A Generalized Method for Generating N-fold Random Joint Distributions from Observations

Frank Robasky, Cindy Engholm

January 2025

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

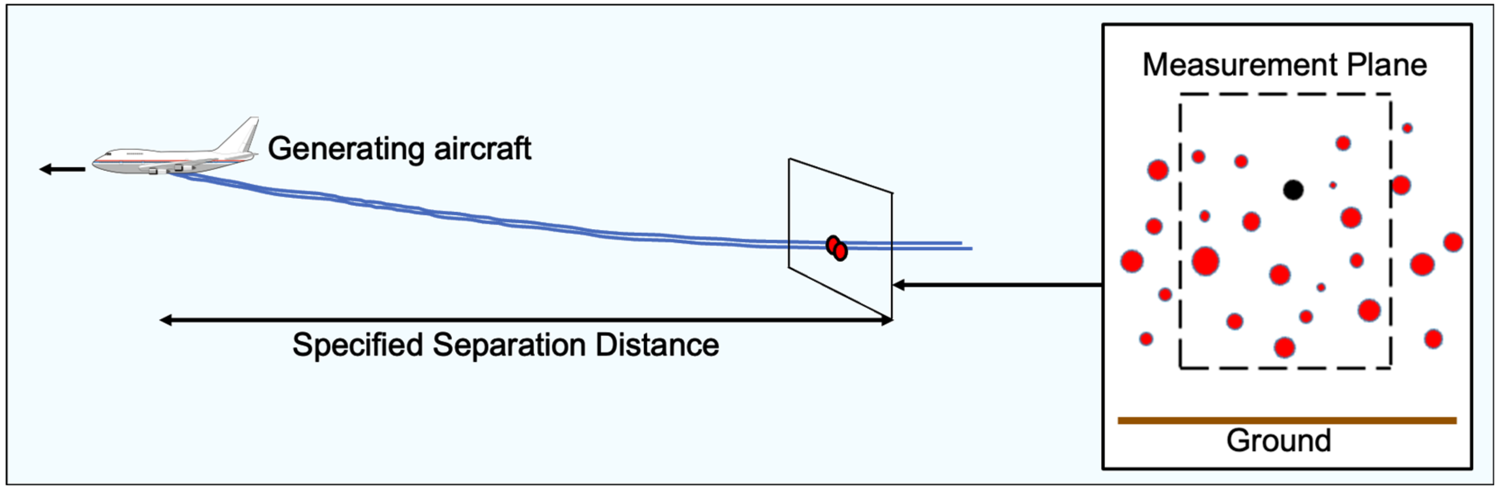
This material is based upon work supported by the Federal Aviation Administration under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Federal Aviation Administration.

Outline

- **Motivation / problem**
- **Existing approaches**
- **Method & implementations**
- **Practical aspects and areas for improvement**
- **Summary**

Motivation

- **Planes encountering wake vortices can experience loss of control / potentially fatal outcomes**
- **FAA investigating using physics-based models to establish safe separations**
  - **Modeling performed on a site-by-site basis for top NAS airports**
  - **Monte Carlo techniques used to assess risk (~10 million runs)**
  - **Requires joint probability distributions of: landing weight, landing speed, winds (headwind, crosswind), temperature (density), stability (dϴ/dz), turbulence (Eddy Dissipation Rate (EDR))**



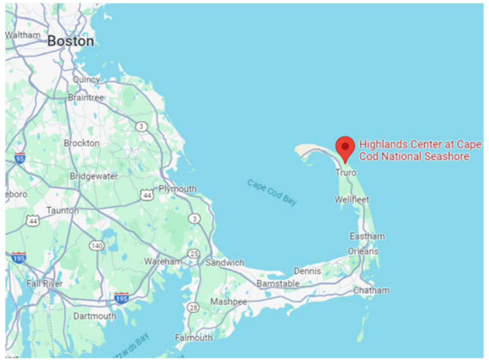
EDR = Eddy Dissipation Rate
FAA = Federal Aviation Association
NAS = National Airspace System

Graphic courtesy of NorthWest Research Association

Base Datasets for EDR Calculations

US Department of Energy Atmospheric Radiation Measurement (ARM) sites

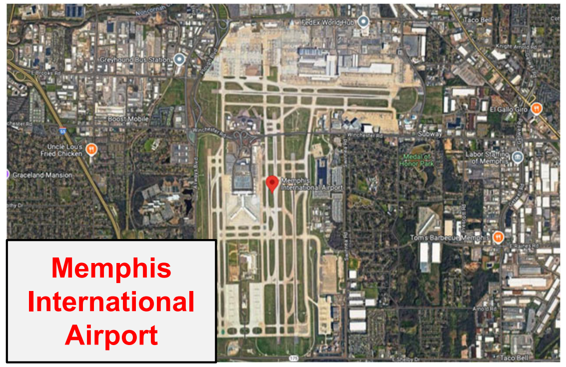
- **Cape Cod**
  - **July 2012 – June 2013**
  - **Lidar observations**
    - **Vertical stares**
    - **Wind profiles**
  - **Surface observations**
  - **8.6 K EDR values (1 / hour)**



