Hybrid Simulation and the Structure of HDF5 for ML Analysis

Shinshu University
M1 Atsushi Saito

2025/10/15
Telescope Array Data Analysis Machine Learning
@IoP, Academia Sinca, Taiwan

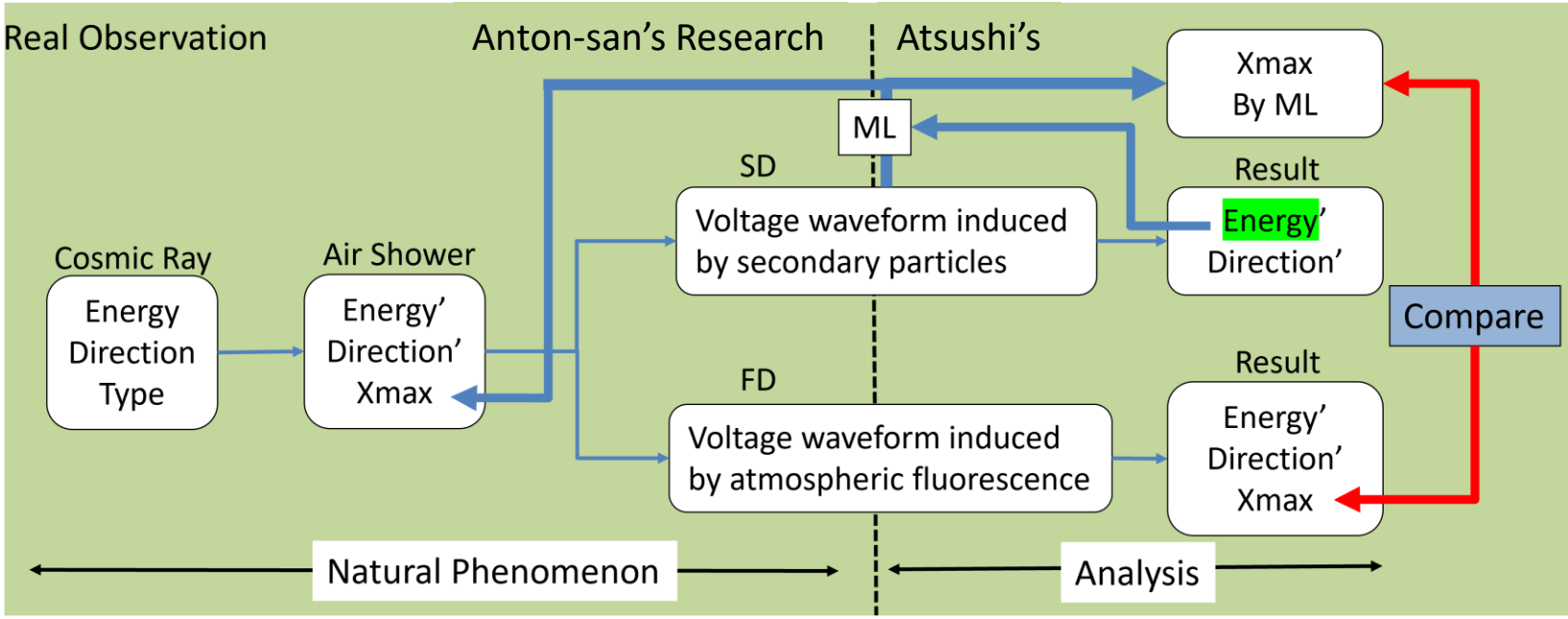
Contents

- Overview and Purpose of this research
- Comparison of SD and FD Observations & Reconstruction Methods
- Comparison of the Analysis Flow: Real Observation vs. MC Simulation
- Why use the FD-reconstructed Xmax as a reference for comparison
- Current Status
- Hybrid Data Set and HDF5 file structure

Overview and Purpose of this research

Anton-san's Purpose : To predict the true Xmax of air showers from SD data using Machine Learning(ML)

