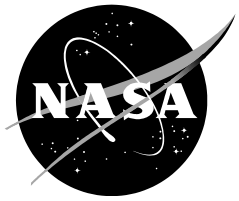
NASA/TM-20220019263



Assessing Several Non-Traditional Data Sources for Value in Aviation Safety

Nikunj C. Oza
NASA Ames Research Center

Chad Stephens
NASA Langley Research Center

Fasil Alemante
Booz Allen Hamilton
NASA Langley Research Center

Rami Houssami
Booz Allen Hamilton
NASA Langley Research Center

Shannon Walker
Booz Allen Hamilton
NASA Langley Research Center

Immanuel Barshi
NASA Ames Research Center

Stephen Casner
NASA Ames Research Center

Ilya Avrekh
KBR, Inc.
NASA Ames Research Center

Erin Flynn-Evans
NASA Ames Research Center

Lucia Arsintescu
San Jose State University
NASA Ames Research Center

Rachel Jansen
NASA Ames Research Center

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, CA 94035-1000

December 2022

Acknowledgments

This report is available in electronic form at
http://

Abstract: The NASA System-Wide Safety (SWS) project and its predecessor projects have been developing Machine Learning (ML) algorithms for commercial aviation safety for many years. These algorithms have been applied to Flight Operations Quality Assurance (FOQA); radar track data (e.g., Threaded Track); and safety reports, including Aviation Safety Reporting System (ASRS) and Aviation Safety Action Plan (ASAP). SWS is working with partners to get access to other data that air carriers provide, such as maintenance data, and has been assisting carriers in working with other data, such as Line Operations Safety Audit (LOSA) data, using manual methods. However, the project has discussed whether there are other data that are not traditionally used in aviation safety analysis that may be useful. This paper discusses four sets of data and models that are not traditionally used in aviation safety but that have shown promise for such use. In the future, we plan to incorporate such data into ML algorithms to use with data that we have used before and determine the additional benefit that is actually achieved under different contexts from the inclusion of these non-traditional data sources.

Introduction

The NASA SWS project and its predecessor projects have developed several Machine Learning algorithms to solve two primary problems---anomaly detection and precursor identification. Anomaly detection is used to perform vulnerability discovery---finding statistical anomalies, of which some may be safety issues that have not been previously identified and characterized in the form of exceedances. Precursor identification looks for precursors, that may possibly represent causes, of safety issues (e.g., initiating descent late may be a precursor to a high-speed exceedance). For these problems, we have worked with FOQA data, radar track data, and safety reports from ASRS and ASAP. For assessing human performance and fatigue, SWS has used data from wearable sensors and a mobile application designed to assess the user's fatigue (PVT+). However, we hypothesized that there are other data sources that have not traditionally been used to assess aviation safety that could turn out to be useful. In 2018, we briefly explored the potential usefulness of pilot blogs for aviation safety. In some preliminary work, we observed spikes in blog posts the day after any incident. We were motivated to explore pilot blogs by the thought that their posts could be written in response to changes in regulations or procedures, such as to complain about the procedures or discuss difficulties in handling the procedures and could therefore be predictive of close calls or incidents.

Based on the preliminary result obtained on pilot blog data, we decided to explore other data sources in the context of problems on which we were already working. To that end, we describe four threads of work in this report:

1. The use of ASRS and associated information, together with traditional/objective data such as SWIM and FOQA, to obtain information beyond what each one could provide on its own.
2. Work done so far on assessing whether the Boeing Alertness Model (BAM) can be used to predict flights that are at increased risk for performance issues due to pilot fatigue. The goal is to make changes to legal operations that reduce the risk of fatigue while maintaining efficiency.
3. Initial explorations of how to use crowdsourced ADS-B data to assess stability of flight approach and landing.

4. An example of insights that a human factors expert can provide on an ASAP report, some key questions about the ASAP Event Review Committee (ERC) debrief process, and what additional value can be gained from the ASAP reports and ERC debrief information.

Use of ASRS and Associated Information

Taken from the ASRS website: "ASRS captures confidential reports, analyzes the resulting aviation safety data, and disseminates vital information to the aviation community."[1] Reports are submitted anonymously through electronic report submission on the website or by mail. Five different report forms are available, depending on the role of the submitter.