- **Southern Great Plains (SGP) network**
  - **Lidar observations: 2010-present**
    - **Vertical stares: 1 Hz, from 105 m**
    - **Wind profiles: 15 min resolution, every 25 m from 90 m**
  - **Meteorological tower observations: 2015-present**
    - **Sonic anemometer winds: 10 Hz at 4, 25, and 60 m**
  - **Surface observations: 1993-present**
    - **1 min resolution**
  - **16.3 K EDR values (2 years @ 1 / hour)**
  - **182.2 K EDR values (2 years @ 1 / 5 min)**

NWRA NorthWest Research Associates Since 1984

- **NASA Memphis Dataset**
  - **May 2013 – March 2015**
  - **Lidar observations**
  - **Met tower**
  - **Surface observations**
  - **175.4 K observation times (limited to aircraft landings)**
  - **60.9 K EDR values**



EDR = Eddy Dissipation Rate

Outline

- Motivation / problem
- Existing approaches
- Method & implementations
- Practical aspects and areas for improvement
- Summary

Existing Approaches

Oversample The Observations

- Easy to implement
- Method lacks robustness, especially if number of desired samples >> number of observations
  - 10M samples from 150K observations nominally results in each observation being repeated 67 times
  - Does not allow for realistic though unobserved scenarios

Treat Variables Independently

- Would potentially result in over-representation of unrealistic combinations of parameters
  - Approach ignores meteorological dependencies and interrelationships
  - E.g., high dissipation rate (turbulence) values are much less likely during very stable conditions

Sample from Idealized / Fitted Distributions

- Can efficiently address the robustness issue
- May be difficult to find the appropriate ideal distributions
- May lose desired small-scale distribution characteristics

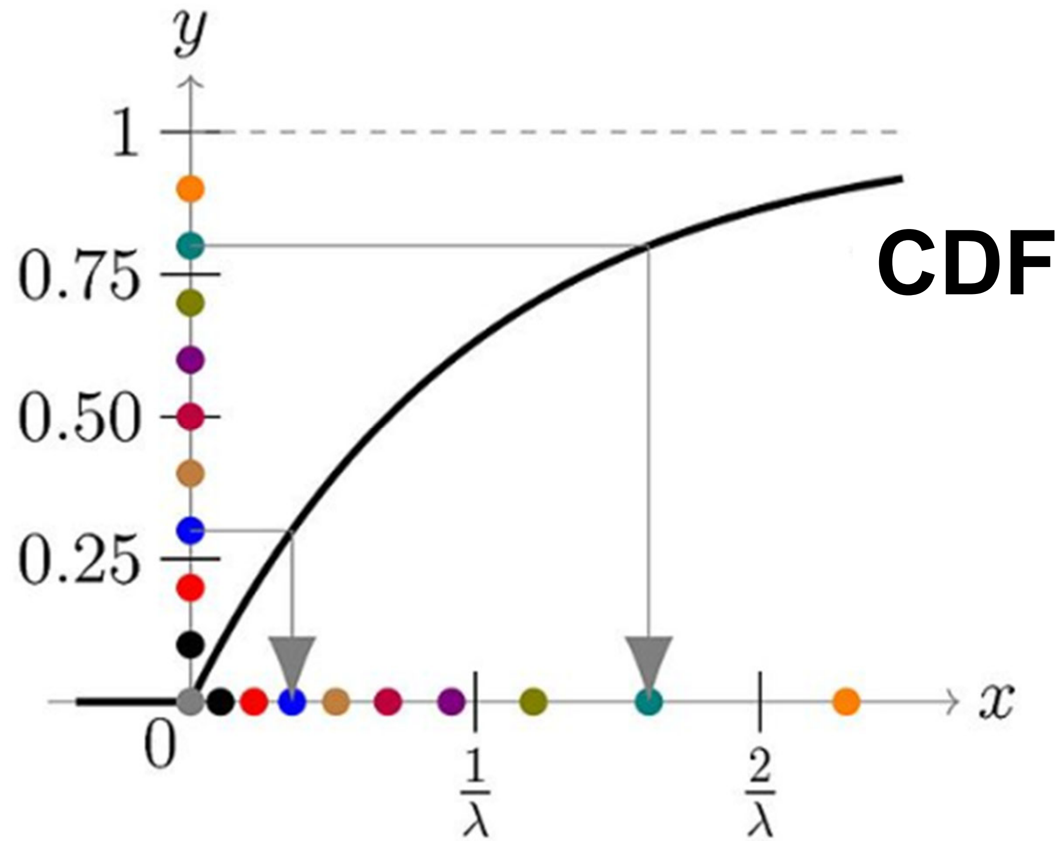
A robust method which preserves the observed joint relationships between variables is needed

Outline

- Motivation / problem
- Existing approaches
- Method & implementations
- Practical aspects and areas for improvement
- Summary