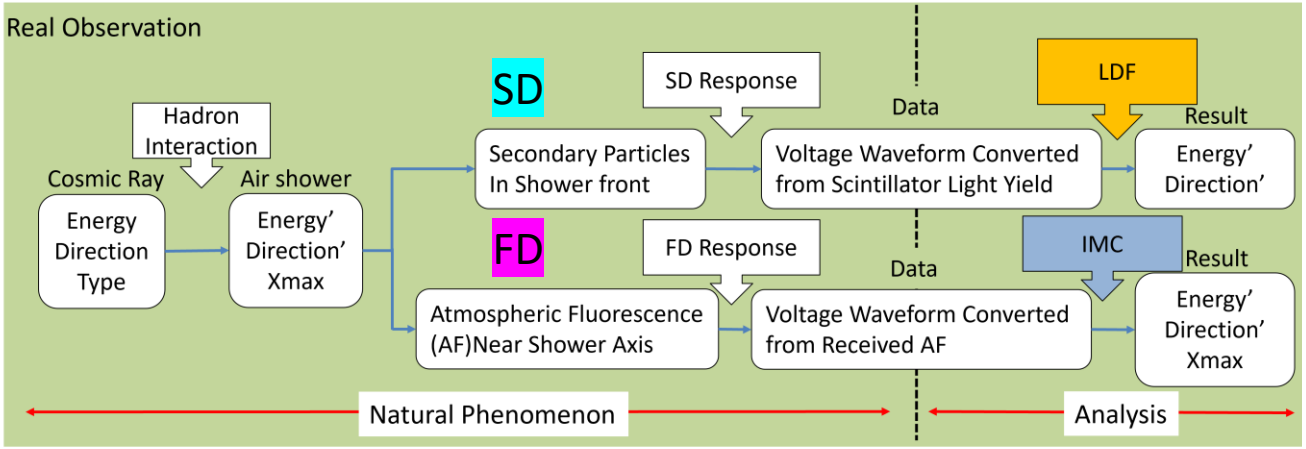
Motivation: How much error can a ML-based analysis method have when predicting the true Xmax?