The anonymous nature of ASRS assures that details such as tail number, location, exact time, and the specifics of the reporter are unavailable. Published databases include only descriptions of the chain of events that led to the filing of the safety report, which are themselves synopses of the original report. This lack of identifying features presents a challenge for joining the data with other traditional data sources, such as System Wide Information Management (SWIM). However, the data do present information that could be useful for analyzing aviation safety trends.

Natural Language Processing

Since the main content of published ASRS reports is prose written by humans, Natural Language Processing (NLP) may be used to analyze the reports. The existing ASRS categories offer a limited number of classes, such as altitude deviations or aircraft equipment problems. The human written free text forms in the reports, though unstructured, offer rich descriptions of the event that may not fall ideally within these categories. Topic modelling, clustering, and other techniques can provide deeper insight beyond the high-level fixed categories available to reporters. To start, a quick analysis of co-occurring keywords (e.g. Latent Dirichlet Allocation) might reveal some unique topics and risk factors. Word association and stacked embedding models may further reveal compound word correlations suggestive of deeper themes present in the data. Further insights on risk factors could also be gleaned by generating report embeddings (doc2vec[2], fasttext[3]), projecting them into lower dimensional space (UMAP[4]), and clustering (HDBscan[5]). Becky Hooey, ASRS director, expressed interest in NLP when the team discussed it with her.

---

[1] https://asrs.arc.nasa.gov/index.html

[2] NLP model based on word2vec used for generating numerical representations of documents.

[3] library used for learning word embeddings and text classification

[4] Uniform Manifold Approximation and Projection – a projection technique used for dimension reduction

[5] extension of DBSCAN – clustering algorithm

Augmenting with Google Analytics (GA) Data

Though identifying information is not stored in the ASRS databases, there is summarized usage data available through the government's Google Analytics (GA) account (also known as Universal Federated Analytics or Digital Analytics Program (DAP)). This data could be invaluable at providing a new independent variable to augment the SWIM dataset, as it would present how frequented the ASRS service was at any given time. With the right setup, GA extends an application programming interface (API) with available Java and Python libraries[6]. This means it can potentially be integrated into another tool, such as Runtime for Airspace Concept Evaluation's (RACE) actor framework. GA data also presents an opportunity for analyzing user-level analytics, which would inform whether ASRS reports are primarily filed often by a small group of reporters or infrequently by a diverse group of reporters. The data could also serve as a proxy for total report counts.

Statistical Relevance for GA data

In addition to aggregation and summary statistics, the GA data can be used for powering relevant statistical tests, for example, to check for seasonality in reporting. If seasonality does not exist, a simple t-test could determine whether there is a significant increase in number of safety incidents reported. Conversely, if the data is highly seasonal, we could use a Chow breakpoint test[7] to determine if there is a change in number of safety incidents (breakpoint). Combining GA data with NLP could also yield further insights into common failure modes and their incidences over time.

Analyzing Server Logs

As a fallback, if GA data turns out to not be usable, we could take advantage of parsing and analyzing server logs to generate similar statistics. This is somewhat of a last-ditch strategy, as it would be time consuming compared to taking advantage of an existing solution.

Map / Informatics

There could be value in displaying historical incident information on a map display for dispatchers to use. Dispatchers could use this historical information to see what sorts of incidents are frequently reported in specific airspace or locations. This display could take the form of caution flags, markers, or tooltips that the dispatcher could toggle on and off.

---

[6] https://developers.google.com/analytics/devguides/reporting/core/v4

[7] statistical test for determining change in trend lines at a (hypothesized as significant) a priori determined point in time

Use of Alertness Model for Crew Fatigue Predictions

Proficient pilot performance is central to the safe operation of an aircraft. However, due to the demand for 24-hour aviation operations, long-, and ultra-long-haul flights, pilots are often scheduled for extended and irregular work hours, with varying workload (Flynn-Evans et al., 2018, Bourgeois-Bougrine et al., 2003, Arsintescu et al., 2020). Such operations can reduce pilot performance, introducing a vulnerability in the operation that has the potential to interact with other factors to cause an incident or accident. There is a need for tools and technology that provide in-time information that maximizes safety, while also maintaining efficiency. Such tools must be scalable to be effective at the level of the national airspace. They must also enable in-time mitigation strategies to reduce risk when identified.