Basis of Method: Inverse Transform Sampling

- Works backwards from a flat sampling of cumulative distribution function (CDF) values to the variable source values
- 1-Dimensional illustration:
  - Compute the CDF of the observations, which has a range [0, 1]
  - Generate a set of random numbers of the desired sample size from a uniform distribution over the range [0, 1]
  - These values can be then mapped to their corresponding data values, yielding a realistic random sample of the original distribution

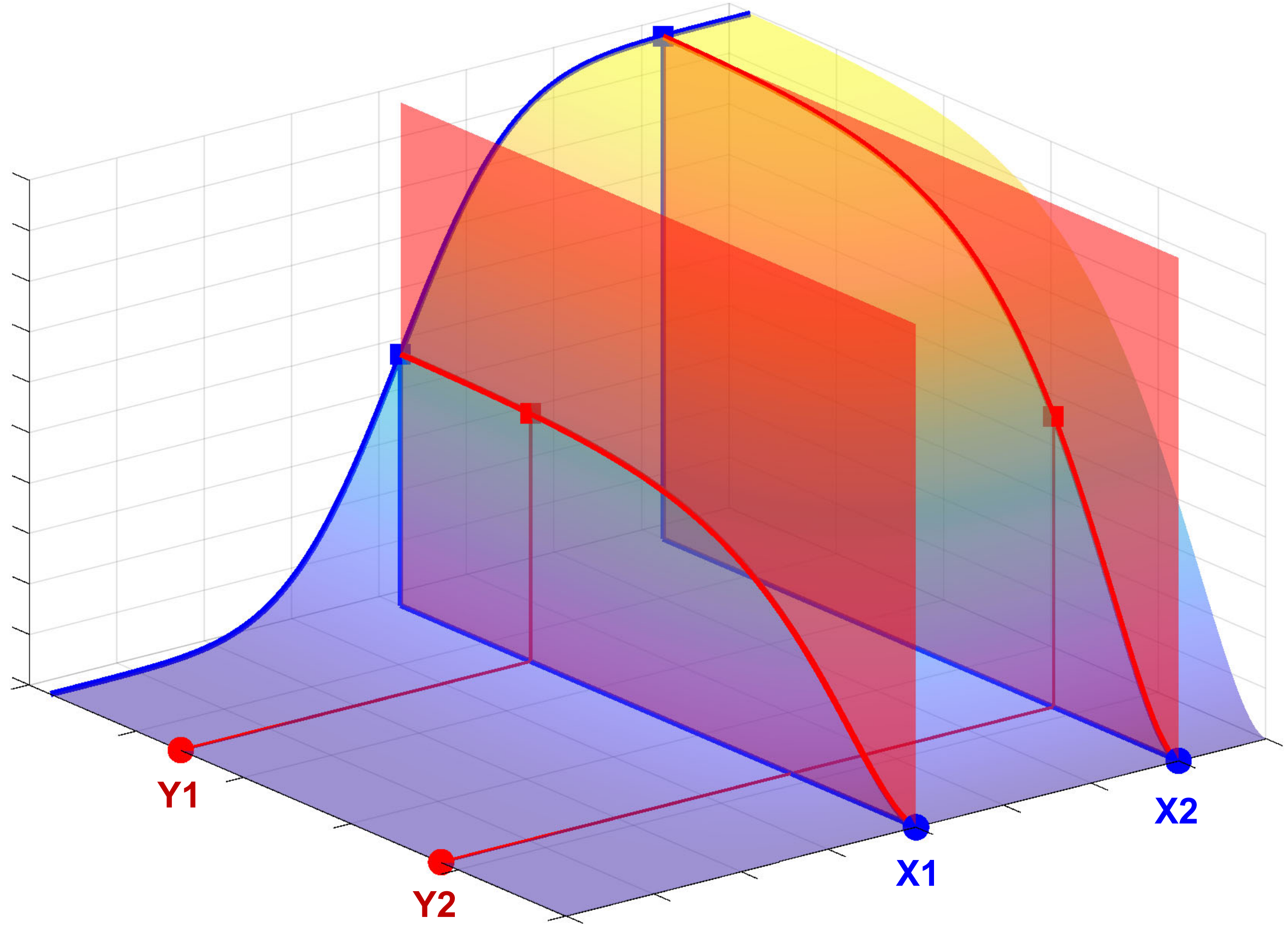


CDF = Cumulative Distribution Function

Extension of Method to Two Dimensions

- Apply the 1D method to one of the marginal distributions to yield random values for that variable (X)
- For each selected random value of X:
  - Take a slice of the 2D CDF at that value to yield a CDF of Y
  - Apply the 1D method to yield random values of Y for each value of X
- This results in random values of X and Y that preserve the original joint relationship