Purpose of this : To estimate the error of ML-based analysis methods

Method: To Compare the Xmax predicted by ML vs the Xmax reconstructed by FD

Comparison of SD and FD Observations & Reconstruction Methods



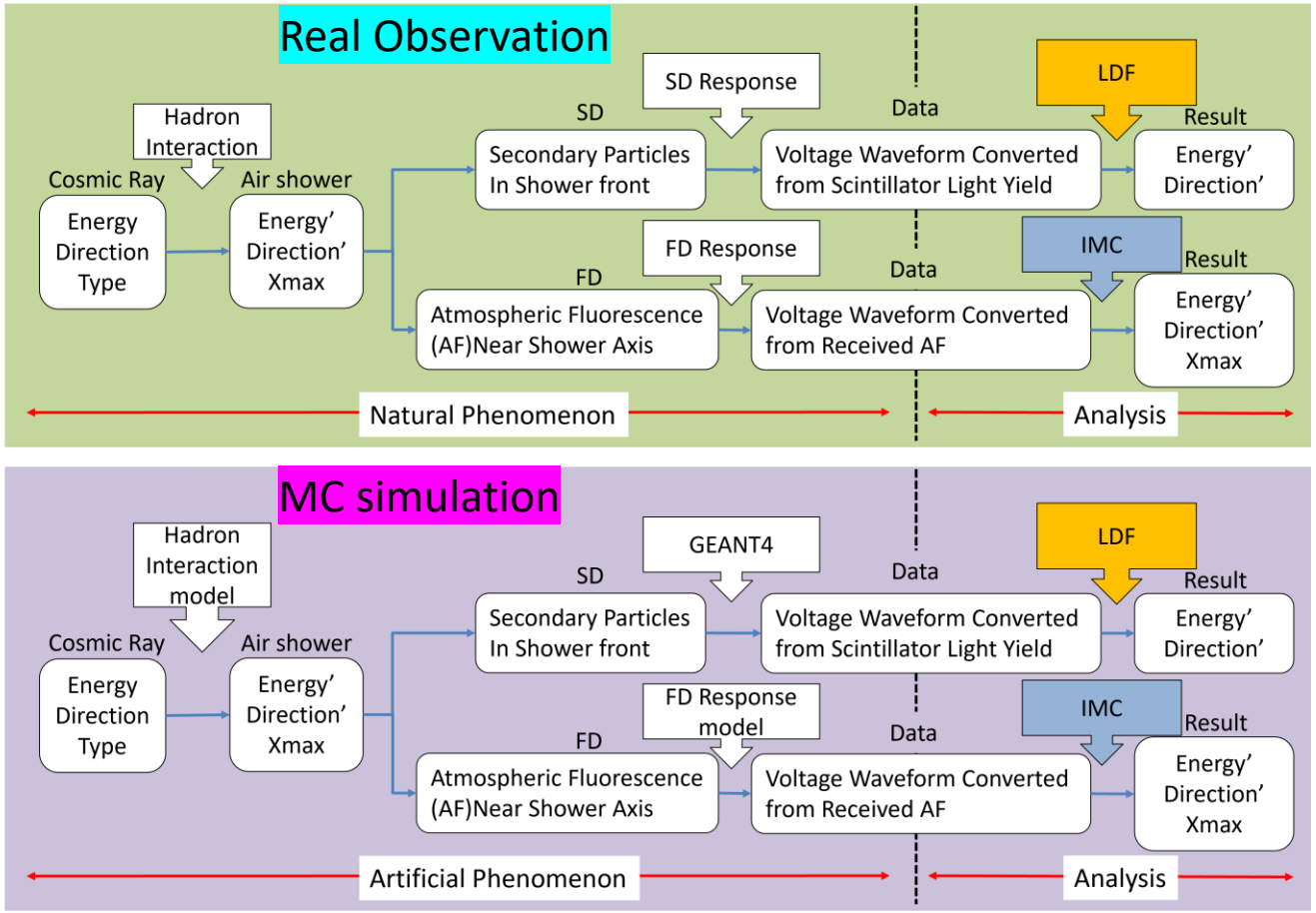
	SD	FD
Target	Shower front	Shower axis
Object	Secondary particle	Photons
Method	Direct/ Sampling	Indirect/ Part of shower development
Measured quantity	Particle density	Received light amount
MC dependence	dependent	independent

SD estimates the primary energy based on the particle density and lateral distribution (= LDF)
→ This method relies on a lookup table and rainbow plot

FD determines the Xmax that best reproduces the observed data along the shower axis (= Invers Monte Carlo)
→ This method is based on observational data

The reconstruction method of the SD depends on MC simulations,
While The reconstruction method of the FD does not depend on MC simulations