There have been many studies conducted to identify the causes and consequences of pilot performance impairment (Flynn-Evans et al., 2018, Arsintescu et al., 2021, Young, 2008, Endsley, 1999, Boril et al., 2020, Hitchcock et al., 2010, Honn et al., 2016), but passive tools to monitor pilot performance changes remain challenging to implement at scale. While it is important to continue to explore objective indicators of pilot performance impairment, exploring alternative, non-traditional methods for assessing changes in pilot performance during flight operations will allow us to scale up these analyses. There are several biomathematical models to predict pilot alertness, fatigue, and performance that show promise as tools that could be integrated into broader risk assessment tools. For biomathematical models to be useful in risk assessment, they must be 1) scalable and 2) they must be validated in operational environments against objective data to support their widespread use and integration with other risk factors.

The mitigations that have typically been employed to counter performance impairment in pilots have not allowed for in-time correction during a flight or even within a day. Historically, such mitigations have involved duty-hour restrictions or augmentation of certain types of flights. While these regulations have been a critical component of risk management, the nature of aviation operations still results in some flights where pilots experience performance degradation. In addition, some pilot schedules are inefficient, with pilots who are eligible for duty sidelined due to sub-optimal scheduling procedures.

Optimizing pilot schedules to sustain pilot performance, while also maintaining efficiency has the potential to not only provide in-time safety information, but also has the potential to introduce significant cost savings to airlines.

This work seeks to achieve three primary aims:

1. Evaluate whether the Boeing Alertness Model (BAM; [Åkerstedt et al., 2004, 2007])
   can be scaled and integrated into flight schedules at the level of all flights that enter an entire country or region.
2. Validate biomathematical models against objective inflight data.
3. Use the scaled BAM to identify legal operations that are associated with reduced pilot alertness and elevated inefficiency in order to introduce best practices that minimize pilot fatigue without sacrificing scheduling efficiency.

Methods

Jeppesen has developed a tool (Concert) that uses actual flights to simulate pilot schedules based on rules governing work and rest times (e.g., Federal Aviation Regulations [FAR, 14 CFR Parts 117, 119, and 121, 2009]) to estimate changes in pilot alertness. This tool can be used to identify legal schedules that introduce risk of pilot alertness degradation in order to introduce mitigations.

Files from OAG that track commercial flights are fed into Jeppesen's pilot scheduling software which generates pilot schedules based on the actual flights flown and makes alertness predictions using BAM for each pilot and flight. We are then able to observe how introducing and taking away constraints impacts alertness/fatigue predictions. More specifically, the planned flight timetables currently loaded into the Concert platform include all flights that were operated around the world for an entire week ($n \approx 3$ million flights). We have access to this data for three separate weeks, including one before the pandemic and one in 2021. These files are imported to the Jeppesen crew pairing software to generate realistic schedules using regional regulations as a guide. These schedules are then imported to the Concert platform where BAM is implemented to generate biomathematical models of alertness which is all accessible in a separate platform (Qlik Sense) that enables quick and user-friendly data visualization and analytics.

This analytics tool contains many variables, from airline-specific information (e.g., airline, aircraft types, trips, working periods, departure and arrival destination duty duration, etc.) to regional information (e.g., destination and departure traffic density). The Qlik platform uses the Karolinska Sleepiness Scale [KSS, Åkerstedt and Gillberg, 1991] alertness ratings as well as Absolute Fatigue Risk (AFR) which represents the likelihood of a fatigue-related incident occurring calculated based on the KSS. This has then been averaged to calculate Normalized Fatigue Risk (NFR), which is our primary fatigue-related outcome.

Work in Progress

We have generated BAM alertness predictions for all of the flights available in the database (more than one million flights). We are currently reviewing the schedules and predictions to identify anomalies. Preliminary visualization suggests that we can identify pilots and flights that are vulnerable to pilot performance impairment. Furthermore, it appears that assessing the data in this manner will allow us to identify regions that may pose an elevated safety risk due to reduced pilot alertness (Figure 1).

