IPSW 2024 IATA Final Presentation

Estimating turbulence duration and the likelihood of turbulence occurring

C. Sean Bohun (OntarioTech U), Alberto Fornaci (IATA), Ahmed Harrabi (Polimi), Pierre Houedry (U de Rennes), Slim Ibrahim (UVic), Wuding Li (UdeM), Mike Lindstrom (UTRGV), Ronnie Liu (HEC), Sudeesha Warunasinghe (UCalgary), Lingyi Yang (Oxford), Ismael El Yassini (UWaterloo)

May 17, 2024

Agenda

Motivation

Turbulence is...

- the **leading cause of injuries** to cabin crew and passengers in non-fatal accidents (FAA)

Motivation

Turbulence is...

- the **leading cause of injuries** to cabin crew and passengers in non-fatal accidents (FAA)
- costing the aviation industry **hundreds of millions of dollars** every year

Motivation

Turbulence is...

- the **leading cause of injuries** to cabin crew and passengers in non-fatal accidents (FAA)
- costing the aviation industry **hundreds of millions of dollars** every year
- causing **brand damage**

Motivation

Turbulence is...

- the **leading cause of injuries** to cabin crew and passengers in non-fatal accidents (FAA)
- costing the aviation industry **hundreds of millions of dollars** every year
- causing **brand damage**
- contributing to the **fear of flying**

Motivation

Turbulence is...

- the **leading cause of injuries** to cabin crew and passengers in non-fatal accidents (FAA)
- costing the aviation industry **hundreds of millions of dollars** every year
- causing **brand damage**
- contributing to the **fear of flying**
- increasing: **149%** - the projected increase in the frequency of severe turbulence (P. Williams, 2017)

Problem Statement #1

What is the problem?

- **How long does turbulence last?**
- Does its duration depend on the time of the year? Is there a relationship between turbulence duration and altitude? What about wind speed and temperature?

Meteorologists have been attempting to answer that question for a very long time, and this is the first time we have enough objective data to address the question.

Problem Statement #2

What is the problem?

- Given the aircraft's location, heading, and speed, we want to

  determine the likelihood of turbulence ahead of the aircraft

  based on live and historical turbulence data
- Can take into account wind speed and direction, temperature and location, history and live data

Such information can then be dispatched to the pilots in real-time so they can make informed tactical decisions and avoid turbulence.

Data

What data we have:

1. Aircraft collect data every 15 minutes (or more often when turbulent)
2. Data include altitude, longitude, latitude, windspeed, wind direction, temperature, and EDR reading for each aircraft with a timestamp
3. From this we can infer other values such as “gradients”

```
  measurement_observationTime  measurement_altitude  measurement_latitude  \
0         2024-02-12 19:08:10                  5200                33.693
1         2024-02-12 19:09:10                  5100                33.694
2         2024-02-12 19:10:10                  4100                33.694

   measurement_longitude  measurement_temperature  measurement_wind_speed  \
0                -84.717                     11.0                      30
1                -84.656                     11.0                      31
2                -84.595                     11.5                      32

   measurement_wind_direction measurement_edr_algorithm  \
0                         166                   NCAR_V3
1                         161                   NCAR_V3
2                         139                   NCAR_V3

   measurement_edr_peak_value  measurement_edr_mean_value  ...  \
0                        0.04                        0.02  ...
1                        0.08                        0.04  ...
2                        0.06                        0.02  ...

                            metadata_id                         metadata_tafi
0  4579b2ba-0000-4e31-a244-ddc6d7cccfa2  001f994f-9a4c-43e4-9273-32bcefb097f1
1  084321ac-93d8-465a-9527-df10f06b07f5  001f994f-9a4c-43e4-9273-32bcefb097f1
2  ecded18a-6526-44e9-8e3b-c44df0aaf9c7  001f994f-9a4c-43e4-9273-32bcefb097f1
```

Agenda

Problem 1 - Clustering Approach

Clustering

- Select data with EDR above 0.13 $m^{2/3}/s$ (felt by small aircraft), and chunk into 6-hour blocks. Spatially kmeans-cluster each chunk, and temporally kmeans-cluster each spatial cluster to identify "**turbulence events**"
- Characterize events with a peak EDR, mean distance from centroid, duration, mean altitude, and mean magnitudes of temperature/velocity directional derivatives