Comparison of the Analysis Flow: Real Observation vs. MC Simulation



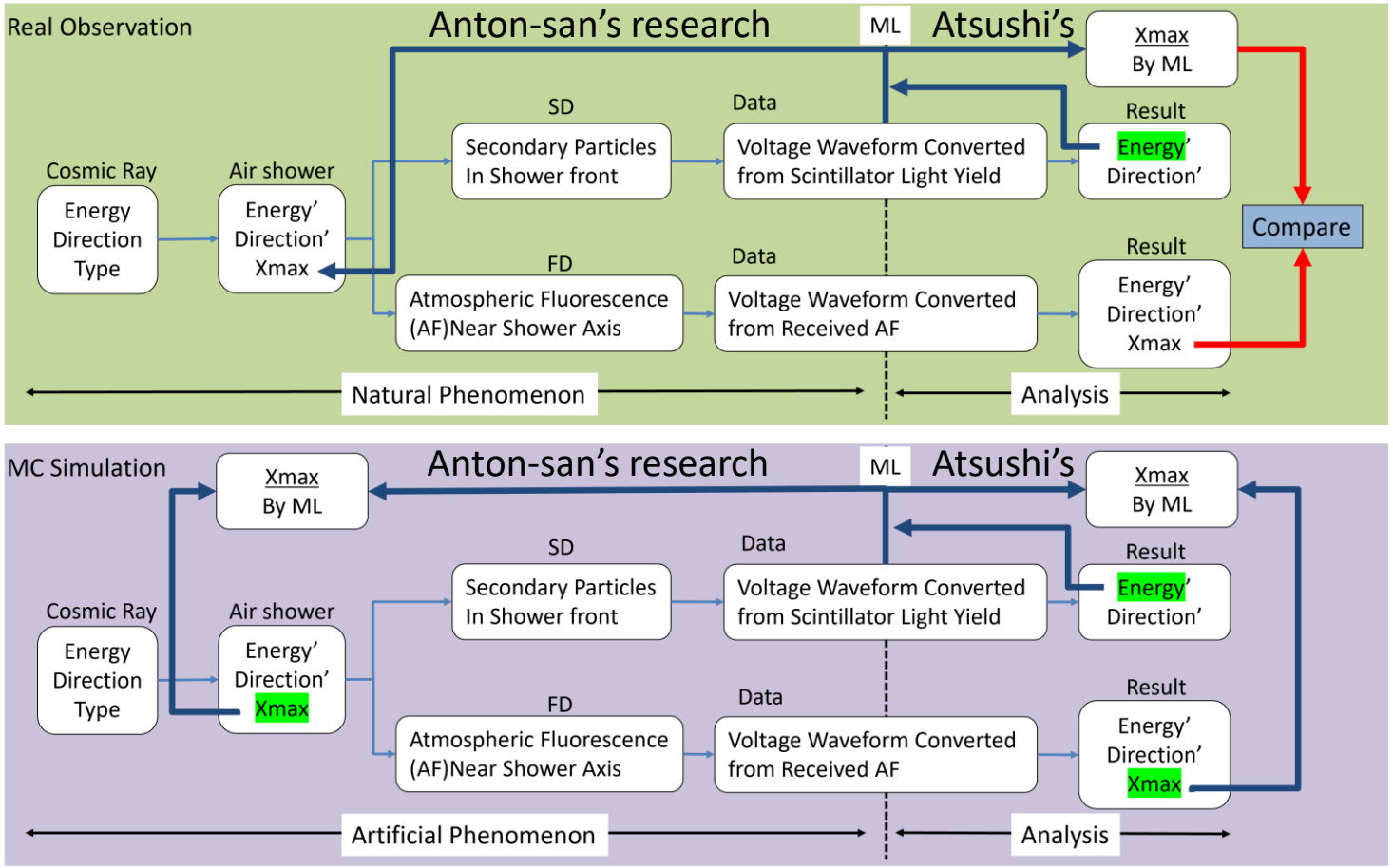
	Real Observation	MC simulation
Experimental setup	Natural/ Continuous	Nature-like/ Discrete
Interaction	Occurs as a natural phenomenon	Carried out Probabilistically
Detector Response	Continuous/ Non-uniform	Sampled/ Uniform
Analysis	LDF, IMC	LDF, IMC

MC simulations cannot fully reproduce the natural phenomena observed in real data

The same reconstruction programs for the SD and FD are used for both real observations and MC simulations

The analysis procedures are minimally model-dependent and capable of reproducing real observations.

Why use the FD-reconstructed Xmax as a reference for comparison



To estimate the error, the following conditions must be satisfied:

①Comparable with real observational data

②Less depends on Models

③Able to reproduce the conditions of real observations

These conditions are satisfied by the **Xmax reconstructed by the FD**.

To use the Xmax reconstructed by FD as a reference for comparison, we need **construct a ML model** that predicts the FD-reconstructed Xmax,

