$

and

$$D_i \sim \mathcal{N}_{[0,\infty)}\left(\mu_i^{\text{dep}}, \sigma^{\text{dep}}_{c_i,d_i,v_i}\right), \qquad \mu_i^{\text{dep}} = \alpha^{\text{dep}}_{c_i,d_i,v_i}\,(1-\beta_i)\,S_i,$$

where $\mathcal{N}_{[0,\infty)}(\mu,\sigma)$ denotes a normal distribution with mean μ and standard deviation σ truncated below at zero. This parametrization is a Bayesian linear regression, where $\alpha^{\text{arr}}_{c,d,v}$ represents the latent variable describing the expected arrival throughput per controller-hour allocated to arrivals ($\beta_i S_i$) under conditions c, d, and v, while $\alpha^{\text{dep}}_{c,d,v}$ has the analogous interpretation for departures per controller-hour allocated to departures, $(1-\beta_i)S_i$. Note that this is a simplifying assumption, and that on-position controllers are not in practice allocated or differentiated between arrivals and departures as such. There is no intercept to the linear model since throughput should approach zero as staffing approaches zero. The standard deviation parameters $\sigma^{\text{arr}}_{c,d,v}$ and $\sigma^{\text{dep}}_{c,d,v}$ capture the variability unexplained by coefficients $\alpha^{\text{arr}}_{c,d,v}$ and $\alpha^{\text{dep}}_{c,d,v}$, respectively.

Weakly informative priors are assigned to the regression coefficients,

$$\alpha^{\text{arr}}_{c,d,v} \sim \mathcal{N}_{[0,\infty)}(20,\ 15), \qquad \sigma^{\text{arr}}_{c,d,v} \sim \text{Exp}\left(\lambda = \frac{1}{7}\right),$$

$$\alpha^{\text{dep}}_{c,d,v} \sim \mathcal{N}_{[0,\infty)}(20,\ 15), \qquad \sigma^{\text{dep}}_{c,d,v} \sim \text{Exp}\left(\lambda = \frac{1}{7}\right),$$

where $\text{Exp}(\lambda)$ denotes an exponential distribution with rate parameter λ. Intuitively, this means that the prior for the mean arrival rate per controller hour is normally distributed with a mean of 20 with a standard deviation of 15, and the standard deviation has a prior that is exponentially distributed with a mean of 7, and the same for the departures. This prior covers the range of throughput observed across the NAS without biasing the site-specific estimates.

The full model structure is illustrated in Figure 3 as a DAG generated by PyMC, shown here for BOS with its $k = 17$ runway configurations considered. Grey nodes represent observed quantities (staffing S_i, arrival load β_i, and the throughput observations A_i and D_i), while white nodes represent the unobserved variables to be inferred based on reasonable priors ($\sigma^{\text{arr}}_{c,d,v}$, $\mu^{\text{arr}}_{c,d,v}$, $\mu^{\text{dep}}_{c,d,v}$, $\sigma^{\text{dep}}_{c,d,v}$, in left-to-right order). Because each latent variable is stratified across configuration, day/night, and IMC/VMC, the model simultaneously infers $4 \times 17 \times 2 \times 2 = 272$ distinct throughput RVs at BOS alone. Even though the 9081 hours in the dataset are spread unevenly across the different $17 \times 2 \times 2 = 68$ different conditions, in practice the model has sufficient samples in each operational condition to make precise throughput predictions, as we will demonstrate in Section VI.

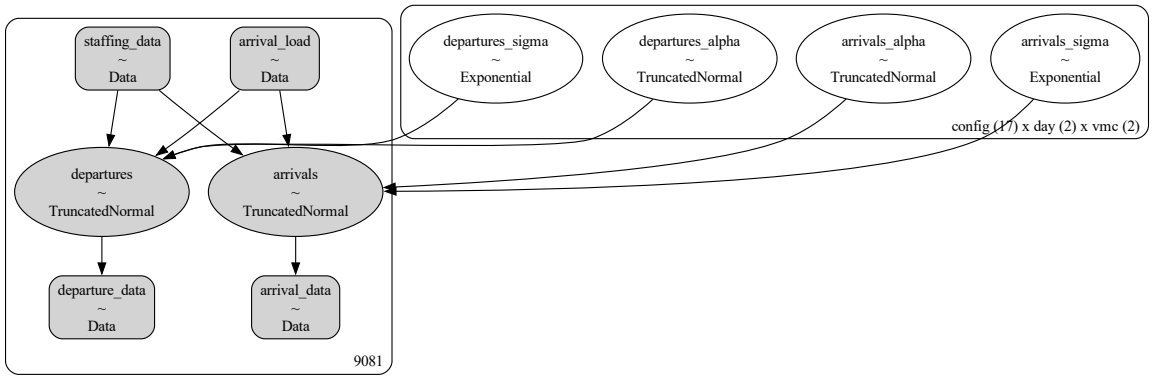


Fig. 3 Directed acyclic graph of the airport throughput model at BOS, shown for the 17 most common runway configurations. Grey nodes are observed quantities; white nodes are unobserved variables inferred from data.

B. Runway incursion likelihood model

The RI likelihood model is a binomial Generalized Linear Model (GLM), where we test the various hypotheses on how supervisors may impact RI rates. As a reminder, a major assumption of the model is that RIs are draws from a binomial RV, whose probability is defined as a combination of base-rate effects and rate-driven effects.

At a high level, the probability of a RI during an hour i is expressed by the following GLM:

$$RI_i \sim \text{Binomial}(N_i, p_i), \quad \text{logit}(p_i) = \psi_i + \boldsymbol{\phi}_i^\top X_i, \tag{1}$$

where $N_i = A_i + D_i$ is the total number of arrival and departure operations observed during hour i, RI_i is the number of RIs observed in that hour, and p_i is the RI probability per operation. The ψ_i term accounts for the base rate of RIs and the $\boldsymbol{\phi}_i$ vector term accounts for the impact of rate-dependent factors, where X_i are rate-based observations during hour i, i.e. arrival rate, departure rate, and workload. The logit function is a nonlinear transformation of the following form,

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right), \tag{2}$$

used to map the linear regression, which has a real number value, into the probability domain [0,1]. This is a common transformation in GLMs in estimating probabilities of discrete events from continuous RVs [28]. Intuitively, a change of x logits in the linear predictor multiplies the odds of the event by e^x. Table 1 makes this more concrete by showing changes of odds for common logit values.

Table 1 Odds multipliers corresponding to logit changes

Logit change	−1.00	−0.75	−0.50	−0.25	0.00	0.25	0.50	0.75	1.00
Odds multiplier	0.37×	0.47×	0.61×	0.78×	1.00×	1.28×	1.65×	2.12×	2.72×

Figure 4 shows the binomial GLM that will be described in detail below.

Indices and observed variables. Let $i = 1, \ldots, N$ be the indices for every hour of operations in the dataset for all 40 airports, with N = 94,960. In each hour i, the observed data are:

$$\begin{aligned}
s_i &\in \{1,2\} && \text{(supervisor category)},\\
d_i &\in \{1,2\} && \text{(daytime category)},\\
v_i &\in \{1,2\} && \text{(VMC category)},\\
c_i &\in \{1,\ldots,586\} && \text{(airport and runway configuration)},\\
a_i &\in \{1,\ldots,40\} && \text{(airport)},\\
w_i &\in \{1,\ldots,5\} && \text{(workload category)},\\
\overline{A}_i &\in \{1,\ldots,5\} && \text{(arrival rate category)},\\
\overline{D}_i &\in \{1,\ldots,5\} && \text{(departure rate category)},
\end{aligned}$$

where the workload, arrival and departure rates have been binned into 5 discrete categories by percentile of historical values observed at each airport and runway configuration, transforming them from continuous to ordered categorical variables. The percentile bins are the 0-10th, 10-25th, 25-75th, 75-90th, and 90-100th percentiles for all ordered categorical rate variables. The workload category for hour i is a mixture of the arrival and departure workloads computed from the airport throughput model, weighted by the arrival proportion β_i, and binned into the appropriate percentile. In addition, each configuration c has a derived rare configuration variable $r_c \in [0,1]$, which quantifies how infrequently a given runway configuration c is used at its airport. Let h_c denote the fraction of total observed hours at each airport during which configuration c was active. The rare configuration variable is then defined as

$$r_c = \min\left(1,\ \frac{1}{100\, h_c}\right),$$

so that a configuration active in 1% of hours or fewer receives $r_c = 1$ (maximally rare), one active in 2% of hours receives $r_c = 0.5$, and so on. Configurations classified as miscellaneous (i.e. those not corresponding to a standard