Problem 1 - Clustering Approach

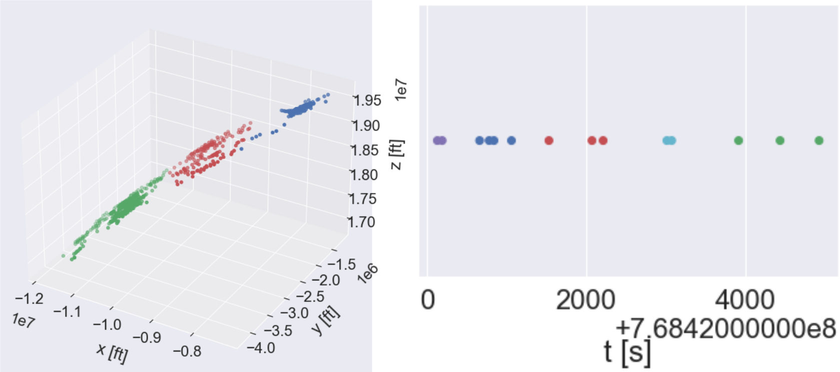


Figure: Left: spatial clusters of random 6-hour time chunk. Right: time clusters of random space cluster based on occurrence time. Each color represents a different cluster.

Problem 1 - Clustering Approach

Distributions

Can find distributions of parameters for our turbulence events; the characteristic **duration is approximately 1700 s = 28 min** (median)

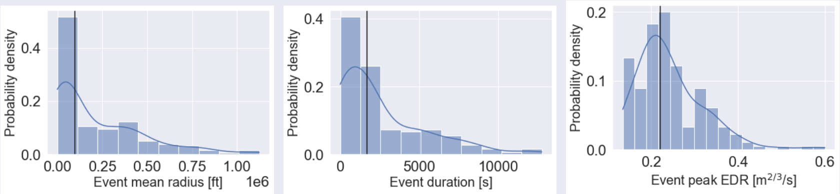
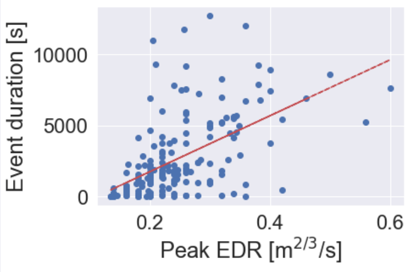


Figure: Distributions of event radius (mean distance to centroid), event duration, and peak EDR. Medians are vertical bars with 100,000 ft (left), 1700 s (middle), and 0.22 $m^{2/3}/s$ (right).

Problem 1 - Clustering Approach

Trends

Can identify trends between the summary statistics of turbulence events, but **there is a lot of noise**



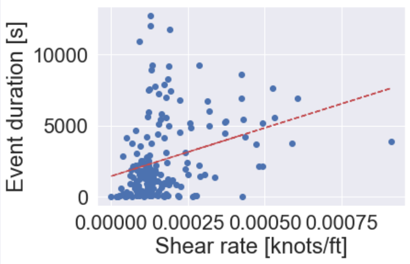


Figure: Scatterplots and trendlines. R^2-values are 0.3 (left) and 0.1 (right).

Problem 1 - Random Forest Classifier

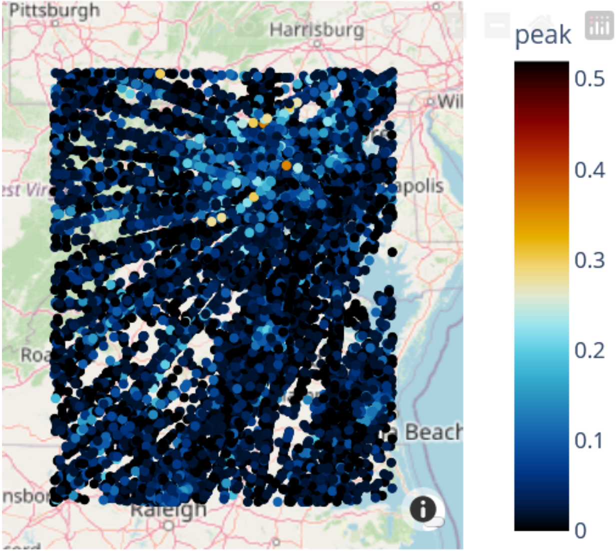
Data Re-structuring - The Grid

1. Considered a location in US mainland (36 to 40 latitude and -76 to -80 longitude)
2. Considered last week of April 2024
3. Considered altitude below 40,000 ft
4. Created a grid where the cell is 0.1×0.1 degrees
5. Any EDR reading ≥ 0.13 considered turbulence
6. EDR reading with 15-minute intervals to be considered as continuous turbulence

Assumption

If EDR reading ≥ 0.13 is recorded in a single square of the grid, the whole square is experiencing turbulence.

