

# A Quantitative Analysis of Exoplanetary Databases

Presented to the faculty of Lycoming College in partial fulfillment  
of the requirements for Departmental Honors in  
Department of Astronomy and Physics

by  
Maha Soojaysingh K. Jhugaroo  
Lycoming College  
5/4/2023

Approved by:



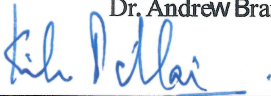
Dr. Emily Wilson



Dr. David Fisher



Dr. Andrew Brandon



Dr. Krish Pillai

# A Quantitative Analysis of Exoplanetary Databases

Soojay Jhugaroo

(Dated: 5/5/2023)

## Abstract

After exoplanets are discovered through sky surveys, such as TESS, Kepler, KELT etc., the parameters of these exoplanets are resolved by different groups of researchers and are published in astronomical journals. Because the methodology used in evaluating exoplanetary parameters are different among research groups, different peer-reviewed articles will often differ in resolved parameters for the same exoplanet. This becomes a concern when individual catalogs, with dissimilar inclusion criteria, enter those parameters in their database. Disagreements in their data displayed is bound to happen. This project aims to quantify the discrepancies in exoplanetary parameters listed among these databases through a comparative analysis followed by a highlight of the reasons behind these differences. This work specifically focuses on the most widely-used publicly available databases which comprises of NASA Exoplanet Archive (NEA), Exoplanet.eu (EU) and Open Exoplanet Catalogue (OEC). The statistical test used for comparisons was a two-sample Kolmogorov-Smirnov test and a two-sample Anderson-Darling test. Preliminary results show that there is no significant disagreement in mass, radius and orbital period parameters among the databases. However, analysis of parameter distributions reveal key differences about the catalogs.

## CONTENTS

List of Figures	2
List of Tables	4
I. Introduction	7
A. Exoplanet Detection & Cataloguing	7
B. Inclusion Criteria	8
II. Methodology	9
III. Data Processing	9
A. NASA Exoplanet Archive (NEA)	9
B. Exoplanet.eu (EU)	12
C. Mass Sin ( $i$ )	13
D. Open Exoplanet Catalogue (OEC)	13
E. Database Planet Matching	16
IV. Statistical Result Interpretation	17
A. Cumulative Distribution Function Plot	18
B. 2-Sample Kolmogorov-Smirnov Statistic	18
C. 2-Sample Anderson-Darling Statistic	19
D. D-statistic, AD-statistic and $p$ -values	20
E. Kernel Density Estimation)	20
V. Results & Discussion	22
A. Biases	23
B. Differences	24
VI. Conclusion	25
VII. Future Work	25

VIII. Acknowledgements	25
IX. Appendix	26
References	28

## LIST OF FIGURES

1	The data in column “pl_refname” that were used to filter the exoplanets to select entries with the most recent resolved parameters. ....	11
2	Only numbers left in the “pl_refname” column (2nd entry in figure) after all characters were removed using the Re library. The first 4 numbers represent the publication year. ....	11
3	Commented code to select resolved exoplanet parameters from the most recent publication. ....	12
4	This figure [9] is a depiction of inclination $i$ which is the angle between an exoplanet’s orbital plane and the plane of the sky [11]. ....	14
5	Commented code to retrieve data files from OEC GitHub repository. ....	14
6	Console showing array with empty strings as well as NONE values. ....	15
7	Code converting all strings to NONE value and converting array to float. ....	15
8	Two databases being matched results in three subsets. The match, unique A and unique B subsets. ....	16
9	CDF plot of NEA and EU with the arrow showing the KS statistic computed.	18
10	CDF plot visually illustrating a 2-sample AD statistic [3]. ....	19
11	Illustration of a KDE plot (right) of a distribution as compared to a histogram (left). [5] ....	21
12	Different types of kernels available on the Scikit-learn Python library. [1] ....	22
13	Representation of a KDE as the average of kernels. [5] ....	22
14	Mass PDFs of each database’s match (red, blue and green) and unique subsets (tan). ....	22
15	Radius PDFs of each database’s match (red, blue and green) and unique subsets (tan). ....	22
16	Orbital Period PDFs of each database’s match (red, blue and green) and unique subsets (tan). ....	23
17	CDF plot for mass distributions of NEA, EU and OEC. ....	26
18	CDF plot for radius distributions of NEA, EU and OEC ....	27

19	CDF plot for orbital period distributions of NEA, EU and OEC .....	27
20	Summary of the dataprocessing procedure for all databases. ....	28

## LIST OF TABLES

I	Summary of statistics of NEA, EU and OEC as of 4/29/2023. ....	8
II	Exoplanet HD 142 b has 4 publications for its resolved parameters with the 2022 publication being the most precise and having less error (NASA Exoplanet Archive, 2022). ....	10
III	Planet Kepler 381b is listed under slightly different name variations in NEA, OEC and EU. ....	16
IV	$p$ -values obtained through the KS and AD tests for all exoplanetary parameters compared between databases for matched subsets. ....	23

## I. INTRODUCTION

### A. Exoplanet Detection & Cataloguing

Exoplanets are planets found orbiting other stars outside the Solar System. Ever since the discovery of the first exoplanet in 1992 by Wolszczan and Frail [10], over 5347 exoplanets have been discovered and confirmed as of May 2023 by both space-led missions and ground-based observatories [7]. The vast diversity of exoplanets discovered over the years have contributed to the improvement of our planetary evolution models while also challenging existing theories. For instance, in 2019, a giant exoplanet of 0.46 Jupiter masses ( $M_J$ ) was detected orbiting a very-low-mass (M-dwarf) star of 0.6 Solar masses ( $M_\odot$ ) which is unusually high based on our prior predictions through simulations for such a small star [6].

As more exoplanets are discovered, several groups of people have compiled and categorized them into publicly available databases to provide the astronomical community with comprehensive data about these extra-solar bodies. These databases include key parameters about many exoplanets such as the planet’s mass, orbital period and eccentricity, radius, method of detection, distance to its host star, and information about the host star among others. Data from these databases have been crucial in the statistical studies of exoplanets. For example, Bean et al. [2] used a powerful statistical marginalization approach in which they constrain a huge exoplanetary survey under parameters and measurements we deem necessary for a planet to harbor life. This statistical approach provides context which helps in understanding individual exoplanets and allows broad conclusions to be drawn. Furthermore, it has advantages over contemporary “planetary systems science,” which consists of analyzing comprehensive data about individual exoplanets to infer its habitability, by overcoming the lack of adequate individual planetary data due to current instrument limitations. Found below are the most widely used exoplanetary databases:

- NASA Exoplanet Archive; [exoplanetarchive.ipac.caltech.edu](https://exoplanetarchive.ipac.caltech.edu)
- Exoplanet.eu; [www.exoplanet.eu](http://www.exoplanet.eu)



- Open Exoplanet Catalogue; [www.openexoplanetcatalogue.com](http://www.openexoplanetcatalogue.com)

## B. Inclusion Criteria

Because each of these databases are updated by different groups of people, each of them has their own criteria when deciding which planet is to be included in their databases (inclusion criteria). The first inclusion criterion is the exoplanet’s mass. The NASA Exoplanet Archive (NEA) only includes planets of less than  $30 M_J$  while Exoplanet.eu (EU) only includes planets of mass less than  $60M_J$ , and no information is provided for Open Exoplanet Catalogue (OEC). The second inclusion criterion is the confidence criteria or literary references of the planet’s confirmation in peer-reviewed journals. While NEA and EU only include exoplanets that have been published in peer-reviewed journals, OEC is open-source which means that anyone, from professional astronomers to the general public, can make contributions to the database. This signifies that it potentially includes objects that have not yet been confirmed to be exoplanets and that have been retracted or are controversial. These inclusion criteria disparities make a difference in what is listed in the different databases and account for exoplanetary parameter inconsistencies. For example, exoplanet ‘51-Eridani-b’ is recorded with a planetary mass of  $2M_J$  in OEC and NEA but as  $9M_J$  in EU. Exoplanet ‘HD181720-b’ is recorded with a mass of  $0.37M_J$  in NEA while on EU it has a recorded mass of  $12.12M_J$ . The planetary radius of ‘CoRoT-21 b’ is listed as  $2R_J$  (Jupiter radius) in OEC and Exoplanet.eu but is not listed in NEA.

