MLARE: ML-Driven Atmospheric Retreival of Exoplanets

Swastik Dewan [1] Tasneem Basra Khan [1]

[1]National Institute of Science Education and Research (NISER) , Bhubaneswar

Introduction

Exoplanets are distant worlds with atmospheres that remain a mystery to us. Studying their atmospheres has long been difficult for astronomers. However, new machine learning techniques are providing a powerful way to unlock the secrets of these alien skies. By training on large amounts of simulated data, machine learning models can decode the information hidden in telescope observations like spectra and light curves. This allows us to discover the gases present, temperature structures, and atmospheric motions on exoplanets light-years away.

In this poster, we present a new ML model which hasn't been used previously in the exoplanet community and evaluate the performance of our model.

Comparative Analysis of ML models

We tested different models with HELA dataset and the R^2 are presented below. The table has model R^2 values and the graphs plotted have individual RvP assessments for the labels in testing dataset.

ML Baseline algorithms	R^2 score
SVR	0.570
XGBoost	0.732
Random Forest	0.747
Gradient Boosting	0.660
kNN	0.744

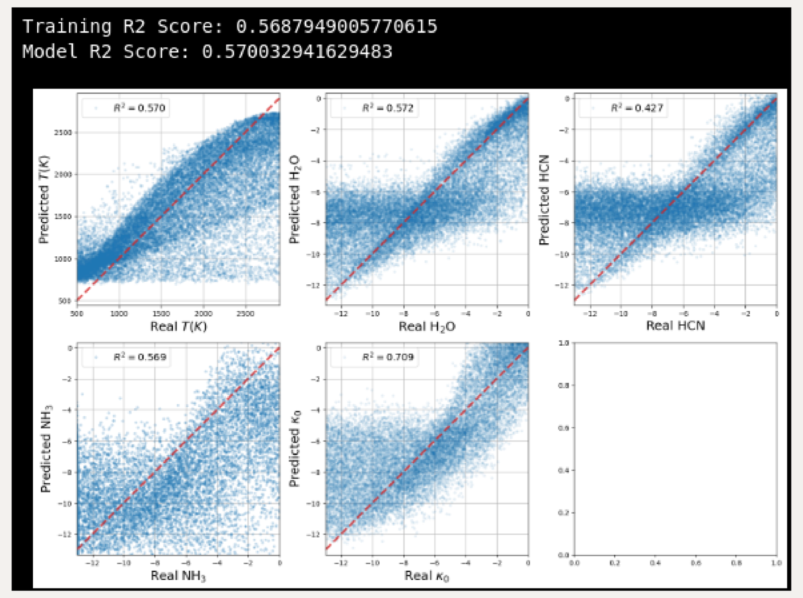


Figure 1. RvP plot for SVR

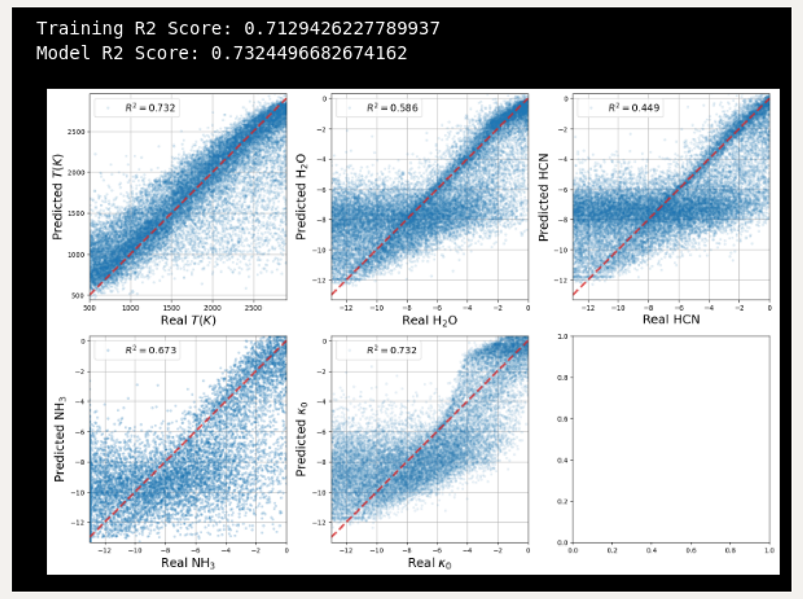


Figure 2. RvP plot for XGBoost

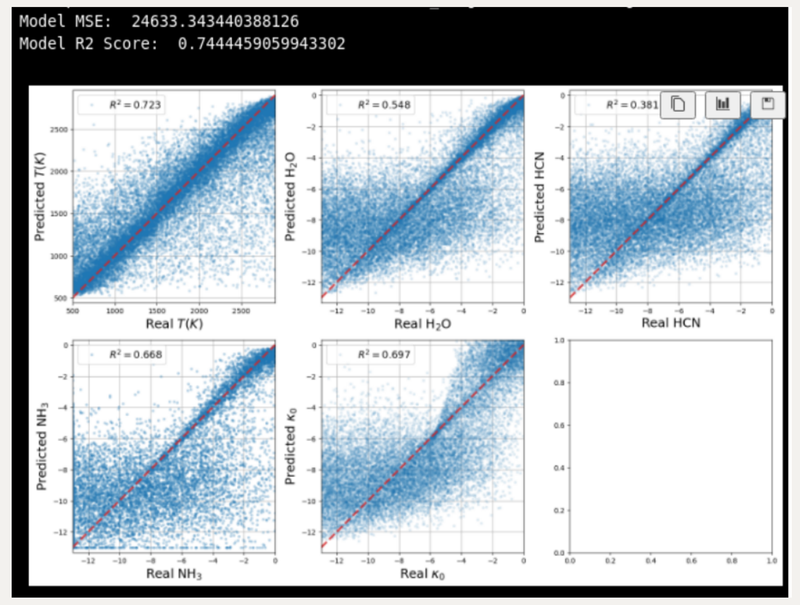


Figure 3. RvP plot for kNN

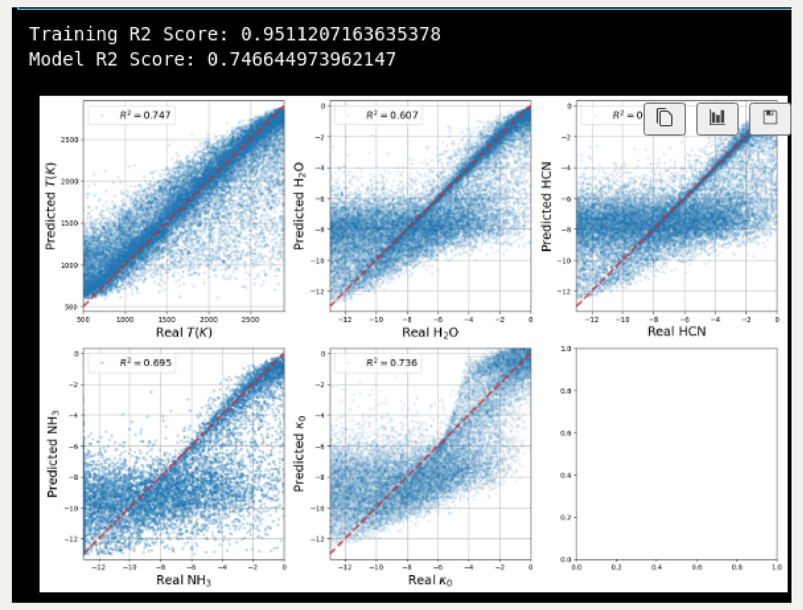
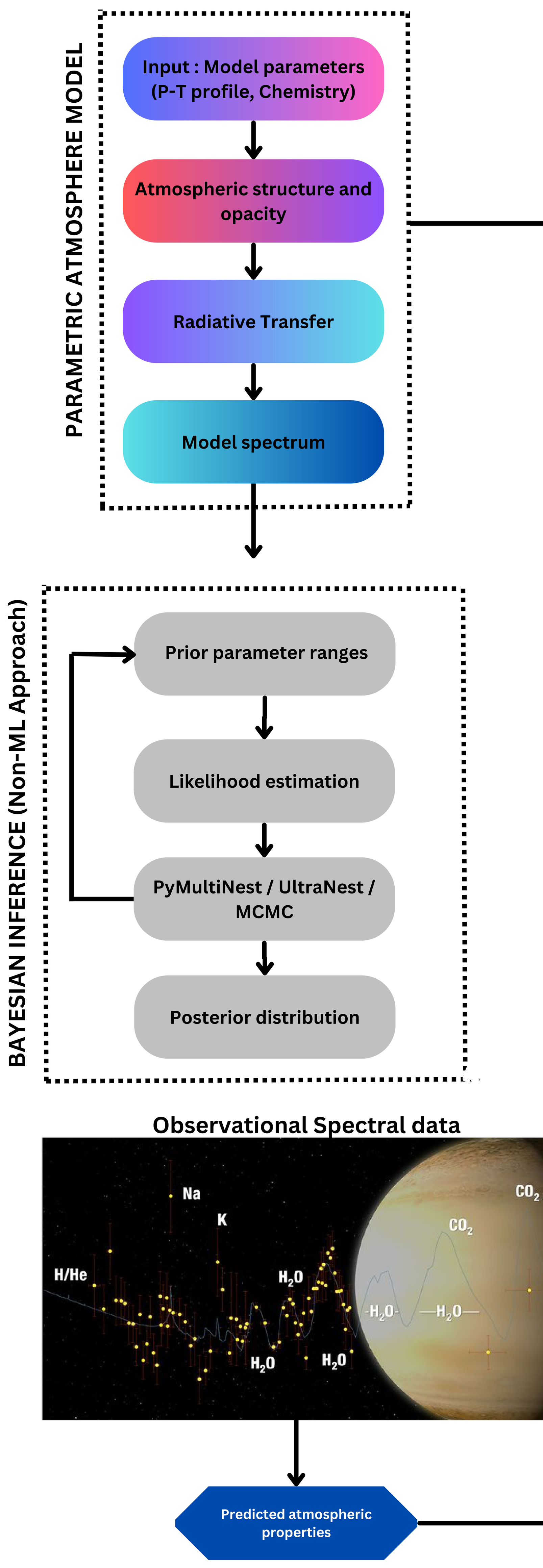