Problem 1 - Random Forest Classifier



Problem 1 - Random Forest Classifier

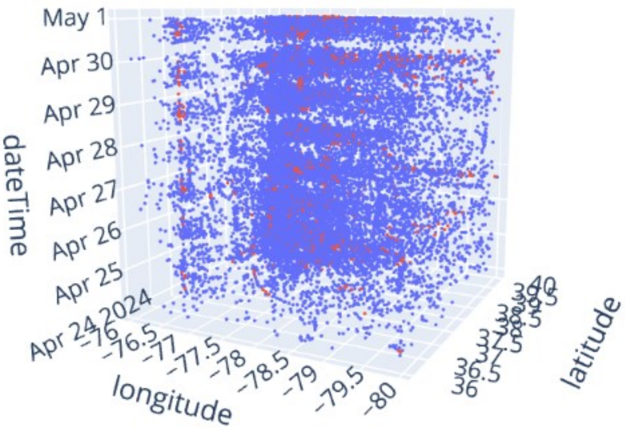


Figure: Time Series plot of EDR readings

Problem 1 - Random Forest Classifier

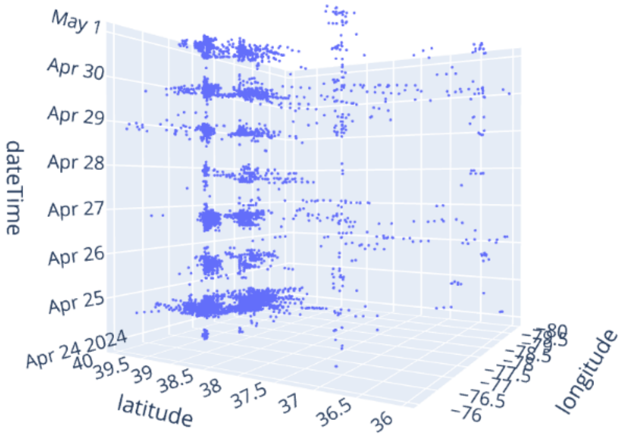


Figure: Time Series plot of EDR readings (Only Turbulent events)

Problem 1 - Random Forest Classifier

Data

- Whole US Mainland
- Altitude between 30,000 ft to 40,000 ft
- Considered last week of April 2024

Model

- Non-linear model
  - Random Forest Classifier

Results

- Dependent variable: turbulence duration less than 1 hour, turbulence duration greater than 1 hour.
- Average accuracy over 5 experiments: 0.83

Problem 1 - Summary

Clustering

- Identified meaningful turbulence events and statistics of interest like time and length scales
- Found interesting trends, such as turbulence duration decreasing with altitude
- Trends need further scrutiny to test for effects of outliers

Random Forest Classifier

- Computed the mean duration of turbulence over different cells based on historical data
- Given temperature and wind speed temporal gradients, etc., can predict whether turbulence will last less than 1 hour or more than 1 hour
- Limited by grid being arbitrary — not informed by data

Agenda

Problem 2

Neural Network - EDR Data Visualization on the West Coast

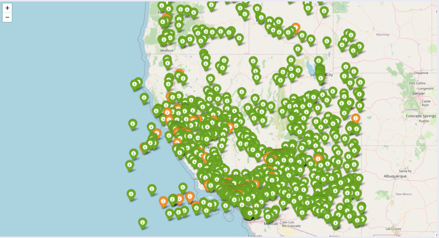


Figure: Map of the West Coast with indexed EDR data points.

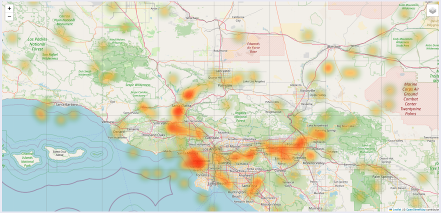


Figure: Heatmap of EDR distribution over the West Coast.