Catalogue	Mass Criteria	Confidence Criteria	Confirmed Exoplanets
NEA	$< 30M_J$	Peer-reviewed Articles	5347
EU	$< 60M_J \pm 1\sigma$	Peer-reviewed Articles	5365
OEC	-	Open-source	5040

**TABLE I:** Summary of statistics of NEA, EU and OEC as of 4/29/2023.

## II. METHODOLOGY

Due to the data being analyzed in this project consisting of mainly unpaired data coming from different ground-based and space-based telescopes (KEPLER, K2, TESS, KELT etc.) with unknown mathematical distribution of the observed planetary property, non-parametric statistical tests are crucial in the statistical comparison of such data. The Kolmogorov-Smirnov test (KS test) is widely-used in astronomy for data analysis for this specific reason. However, in literature, statistical tests that are much more powerful than the KS test have been tested such as the Anderson-Darling test (AD test) [8]. In the scope of this project, a two-sample KS test and a two-Sample AD test will be used since two sets of data from two databases will be compared at once. This research project was initiated by starting a pipeline in Python that includes a two-sample KS test. Afterwards, all the data from the publicly available databases were downloaded and retrieved from the databases mentioned above. These data were then carefully processed before being subjected to the KS and AD tests. For instance, to be safe, for a mass comparison, any planets above  $10M_J$  will be discarded from this analysis because they could potentially be brown dwarfs. This is because beyond  $10M_J$  the boundary between super-planet and brown dwarfs gets blurry despite the agreed lower mass boundary for brown dwarfs generally being agreed to be  $13M_J$  [12]. After processing, the different parameters from all the databases (mass, radius, orbital period etc.) were subjected to the KS test, The AD test and the results were be interpreted accordingly.

## III. DATA PROCESSING

### A. NASA Exoplanet Archive (NEA)

After having studied the statistics needed for my project and downloading the databases that I will be using for my comparisons as CSV files, I quickly realized that the data in those CSV files were not in a format ready for comparison. Each database needs to have the data to be compared cleaned before being subjected to the KS test. Because each

database catalogs exoplanets differently, they each came with their own sets of challenges when cleaning up the data. The first two databases that I wanted to compare were NASA Exoplanet Archive and Exoplanet.eu for the exoplanetary mass parameter. After retrieving data from NASA Exoplanet Archive (NEA) as a CSV file, the first issue was that there were 33376 confirmed exoplanet entries which is alarming considering that as of 10/11/2022 the database states having 5338 confirmed exoplanets. Upon closer inspection, it seemed that the same exoplanet appeared multiple times as they had multiple papers attempting to calculate their exoplanetary parameters and NEA listed all of them. I decided to use the mass parameter from the most recent publications for each exoplanet. This is because I noticed that for the most part, the most recent publications for planets were more comprehensive with parameter calculations, had the least error variability and had increased precision especially for planetary mass (see TABLE II).

Exoplanet Name	Publication Date	Mass ( $M_J$ )
HD 142 b	2002	$1 \pm 0.1$
HD 142 b	2006	$1.045 \pm 0.061$
HD 142 b	2012	$1.02 \pm 0.03$
HD 142 b	2022	$1.03838^{+0.4682}_{-0.050864}$

**TABLE II:** Exoplanet HD 142 b has 4 publications for its resolved parameters with the 2022 publication being the most precise and having less error (NASA Exoplanet Archive, 2022).

However, there was no easy way to filter for only most recent publications prior to downloading the database. Even after downloading the data as a CSV file, the column containing information about the data of the most recent papers resolving the planets’ parameters (namely “pl\_refname”) looks like gibberish. It only contains a publication link and the authors but nothing that could be directly useful in helping to sort, filter and select the most recently calculated exoplanetary parameters (see FIG. 1).

To resolve the issue, I first uploaded the CSV file onto Python (PyCharm IDE) as

```

pl_refname
<a refstr=LIU_ET_AL_2008 href=https://ui.adsabs.harvard.edu/abs/2008ApJ...672..553L/abstract target=ref> Liu et al. 2008 </a>
<a refstr=KUNITOMO_ET_AL_2011 href=https://ui.adsabs.harvard.edu/abs/2011ApJ...737..66K/abstract target=ref> Kunitomo et al. 2011</a>
<a refstr=DOLLINGER_ET_AL_2009 href=https://ui.adsabs.harvard.edu/abs/2009A&A...505.1311D/abstract target=ref> Dollinger et al. 2009 </a>
<a refstr=KUNITOMO_ET_AL_2011 href=https://ui.adsabs.harvard.edu/abs/2011ApJ...737..66K/abstract target=ref> Kunitomo et al. 2011</a>
<a refstr=STASSUN_ET_AL_2017 href=https://ui.adsabs.harvard.edu/abs/2017AJ....153..136S/abstract target=ref>Stassun et al. 2017</a>
<a refstr=SATO_ET_AL_2008 href=https://ui.adsabs.harvard.edu/abs/2008PASJ...60.1317S/abstract target=ref> Sato et al. 2008 </a>
<a refstr=KUNITOMO_ET_AL_2011 href=https://ui.adsabs.harvard.edu/abs/2011ApJ...737..66K/abstract target=ref> Kunitomo et al. 2011</a>
<a refstr=NAEF_ET_AL_2004 href=https://ui.adsabs.harvard.edu/abs/2004A&A...414..351N/abstract target=ref> Naef et al. 2004 </a>
<a refstr=ROSENTHAL_ET_AL_2021 href=https://ui.adsabs.harvard.edu/abs/2021ApJS...255....8R/abstract target=ref>Rosenthal et al. 2021</a>
<a refstr=GOZDZIEWSKI_ET_AL_2008 href=https://ui.adsabs.harvard.edu/abs/2008MNRAS.385..957G/abstract target=ref> Gozdziowski et al. 2008</a>

```

**FIG. 1:** The data in column “pl\_refname” that were used to filter the exoplanets to select entries with the most recent resolved parameters.

a Pandas library dataframe. For easier manipulation, I converted the database into a giant NumPy array so that I could easily index relevant data. In column “pl\_refname” the year and the date of publication of the paper was included and that was something that could be used to select the exoplanet mass of interest for each unique entry. I then proceeded to iterate through all “pl\_refname” entries using and eliminated all alphabetic characters leaving behind only numbers using the Re library. This left me with only numbers in “pl\_refname.” I could then select the planet having the highest “pl\_refname” number to select the most recent planetary mass parameter for any given planet in the database. Since the publication year was the first number to appear every time, any numbers that followed were extraneous.

```

[['xi Aql b' '20112011737662011' 2.02]
 ['xi Aql b' '20082008605392008' 2.8]]

```

**FIG. 2:** Only numbers left in the “pl\_refname” column (2nd entry in figure) after all characters were removed using the Re library. The first 4 numbers represent the publication year.