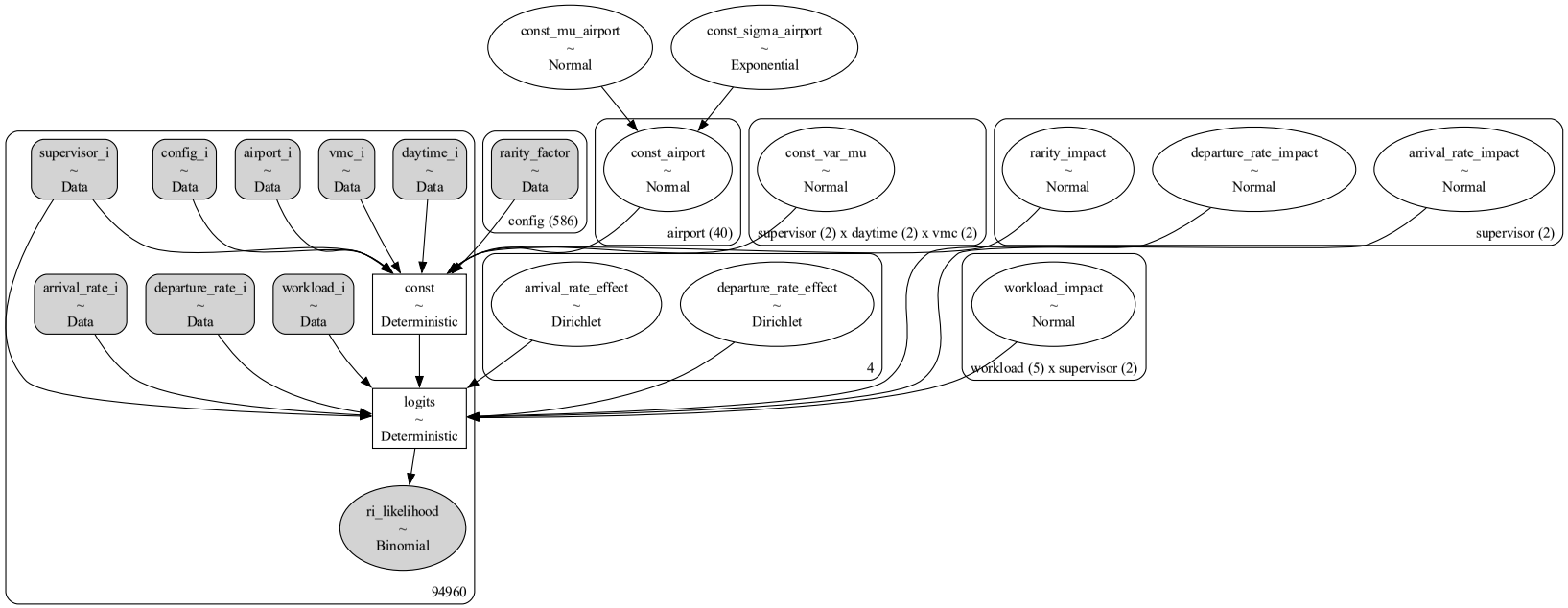


Fig. 4 DAG of the RI likelihood model in PyMC.

named runway layout) are always assigned $r_c = 1$, since by definition they represent atypical operations. Recall that we include the rare configuration variable to account for the base-rate effect of supervisors that is mediated by that variable.

Recall that the RI rate is given by the following equation,

$$RI_i \sim \text{Binomial}(N_i, p_i), \quad \text{logit}(p_i) = \psi_i + \boldsymbol{\phi}_i^\top X_i, \tag{3}$$

where $N_i = A_i + D_i$ is the total number of operations during hour i.

Base-rate effects: The intercept ψ_i, the base rate of RIs during hour i, is composed of three components, given by

$$\psi_i = b_{a_i}^{(\text{airport})} + b_{s_i,d_i,v_i}^{(\text{sdv})} + b_{s_i}^{(\text{rar})}\, r_{c_i}, \tag{4}$$

where

- $b_{a_i}^{(\text{airport})}$ is the baseline incursion rate at the specific airport,
- $b_{s_i,d_i,v_i}^{(\text{sdv})}$ is the rate-independent impact of the combination of supervisor category, daytime category, and VMC category, capturing differences in baseline risk across different operating conditions, and
- $b_{s_i}^{(\text{rar})}\, r_{c_i}$ models the effect of configuration rarity, which depends on whether a supervisor is present.

This representation of the base rate of RIs enables us to measure the direct effect of a supervisor on the RI rate, as well as the supervisor effect that is mediated by rare configurations. The supervisor direct effect is evaluated through the $b_{s_i,d_i,v_i}^{(\text{sdv})}$ variable, by comparing its estimate with the same daytime category d_i and meteorological condition v_i, but with different supervisor category s_i. The supervisor effect mediated through rare configurations is evaluated through $b_{s_i}^{(\text{rar})}\, r_{c_i}$, by comparing the estimate of the rare configuration effect under different s_i.

Rate-dependent effects: The modeling of rate-dependent effects in the $\boldsymbol{\phi}_i^\top X_i$ term is more complex. We chose to model arrival and departure rate effects using an ordered categories approach with Dirichlet-distributed increment weights, because their effects are expected to be *monotonic*: each additional unit of traffic is expected to consistently push the RI likelihood either higher or lower, but the impact of each unit may be nonlinear and thus unequal. A Dirichlet prior on the $K - 1$ increments between K ordered bins encodes this ordinal structure [28] while capturing the fact that the operational rate by definition can't be lower than the lowest quantile.

Workload, by contrast, is modeled with a normal prior for each workload bin and supervisor combination, without enforcing monotonicity. There is reason to believe that the relationship between workload and RI risk is non-monotonic; it is possible for example that at low workloads, controllers may be less vigilant than at moderate workloads due to lower focus, while also being less capable of safe operations during high workloads. Thus imposing a monotonic constraint would be overly restrictive.

We will build the $\boldsymbol{\phi}_i^\top X_i$ term incrementally. The first component is the workload term,

$$\gamma_{w_i,s_i},$$

which models the effect of workload category, allowing this effect to vary by supervisor. Given that the effect is binned by w_i, note that there is no workload multiplier. Then we express the following terms

$$\xi_{s_i}^{\text{dep}}\, M_{\text{dep}}\left(\overline{D}_i\right) \qquad \text{and} \qquad \xi_{s_i}^{\text{arr}}\, M_{\text{arr}}\left(\overline{A}_i\right)$$

which capture the effects of departure rate and arrival rate. These are modeled as monotonic effects through M_{dep} and M_{arr} as broken down below, where $\xi_{s_i}^{\text{dep}}$ and $\xi_{s_i}^{\text{arr}}$ estimate the magnitude of the rate effects, stratified by the presence of a supervisor.

Monotonic effects for ordered arrival and departure rate. Both arrival and departure rates are each binned into $K = 5$ ordered categories. For departure rate we introduce four simplex weights

$$\boldsymbol{\zeta}^{\text{dep}} = \left(\zeta_1^{\text{dep}}, \zeta_2^{\text{dep}}, \zeta_3^{\text{dep}}, \zeta_4^{\text{dep}}\right),$$

with a uniform prior over the simplex,

$$\boldsymbol{\zeta}^{\text{dep}} \sim \text{Dirichlet}(\mathbf{1}_4)\,,$$

where $\mathbf{1}_4$ is the all-ones concentration vector, reflecting no prior belief about which increments are larger.

The monotonic score for departure-rate bin $\overline{D}$ is then

$$M_{\text{dep}}(\overline{D}) = \sum_{k=1}^{\overline{D}-1} \zeta_k^{\text{dep}}, \qquad \overline{D} \in \{1, \ldots, 5\},$$

implementing a cumulative sum of increments prepended by zero. By construction, we have

$$M_{\text{dep}}(1) = 0, \qquad M_{\text{dep}}(5) = \sum_{k=1}^{4} \zeta_k^{\text{dep}} = 1,$$

with intermediate categories placed between 0 and 1. Similarly for arrival rate,

$$\boldsymbol{\zeta}^{\text{arr}} = \left(\zeta_1^{\text{arr}}, \zeta_2^{\text{arr}}, \zeta_3^{\text{arr}}, \zeta_4^{\text{arr}}\right) \sim \text{Dirichlet}(\mathbf{1}_4)\,,$$

and

$$M_{\text{arr}}(\overline{A}) = \sum_{k=1}^{\overline{A}-1} \zeta_k^{\text{arr}}, \qquad \overline{A} \in \{1, \ldots, 5\}.$$

These allow us to capture the arrival and departure rate effects as ordered categories, where the supervisor can reduce the RI risk due to operational rate through the ζ^{arr} and ζ^{dep} variables, respectively.