Problem 2

Network Overview - Probabilistic Neural Network Model

Our network utilizes a probabilistic approach where outputs are sampled from a Gaussian distribution, allowing for uncertainty modelling in atmospheric conditions so that

$$F : (\mathbf{X}, \mathbf{Y}) \mapsto (\mu_1, \sigma_1^2, \ldots, \mu_n, \sigma_n^2).$$

Model Training and Evaluation

The model is optimized using a loss function that incorporates prediction variance, enhancing reliability with

$$\mathcal{L} = -\sum_{i,j} \log p_j^i,$$

and training utilizes the Adam optimizer with a learning rate of 0.005.

Problem 2

Data Handling Preprocessing with Grid System

Data is segmented using a 50 km x 50 km grid system, crucial for localizing predictions and identifying spatial patterns in turbulence occurrences.

Feature Engineering

Features such as temperature and wind speed are normalized to suit machine learning applications, highlighting the importance of accurate, and extensive meteorological data.

Data Requirements

While the model shows promise, it would be beneficial having additional data, like atmospheric pressure and more meteorological conditions, to improve prediction capabilities.

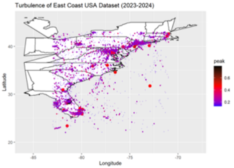
Problem 2 - Logistic Regression (Context and Results)

DATASET

East Coast USA Random Sample for 2023-2024

Training: 350 000 observations
Test: 75 000 observations

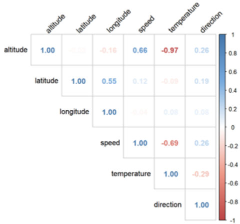
15% of the data: Presence of Turbulence (Peak EDR > 0.13)



VARIABLES

Response Variable: Contains or not a turbulence which Peak EDR > 0.13

Covariables: Aircraft's position (longitude, latitude), speed, temperature, direction and seasonality



RESULTS FOR TEST DATASET

Accuracy	69.8%
Sensitivity	68.7%
Specificity	70.0%
AUC	74.5%

Confusion Matrix

	Observation 0	Observation 1
Prediction 0	44420	3615
Prediction 1	19034	7931

Pros and Cons

+ Simple, interpretable with covariables

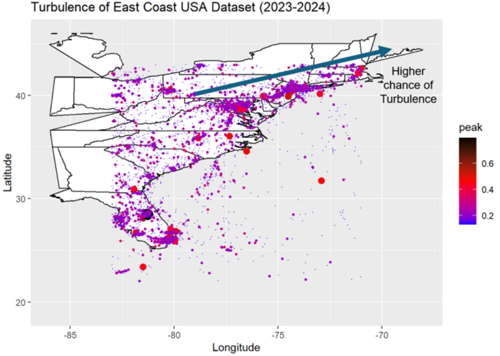
- Doesn't take account time and atmospheric pressure. We only used seasonality.

Problem 2 - Logistic Regression (Interpretations)

Covariable	% of difference in odds of Turbulence when the covariable increases by 1 unit
Latitude (towards North)	+3.0
Longitude (towards East)	+1.6
Speed	+0.6
Temperature	+4.3
Direction	+0.01
Season: Fall to Spring	+77
Season: Fall to Summer	-12
Season: Fall to Winter	+47

We use odds in logistic regression: higher odds → higher chance of turbulence than not having one.

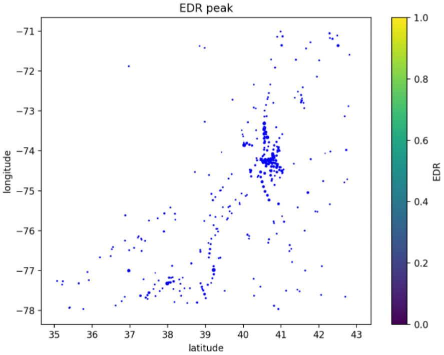
$$\hat{odds}(Y|X) = \frac{\hat{P}(Y=1|X)}{1-\hat{P}(Y=1|X)} \longleftrightarrow \hat{P}(Y=1|X) = \frac{\exp(X\hat{\beta})}{1+exp(X\hat{\beta})}$$