AND THEN,

compare the Xmax predicted by ML

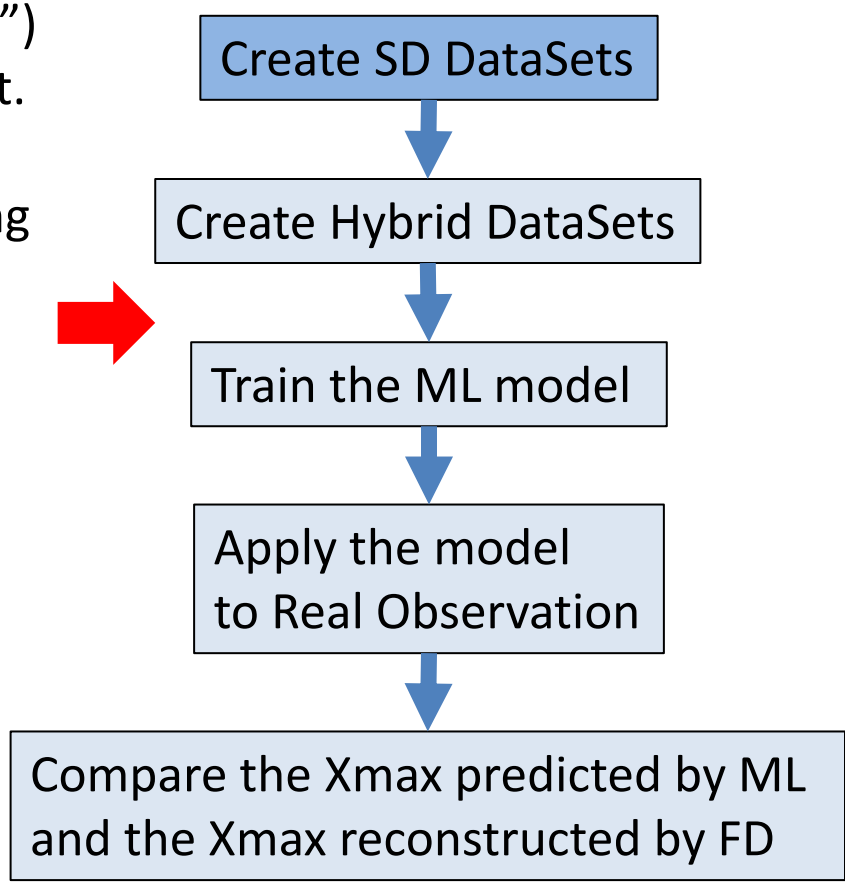
And the Xmax reconstructed by FD

Current Status

- Created SD DataSets → Anton-san's SD DataSet(includes "Bad SD info")
- Created Hybrid DataSets → The amount of data Sets is sufficient for a Test. (have been creating)
- Created HDF5 files includes SD-FD event parameters for ML model training

Current:

Checking whether the Hybrid DataSets are appropriate



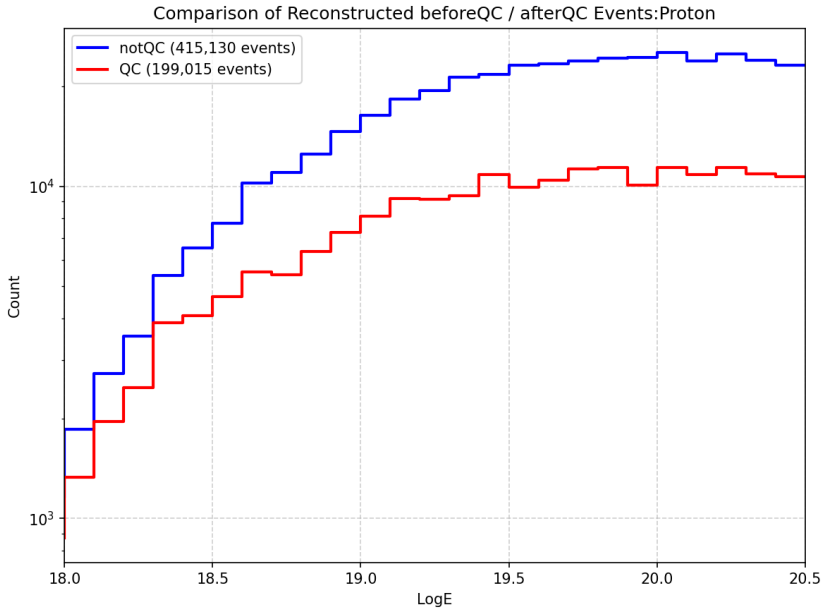
MC DataSets (Thrown)

Particle:	Proton, Iron
Interaction model:	QGSJET-II-04
LogE:	18.0 ~ 20.5 (0.1 energy bin)
Shower Types:	1,000 showers / bin
Energy Distribution:	E^{-1}
Observation period:	2008/04/17 ~ 2016/06/03 (3,000 days)
FD Station:	BR