To select the planets with the highest number under the “pl\_refname” column, I again iterated all the entries in the database through a for loop. The code goes down the list and checks whether the name of the entry below the current one is similar or not. If not, it looks for all similar entry names in the database and compiles them in an array. For example, in FIG. 2 the code looked and for all planets named “Xi Aql b” and compiled all the entries in an array. It then selects the entry with the highest “pl\_refname” number in that array to select the most recent publication date. I then appended the entry of interest

to a list called “BestPar\_NEA.” Due to the way I have written the code, the loop will not consider the last unique planet entry in the database because it must check whether the name of an entry is not similar to the entry below it to consider it a unique entry. The last planet is as such not accounted for since there are no more entries below it, and I needed to consider it separately. After filtering the data through the loop, the “BestPar\_NEA” array was exactly 5187 entries in length which reflects the number of unique entries mentioned by NEA. The code in FIG. 3 is for the filtering procedure.

```

16 # Selecting most recent mass parameters from NEA
17 BestPar_NEA = [] # Initializing array to store planets with most recent publication
18 for i in range(0, len(df_np)-1):
19     if df_np[i,0] != df_np[i+1,0]: # Checking if planet name below is similar to selected planet name
20         x = df_np[np.where(df_np[:,0] == df_np[i,0])] # Finds all entries for planet selected and stores in an array
21         BestParIdx = np.argmax(x, axis=0) # Chooses the highest pl_refname number in the array
22         BestPar_NEA.append(x[BestParIdx[1]]) # Appends the planet entry selected above to our final array
23
24 # most recent Parameter for last object
25 x = df_np[np.where(df_np[:,0] == df_np[len(df_np)-1,0])] # Finds all entries for the last planet in database
26 BestParIdx = np.argmax(x, axis=0) # Chooses highest pl_refname number for last planet
27 BestPar_NEA.append(x[BestParIdx[1]]) # Appends planet entry selected to our final array

```

**FIG. 3:** Commented code to select resolved exoplanet parameters from the most recent publication.

## B. Exoplanet.eu (EU)

For exoplanet.eu (EU), each exoplanet only has one entry. No information is provided on whether the mass it is displaying for exoplanets comes from a most recent publication. While they do cite the parameters they display, upon checking the mass of a dozen exoplanets, most of them are not displaying and citing the most current publications for that specific planet unlike NEA which displays everything. There is also no information provided on the reason they picked to display a specific parameter value for one paper over another for a specific planet. EU however also lists the “mass  $\sin(i)$ ” which is a lower mass limit of the planets along with the nominal mass of the planets in a separate column in the database. This is when I realized that the mass and “mass  $\sin(i)$ ” of planets in NEA were all put under the mass column. In NEA, for planets that did not have a nominal mass in a publication,

the “mass  $\sin(i)$ ” was listed instead. This meant that the final mass array to be used in the comparison for NEA contained both values for planetary nominal mass and mass  $\sin(i)$ .

### C. Mass Sin ( $i$ )

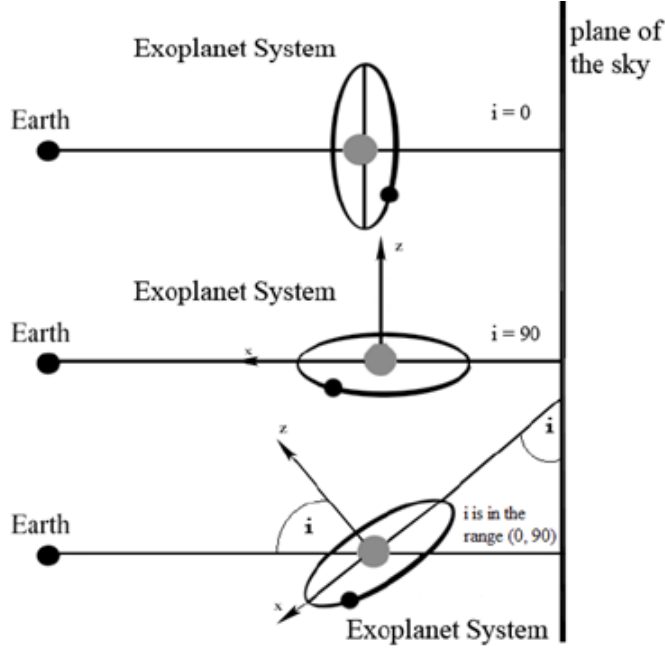
Using both values of mass and mass  $\sin(i)$  is not the optimal way to compare masses because this would be introducing unnecessary inconsistencies that would affect our mass distributions. Ideally, if the planet’s orbital inclination  $i$  is listed in the database, a nominal or true mass value can be calculated through the equation below:

$$M_{True} = \frac{M_{Min}}{\sin(i)} \quad (1)$$

$M_{True}$  is the actual mass of the planet,  $M_{Min}$  is the lower mass limit of the planet and  $i$  is the inclination of the exoplanet’s orbit. The inclination  $i$  is the angle between the plane of an exoplanet’s orbit and the line of sight of an observer from Earth or the plane of the sky (refer to Figure 6) [11]. All planets obtained through the radial velocity method list the “mass  $\sin(i)$ ” value or their minimum mass ( $M_{True} \cdot \sin(i)$ ) also known as the  $\sin(i)$  degeneracy. Only when the same exoplanet has been observed through the transit method in combination with the radial velocity method can a true mass value be calculated. This is because in the radial velocity method, the wobble of the parent star due to the exoplanet’s gravity is used to detect the exoplanet’s existence. If the planet is not passing directly in the middle of its star (from the point of view from Earth), only a lower mass limit can be estimated since the actual inclination of the planet’s orbit is unknown.

### D. Open Exoplanet Catalogue (OEC)

Unlike NEA and EU, OEC’s website does not provide a way to download the data from the database as a CSV file. Instead, the data must be directly retrieved from its GitHub repository that is updated daily. On the GitHub, the data files are stored in both XML and



**FIG. 4:** This figure [9] is a depiction of inclination  $i$  which is the angle between an exoplanet's orbital plane and the plane of the sky [11].