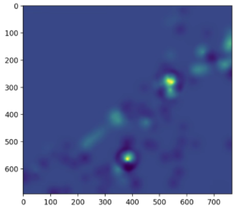


Problem 2 - Kriging

Kriging

- $\{Z_s, s \in D\}$ is the random field of EDR values in the space $D \subset \mathbb{R}^2$
- From knowing value in certain location, interpolate the values in unobserved locations

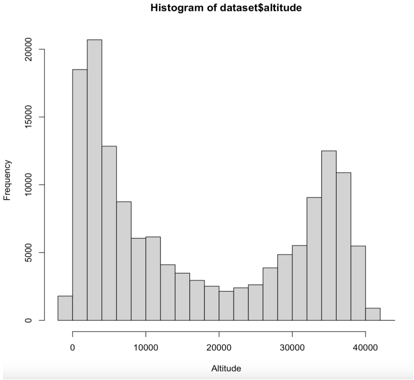




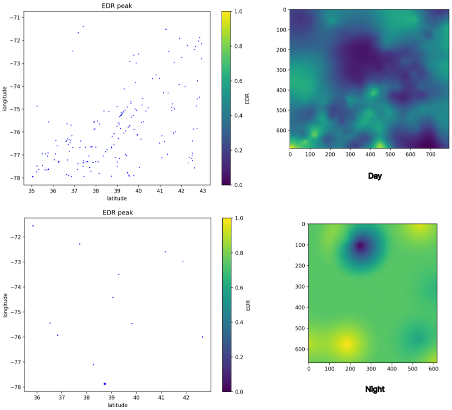
Problem 2 - Kriging

Kriging Assumption

- Stationary: $\mathbb{E}[Z_s] = m_z$
- Spatial Homogeneity: $\text{Cov}(Z_{s_i}, Z_{s_j}) = C(h)$, for all $s_i, s_j \in D$, $h = s_i - s_j$ where C is the covariogram



Problem 2 - Kriging



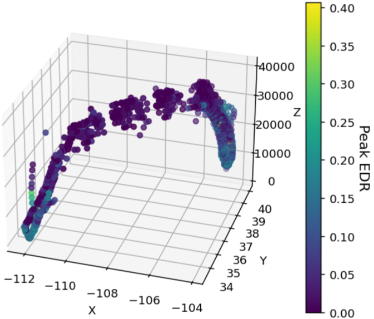
Adjustment to Stationarity

- Data imputation
- Detrend the data

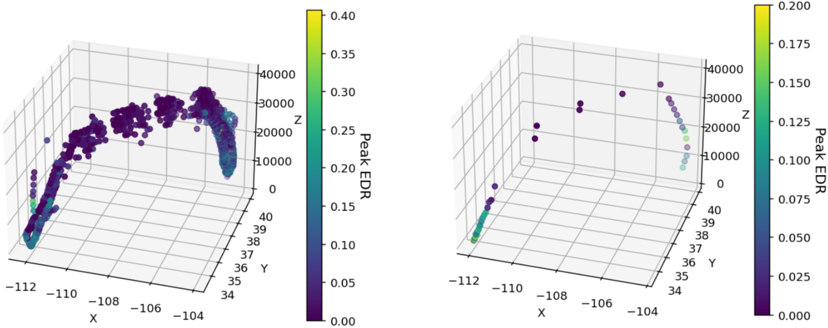
Future work

- Train a neural network using kriging predictions to forecast when a turbulent region will vanish
- Explore higher kriging dimension, despite heavy computation

Following trajectories of aircraft

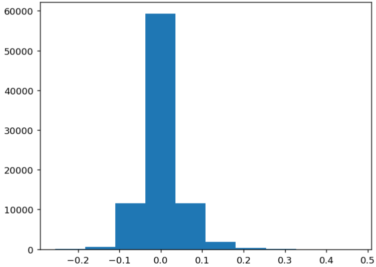


Following trajectories of aircraft

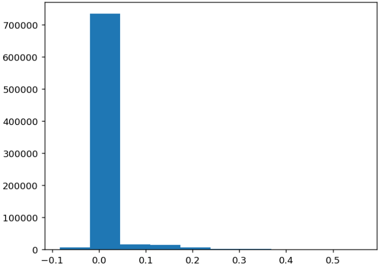


Given the timeseries of information (location, previous turbulance reports) from individual flights, can we predict the peak EDR at their next report?