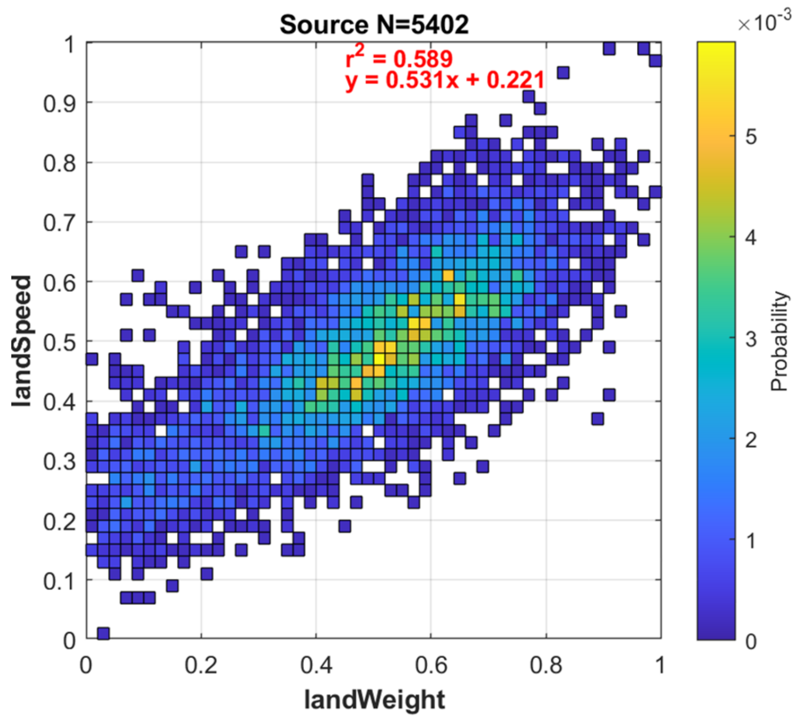
Illustration for Two Random Samples

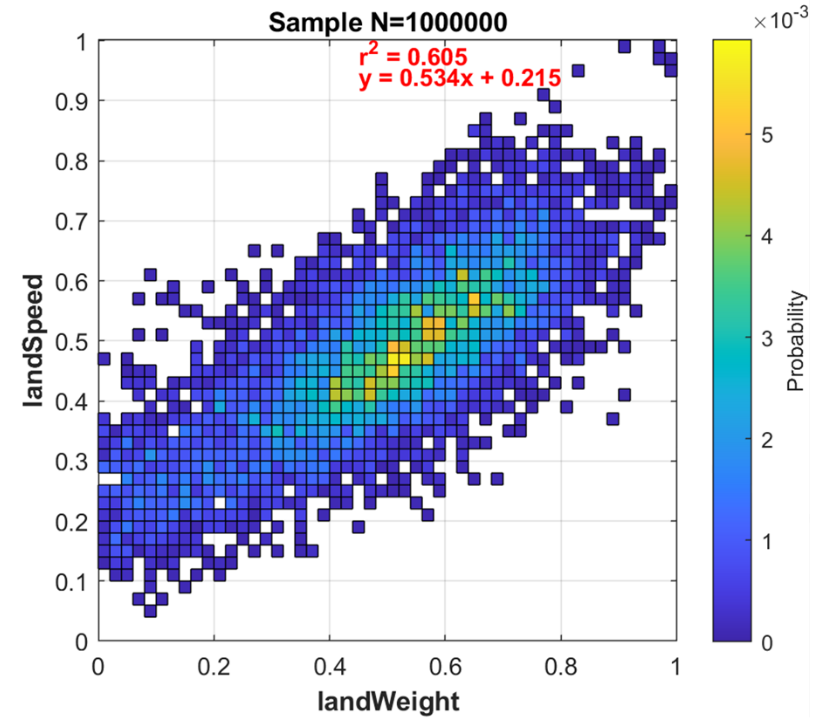


CDF = Cumulative Distribution Function

Two-Dimensional Application

- **Modeling task required random distributions of landing weight and landing speed for selected aircraft**
  - **Joint relationship is required**
- **Data is sampled on a discrete basis, by binning across the available ranges**
  - **~100 bins provide adequate resolution**
- **Generated joint distribution of 1M pairs closely matches the characteristics of the source distribution**