Prior distributions: The RVs in the model have the following weakly informative priors that are based on the standard normal and exponential distributions whenever possible. The airport intercepts are hierarchical, with

$$\mu_{\text{airport}} \sim \mathcal{N}(-10,\ 1),$$

$$\sigma_{\text{airport}} \sim \text{Exponential}(\lambda = 1),$$

$$b_a^{(\text{airport})} \sim \mathcal{N}\left(\mu_{\text{airport}},\ \sigma_{\text{airport}}\right), \qquad a = 1, \ldots, 40,$$

which makes the assumption that RI rates across the airports are not independent and represent some structural variation in RIs observed across the NAS. The mean of -10 in logit space corresponds to a baseline RI probability of 4.5×10^{-5} per operation, consistent with the overall rarity of incursions. Each airport-specific intercept is partially pooled toward this common mean, with pooling strength governed by σ_{airport}.
The supervisor × day/night × IMC/VMC stratified base-rate intercepts are given independent normal priors

$$b_{s,d,v}^{(\text{sdv})} \sim \mathcal{N}(0,\ 1), \qquad s = \{1, 2\},\ d = \{1, 2\},\ v = \{1, 2\}.$$

The supervisor-stratified rare configuration base-rate intercepts are

$$b_s^{(\text{rar})} \sim \mathcal{N}(0,\ 1), \qquad s = \{1, 2\}.$$

The supervisor-stratified workload effects are

$$\gamma_{w,s} \sim \mathcal{N}(0,\ 1), \qquad w = \{1, \ldots, 5\}, \quad s = \{1, 2\}.$$

The supervisor-stratified scaling coefficients for departure-rate and arrival-rate ordered categorical effects are

$$\xi_s^{\text{dep}}, \xi_s^{\text{arr}} \sim \mathcal{N}(0,\ 1), \qquad s = \{1, 2\}.$$

Finally, the increment weights both receive uniform Dirichlet priors

$$\boldsymbol{\zeta}^{\text{dep}}, \boldsymbol{\zeta}^{\text{arr}} \sim \text{Dirichlet}(\mathbf{1}_4)\,.$$

Model summary: Putting everything together, the RI likelihood model is summarized as

$$RI_i \sim \text{Binomial}(N_i, p_i),$$

$$\text{logit}(p_i) = b_{a_i}^{(\text{airport})} + b_{s_i,d_i,v_i}^{(\text{sdv})} + b_{s_i}^{(\text{rar})}\, r_{c_i} + \gamma_{w_i,s_i} + \xi_{s_i}^{\text{dep}}\, M_{\text{dep}}\left(\overline{D}_i\right) + \xi_{s_i}^{\text{arr}}\, M_{\text{arr}}\left(\overline{A}_i\right).$$

This is a hierarchical binomial GLM for modeling RI rates as seen in Figure 4. The baseline risk varies across airports and across supervisor/daytime/meteorological condition combinations, while the effects of rarity, workload, departure rate, and arrival rate differ based on supervisor availability. The workload effects are assumed to be non-monotonic. The arrival rate and departure rate effects are handled through ordered categorical representation, which preserves ordinal structure while allowing the spacing between categories to be estimated from the data.

C. Limitations

The models make several simplifying assumptions worth noting. The airport throughput model represents arrival and departure throughput with separate RVs ($\alpha^{\text{arr}}_{c,d,v}$, $\alpha^{\text{dep}}_{c,d,v}$, $\sigma^{\text{arr}}_{c,d,v}$, $\sigma^{\text{dep}}_{c,d,v}$), and the RI likelihood model represents the arrival and departure rate effects with separate scaling coefficients $\xi^{\text{arr}}_{s_i}$ and $\xi^{\text{dep}}_{s_i}$. These implicitly assume different efficiencies and safety risks in handling arrivals versus departures in terms of throughput per controller hour, which are important to be able to test our hypotheses. An important simplifying assumption in this representation is the β_i arrival proportion variable, which means that the *controller time input* is scaled by the arrival / departure proportion β_i before being put through the airport throughput model, thereby assuming that controllers split their time equally between arrivals and departures. This is almost certainly not the case in the real world. An alternative formulation we tested attempted to rectify this issue by modeling the proportion of ATC capacity allocated to arrivals versus departures as a latent beta-distributed variable. However, this caused convergence issues for the Markov Chain Monte Carlo (MCMC) across most airports and was replaced by the simpler β_i approach.

The workload metric used in the RI likelihood model is relative to historical observations of staffing and throughput. It cannot account for *taskload* differences across sites. For example, it does not assert that controllers have to do similar types and quantities of tasks at two different sites when the ATC workload value is the same. The workload metric is only sufficient to address the workload-related effects on risk at each site separately.

Multilevel modeling was considered for estimating site-level variance for the various supervisor effects, but two challenges made this impractical. The first and most important was the lack of sufficient data covering the different operational conditions across all of the sites; this made it a requirement to use complete pooling of the supervisor effects across the sites. The second was the difficulty in finding groupings of airports over which to perform partial pooling, to account for variation across airports with specific characteristics. While one could consider grouping airports with specific runway layouts, or specific kinds of independent or dependent operations, we were not able to find a compelling pooling metric within the limited time available for the project.

At a higher level of abstraction, it can be debated whether all important confounders have been included in the study, and whether the confounders in the DAG in Figure 1 have been adequately captured in the models. For example, one subtle design choice was to stratify the base effects of supervisors on risk by environmental conditions, i.e. day/night and IMC/VMC, whereas the same stratification was not performed for the rate-dependent effects. We welcome engagement and critique on these topics. We have done our best to resolve these questions by discussing the models with different subject matter experts, and we have accounted for the confounders most obvious to us using universally accepted causal approaches from [28]. As we will show in Section VI, these models have explanatory power.

VI. Results

We developed both the airport throughput and RI likelihood models in PyMC, a probabilistic modeling library in Python [31], and sampled them using the No U-Turn Sampler (NUTS) [32]. We sampled 4 Markov chains on 4 separate processors to ensure convergence. We used 1000 tuning samples in the MCMC algorithm, and then made 1000 draws from the converged posterior distribution, which were used to generate the results below.

A. Airport throughput model outputs

Given that the throughput model was not the primary aim of the project, we provide some abbreviated results below. Using the posterior outputs of the throughput model, we quantified the probabilistic R^2 of the throughput predictions via the methods in [33]. The R^2 estimates are shown in Figure 5 where the colored bars show the mean R^2, and the error bars in black show its standard deviation. Based on these results, the throughput model can explain a median of 65% of the variation in arrival and departure throughput across the 40 airports considered, which is impressive given that the model considers only staffing and 3 categorical variables as its independent variables.

The model allows us to understand how operational rates vary by runway configuration, IMC/VMC and day/night conditions. Table 2 shows throughput estimates at two example airports and runway configurations during night IMC, where the Highest Density Interval (HDI) gives the bounds of the most likely estimates for each variable. These kinds of metrics can be used in the future for helping inform staffing decisions based on expected conditions and traffic levels. In this study, they are used to inform, in a normalized way, the level of controller workload during a particular historical hour of operations.

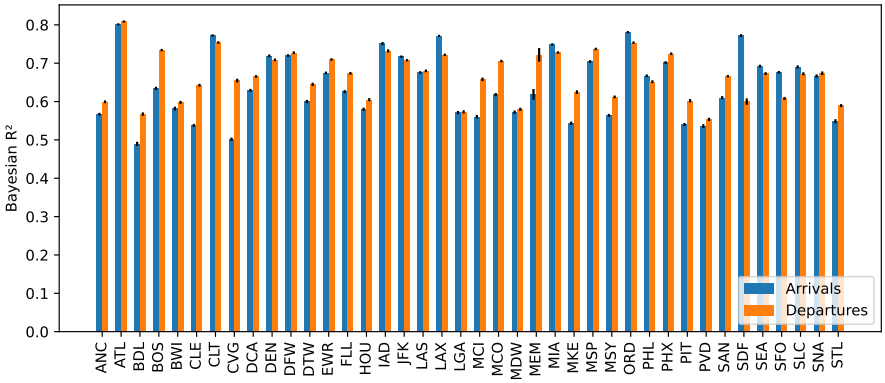


Fig. 5 Bayesian R^2 of airport throughput models for arrivals and departures across 40 ASDE-X or ASSC airports.