Following trajectories of aircraft — results

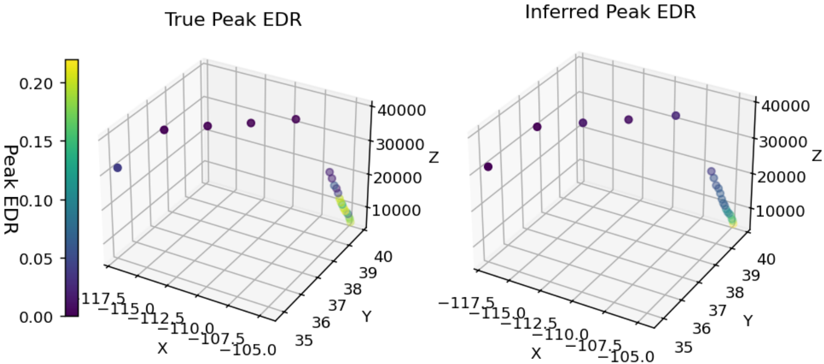


(a) Regression residuals, RMSE:0.045



(b) GRU residuals, RMSE:0.029

Following trajectories of aircraft — reconstructed



Problem 2 - Summary

Neural Network

- The model aims to enhance navigational safety and operational efficiency, with further improvements anticipated through the integration of additional meteorological data.

Logistic Regression

- Identify interpretable trends in how the covariables change the odds ratio of turbulence
- However, may not capture all factors affecting the turbulence since it doesn't take account the time and the atmospheric pressure

Problem 2 - Summary

Kriging

- Can use Kriging to infer EDR values at fixed altitudes
- The lack of stationarity may limit its accuracy

Following Trajectories

- We can use information from timeseries to infer whether there is still turbulance when the aircraft moves to a new location.
- Further investigation to refine models and quantify uncertainty.

Agenda

Toy Model - Formulation

Goals

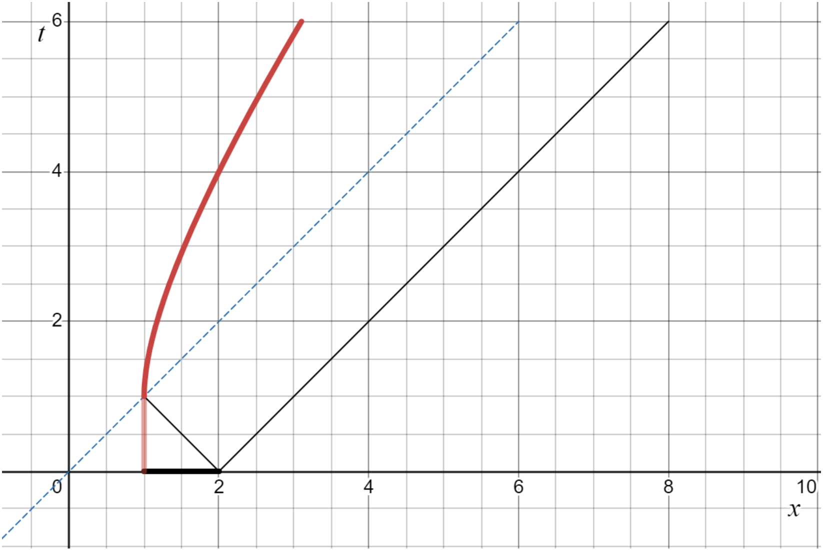
- Construct a simple dynamical model that can mimic the observed behaviour of turbulence
- Base the model in physics
- Characterize the property of 'turbulence'

Assumptions

- Air masses with less density have a higher intrinsic speed
- Evolution of density $\rho(x, t)$ as the variation along a flight path
- Mass is conserved
- Turbulence is associated with discontinuities in ρ

$$\frac{\partial \rho}{\partial t} + \frac{\partial j}{\partial x} = 0, \quad j = \rho v(\rho), \quad v(\rho) = 1 - \rho, \quad \rho(x, 0) = \begin{cases} 1, & 1 \leq \ x \leq 2 \\ 0, & \text{otherwise.} \end{cases}$$