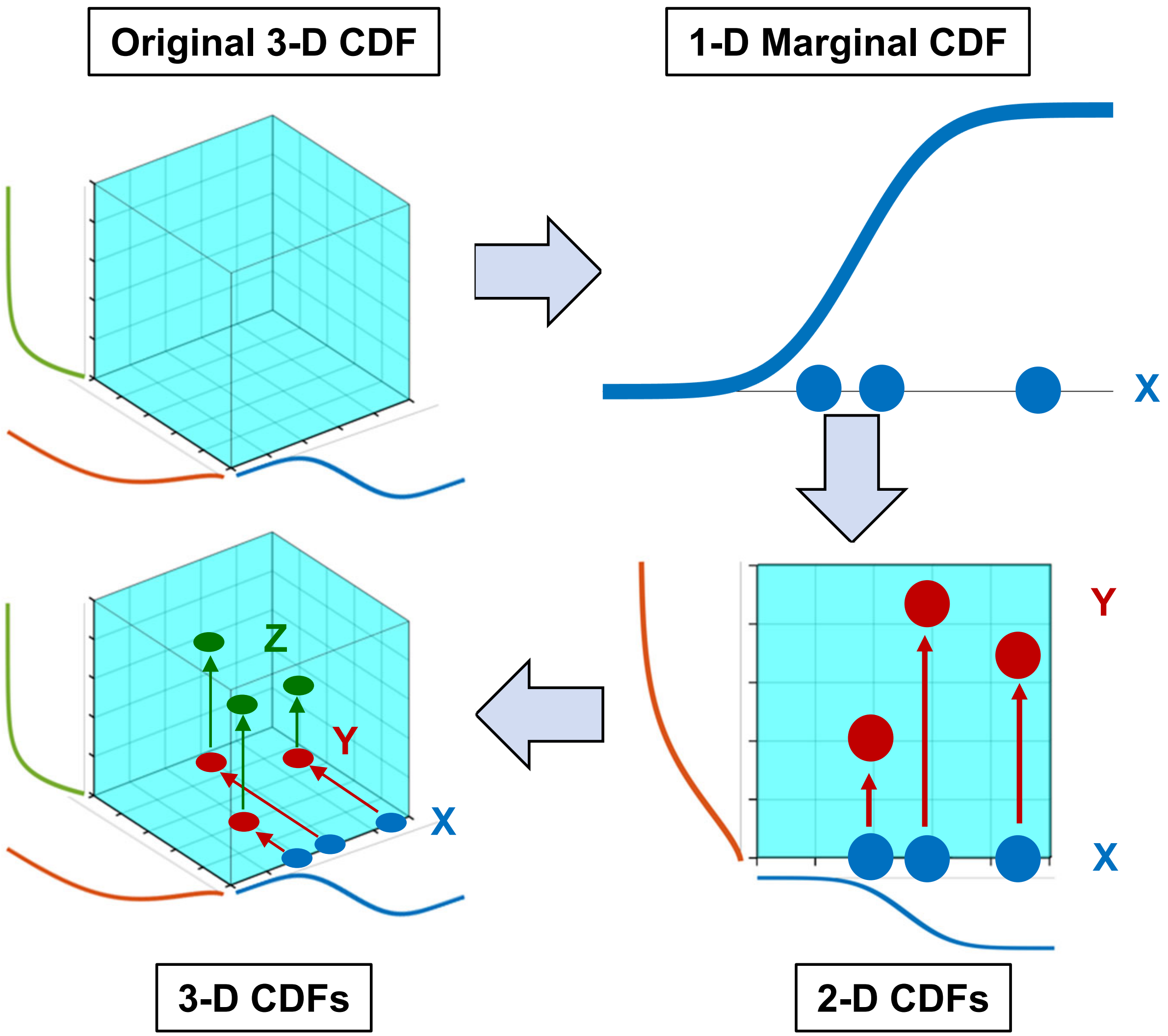




(Data is scaled for display purposes)

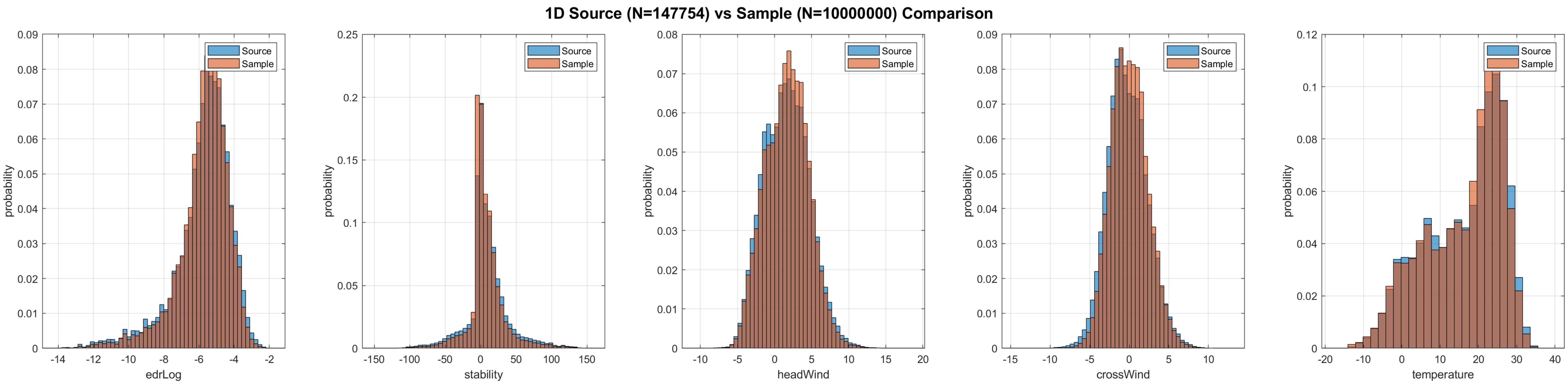
Extension to Three Dimensions and Beyond

- Generate matrix of random index values from a uniform distribution of size (N desired samples) x (M variables)
- Compute marginal CDF along 1 of the dimensions, determine random sample for variable 1
- Expand the CDF to 2 dimensions, employ slices at each value of variable 1 to determine random sample for variable 2
- Expand the CDF to 3 dimensions, employ slices at each combination of variable 1 and variable 2 values to determine random sample for variable 3
- Extend as needed to desired number of dimensions