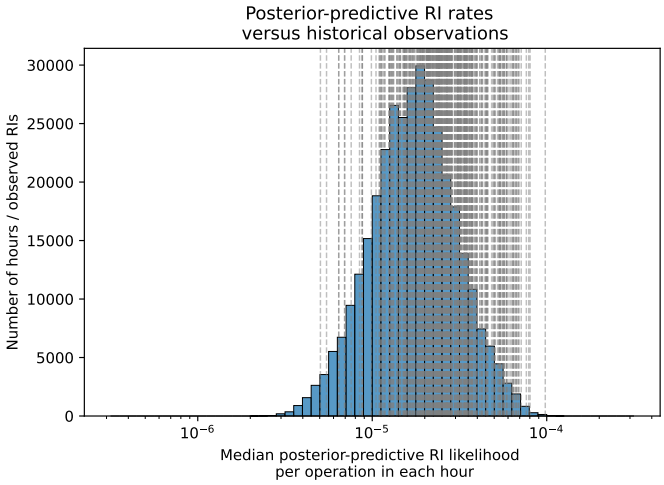
Table 2 Example throughput model outputs: arrivals per hour per controller during night IMC

Airport + Runway Configuration	Mean arrivals/hr/controller				Std. dev. arrivals/hr/controller			
	Mean	Std.	HDI 3%	HDI 97%	Mean	Std.	HDI 3%	HDI 97%
ATL 26R 27L | 26L 27R	17.7	0.6	16.5	18.7	10.3	0.8	8.8	11.7
SAN 27 | 27	13.1	0.2	12.8	13.5	4.9	0.2	4.7	5.2

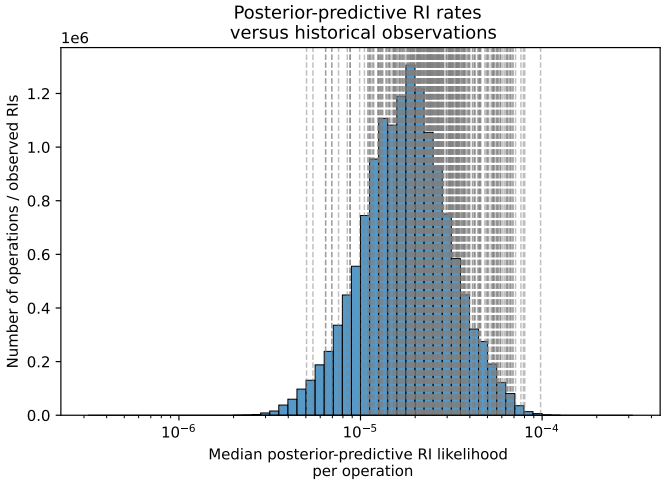
B. RI likelihood model outputs

The RI likelihood model predicts the probability of a RI per operation in each hour, given the sparse observations of RIs. Figure 6 shows the estimated median RI probabilities, aggregated over all hours of operations over all sites. Figure 6a shows the RIs and likelihoods grouped on an hourly basis, and Figure 6b shows the same RIs and likelihoods on a per-operation basis. Based on the model, the median probability of a RI is around 2×10^{-5} RIs per operation. The vertical gray lines show the actual RIs, and the estimated RI probabilities for the hours during which they occurred. In both plots, we expect and see that the RI observations skew to the right, since more event observations are expected during high probability periods. It is important to note that these outputs are *mixtures* of the RI likelihoods across all sites considered, under different runway configurations, IMC/VMC conditions, runway rates, staffing levels, etc. The key question in Section VI.C is whether we can use these likelihoods to identify situations during which a policy intervention (the addition of a dedicated supervisor) can substantially reduce risk.

Figure 7 shows the R^2 of the RI likelihood model, with the 50% and 95% HDIs. The overall R^2 of the model is 0.342, with a 50% HDI of $[0.336, 0.349]$. This means that the model can explain around 34% of the variation in RI rates across the NAS. This is not particularly high given that the model claims to have accounted for the primary causal factors and confounders. However, the results align with the fact that RIs are caused by a combination of factors many of which are not in the model, and can occur due to random chance. The R^2 does vary substantially across the different sites, so it is important to note that this uncertainty in the RI likelihood estimates is accounted for in the hypothesis testing in Section VI.C.



(a) Estimated RI probabilities, grouped on an operational-hour basis.



(b) Estimated RI probabilities, grouped on a per-operation basis.

Fig. 6 Estimated RI probabilities per operation across 40 ASDE-X and ASSC airports.

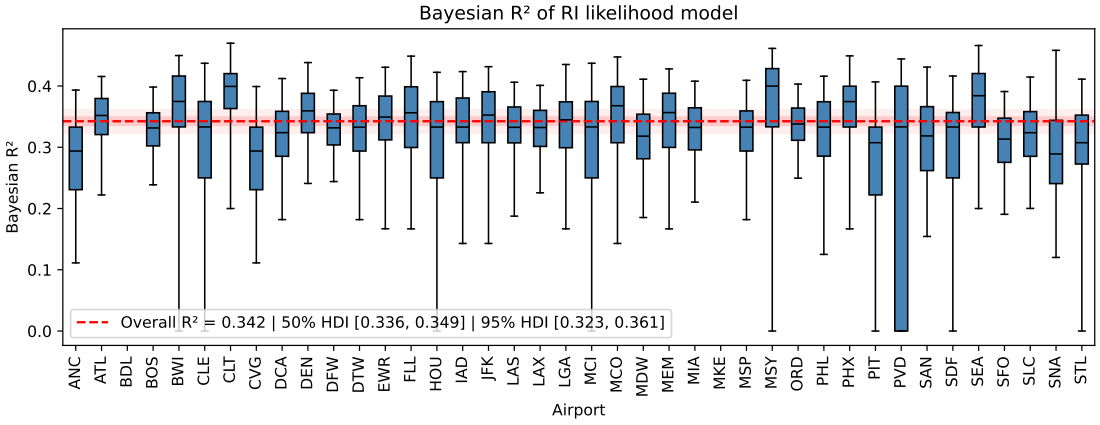


Fig. 7 Bayesian R^2 of the RI likelihood model per airport. Boxplots show the 50% and 95% HDIs.

C. Hypothesis testing

In this section we test whether supervisors have an impact on the RI rate using the posterior predictive estimates from the RI model. While the primary claim is that supervisors help reduce RIs, we consider how the effect of a supervisor can be separated into base-rate and rate-dependent components. When evaluating the impact of any particular variable, the results will be presented in logits, as described in Section V.B. Please refer to Table 1 for the exact impact of a logit change on the odds of a RI. But as a refresher, if the strength of an effect is 1 logit, that means that this effect increases the odds of an event by e, or approximately 2.7. In the following boxplots, the boxes represent the 50% HDI of the estimate, and the whiskers represent the 95% HDI. The outliers are included in a subset of the plots as scatter points outside of these ranges.

1. Base rate effects of a supervisor

The base-rate effect of supervisors is through two pathways, as shown in Figure 1:

- The direct effect of the supervisor on the base rate of RIs at a given airport, stratified by day/night and IMC/VMC conditions, and
- The effect of the supervisor on the base rate mediated by rare configurations.

We first consider the direct effect of supervisors by looking at how differently the three categorical variables in the study (supervisor presence, IMC/VMC and day/night) influence the RI probabilities. These are shown in Figure 8, in 8 boxplots corresponding to the different possible combinations of these binary categories. The blue boxplots show the risk level variation across categories without a supervisor, and the orange boxes show the same risk levels with a supervisor. In general, the consistent trend is for RI risk to increase from left to right for boxes with the same supervisor

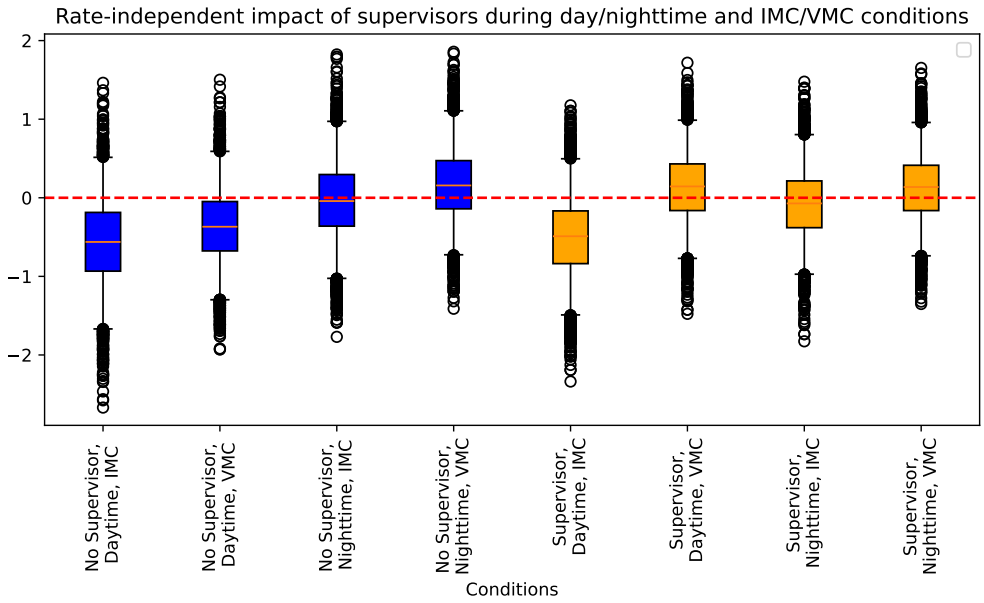


Fig. 8 Impact of categorical variables on the base rate of runway incursions across the airports.