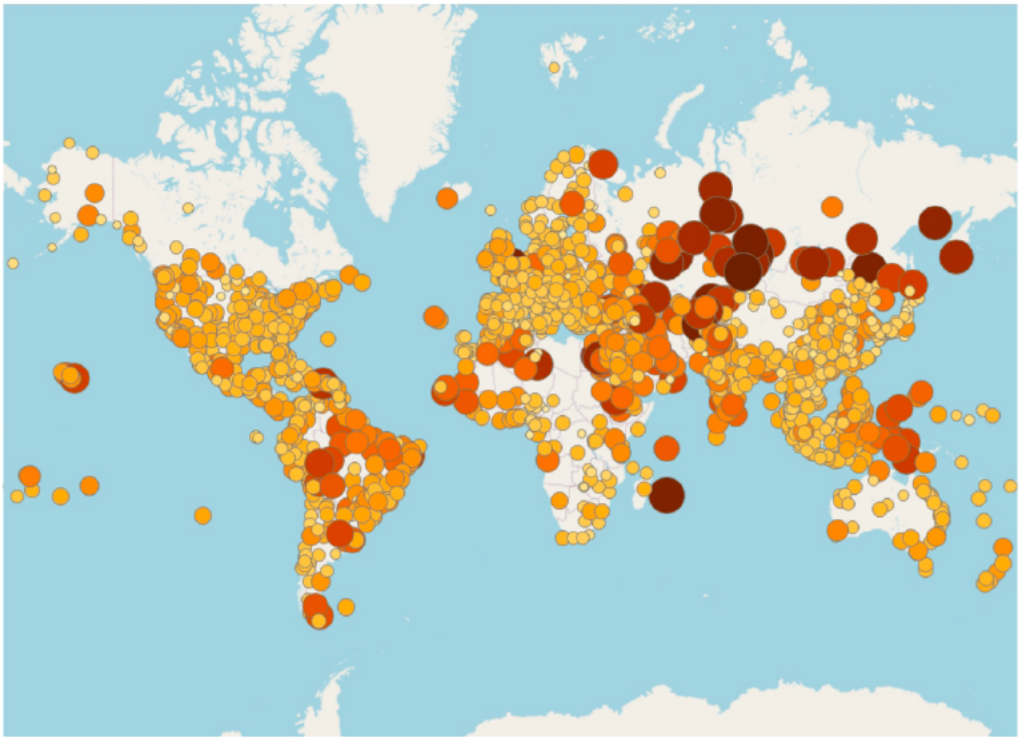


Figure 1: Map showing average modeled alertness (CAS) by arrival location. Darker, larger red dots show worse predicted performance, smaller yellow dots show better predicted performance.

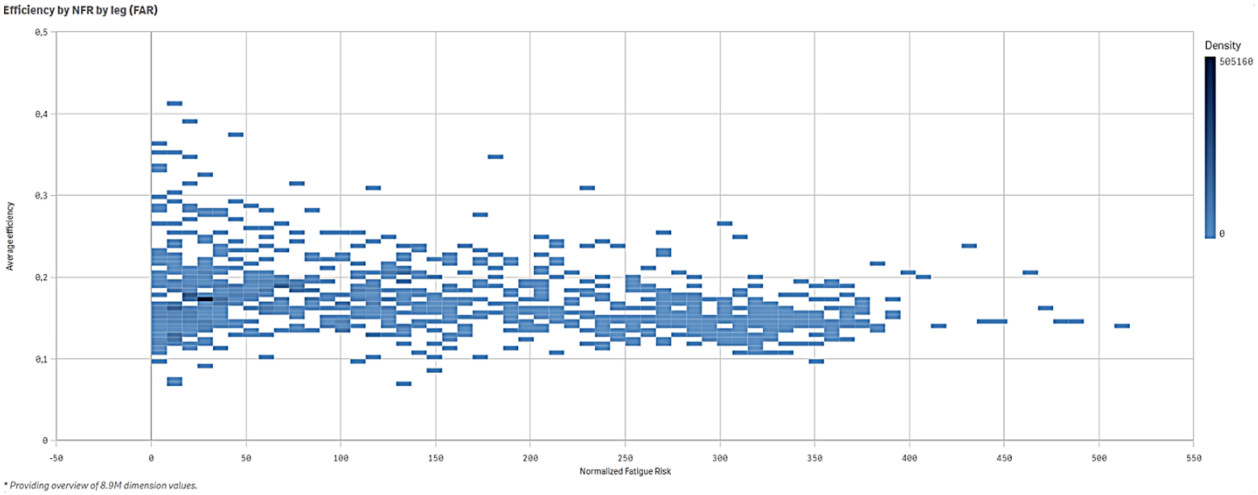


Figure 2: Scatter plot of average efficiency as a function of fatigue risk for each individual leg under FAR

We are currently collecting KSS and performance data from short-haul pilots to validate against the BAM predictions. We have previously collected KSS and performance data from long-haul
pilots to compare against BAM predictions, and we aim to initiate additional studies for this
purpose in the future (we have experienced delays due to the pandemic). We have started to identify flight operations that are both fatigue-inducing and inefficient (Figure

2). We aim to segregate the flights that fall in this category in order to determine whether the schedules for
such flights could be changed to improve pilot alertness predictions and also to improve efficiency.