CDF = Cumulative Distribution Function
PDF = Probability Density Function

Fidelity of Individual Aspect of Distributions

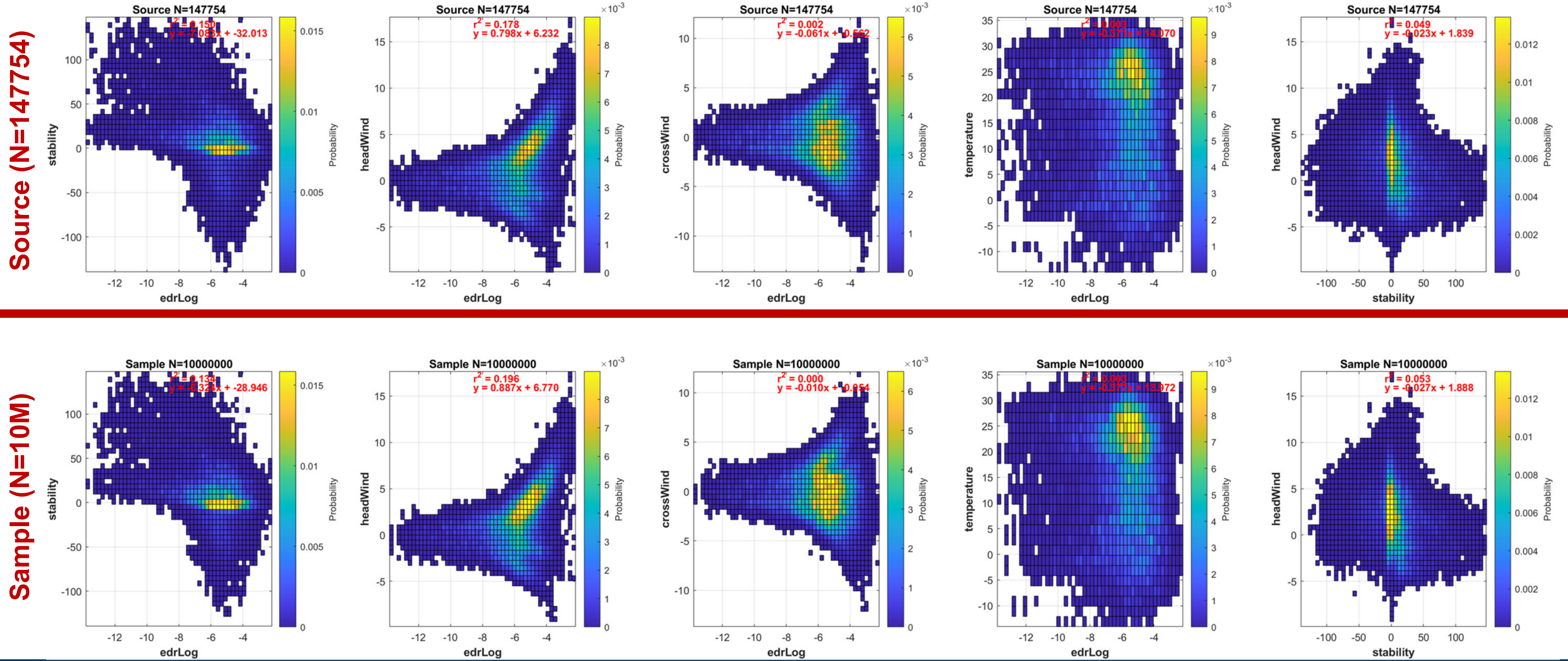


Marginal distributions show generally very good agreement, some discrepancies are noted, especially sample over-emphasis near the distribution peaks

Fidelity of Joint Aspect of Distributions

Visualized through examining the 10 2-way combinations of the 5 environmental variables

Generally, very good agreement seen across all variable combinations



Outline

- Motivation / problem
- Existing approaches
- Method & implementations
- Practical aspects and areas for improvement
- Summary

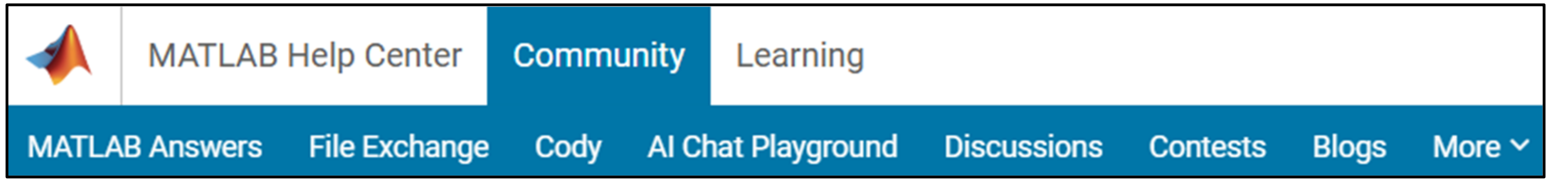
Operational Aspects and Improvements

- Specs
  - 10M joint 5-fold environmental and 2-fold operational distributions
  - Largest CDF matrix (100 x 100 x 100 x 100 x 50): 37.3 GB
  - ~20 min clock (wall) time
    - Intel Core i9-11950H @ 2.60GHz (8 cores), 64 GB RAM