category, with daytime IMC having the lowest relative risk, and nighttime VMC having the highest. By comparing the supervisor figures to no supervisor figures, we can see that supervisor presence has a negligible direct effect on the base rate of RIs under the model. The only exception is under daytime VMC conditions, where the presence of a supervisor is linked to slightly higher RI risk. However, this effect is relatively weak, and given that no effect is seen across the other conditions, the model suggests that the presence of a supervisor does not significantly impact the base rate of RIs directly.

However, we do see a supervisor effect when we consider the impact of rare configurations on the likelihood of a RI, which is shown in Figure 9. Without supervisors, the impact of a rare configuration is a median increase of 0.7 logits in RI likelihood, or approximately a doubling in the probability of a RI. The rare configuration effect is reduced to a median of 0.3 logits if a supervisor is present. This means that, for the rarest configurations (a rarity score of 1), a supervisor

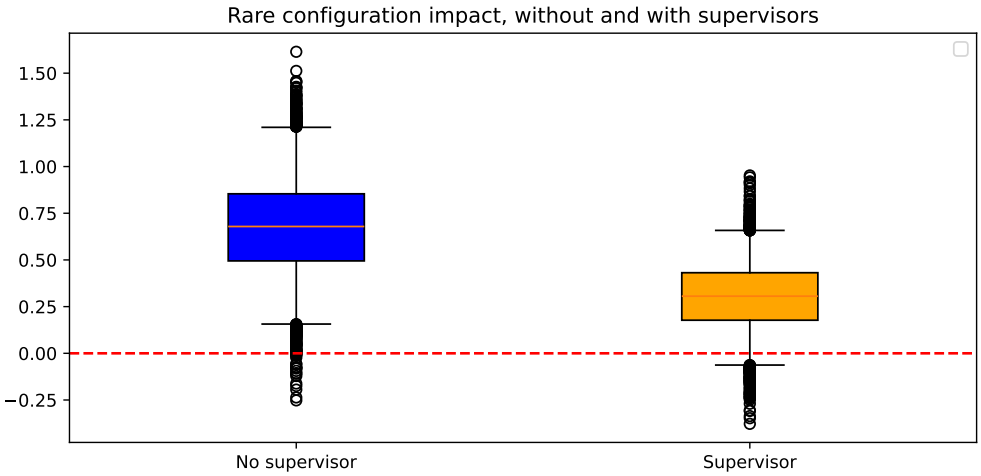


Fig. 9 Impact of rare configurations on RI likelihoods, with and without supervisors.

reduces the likelihood of RIs by a factor of $1 - e^{-0.4} = 0.33$, or 33%. This effect obviously scales with the rarity of the configuration, where no effect would be observed for the most common configurations. But from a policy standpoint, it suggests that supervisor presence meaningfully decreases the risk of RIs under seldom observed runway configurations.

2. Rate-dependent effects of a supervisor

One of our hypotheses was that supervisors may mitigate rate-dependent RI likelihood increases. RIs may be more likely during higher operational rates either due to greater congestion on the taxiways and airspace, or due to greater ATC workload. In the following plots, the effect of increasing workload, and arrival and departure rates are shown, without (blue) and with (orange) a supervisor, considering fixed runway configuration and base rate effects. The effects are in logits as in Section VI.C.1. For the workload plots, we assume a constant arrival and departure rate percentile; thus, increasing workload means that fewer controllers are available to perform the same number of operations. For the arrival and departure rate plots, we assume that the staffing is maintained in proportion to operational load, thus trying to isolate the impact of operational rate on its own.

We first examine the relationship between workload and RI rate in Figure 10, and whether supervisors have any effect on risk. The first finding is that there is not a clear relationship between workload and RI likelihood. Our best explanation is that we are seeing random variation in the data play out, given that the logit impacts of the estimates are low. Additionally, even though there is variation between workload impacts with and without supervisors, the effects are not strong, given that the 50% HDIs of the estimates overlap for all 5 workload categories.

The clearest rate-driven effect on risk is shown in Figure 11 by the relationship between arrival rate and risk. Note, as described in Section V.B, that the lowest arrival rate percentile is missing because it is taken to be the baseline case. This clearly shows that there is a monotonically increasing relationship between arrival rate and RI likelihood, and that supervisors have a weak impact in mitigating risk, showing a less precipitous rise in risk as arrival rate increases. But given the uncertainty in the estimate and the magnitude of the effect, this effect is also possibly an artifact of random variation in the data.

The most unexpected result in the study was in the departure rate and risk relationship, seen in Figure 12. Note that, similar to Figure 11, the lowest percentile is considered to be the baseline and is thus missing. In this particular case, the presence of a supervisor *inverts* the relationship between rate and RI likelihood. When a supervisor is present (orange), increased departure rate weakly increases the likelihood of RIs. However, when there is no supervisor (blue), increased departure rate actually *decreases* the likelihood of a runway incursion.

This is a puzzling result that requires further investigation. There are a few likely explanations for this effect. One possibility is that what we are observing is a real effect, and that supervisors actually increase risk in certain situations. This in itself could be due to different mechanisms. It is conceivable that the presence of a supervisor causes controllers additional stress or distraction in a way that makes them less effective in their duties, especially during departure operations. Another explanation is the one proposed by Schopf et al. [26], where a trusting relationship between the

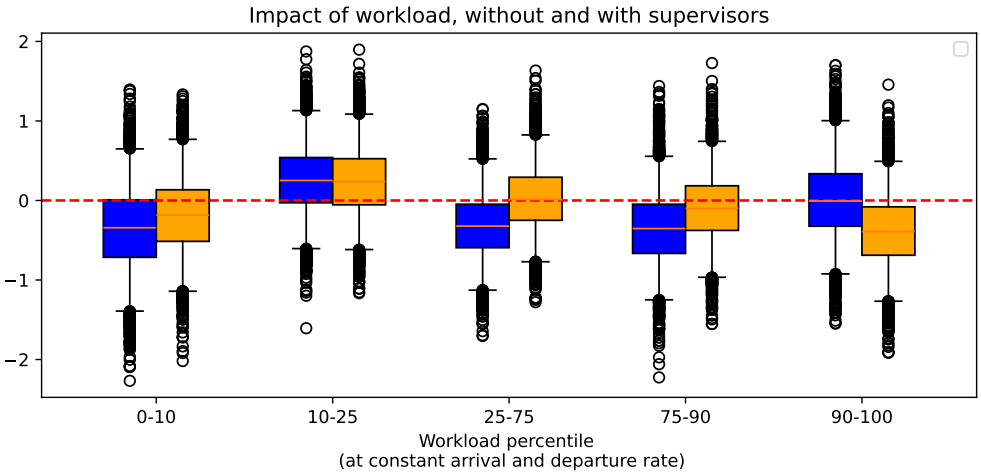


Fig. 10 Impact of higher ATC workload on the runway incursion rate, without (blue) and with (orange) supervisors.