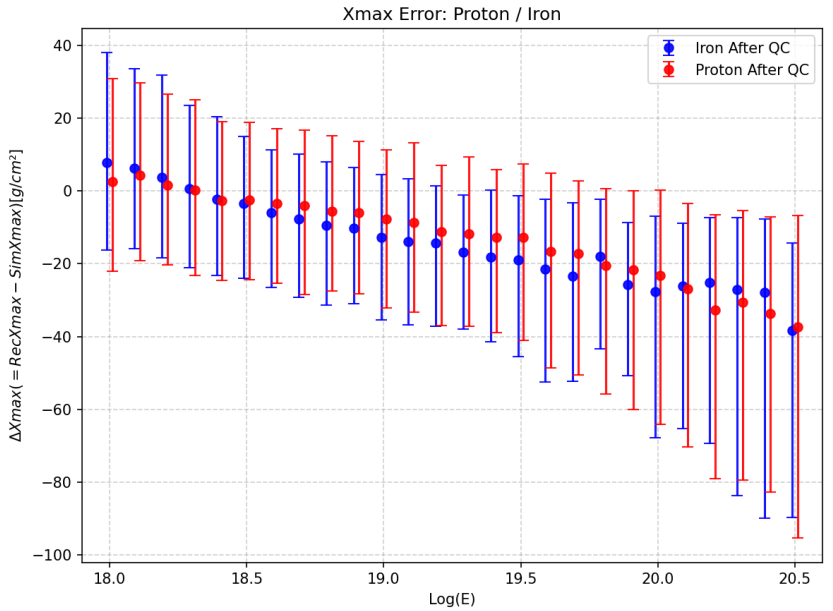
Hybrid Data Check

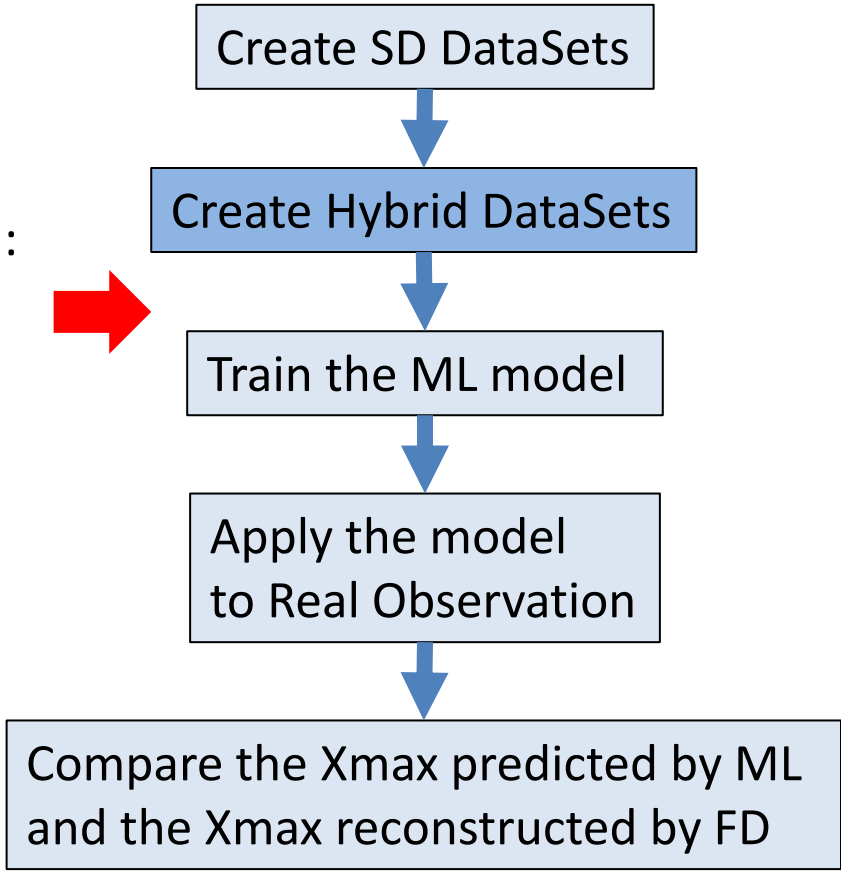
Quality Cut for FD

- Number of PMT > 20
- Shower track length > 15 deg
- Zenith angle < 55 deg
- Minimum viewing angle > 20 deg
- Xstart < Xmax < Xend



The error bar plot below is made from: median and 68% percentile.