Toy Model - Base case



Toy Model - Base case

For $0 \leq t < 1$,

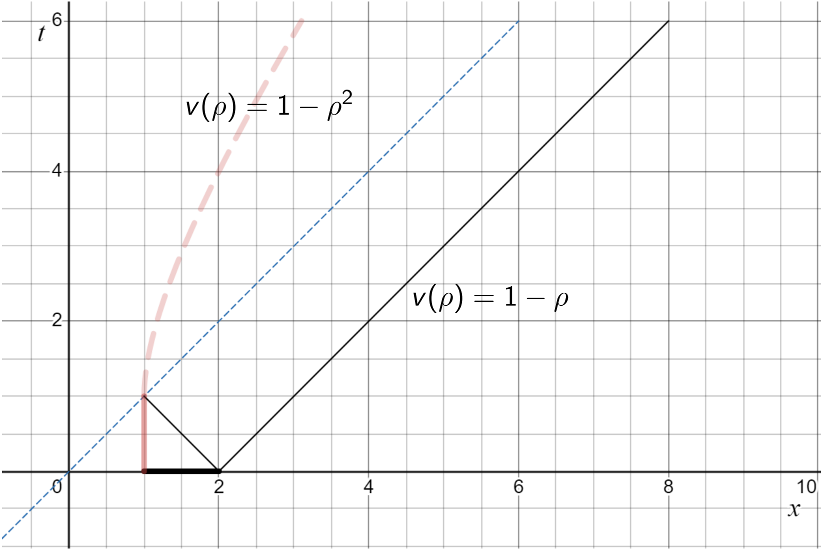
$$\rho(x,t) = \begin{cases} 0, & x < 1, \\ 1, & 1 \leq x < 2 - t, \\ \frac{1}{2}\left(1 - \frac{x-2}{t}\right), & 2 - t \leq x \leq 2 + t, \\ 0, & x > 2 + t. \end{cases}$$

For $t \geq 1$,

$$\rho(x,t) = \begin{cases} 0, & x < \sigma(t), \\ \frac{1}{2}\left(1 - \frac{x-2}{t}\right), & \sigma(t) \leq x \leq 2 + t, \\ 0, & x > 2 + t, \end{cases}$$

with $\sigma(t) = 2 + t - 2\sqrt{t}$. Note that $\rho(\sigma(t), t) = 1/\sqrt{t}$ for $t \geq 1$.
The discontinuity weakens in time as the inital air mass dissipates in space.

Toy Model - Interaction with another air mass



Toy Model - Interaction with another air mass

For $t < x$, $\rho(x,t)$ is unchanged with

$$\frac{\partial \rho}{\partial t} + (1-2\rho)\frac{\partial \rho}{\partial x} = 0, \qquad \rho(x,0) = \begin{cases} 1, & 0 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

For $t > x$, there is now a region, $\sigma_2(t) < x < t$, with a different PDE to reflect the different air mass. In particular, for $t > 1$ the air mass overtakes the rarefaction fan with

$$\frac{\partial \rho}{\partial t} + (1-3\rho^2)\frac{\partial \rho}{\partial x} = 0, \qquad \rho(x=t,t) = \begin{cases} 0, & t < 1, \\ 1/t, & t \ge 1. \end{cases}$$

In this interaction region

$$\rho(x,t) = \frac{1}{2t}\left(1 + \left(1 + \frac{4}{3}(x-t)\right)^{1/2}\right).$$

What about the bounding shock?

Toy Model - Homework

To solve for the location of the shock $\sigma_1(t)$, is satisfies the Rankine-Hugonoit condition

$$\frac{\mathrm{d}\sigma_1}{\mathrm{d}t} = 1 - \left(\frac{1}{2t}\left(1 + \left(1 + \frac{4}{3}(\sigma_1 - t)\right)^{1/2}\right)\right)^2, \qquad \sigma_1(1) = 1.$$

- The differing air mass evolves the rarefaction fan.
- It is the velocity dependence in this new region that determines the leading edge of the turbulence event.

Agenda

Concluding Remarks

Other Data Considerations

1. Additional datasets such as weather data could supplement our work
2. Additional work can be done by studying the inputs used in the aircraft summary statistics such as upward/downward wind speed, etc.

Concluding Remarks

Future Work

1. The methods explored can be validated further and perhaps combined, e.g., clustering to inform the locations of the classifier regions
2. The models developed could be further understood/validated by testing them on all available datasets — each model tended to focus on just one
3. More work could be done to explore turbulence in the descent phase