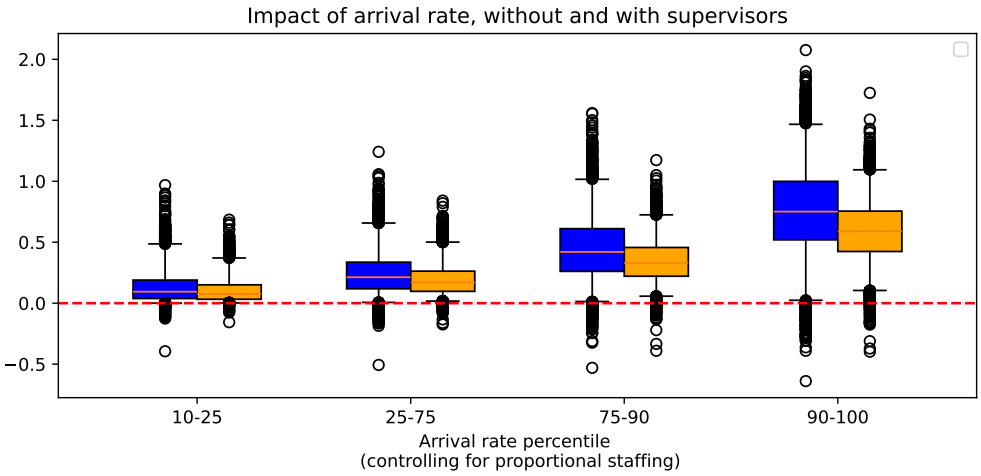


Fig. 11 Impact of arrival rate on the runway incursion rate, without (blue) and with (orange) supervisors.

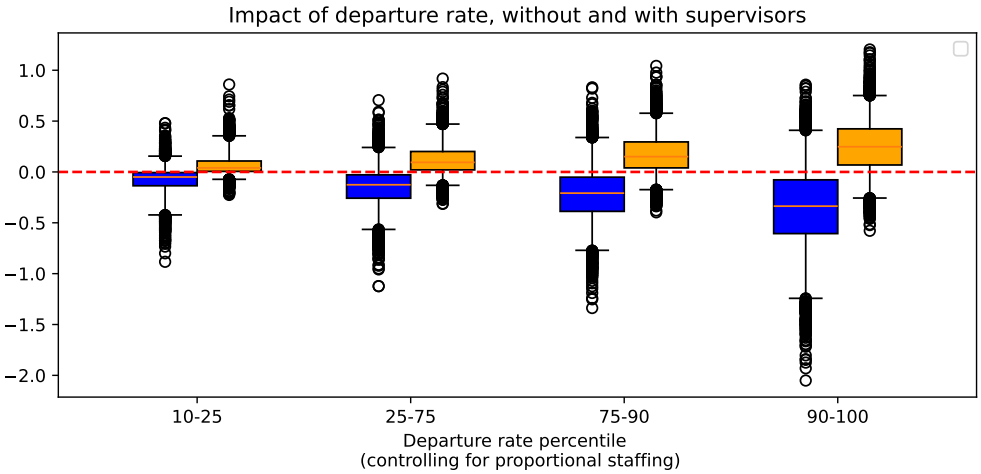


Fig. 12 Impact of departure rate on the runway incursion rate, without (blue) and with (orange) supervisors.

supervisors and on-position controllers can undermine safety culture if there is a belief that the supervisor may intervene to prevent mistakes.

Another explanation is that there is a confounding factor related to supervisor availability and RI risk that we have yet to account for. In this explanation, the confounder is both associated with the presence of a supervisor and higher departure rates, and also substantially increases RI risk. If we fail to account for this confounder, we would find that the presence of the supervisor during high departure rate periods actually increases risk, as observed in Figure 12. It is also possible that, because the effects observed are relatively weak across the board and because of the low data count, this outcome is borne of the inherent stochasticity of RIs rather than any underlying causal relationship. However, the supervisor impact during high departure rates is too strong to ignore, so more work is required to understand the complex risk dynamics present in this paper.

D. Exceptional responders analysis

Given the impacts of supervisors described in Sections VI.C.1 and VI.C.2, the FAA may want to make decisions about when and where it would be beneficial to change supervisor staffing to improve safety. Different sites will have different responses to the presence of a supervisor. We try to amplify this signal by figuring out which sites are *exceptional responders*. These sites have three characteristics: relatively higher levels of RI risk, outsized benefits from supervisors, and low supervisor presence during periods when they are predicted to effectively reduce risk.

To understand where and when these criteria are met, we consider Figure 13, where the orange bars show the availability of a dedicated supervisor, on a per-hour basis, during all operations. For example, we can see that, at Hartsfield-Jackson Atlanta International Airport (ATL), a dedicated supervisor was available during 60% of the operational hours in the study. In contrast, the blue bars show the availability of a supervisor during the top 25% riskiest hours at a given airport, when a supervisor would have decreased the RI risk with at least 75% confidence. During those hours, we see that ATL had a supervisor 43% of the time, lower than the 60% observed across all operations. This is the trend at a majority of sites, where supervisors are less likely to be present during conditions with higher risk, as predicted by the model.

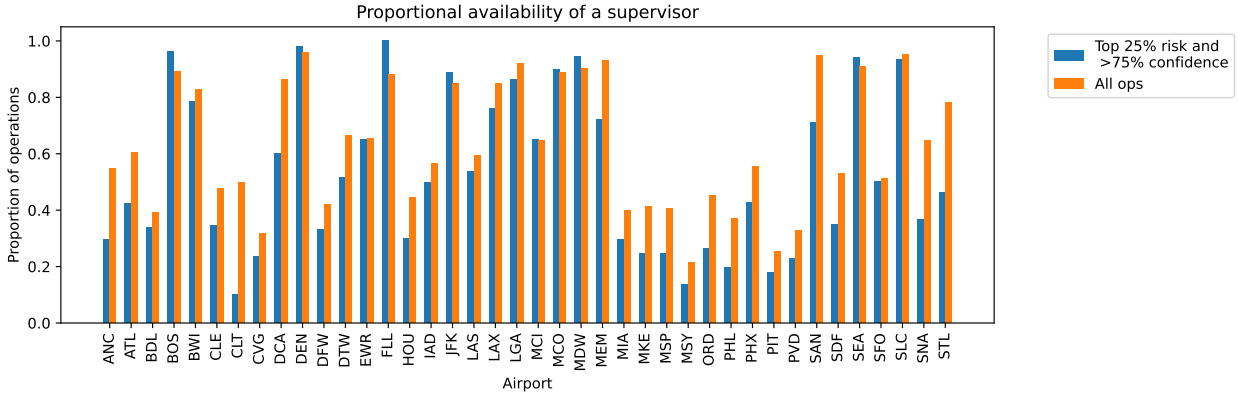


Fig. 13 The availability of a supervisor during operational hours with higher RI risk and higher supervisor benefit (blue), versus supervisor availability during all operations (orange) at the 40 sites studied.

E. Counterfactual supervisor simulation

Now that we have an idea of when supervisors may be effective in reducing risk, we can simulate the addition of supervisors during the aforementioned 25% riskiest hours, when a supervisor has more than a 75% likelihood of reducing risk according to the model. We also restrict our simulation to sites where at least 60 such hours were observed with 30 or more operations per hour, to focus on sites that had consistent estimated risk reductions from a supervisor. Note that 25% and 75% are tunable cutoffs for evaluating high risk and high benefit time periods, and can be changed to adapt to staffing constraints and accommodate different tolerances for risk. However, widening the parameter range (considering lower risk and lower efficacy periods) will reduce the marginal impact of supervisors, so the estimates from this study should be considered to be an upper bound on the benefits of supervisors in reducing RI risk, based on the model assumptions. Additionally, we could have added supervisors during consequential periods by reallocating

them from times when they had potentially lower impact in reducing risk. However, for now we assume that we have the ability to add supervisor hours without impacting other historical time periods when a supervisor was present.

Table 3 Summary of counterfactual supervisor simulation.

Metric	9 airports
Supervisor hours observed in historical data	33,893
Additional supervisor hours simulated	988 (+3%)
RIs observed (14-month period)	87
Simulated decrease in RIs	Median 1.0 (-1%)
	50% HDI [0.3, 1.9] (0% to -2%)

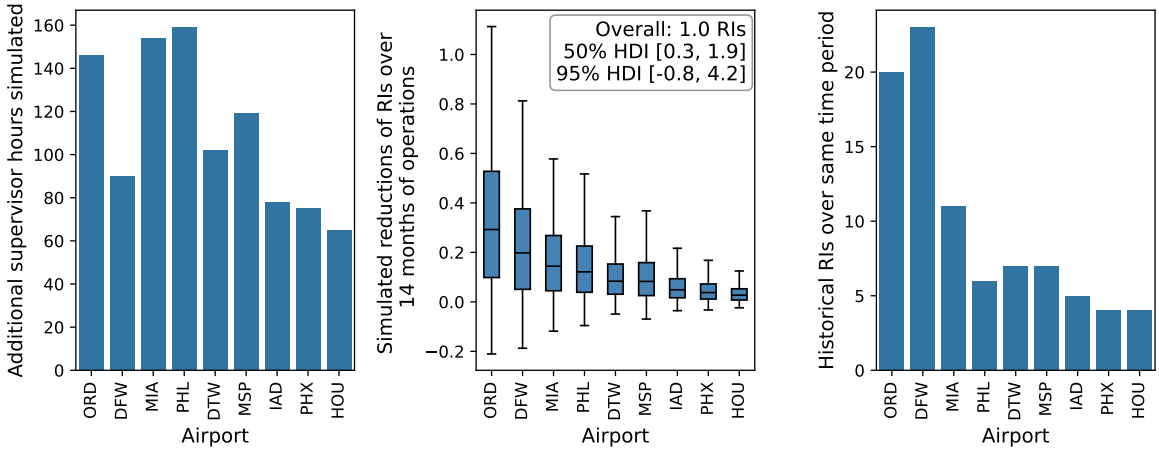


Fig. 14 Results of counterfactual simulations of supervisors at 9 exceptional responders, with boxplots showing the median, and the 50% and 95% HDI responses.