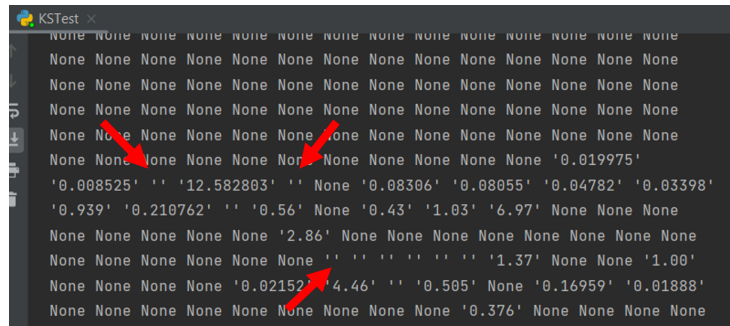
ASCII format. To retrieve and directly upload data from the XML files onto Python, the module “XML.Etree.ElementTree” was used. It retrieves the file through a specified URL after which we can simply iterate through the files using a for loop to find relevant parameters (mass, radius etc.) for all objects. This procedure is described by the code snippet in FIG. 5.

```
# Read data from OEC git repository
url = "https://github.com/OpenExoplanetCatalogue/oec_gzip/raw/master/systems.xml.gz"
oec = ET.parse(gzip.GzipFile(fileobj=io.BytesIO(urllib.request.urlopen(url).read()))

# Output mass and radius of all planets
OEC = []
for planet in oec.findall("./planet"):
    A = planet.findtext("name")
    B = planet.findtext("mass")
    OEC.append(A)
    OEC.append(B)
    #print(planet.findtext("name"), planet.findtext("mass"), planet.findtext("radius"))
```

**FIG. 5:** Commented code to retrieve data files from OEC GitHub repository.

Once the data were retrieved, inconsistencies in the way the data were stored in the array created errors when computing the KS test and associated CDF plot. This is because the KS test in SciPy only accepts data in float form and the data in the array were stored as strings. Converting the array from string to float type caused further errors. Upon closer inspection and after detruncating the array, objects that did not have a mass value listed were sometimes assigned the value 'NONE' and sometimes listed as an empty string (see FIG. 6). This is a problem because when trying to convert the array from string to float type, Python cannot handle the empty strings.



```

None None None None None None None None None None None None None None
None None None None None None None None None None None None None None
None None None None None None None None None None None None None None
None None None None None None None None None None None None None None
None None None None None None None None None None None None None None
None None None None None None None None None None None None '0.019975'
'0.008525' '' '12.582803' '' None '0.08306' '0.08055' '0.04782' '0.03398'
'0.939' '0.210762' '' '0.56' None '0.43' '1.03' '6.97' None None None
None None None None None '2.86' None None None None None None None None
None None None None None None '' '' '' '' '' '1.37' None None '1.00'
None None None None '0.02152' '4.46' '' '0.505' None '0.16959' '0.01888'
None None None None None None None None None '0.376' None None None None

```

**FIG. 6:** Console showing array with empty strings as well as NONE values.

To solve the problem, through a list comprehension, I changed all strings to a NONE value (see FIG. 7) after which, the array successfully converted from string type to float type. Afterwards the KS test was successfully run, and the CDF plot was generated.

```

# converts all empty strings to None
conv = lambda i: i or None
res = [conv(i) for i in OEClustMass]

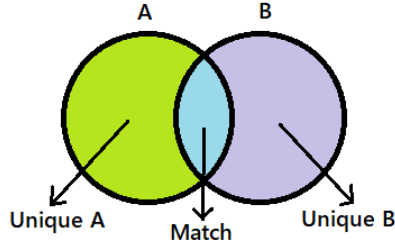
# Converts array to float
res_arr = np.array(res)
res_float = res_arr.astype(float)
CD_OEC = res_float.tolist()

```

**FIG. 7:** Code converting all strings to NONE value and converting array to float.



## E. Database Planet Matching



**FIG. 8:** Two databases being matched results in three subsets. The match, unique A and unique B subsets.

Before comparisons can be made, it is imperative that the parameters from the databases that we are comparing are for the same planets across the board. If some planets appeared in one database and not the other, a fair comparison of the parameter distributions cannot be made. As seen in TABLE I, the databases are not equal in the number of exoplanets they contain and with different inclusion criteria, they also differ in which exoplanets they contain.

To resolve this issue, a method to match planets across the databases was needed. Planetary names can be used. However, this is not the most efficient method because of inconsistencies in planetary naming conventions. This often includes differences in capitalization, word order and numbering. For example, see the table below.

NEA	EU	OEC
Kepler-381-b	381 kepler b	KEPLER 381 B

**TABLE III:** Planet Kepler 381b is listed under slightly different name variations in NEA, OEC and EU.

In order to match the names, all letters were uncapitalized, all non-alphabetical and numeric characters were removed and words were reordered so that numbers come first. A more robust approach for similar planets with completely different names consisted of

using planetary ID (IDs assigned by telescopes that discovered the planet), if available, to then query on the SIMBAD Astronomical database. (SIMBAD is a collection of databases containing objects outside the solar system that can be easily queried through object ID, coordinates, name etc. [4])

After the implementation of these matching methods, about 500 planets remained unmatched for each database either because they were unique planet entries for that specific database, they had no ID, or they were controversial objects. These objects were grouped under the unique subsets category for each of the 2 databases being compared (unique A and unique B) while obtaining the match subset (see FIG. 8). This procedure was repeated for all combination of databases resulting in the following subsets to be statistically compared:

- NEA match
- EU match
- OEC match
- NEA-EU unique\*
- NEA-OEC unique\*
- EU-NEA unique\*
- EU-OEC unique\*
- OEC-NEA unique\*
- OEC-EU unique\*

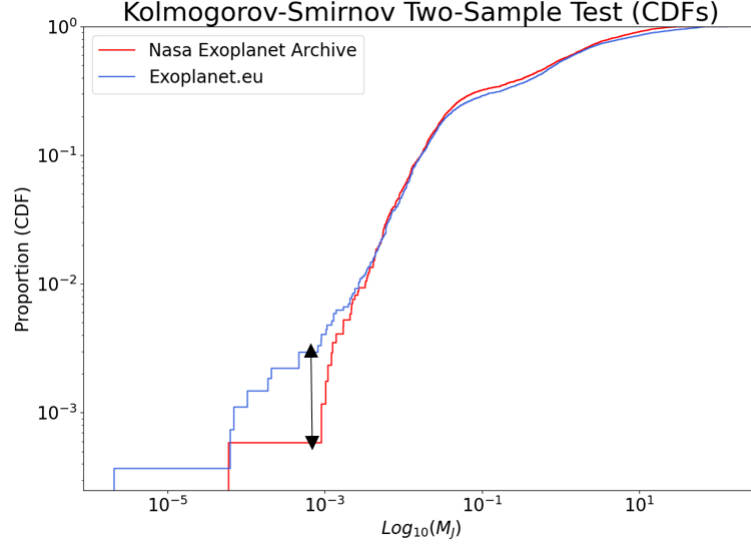
\*Note: A-B unique corresponds to planets in database A but not in database B.

#### IV. STATISTICAL RESULT INTERPRETATION

As mentioned in the section II, a two-sample Kolmogorov-Smirnov test and a two-sample Anderson-Darling test are used to quantify the differences in the parameter distributions between the databases because the distribution of exoplanetary parameters is unknown.

### A. Cumulative Distribution Function Plot

A Cumulative Distribution Function plot (CDF) shows the empirical cumulative distribution function of sample data. In other words, it graphs the cumulative probability of the data in a sample distribution against the data found in the distribution. It allows us to determine how many data points of similar value is in our distribution. FIG. 9 is an example of two CDF plots.