Crowdsourced ADS-B Data to Estimate Stability of Flight Approach and Landing

Stabilized approach criteria have been established to ensure safe approach and landing. In particular, the Flight Safety Foundation (FSF) established the following criteria for stabilized approach:

- Profile:
    - Only small changes in heading/pitch are required to maintain the correct flight path profile.
    - Specific types of approaches are stabilized if they fulfill the following:
        - CAT I ILS: within 1-dot deviation of glide path and localizer.
        - RNAV: within ½-scale deflection of vertical and lateral scales and within RNP requirements.
        - LOC/VOR: within 1-dot lateral deviation.
        - Visual: within 2.75 and 3.25 degrees of visual approach path indicators and lined up with the runway centerline no later than 300 ft AGL.
- Configuration: Aircraft is in the landing configuration (gear and flaps set, speed brakes retracted).
- Energy:
    - Airspeed is stabilized within VREF +10 kt to VREF (without wind adjustments).
    - Thrust is stabilized to maintain the target approach airspeed.
    - Sink rate is no greater than 1,000 fpm.
- General:
    - The stabilized approach gates should be observed.
    - Occasional momentary “overshoots” made necessary by atmospheric conditions are acceptable. Frequent or sustained “overshoots” are not.

FOQA data and radar track data can be used to determine whether the stabilized approach criteria are being satisfied. However, Automatic Dependent Surveillance-Broadcast (ADS-B) also can be used for this purpose and has several advantages. In particular, ADS-B has wider coverage than radar surveillance and is not proprietary, unlike FOQA data. ADS-B is also more accurate than radar track data. There are also free crowdsourced repositories of ADS-B data that are available for research. For this reason, we decided to investigate the use of ADS-B data and Machine Learning to determine whether aircraft satisfy stabilized approach criteria.

ADS-B data is broadcast by aircraft transponder without radar interrogation. Common aircraft state parameters included in ADS-B message are aircraft callsign, transponder ID, aircraft operational status, position, altitude, track, and groundspeed. The position is determined by GPS. The velocity is derived from the GPS position and the inertial measurement system on the aircraft. The altitude information includes both barometric

altitude and GPS altitude “ADS-B out” equipment is mandated by FAA and Eurocontrol. ADS-B data can be captured by widely available inexpensive receivers.

After looking into several sources of ADS-B data (see table below), we decided to investigate using OpenSky data.

Name	Type	Founded	Number of receivers
FlightAware	Commercial	2005	30,000+*
FlightRadar24	Commercial	2006	25,000+*
RadarBox24	Commercial	2007	15,000+
OpenSky network	Non-profit/research	2013	3,500+*
ADS-B Exchange	Non-profit	2016	5,000+*

* Number of receivers obtained from each company’s website, as of January 2021.

In addition to OpenSky data repository being available free for research purposes, OpenSky has the advantage of associated software. In particular, there is an open-source Python library called `Traffic` that is available under the MIT license. The library provides:

- Download assistance tool for the OpenSky historical data.
- Functionality to process flight trajectories, including resampling, filtering faulty data, projecting, querying, and intersecting with geospatial objects.
- Exporting facilities to common visualization tools such as `Matplotlib` or `Cartopy` and Google Earth.
- The `Traffic` library is based on three main core classes for handling: aircraft trajectories through `Flight` class, collections of aircraft trajectories through `Traffic` class and airspaces through `Airspace` class. `Flight` and `Traffic` classes are wrappers around pandas DataFrame.
- Additional databases: Airports, Navaids, Aircraft.

We follow the following processing steps to prepare OpenSky data for use in detecting unstabilized approach:

- Download from OpenSky flight trajectories within specified geographic area for flights landing at specified airport during specified time period (e.g., day).
- For each flight:
    - Truncate trajectory to the altitude below 4000 ft.
    - Determine landing runway (using runways/ILS info) and calculate distance to threshold.
    - Calculate localizer deviations (contiguous intervals longer than certain minimal duration for different altitude ranges).
    - Calculate glideslope deviations.
    - Calculate vertical rate excursions.
    - Calculate flight energy metrics.