In Table 3, we show the results of the counterfactual study. At a high level, a 3% marginal increase in supervisor hours results in a posterior median decrease of approximately 1% in RIs, with a 50% HDI of 0 to 2%. The large uncertainty in these results can be attributed to the low sample count, suggesting that RIs are difficult events to study even with Bayesian approaches. We report the results on a per-airport basis, with the median, the 50%, and 95% HDIs in Figure 14. In addition, we show the estimated reductions in RI probability per operation across the sites in Figure 15, where the presence of a supervisor is estimated to reduce the risk by a median of 20% to 40%, depending on the airport.

The primary takeaway is that, though supervisors can help reduce RI rates under certain conditions, these conditions are relatively rare, and thus changes to supervisor availability alone will likely have a relatively small impact on RI rates. This is due to a combination of factors. According to the model, supervisors tend to reduce RIs most during rare configurations, which by definition do not occur often. Additionally, supervisors may be expected to have greater impacts when a set of specific conditions is met. However, for any single condition that would be expected to cause higher risk according to this particular model, the strength of the safety effect is not strong.

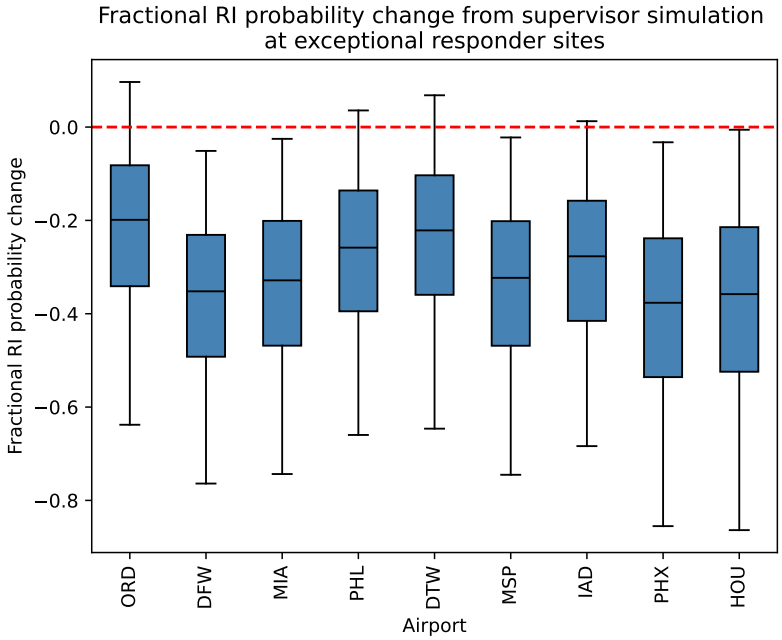


Fig. 15 Simulated RI probability reductions due to supervisor presence at 9 exceptional responder sites, with boxplots showing the median, and the 50% and 95% HDI responses.

VII. Conclusion

The research in this paper takes an important step toward modeling how ATC staffing practices impact safety risk in the NAS, and offering policy prescriptions for improving safety. The understaffing of ATC facilities across the NAS is a well-known phenomenon, but research quantitatively linking understaffing to safety is limited. We need to continue to use state-of-the-art tools to inform the discussion of how, where, and when staffing changes can help improve the efficiency and safety of operations in the NAS.

This paper has presented one powerful approach, Bayesian causal inference, for simulating counterfactuals and evaluating policy prescriptions in the absence of the ability to run randomized experiments. While computationally and technically more complex than a simpler predictive approach, the approach gives insights that would otherwise be out of reach due to its ability to account for confounders. As a policy prescription, we tested whether the presence of a supervisor in an ATC tower helps reduce the likelihood of RIs at the associated airport. In summary, we found a weak overall relationship between supervisor presence and RI risk. The strongest evidence of benefits was found when supervisors were present during rare runway configurations. The counterfactual simulation showed that a 3% marginal increase in supervisor-hours at the exceptional responder sites and hours of operation is estimated to result in a 1% reduction in incursion rates. From another perspective, supervisors are estimated to reduce RI probability by a median of 20% to 40% during time periods when the model predicts they are most effective at promoting safety. This relatively small impact is due to the rarity of the conditions under which supervisors reduce the risk of RIs, according to the model.

It is important to note that this work is a proof of concept, and that the results and claims must be taken in that context. We believe our hypotheses are reasonable, and that our model captures the known confounders in how supervisor staffing may impact RI likelihoods. However, this is a single causal model that considers a single kind of encounter and one specific policy intervention. We hope this work demonstrates the need to continue to study how ATC staffing impacts safety, or how different factors may be causally linked to incident rates. Given our narrow focus on a single kind of risky encounter, it is key that this kind of analysis be expanded to consider other kinds of encounters so we can better understand the risk landscape in the NAS. We hope that this work motivates the use of causal modeling in policymaking, especially towards improving safety.

VIII. Future work

There are some direct extensions of this study to improve the findings on how supervisors impact RI risk. We may want to study whether the impact of supervisors is sensitive to whether RIs due to pilot deviation are omitted from the study. This would come with several challenges, like further reducing sample size and leading to some subjectivity due to the judgment call that is whether or not a RI was a pilot error. However, this may clarify the supervisor picture because it is reasonable to think that pilot deviations are somewhat independent of staffing and/or the presence of a supervisor. Another possible angle to consider is whether and how much operational dependencies, such as dependent arrival and/or departure operations or intersecting runways, impact the rates of RIs, as suggested by Omosebi et al. [19]. While some of these dependencies are captured in the runway configuration variable, it is possible that the same configuration is operated differently under different staffing or weather conditions, making this an interesting but complex line of inquiry. In the context of the policy prescription, we may want to consider a case where, instead of adding supervisors, we reallocate supervisors from hours with the least impact to those with the greatest impact with respect to RIs. This would allow us to more realistically consider the human capital constraints that ATC towers may be facing.

Looking at the bigger picture of safety in the NAS, the study requires extension to other kinds of risky incidents. In hindsight, RIs are complex and rare encounters and it is possible that the signal-to-noise ratio in our study was weak due to this complexity and the low sample counts. NMACs or loss of standard separation events are good candidates for other incidents to study. It would also help to extend this work to study incident precursors, such as operational errors, miscommunications, or Traffic Collision Avoidance System (TCAS) traffic or resolution advisories. Studying precursors would enable one major improvement; due to greater sample counts, they may enable partial pooling to better understand how the efficacy of different policy prescriptions varies across different sites. Then, they can be linked to actual incident rates to see how observations of precursors are linked to incident rates.

In terms of staffing policy prescriptions, our framework can be used to study other relevant staffing factors, such as the chronic use of overtime and the combination of controller positions. The same causal modeling approaches can also be used to study non-staffing factors that may be related to safety, such as the availability of certain technologies and the use of specific kinds of operations.

IX. Acknowledgements

We thank the following MITRE collaborators: Jeff Shepley and Dr. Joe Hoffman for their help in the data fusion and analysis; Toby Jones, for helping us understand the supervisor role and validate hypotheses with his expertise in ATC tower operations; Karl Meyer, for his mentorship, and his championing of this research and additional efforts for improved safety; Dr. Craig Wanke, Dr. Anthony Santago II, Jason Reinhart and Dennis Sawyer for their support for this research effort and their reviews, which elevated the quality of this work.

X. Notice

References

[1] Federal Aviation Administration, "Runway Incursions," , 2026. URL `https://www.faa.gov/airports/runway_safety/resources/runway_incursions`, accessed April 13, 2026.

[2] Library of Congress, "Collision at LaGuardia Airport Spotlights Long-standing Concerns Over Runway Safety," , 2026. URL `https://www.congress.gov/crs-product/IN12675`, accessed April 13, 2026.