**FIG. 9:** CDF plot of NEA and EU with the arrow showing the KS statistic computed.

### B. 2-Sample Kolmogorov-Smirnov Statistic

A 2-sample KS test measures the probability that two datasets come from the same distribution. Visually, it is the maximum distance between the two CDF plots of the distributions being compared. The 2-sample KS test is given by the following formula:

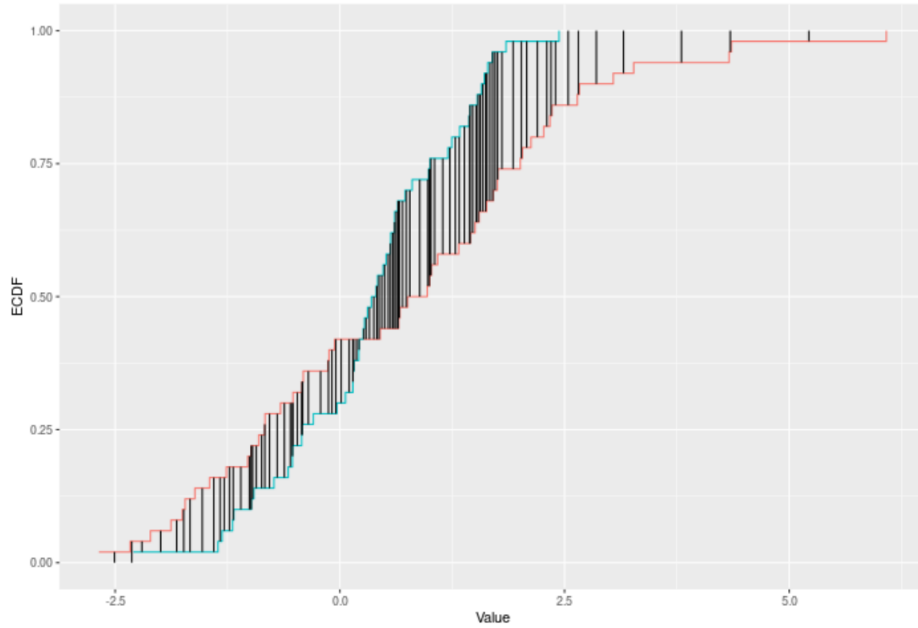
$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)| \quad (2)$$

where  $F_{1,n}$  and  $F_{2,m}$  are the distribution functions of the two samples being compared. The variables  $n$  and  $m$  are the sizes of the respective distributions and “sup” is the supremum

function which takes the maximum difference between the two distribution functions, i.e the maximum vertical distance between two CDF plots associated with the respective distribution functions (see arrow on FIG. 9).

### C. 2-Sample Anderson-Darling Statistic

Similar to the 2-sample KS test, the 2-sample AD test measures the probability that two datasets come from the same distribution. However, it is more powerful than the KS test as it requires less data to achieve the same statistical results. It is also more sensitive to differences if data samples have the same mean and standard deviation but differ on the tail ends only. Visually, the 2-sample AD statistic is the weighted sum of the height of the vertical lines as seen in FIG. 10 where more weight is given to lines that are closer together.



**FIG. 10:** CDF plot visually illustrating a 2-sample AD statistic [3].

The 2-sample KS test is given by the following formula:

$$AD = \frac{1}{mn} \sum_{i=1}^{n+m} (N_i Z_{(n+m-i)})^2 \frac{1}{i Z_{(n+m-i)}} \quad (3)$$

where  $Z_{n+m}$  is the combination of the ordered samples  $X_n$  and  $Y_m$  having size  $n$  and  $m$  respectively.  $N_i$  is the number of data points that are equal or greater than the  $i^{th}$  data point in  $Z_{n+m}$ .

#### **D. D-statistic, AD-statistic and $p$ -values**

The raw value obtained from the KS test that quantifies the difference between two distributions is known and reported as the D-statistic (KS statistic). The bigger the value of the D-statistic, the higher the probability of the 2 distributions being different becomes. Conversely, if the D-statistic is small, the probability that the two distributions are different is low.

For the AD test, the reported value quantifies the weighted sum of the lines as seen on FIG. 10 and is known as the AD-statistic (sometimes  $A^2$ ). It is interpreted similarly to the D-statistic.

In statistics,  $p$ -values are used to determine the probability that an observation made was by complete chance. A small  $p$ -value therefore indicates that our observation is statistically significant. In the case of the 2-sample KS test and AD-test for a small  $p$ -value, we then reject our null hypothesis, that is the hypothesis that there is a difference between the two samples. In other words, for a 2-sample KS-test and AD-test, if the  $p$ -value is below a given threshold (typically  $<0.05$ ), there is a high likelihood that both distributions are similar. A  $p$ -value above the given threshold would mean that both distributions are highly likely to be different.

#### **E. Kernel Density Estimation)**

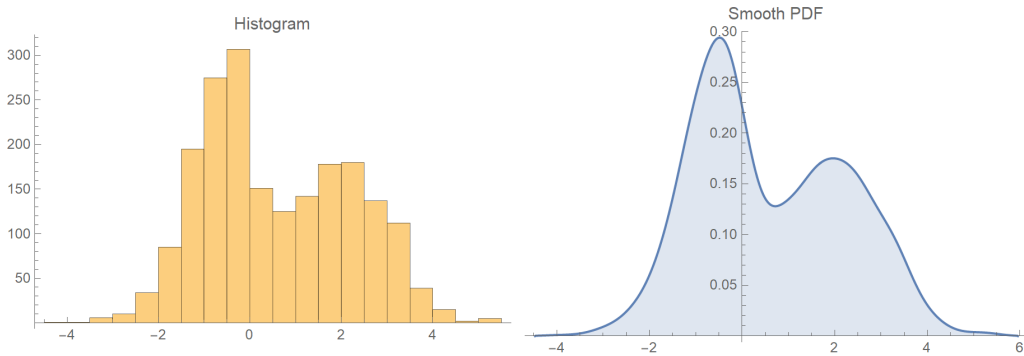
Kernel Density Estimation (KDE) Function is a non-parametric probability that can visually display the shape of a distribution. Another typical way of doing this is by a histogram (see FIG. 11). However, when comparing distributions a histogram can become quite visually overwhelming and make it difficult to observe any potential differences hence the use of a KDE.

In a KDE, individual data points are represented as kernels. A kernel is a function that is typically symmetric reaching a maximum value in the middle and tapering away from the center. Many kinds of kernels exist to represent probability densities (see FIG. 12). For this work, a Gaussian distribution kernel is used. To form the KDE function, the individual heights of the kernels at each point on the x-axis are added, i.e, it is the average shape of the the kernels added together (see FIG. 13). The KDE function is defined as follows:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) \quad (4)$$

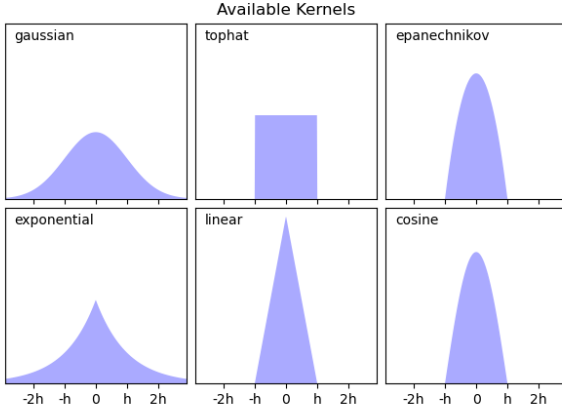
where  $h$  is the bandwidth or predetermined width of the kernels,  $n$  is the number of data points  $x_i = \{x_1, x_2, \dots, x_n\}$ ,  $x$  is a point on the x-axis at which the KDE is to be evaluated and  $K$  represents the kernel function.  $K$  for a Gaussian kernel is defined as:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u^2}{2}\right) \quad (5)$$