Figure 4. RvP plot for Random Forest

Methodology

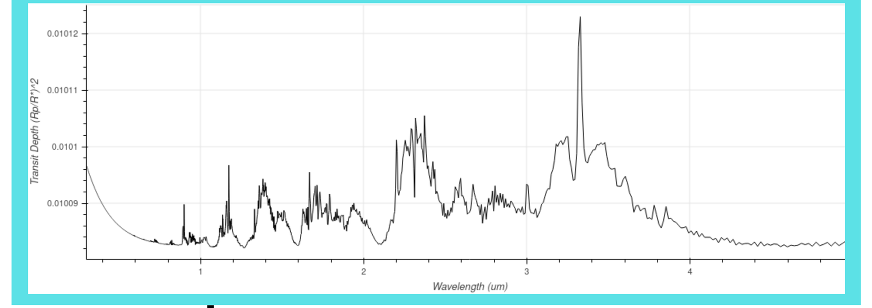


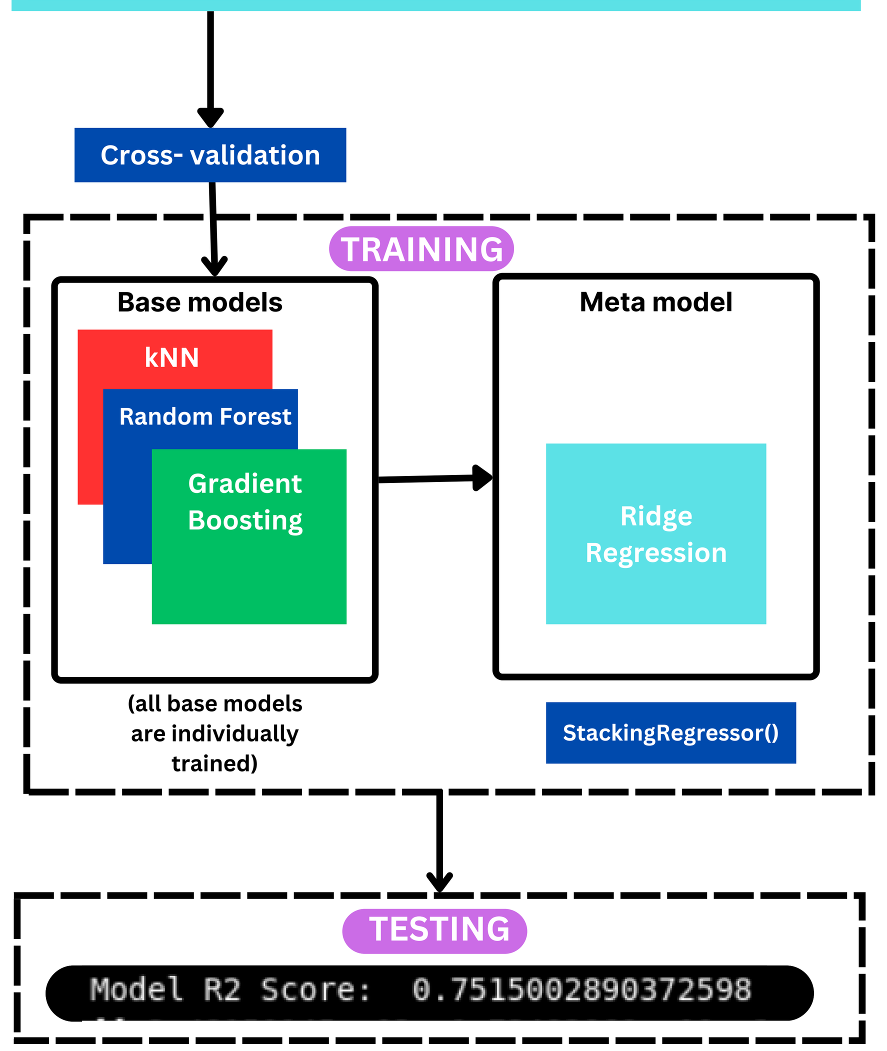
HELA DATASET

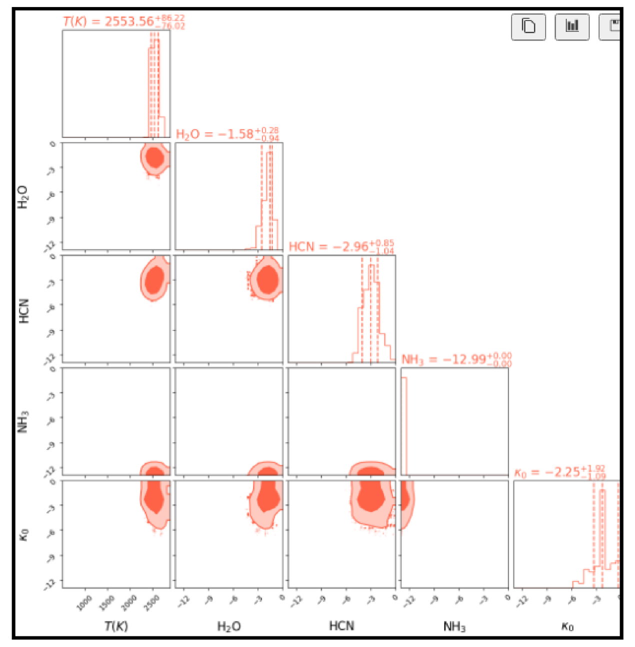
- Synthetically generated WFC3 Transmission spectra : **80,000 training spectra** and **20,000 testing spectra**
- 5 parameters/labels : **Temperature** (T), logarithmic volume mixing ratios of **water**(X_{H_2O}), **ammonia**(X_{NH_3}), **hydrogen cyanide**(X_{HCN}) and **constant cloud opacity** (κ_0)
- Considered a standard dataset in atmospheric retrieval community

CURATED DATASET from ExoCTK generic grid

- Synthetically generated **112640 synthetic transmission spectra** from **ExoCTK generic grid** through **web-scrapping**
- 8 parameters/labels : **Temperature** (T), **gravity** (g), **radius of planet and star** (R_p and R_s), **metallicity**, **C/O ratio**, **cloud** and **haze**
- Parameter ranges for each are taken from this paper.[1]







Results

Our model outperforms HELA model in terms of accuracy (R^2) score and it can be clearly seen in the following RvP assessment of the five labels.