I am in the process of investigating the cause of a potential problem in the simulation.

HDF5 for ML model training

Dstparser:

The tool to convert files to HDF5(.h5)

HDF5 is binary file for ML model training

HDF5

SD's detector_info(7 x 7) -> compress
- arrival_times_[up/low]
- detector_position
- detector_status…etc.

SD's standard_reconstructed_data()
- reconst_Energy
- reconst_S800
- shower_core
- shower_axis…etc.

Anton-san made the HDF5 file format

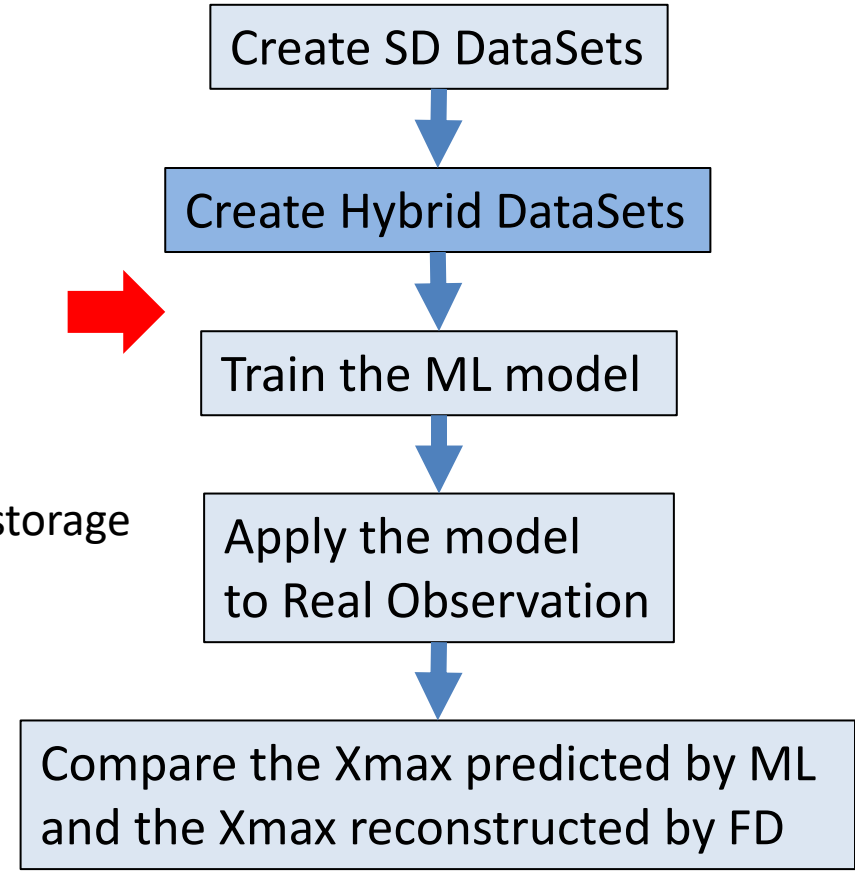
The HDF5 file includes 2types :

・7 × 7 tile-format

・the scalar quantities

The 7 × 7 SD tile is further compressed for storage

Step 1. Coordinate Transformation
Step 2. Binning
Step 3. Physical Quantity Mapping



HDF5 includes FD parameters for ML model training

Dstparser:

The tool to convert files to HDF5(.h5)

HDF5 is binary file for ML model training

HDF5

SD's detector_info(7 x 7) -> compress

- arrival_times_[up/low]
- detector_position
- detector_status…etc.

SD's standard_reconstructed_data()

- reconst_Energy
- reconst_S800
- shower_core
- shower_axis…etc.