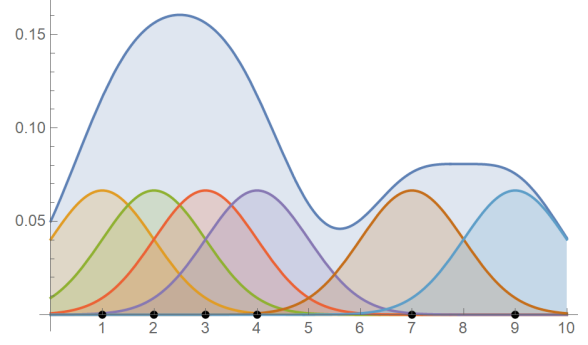


**FIG. 11:** Illustration of a KDE plot (right) of a distribution as compared to a histogram (left).

[5]



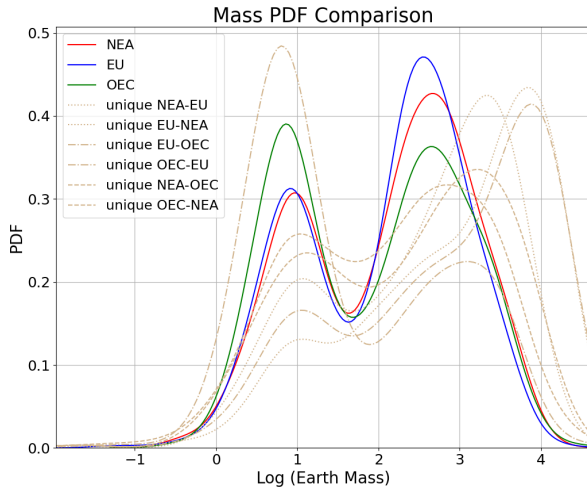
**FIG. 12:** Different types of kernels available on the Scikit-learn Python library. [1]



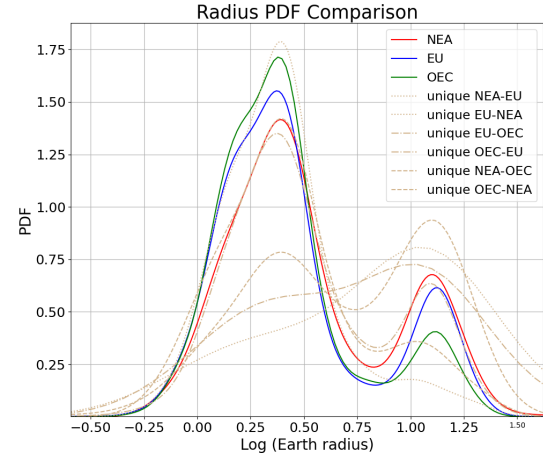
**FIG. 13:** Representation of a KDE as the average of kernels. [5]

## V. RESULTS & DISCUSSION

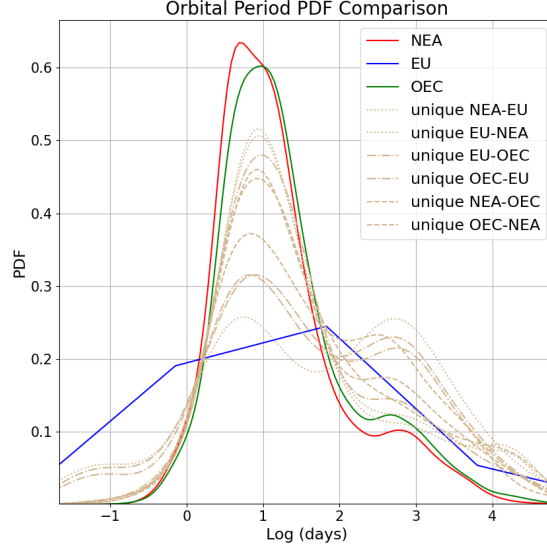
After data processing was done, the results of mass distributions and radius distributions between NEA, EU and OEC were obtained.



**FIG. 14:** Mass PDFs of each database's match (red, blue and green) and unique subsets (tan).



**FIG. 15:** Radius PDFs of each database's match (red, blue and green) and unique subsets (tan).



**FIG. 16:** Orbital Period PDFs of each database’s match (red, blue and green) and unique subsets (tan).

Comparisons	KS test ( $p$ -values)			AD test ( $p$ -values)		
	Mass ( $M_{\oplus}$ )	Radius ( $R_J$ )	Orbital Period (d)	Mass ( $M_{\oplus}$ )	Radius ( $R_J$ )	Orbital Period (d)
NEA-EU	1.678e-34	1.084e-7	2.217e-07	<0.001	<0.001	<0.001
NEA-OEC	2.864e-22	0.0	1.010e-6	<0.001	<0.001	<0.001
EU-OEC	0.013	0.018	7.079e-6	<0.001	<0.001	<0.001

**TABLE IV:**  $p$ -values obtained through the KS and AD tests for all exoplanetary parameters compared between databases for matched subsets.

### A. Biases

The first thing that can be noticed from these plots are detection biases. For example in FIG. 14 we are more likely to detect masses at about 10 Earth masses and about  $316 M_{\oplus}$  ( $10^{2.5}$ ) than at about  $56 M_{\oplus}$  ( $10^{1.75}$ ). These biases are a direct result of our current detection methods. For instance, planets found around  $316 M_{\oplus}$  have a high likelihood of having been



detected through the radial velocity method since this detection method is biased towards more massive planets. The high density of smaller planets at  $10 M_{\oplus}$  are planets mostly detected through the transit method which is reasonable since these planets mostly have relatively shorter orbital periods.

For radius (FIG. 15), it seems that we mostly detect planets at about  $2.37 R_{\oplus}$  ( $10^{0.375}$ ) with a sudden rise in density at  $13.3 R_{\oplus}$  ( $10^{1.125}$ ). The planets around  $13.3 R_{\oplus}$  are mostly made up of objects discovered through the transit method. This is nothing surprising since planets with larger radii produce appreciable light curves when passing in front of their host stars and are thus more easily detected.

For Orbital period, so far, we only seem to be detecting planets with orbital periods of about 10 days. This is because for longer periods, it is difficult to confirm the discovery of a planet using solely indirect detection methods which are the methods that make up the majority of our exoplanet discoveries.

## B. Differences

For mass comparison, there is a high number of mostly transit planets at  $10 M_{\oplus}$  for all unique subsets but especially EU. This is an interesting finding because the transit method does not allow for the mass of an object to be resolved directly. Only an upper mass estimate can be calculated and NEA, EU and OEC all treat this estimate differently. EU inputs the value as a nominal mass value while NEA and OEC generally treat it as an upper mass limit or completely omit it. This difference is a direct consequence of the different inclusion criteria of the databases.

For radius comparison, the NEA-OEC unique subset has a relatively higher density of planets at  $13.3 R_{\oplus}$  ( $10^{1.125}$ ). For this unique subset specifically, multiple papers revealed that these planets have had their radii calculated using a theoretical mass-radius relation. These planets also have surprisingly short periods which means that there could be tidal forces affecting the radii of the planets. While NEA chooses to include these theoretical values, OEC and EU do not.

It is also interesting to note that the distribution of EU (blue) on FIG. 16 appears

very different compared to NEA and OEC despite  $p$ -values for NEA-OEC and NEA-EU in TABLE IV being less than 0.05. This was likely caused by an issue with Python library “Seaborn” that was being used to generate the plots. This issue warrants further investigation.

## VI. CONCLUSION

While findings suggest that there are no significant disagreements in parameter distributions between the databases, the analysis of the probability density plots of unique subsets do reveal differences in the data that the databases display. While these differences are not alarming, it does highlight the inconsistency existing in the exoplanet cataloging process between NEA, EU and OEC.