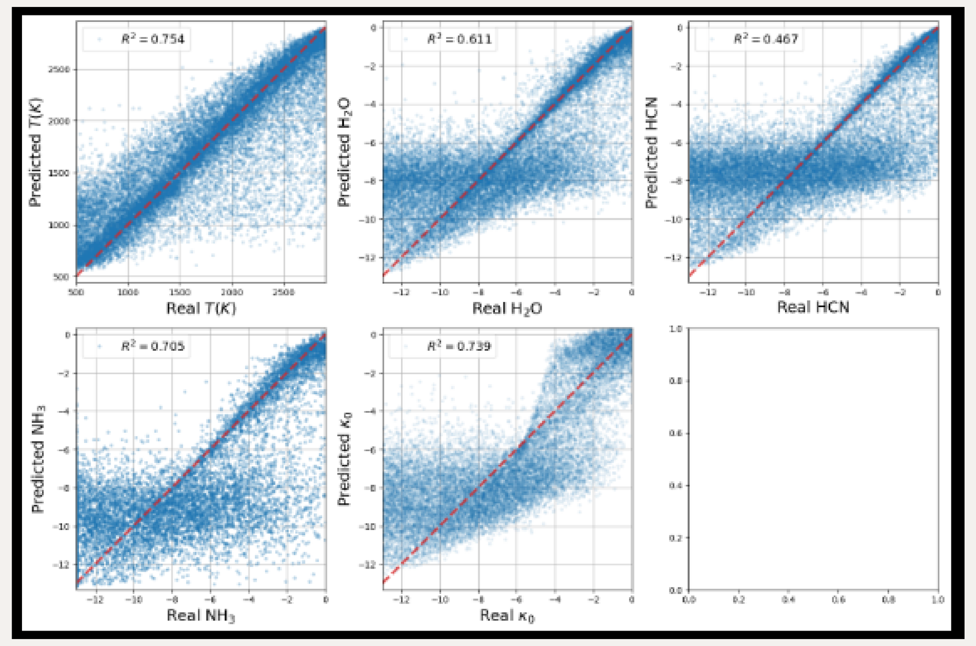


Figure 5. RvP plot for ensemble Stacking model

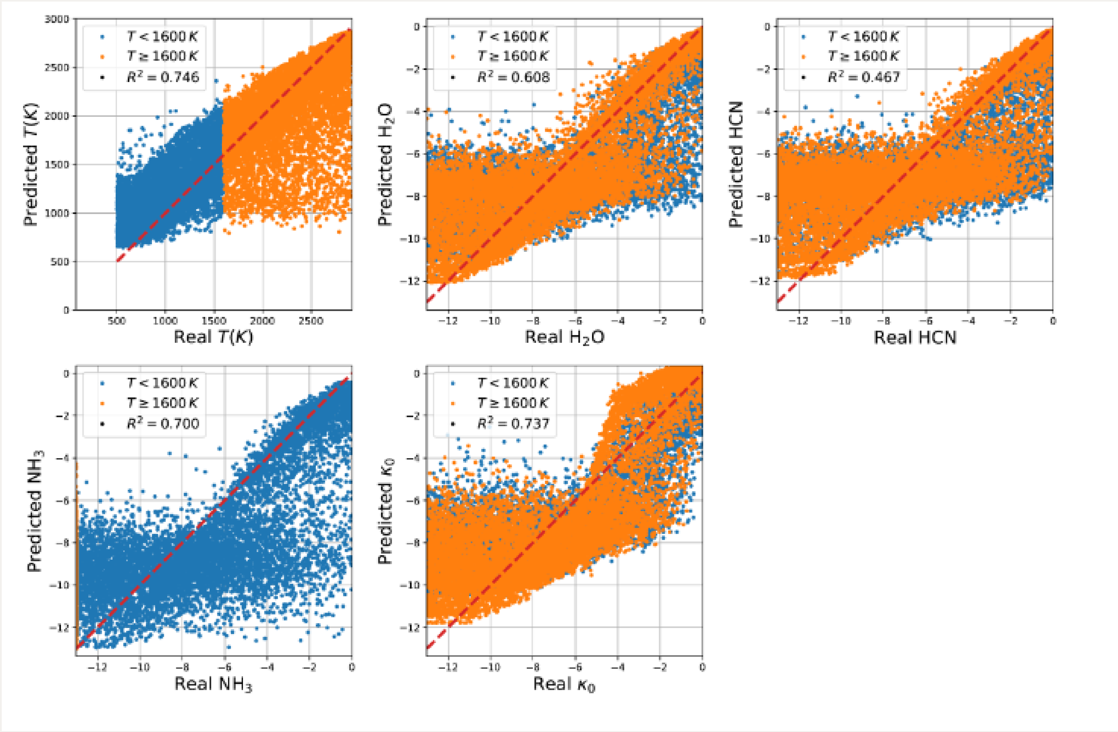


Figure 6. RvP plot as published in HELA paper

Other than this, various ML models were run on our dataset (Generated using ExoCTK) like CNN, KNN, RF , Gradient Boosting. Work has to be done on improving the R^2 score using the optimal algorithm.

Future Plans

1. A more complex model, or feature extraction on the generated ExoCTK dataset may lead to better evaluation score. That will be our primary motive heading on.
2. The Stacking model used is novel in the atmospheric retrieval community, and MLARE will be developed to provide a comparative retrieval study of different ML algorithms never done before in exoplanetary research, which will be run simultaneously on the same exoplanetary spectra to predict atmospheric parameters.
3. Fully developed MLARE will be integrated into an in-house atmospheric retrieval model. MLARE can be used on its own to retrieve atmospheric parameters or combined with general sampling methods of the cumulative model, to generate better and faster results.

References

Scan the QR to get all the references used in the poster.