Upcoming work includes testing the pipeline and testing ML algorithms with these data to determine how much more effective ADS-B data is at detecting unstabilized approach.

Assessment of additional data from Aviation Safety Action Program (ASAP) Event Review Committee (ERC) beyond ASAP Reports

The goal of this work was to assess the benefits gained from having a non-airline human factors expert (i.e., NASA) observe an Aviation Safety Action Program (ASAP) Event Review Committee (ERC) debrief of the crew that submitted an ASAP report. A second goal was to assess the potential benefit of review and annotation of an ASAP report by a non-airline human factors expert.

Part of the motivation of having a human factors expert perform a debrief and report annotation is due to problems inherent with ASAP reports. One problem is that each report is a retrospective reconstruction of the events---every time the reporter replays the event (re-remember), it will change. Our mind creates stories to lend coherence to our experience; we are sense-making machines; therefore, the report may never be completely accurate. Additionally, the report normally describes the reporter's observations, which may be symptoms of general systemic issues. A human factors expert can identify these underlying issues/causes that may be common across multiple events even though they may have been written up differently in the different reports. Such identification of causes can contribute to improved training, procedures, LOSA, FOQA, and other aspects of aviation operations.

As part of this effort, three NASA human factors experts observed two ERC debriefs each, for a total of six debriefs and associated reports. As such, this is preliminary work and there is much potential for future work. Performing and observing additional debriefs is an obvious task to be done that is likely to yield significant additional insights. However, a process to handle the information that is collected, and identify and calculate relevant metrics is needed. A systematic process to assess these results to determine where within carrier operations the learned information can provide benefit and how to properly change those aspects of operations are also needed.

References

Åkerstedt, T., Folkard, S., & Portin, C. (2004). Predictions from the three-process model of alertness. Aviation, Space and Environmental Medicine, 75, A75-A83.

Åkerstedt T., Gillberg M. (1990). Subjective and objective sleepiness in the active individual. International Journal of Neuroscience, 52, 29–37.

Arsintescu, L., Chachad, R., Gregory, K. B., Mulligan, J. B., & Flynn-Evans, E. E. (2020). The relationship between workload, performance and fatigue in a short-haul airline. Chronobiology International, 37(9-10), 1492-1494.

Arsintescu, L., Pradhan, S., Chachad, R. G., Gregory, K. B., Mulligan, J. B., & Flynn-Evans, E. E. (2022). Early starts and late finishes both reduce alertness and performance among short-haul airline pilots. Journal of sleep research, 31(3), e13521.

Boril, J., Smrz, V., Blasch, E., & Lone, M. (2020). Spatial disorientation impact on the precise approach in simulated flight. Aerospace Medicine and Human Performance, 91(10), 767-775.

Bourgeois-Bougrine S., Cabon, P., Mollard, R., Coblentz A., Speyer J. 2003. Fatigue in aircrew from shorthaul flights in civil aviation: the effects of work schedules. Human Factors Aerospace Safety: An International Journal, 3:177–187.

Endsley, M. R. (1999). Situation awareness in aviation systems. Handbook of aviation human factors, 257, 276.

Flynn-Evans, E.E., Arsintescu, L., Gregory, K., Mulligan, J., Nowinski, J.L., Feary, M. 2018. Sleep and neurobehavioral performance vary by work start time during non-traditional day shifts. Sleep Health, 4(5):476–484. doi:10.1016/j.sleh.2018.08.002

Hitchcock, L., Bourgeois-Bougrine, S., & Cabon, P. (2010). Pilot performance. Handbook of aviation human factors, 14-1.

Honn, K. A., Satterfield, B.C., McCauley, P., Caldwell, J.L., Van Dongen, H.P. 2016. Fatiguing effect of multiple take-offs and landings in regional airline operations. Accident Analysis and Prevention, 86:199–208. doi:10.1016/j.aap.2015.10.005

Federal Aviation Administration, Department of Transportation (2009). Flightcrew member duty and rest requirements. 14 CFR Parts 117, 119, and 121. Docket No.: FAA-2009-1093; Amdt. Nos. 117-1,119-16, 121-357 RIN 2120–AJ58.

Young, J. A. (2008). The effects of life-stress on pilot performance. Moffett Field, Calif.: Ames Research Center.