## VII. FUTURE WORK

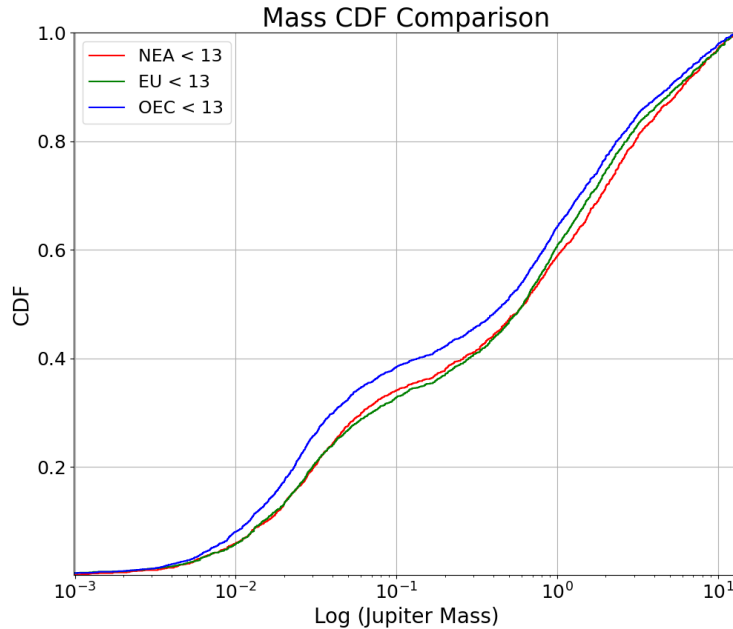
Further analysis of unique subsets is necessary to see if other differences exist within the data displayed by the databases. While an attempt was made at matching planets between the databases prior to analysis the amount of unmatched planets obtained and well as shapes of the densities of the unique subsets suggests that more planets from the unique subsets could have been matched. To obtain more conclusive results, a better planet matching procedure needs to be implemented. It would also be interesting to see if any discrepancies exist in host star parameters listed on the databases such as surface temperature, stellar metallicity and mass. The references of the planets could also be compared between databases which would likely entail using packages available on R.

## VIII. ACKNOWLEDGEMENTS

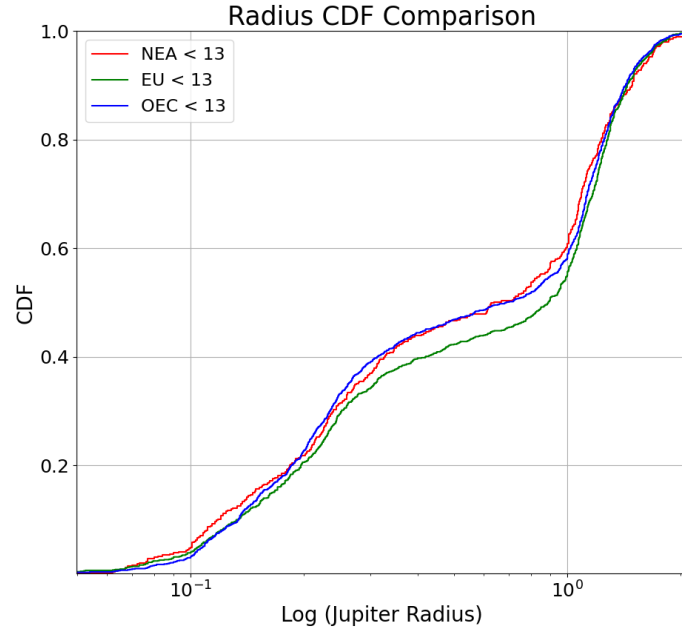
I would like to thank Dr. Wilson, my advisor, for constantly supporting me through the challenges encountered in this project as well as throughout my senior year. I am also very appreciative of Dr. Fisher who believed in my project, when applying for the

Haberberger fellowship junior year, and initially signed my proposal even if this work does not completely align with his area of expertise. I would also like to thank committee members Dr. Brandon and Dr. Pillai for taking the time to review this work. Another thank you goes to all my close friends for listening, providing feedback and advice when I encountered roadblocks and for continuously cheering me up. I am incredibly grateful for my beautiful family's continuous support namely my parents Assokumar and Sookrita as well as my little sister Akriti. I would like to thank Joanne and Arthur Haberberger for funding this project through Lycoming College's Center for Enhanced Academic Experiences, without whom, this project would not have been possible. Lastly, I would like to thank the Astronomy and Physics Department at Lycoming College for nurturing my passion for Astrophysics.

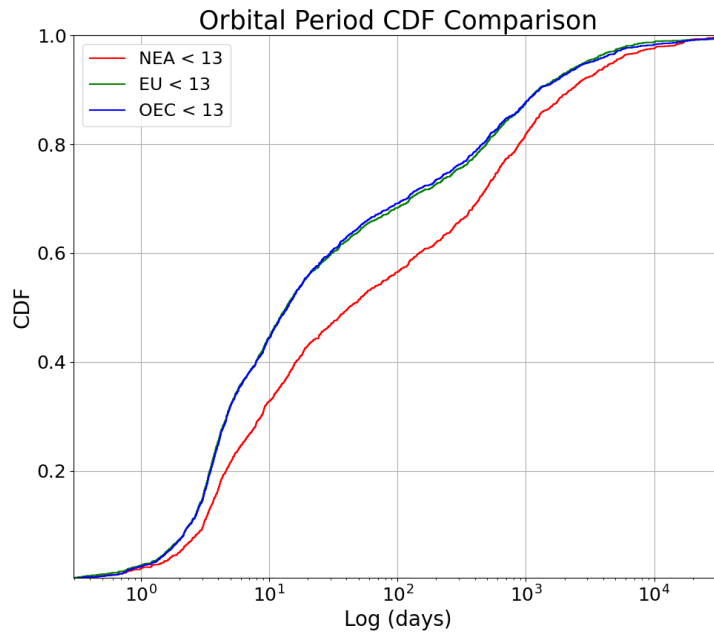
## IX. APPENDIX



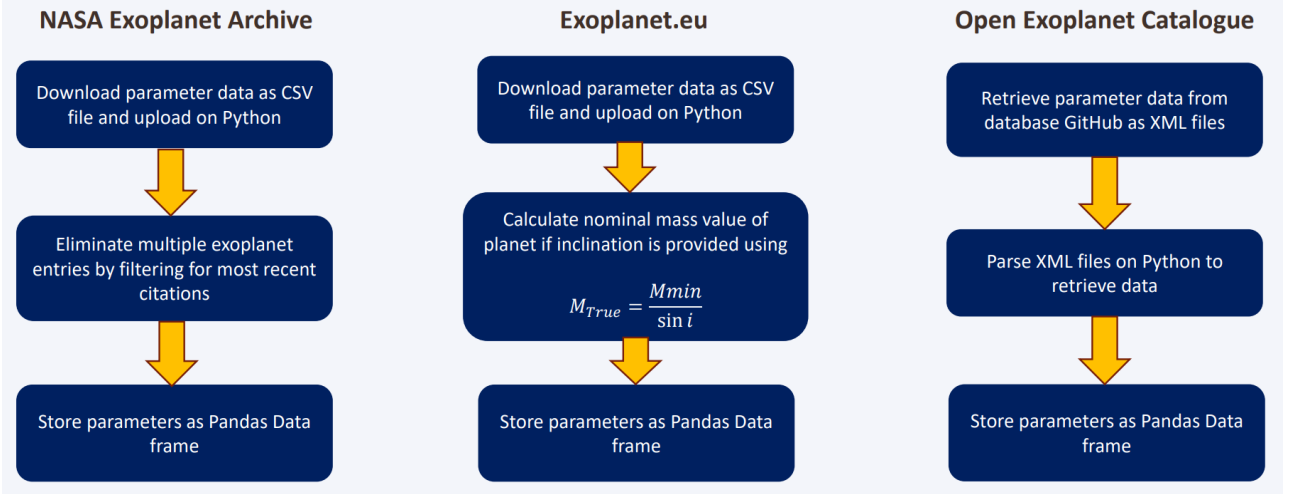
**FIG. 17:** CDF plot for mass distributions of NEA, EU and OEC.