FD's standard_reconstructed_data()

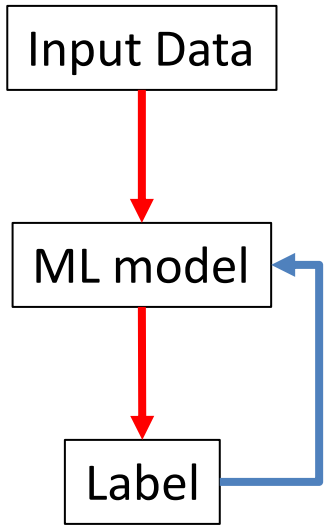
- thrown_Xmax
- **reconst_Xmax**
- thrown_LogE0
- reconst_LogE0

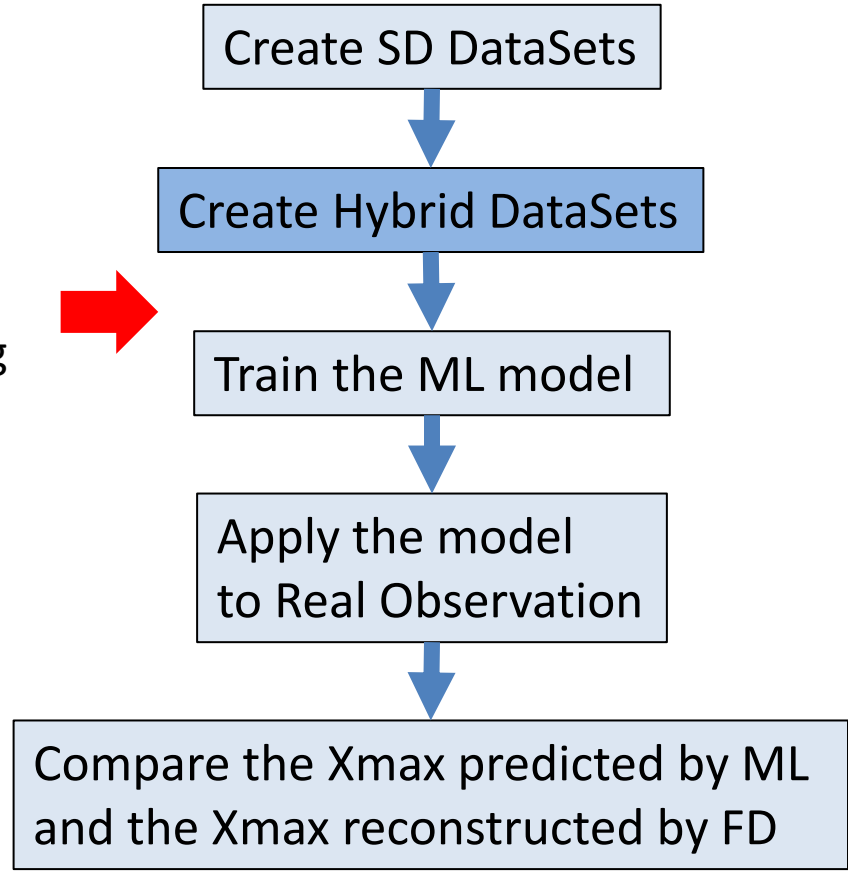
Input Data

Label

In accordance with the same format,

- Perform SD-FD event-by-event matching
- Append the FD-reconstructed to the file





Summary

Summary:

・The Xmax reconstructed byFD satisfies the necessary conditions to serve as a reference for comparison.

①Comparable with real observational data

②Less depends on Models

③Able to reproduce the conditions of real

・Currently validating the quality of the Hybrid Dataset.

An issue has been identified where the error is larger than expected, suggesting a potential problem in the simulation.

Future:

Investigate and resolve the cause of the Xmax error.

Thoroughly check the simulation and reconstruction process.

Proceed to train the Machine Learning model.

Once the dataset quality is confirmed, I will begin training the model.

Evaluate the model performance.

Apply the trained model to real observation data

and compare its predictions with FD-reconstructed Xmax to **the perspective of uncertainty**

Finally, I would like to thank all the members who supported my two-month research stay and made this work possible.
Thank you very much.

SHINSHU UNIVERSITY

Back Up