- Potential Improvements
  - Parallelization and other efficiency upgrades
    - Exploit MATLAB “big data / tall array” functionality
  - Implement quantitative assessments of closeness
  - Reduce binning artifacts; enable continuous samples

Public Availability

- Tool is being made available for public use
- Is currently in the approval process
- Will be found on the Matlab File Exchange / github
  - Check for “N-Dimensional Joint Distribution Simulator”
  - Or email cde@ll.mit.edu or frankr@ll.mit.edu to get the link when ready



Outline

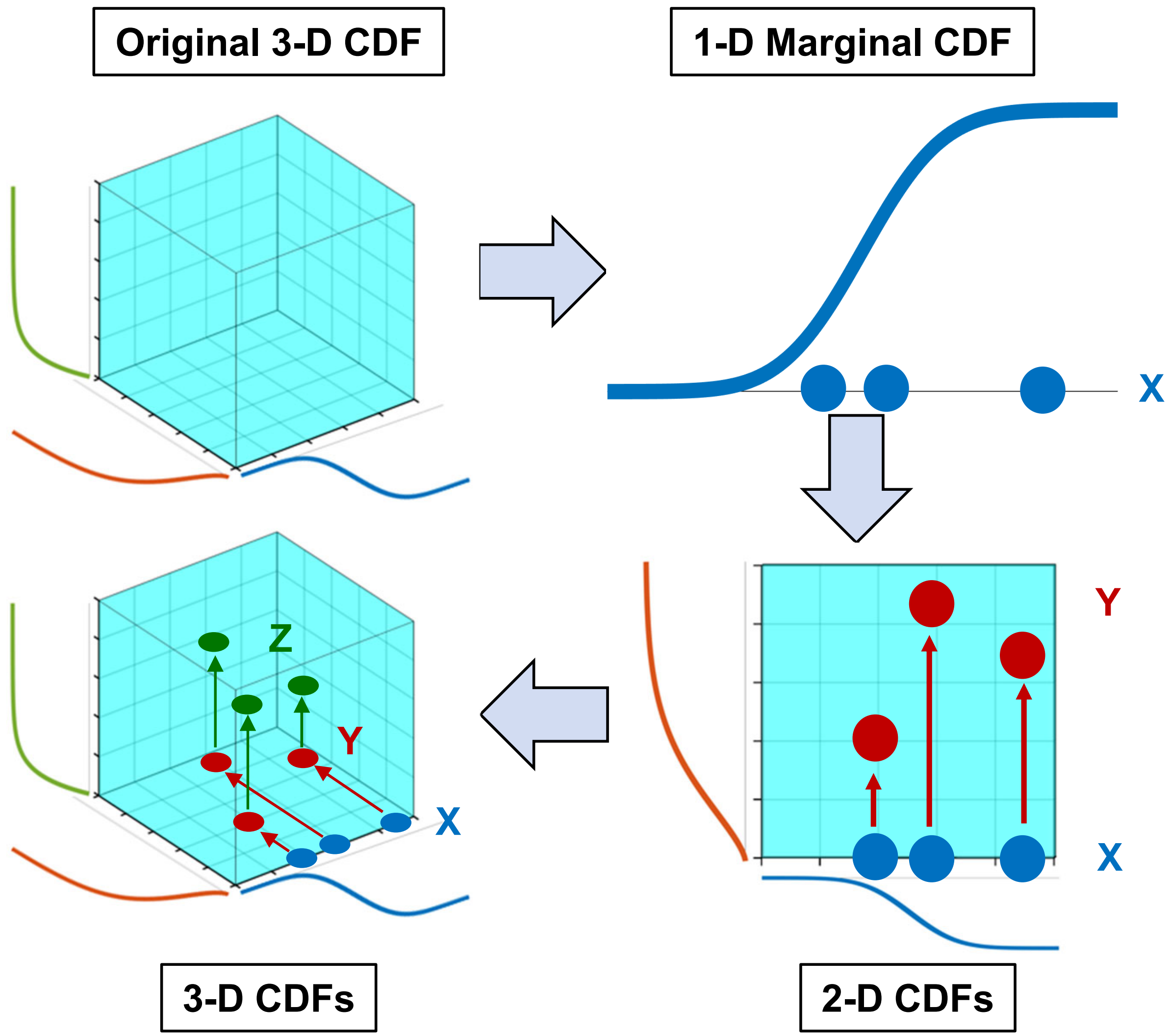
- Motivation / problem
- Existing approaches
- Method & implementations
- Practical aspects and areas for improvement
- Summary

Summary

- In response to Monte Carlo aircraft wake modeling needs, a generalized tool was developed to provide large joint N-variable random distributions from observations
- The basis of the tool is inverse transform sampling, whereby samples from a uniform distribution are used to index into an empirical CDF to yield realistic distributions of the base variable(s)
- The tool was successfully used to generate 10M operational (2 variables) and environmental (5 variables) joint random samples whose distributions matched closely those of the originating observations
- The tool has been packaged for general use and is publicly available