**FIG. 18:** CDF plot for radius distributions of NEA, EU and OEC



**FIG. 19:** CDF plot for orbital period distributions of NEA, EU and OEC



**FIG. 20:** Summary of the dataprocessing procedure for all databases.

- 
- [1] Density estimation. URL <https://scikit-learn.org/stable/modules/density.html>.
  - [2] Jacob L. Bean, Dorian S. Abbot, and Eliza M.-R. Kempton. A statistical comparative planetology approach to the hunt for habitable exoplanets and life beyond the solar system. *The Astrophysical Journal Letters*, 841(2):L24, may 2017. doi:10.3847/2041-8213/aa738a. URL <https://dx.doi.org/10.3847/2041-8213/aa738a>.
  - [3] Connor Dowd. Fast permutation based two sample tests. <https://cran.r-project.org/web/packages/twosamples/twosamples.pdf>.
  - [4] D. Egret. S.I.M.B.A.D story: a description of the data base of the Strasbourg stellar data center. *Bulletin d'Information du Centre de Donnees Stellaires*, 24:109–123, March 1983.
  - [5] Stathis Kamperis. A gentle introduction to kernel density estimation, Dec 2020. URL <https://ekamperi.github.io/math/2020/12/08/kernel-density-estimation.html>.
  - [6] J. C. Morales, A. J. Mustill, I. Ribas, M. B. Davies, A. Reiners, F. F. Bauer, D. Kossakowski, E. Herrero, E. Rodríguez, M. J. López-González, C. Rodríguez-López, V. J. S. Béjar, L. González-Cuesta, R. Luque, E. Pallé, M. Perger, D. Baroch, A. Johansen, H. Klahr, C. Mordasini, G. Anglada-Escudé, J. A. Caballero, M. Cortés-Contreras, S. Dreizler, M. La-

farga, E. Nagel, V. M. Passegger, S. Reffert, A. Rosich, A. Schweitzer, L. Tal-Or, T. Trifonov, M. Zechmeister, A. Quirrenbach, P. J. Amado, E. W. Guenther, H.-J. Hagen, T. Henning, S. V. Jeffers, A. Kaminski, M. Kürster, D. Montes, W. Seifert, F. J. Abellán, M. Abril, J. Aceituno, F. J. Aceituno, F. J. Alonso-Floriano, M. Ammler von Eiff, R. Antona, B. Arroyo-Torres, M. Azzaro, D. Barrado, S. Becerril-Jarque, D. Benítez, Z. M. Berdiñas, G. Bergond, M. Brinkmüller, C. del Burgo, R. Burn, R. Calvo-Ortega, J. Cano, M. C. Cárdenas, C. Cardona Guillén, J. Carro, E. Casal, V. Casanova, N. Casasayas-Barris, P. Chaturvedi, C. Cifuentes, A. Claret, J. Colomé, S. Czesla, E. Díez-Alonso, R. Dorda, A. Emsenhuber, M. Fernández, A. Fernández-Martín, I. M. Ferro, B. Fuhrmeister, D. Galadí-Enríquez, I. Gallardo Cava, M. L. García Vargas, A. Garcia-Piquer, L. Gesa, E. González-Álvarez, J. I. González Hernández, R. González-Peinado, J. Guàrdia, A. Guijarro, E. de Guindos, A. P. Hatzes, P. H. Hauschildt, R. P. Hedrosa, I. Hermelo, R. Hernández Arabi, F. Hernández Otero, D. Hintz, G. Holgado, A. Huber, P. Huke, E. N. Johnson, E. de Juan, M. Kehr, J. Kemmer, M. Kim, J. Klüter, A. Klutsch, F. Labarga, N. Labiche, S. Lalitha, M. Lampón, L. M. Lara, R. Launhardt, F. J. Lázaro, J.-L. Lizon, M. Llamas, N. Lodieu, M. López del Fresno, J. F. López Salas, J. López-Santiago, H. Magán Madinabeitia, U. Mall, L. Mancini, H. Mandel, E. Marfil, J. A. Marín Molina, E. L. Martín, P. Martín-Fernández, S. Martín-Ruiz, H. Martínez-Rodríguez, C. J. Marvin, E. Mirabet, A. Moya, V. Naranjo, R. P. Nelson, L. Nortmann, G. Nowak, A. Ofir, J. Pascual, A. Pavlov, S. Pedraz, D. Pérez Medialdea, A. Pérez-Calpena, M. A. C. Perryman, O. Rabaza, A. Ramón Ballesta, R. Rebolo, P. Redondo, H.-W. Rix, F. Rodler, A. Rodríguez Trinidad, S. Sabotta, S. Sadegi, M. Salz, E. Sánchez-Blanco, M. A. Sánchez Carrasco, A. Sánchez-López, J. Sanz-Forcada, P. Sarkis, L. F. Sarmiento, S. Schäfer, M. Schlecker, J. H. M. M. Schmitt, P. Schöfer, E. Solano, A. Sota, O. Stahl, S. Stock, T. Stuber, J. Stürmer, J. C. Suárez, H. M. Tabernero, S. M. Tulloch, G. Veredas, J. I. Vico-Linares, F. Vilardell, K. Wagner, J. Winkler, V. Wolthoff, F. Yan, and M. R. Zapatero Osorio. A giant exoplanet orbiting a very-low-mass star challenges planet formation models. *Science*, 365(6460):1441–1445, 2019. doi:10.1126/science.aax3198. URL <https://www.science.org/doi/abs/10.1126/science.aax3198>.

- [7] NASA Exoplanet Science Institute. Planetary systems composite table, 2020. URL <https://catcopy.ipac.caltech.edu/doi/doi.php?id=10.26133/NEA13>.
- [8] Nornadiah Mohd Razali and Yap Bee Wah. Power comparisons of shapiro-wilk , kolmogorov-smirnov , lilliefors and anderson-darling tests. 2011.
- [9] Kamen Todorov. Determining the temperature of exoplanet hat-p-1b. *Physics, Astronomy and Geophysics Honors Papers. Paper*, 1, 01 2008.
- [10] A. Wolszczan and D. A. Frail. A planetary system around the millisecond pulsar psr1257 + 12. *Nature*, 355(6356):145–147, 1992. ISSN 0028-0836. doi:10.1038/355145a0.
- [11] M. Zeilik and S. A. Gregory. *Introductory Astronomy and Astrophysics*. Thomson Learning, Inc, 1998. ISBN 978-0030062285.
- [12] Ben Zuckerman. Brown dwarfs: At last filling the gap between stars and planets. *Proceedings of the National Academy of Sciences*, 97(3):963–966, 2000.