[3] National Transportation Safety Board, "Aviation Investigation Preliminary Report," Tech. Rep. DCA26MA161, National Transportation Safety Board, Washington, DC, April 2026. Preliminary Report, Accident Date: March 22, 2026. Location: LaGuardia Airport, New York, NY.

[4] Arel, T. L., "Addressing Close Calls to Improve Aviation Safety," Statement before the U.S. Senate Committee on Commerce, Science, and Transportation, Subcommittee on Aviation Safety, Operations, and Innovation, November 2023. URL `https://www.transportation.gov/addressing-close-calls-improve-aviation-safety`, chief Operating Officer, Air Traffic Organization, Federal Aviation Administration. Published on the U.S. Department of Transportation Testimony Documents page. Accessed April 20, 2026.

[5] Netjasov, F., and Janic, M., "A review of research on risk and safety modelling in civil aviation," *Journal of Air Transport Management*, Vol. 14, No. 4, 2008, pp. 213–220. https://doi.org/10.1016/j.jairtraman.2008.04.008.

[6] Ale, B. J. M., Bellamy, L. J., van der Boom, R., Cooper, J., Cooke, R. M., Goossens, L. H. J., Hale, A. R., Kurowicka, D., Morales, O., Roelen, A. L. C., and Spouge, J., "Further development of a Causal model for Air Transport Safety (CATS): Building the mathematical heart," *Reliability Engineering & System Safety*, Vol. 94, No. 9, 2009, pp. 1433–1441. https://doi.org/10.1016/j.ress.2009.02.024.

[7] Borener, S., Trajkov, S., and Balakrishna, P., "Design and development of an integrated safety assessment model for NextGen," *International Annual Conference of the American Society for Engineering Management*, 2012, pp. 5–9.

[8] Nikdel, S., Noh, S., and Shortle, J., "Common cause failure analysis for aviation safety assessment models," *2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC)*, IEEE, 2021, pp. 1–10.

[9] Ud-Din, S., and Yoon, Y., "Analysis of Loss of Control Parameters for Aircraft Maneuvering in General Aviation," *Journal of Advanced Transportation*, Vol. 2018, No. 1, 2018, p. 7865362. https://doi.org/https://doi.org/10.1155/2018/7865362, URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2018/7865362.

[10] Valdés, R. M. A., Comendador, V. F. G., Sanz, L. P., and Sanz, A. R., "Prediction of aircraft safety incidents using Bayesian inference and hierarchical structures," *Safety science*, Vol. 104, 2018, pp. 216–230.

[11] Ziegler Haselein, B., da Silva, J. C., and Hooey, B. L., "Multiple machine learning modeling on near mid-air collisions: An approach towards probabilistic reasoning," *Reliability Engineering & System Safety*, Vol. 244, 2024, p. 109915. https://doi.org/https://doi.org/10.1016/j.ress.2023.109915, URL https://www.sciencedirect.com/science/article/pii/S0951832023008293.

[12] Ayra, E. S., Rios Insua, D., and Cano, J., "Bayesian network for managing runway overruns in aviation safety," *Journal of aerospace information systems*, Vol. 16, No. 12, 2019, pp. 546–558.

[13] Federal Aviation Administration, "Runway Safety Statistics," , 2026. URL https://www.faa.gov/airports/runway_safety/statistics, accessed April 28, 2026.

[14] Ison, D. C., "Empirical Analysis of Trends in Runway Incursions in the United States," *Journal of Aviation Technology and Engineering*, Vol. 9, No. 1, 2020. URL https://docs.lib.purdue.edu/jate/vol9/iss1/1/.

[15] Aero-News Network, "FAA Adopts ICAO Definition For Runway Incursions," , October 2007. URL https://www.aero-news.net/index.cfm?do=main.textpost&id=daa0d4c6-1920-4781-a8a9-933450ed4369, accessed April 28, 2026.

[16] Knott, B., Gannon, A., and Rench, M., "Runway Incursion: Human Factors In Runway Incursions," Tech. Rep. ADA402929, Crew System Ergonomics Information Analysis Center (CSERIAC), Wright-Patterson AFB, OH, June 2000. URL https://apps.dtic.mil/sti/pdfs/ADA402929.pdf, final Report. Approved for public release.

[17] Kelley, D. R., and Adam, G. L., "The human factors of runway incursions caused by'pilot error'- A survey of U. S. airline pilots," *International Symposium on Aviation Psychology, 9 th, Columbus, OH*, 1997, pp. 911–917.

[18] Bhargava, D., and Marais, K., "Narrative Analysis of Runway Incursion Reports in the National Transportation Safety Board Database To Identify Contributing," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 64, SAGE Publications Sage CA: Los Angeles, CA, 2020, pp. 144–148.

[19] Omosebi, O., Azimi, M., Olowokere, D., Wanyan, Y., Zhao, Q., and Qi, Y., "Investigating Runway Incursion Incidents at United States Airports," *Future Transportation*, Vol. 3, No. 4, 2023, pp. 1209–1222. https://doi.org/10.3390/futuretransp3040066.

[20] USAFacts, "Is Flying Safer Than Driving?" https://usafacts.org/articles/is-flying-safer-than-driving/, 2026. Accessed April 30, 2026.

[21] USAFacts, "Is There a Shortage of Air Traffic Controllers?" https://usafacts.org/articles/is-there-a-shortage-of-air-traffic-controllers/, 2025. Accessed April 30, 2026.

[22] Federal Aviation Administration, "Air Traffic Controller Workforce Plan (2025-2028)," Tech. rep., Federal Aviation Administration, 2025. URL https://www.faa.gov/sites/faa.gov/files/fy25-air-traffic-controller-workforce-plan_0.pdf, accessed April 30, 2026.

[23] National Academies of Sciences, Engineering, and Medicine, *The Air Traffic Controller Workforce Imperative: Staffing Models and Their Implementation to Ensure Safe and Efficient Airspace Operations*, The National Academies Press, Washington, DC, 2025. https://doi.org/10.17226/29112, URL https://www.nationalacademies.org/read/29112.

[24] U.S. Government Accountability Office, "Air Traffic Control: FAA Enhanced the Controller-in-Charge Program, but More Comprehensive Evaluation Is Needed," Tech. Rep. GAO-02-55, U.S. Government Accountability Office, 2001. URL https://www.gao.gov/products/gao-02-55.

[25] Schroeder, D. J., Bailey, L. L., Pounds, J. C., and Manning, C. A., "A Human Factors Review of the Operational Error Literature," Tech. Rep. DOT/FAA/AM-06/21, Federal Aviation Administration, Civil Aerospace Medical Institute, Oklahoma City, OK, 2006.

[26] Schopf, A. K., Stouten, J., and Schaufeli, W. B., "The role of leadership in air traffic safety employees' safety behavior," *Safety Science*, Vol. 135, 2021, p. 105118. https://doi.org/10.1016/j.ssci.2020.105118.

[27] Pearl, J., *Causality: Models, Reasoning, and Inference*, 2nd ed., Cambridge University Press, 2009.

[28] McElreath, R., *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, 2nd ed., CRC Press, 2020.

[29] Textor, J., van der Zander, B., Gilthorpe, M. K., Liskiewicz, M., and Ellison, G. T. H., "Robust causal inference using directed acyclic graphs: the R package 'dagitty'," *International Journal of Epidemiology*, Vol. 45, No. 6, 2016, pp. 1887–1894.

[30] Eckstein, A., Kurcz, C., and Silva, M., "Threaded Track: Geospatial Data Fusion for Aircraft Flight Trajectories," Tech. Rep. Product 10-2.2-1, The MITRE Corporation, August 2012. URL https://apps.dtic.mil/sti/pdfs/AD1107970.pdf, approved for Public Release; Distribution Unlimited. Case Number 17-3649.

[31] Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesbeck, C. J., Kochurov, M., Kumar, R., Lao, J., Luhmann, C. C., Martin, O. A., Osthege, M., Vieira, R., Wiecki, T., and Zinkov, R., "PyMC: A Modern and Comprehensive Probabilistic Programming Framework in Python," *PeerJ Computer Science*, Vol. 9, No. e1516, 2023. https://doi.org/10.7717/peerj-cs.1516.

[32] Hoffman, M. D., Gelman, A., et al., "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.*, Vol. 15, No. 1, 2014, pp. 1593–1623.

[33] Gelman, A., Goodrich, B., Gabry, J., and Vehtari, A., "R-squared for Bayesian Regression Models," *The American Statistician*, Vol. 73, No. 3, 2019, pp. 307–309. https://doi.org/10.1080/00031305.2018.1549100.