Extension to Three Dimensions and Beyond

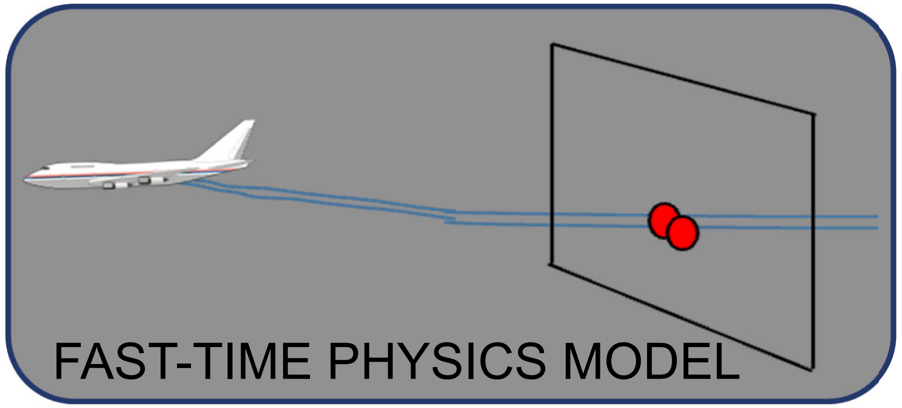
- Generate matrix of random values R from a uniform distribution of range [0,1] and size (N desired samples) x (M variables)
- Compute marginal CDF for 1 of the dimensions. Use random indices for the 1st variable of R (R1) to find corresponding random values X.
- Determine CDF slices across the remaining (2nd & 3rd) dimensions of the original PDF for each of the values in X. Use the corresponding random indices in R2 to determine Y.
- Repeat with the PDF/CDF across the last (3rd) dimension of the original PDF for each of the value pairs (X,Y). Use the corresponding random indices in R3 to generate Z
- The set {X,Y,Z} comprise the final joint distribution of random values
- Method is extensible to M>3 variables



CDF = Cumulative Distribution Function
PDF = Probability Density Function

5-Dimensional Implementation: Problem

- **Wake behavior modeling task required joint random distributions of turbulence dissipation rate, stability, headwind and crosswind (on arrival), and temperature**
- **Binning/resolution adjustable to the requirements of each variable**
  - **100 bins for all except temperature (50)**
  - **Resolution was deemed adequate for this task**
- **Turbulence has an exponential distribution, sampling is done on a log basis for improved resolution**
- **Generated joint distribution of 10M instances closely matches the characteristics of the source distribution when viewed across all 2-way combinations**



Variable	N Bins	Resolution	Units
landing weight	100	1010	lbs
landing speed	100	0.23	kts
eddy dissipation rate	100	1.29E-03	m^2 / s^3
stability	100	2.88	K / km
headwind, crosswind	100	0.27	kts
temperature	50	0.99	°C

GE0

Slide 21

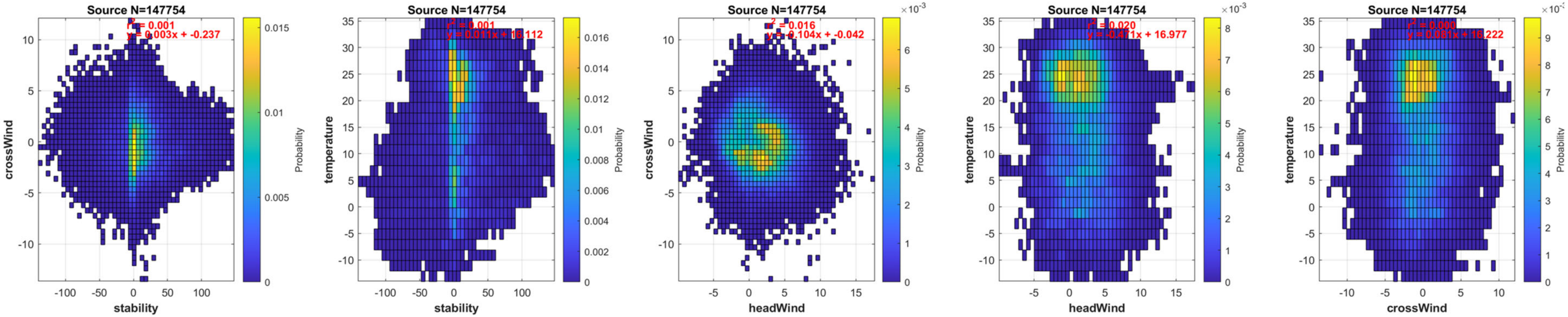
GE0	The font in this table should be Arial for consistency. Do we have the original? Enea, Gabriele - 0443 - MITLL, 2025-01-06T20:29:53.414

Fidelity of Joint Aspect of Distributions (2/2)

Visualized through examining the 10 2-way combinations of the 5 environmental variables

Generally, very good agreement seen across all variable combinations

Source (N=147754)



Sample (N=10M)