Application of Data Mining to Search for Potentially Habitable Exoplanets

Allen Chen[1], Hrithik Pai[1], Aaryan Rustagi[1], Rishab Pangal[2], Aditya Iyengar[3], Aaryan Divate[4], Akhil Deshmukh[5], Stanley Luo[6], Aaron Li[7], Aarav Sharma[8], and Robert Downing[9]

[1]Aspiring Scholars Directed Research Program, 43505 Mission Blvd, Fremont, CA 94539

Many light years away from our own solar system, over four thousand confirmed planets orbit stars in a fashion similar to our own eight planets and the sun. With the discovery of these planets, called "exoplanets," comes the question of extraterrestrial life, a concept scientists have been exploring for years. The possibility of exoplanetary habitability relies on a number of factors, such as spectral type, density, and eccentricity, but most importantly: whether the exoplanet in question contains water, the fundamental requirement for life, as we know it, to exist. To determine whether an exoplanet provides the ideal conditions for sustaining this vital ingredient for life, we considered the concept of the Goldilocks Zone, or the circumstellar habitable zone (CHZ)—the range of orbits around a star where liquid water is capable of existing. The research we have been conducting this summer utilizes the public dataset provided by NASA and Caltech and data mining methods, including Python and Microsoft Excel, to identify exoplanets with potentially habitable conditions. The discovery of the exoplanet K2-18b's water vapor-containing atmosphere was a major part of our research, in which we focused on identifying exoplanets with similar attributes to that of K2-18b, in hopes that they too may be able to retain atmospheric water vapor. After a two-month period, we discovered that 59 exoplanets orbit in the CHZ of their host star. As for the K2-18b ruleset, only 1 planet, K2-3d, satisfies the conditions. We believe K2-3d to have a high degree of similarity to K2-18b, but more in-depth analysis will have to be conducted to conclude its potential to support atmospheric water vapor and life as we know it.

Habitable Zone | Habitable Exoplanets | Orbital Period | Density | Sustainable Atmosphere

Correspondence: ***robert.downing@fremontstem.com***

Introduction

Background. The possibility of life on planets outside of our solar system has intrigued humanity for centuries. Since the first exoplanet, or planet that exists outside of our solar system, was discovered in 1992, astronomers have confirmed the existence of over four thousand exoplanets(1). One of the earliest methods of exoplanet detection was the measure of the radial velocity of a star using Doppler Spectroscopy, which indicates the presence of an exoplanet by observing Doppler shifts, or changes in the wavelength of light, in the spectrum of the planet's host star(2). A more efficient method is to observe transiting planets, which allows astronomers to identify planets through the apparent decline of the brightness of a star as a planet "transits" between the star and observer(3). An exoplanet's transit can reveal many important properties, such as its mass, orbital period, semi-major axes, etc. In 2009, the Kepler Space Telescope, using the transit method, was launched into space, positioning itself at a small patch of the sky. It was able to observe tiny differences in the stars' light when planets moved between the telescope and the star. The observations of different light shifts from the Kepler Space Telescope were used to confirm around 2000 exoplanets, along with 2400 other planetary candidates. After 9 years of service, Kepler was replaced by the Transiting Exoplanet Survey Satellite (TESS) to explore areas 400 times larger than that covered by the Kepler Mission.

Habitability. There are many approaches to determine whether an exoplanet can support life. However, arguably the most important factor for exoplanetary habitability is the planet's capability to support liquid water, which is essential to carbon-based life(4, 5). In order for liquid water to exist, the surface temperature of an exoplanet must be between 273.15 and 373.15 °K. Although it is not currently possible to directly observe an exoplanet's surface temperature, there are ways to model the area around a star in which liquid water can exist—the most common of which is to determine the CHZ of the host star(6). Using the planet's semi-minor and semi-major axes (shortest and longest radii of an elliptical orbit, respectively), we are able to determine whether the exoplanet stays within its host star's CHZ and can support liquid water.

K2-18b is an exoplanet first discovered in 2015 by the NASA-owned Kepler spacecraft. It orbits the M-type red dwarf K2-18, approximately 124 light-years away from Earth. In September 2019, independent studies from Benneke et al. and Tsiaras et al. reported the detection of water vapor in K2-18b's atmosphere(7, 8). It was the first time that water vapor was identified in the atmosphere of a smaller exoplanet located in the CHZ of its host star. Despite this finding, K2-18b was still widely regarded to be uninhabitable for life as we know it—it is 2.3 times wider and 8 times more massive than Earth and according to Benneke et al., the exoplanet is more like Neptune than Earth(7). Because of its massive size, K2-18b was thought to have a hydrogen-rich atmosphere, with much higher temperatures and pressures than that of Earth—conditions too extreme for life to exist. However, recent studies have shown that K2-18b may still have the potential to support liquid water and habitable conditions, even though it is classified as a mini-Neptune. According to Madhusudhan et al., K2-18b's hydrogen atmosphere may not

be as thick as previously thought, and its water layer may be under temperatures and pressures similar to those in Earth's oceans(9). They also reported lower-than-expected levels of methane and ammonia in K2-18b's atmosphere, suggesting evidence of chemical disequilibrium, and possibly biological activity. These findings suggest that the potential for habitable conditions are not necessarily restricted to Earth-like rocky planets.

As of August 18, 2020, there are 4,201 confirmed exoplanets, with up to 365 variables describing each planet. There is simply too much data to sort through manually—so we turned to Microsoft Excel and Python, using tools like Pandas to analyze the dataset and identify potentially habitable exoplanets. Our research consisted of two methods: one, we found exoplanets orbiting within their host star's CHZ; second, we created a rule set aiming to find exoplanets with similar attributes to that of K2-18b.

Methods

Calculation of the CHZ. Our first approach to determine exoplanetary habitability was to check if the planet lies within the CHZ of its host star. A method provided by Tom E. Morris uses the absolute luminosity of a star to calculate the radius of the inner and outer boundaries of its respective CHZ(10). A star's luminosity is the measure of radiated power, or light. Luminosity is useful in determining the CHZ because it affects the temperature of surrounding planets; a star with low luminosity will have a much closer CHZ than a star with a high luminosity.

The first step in finding the CHZ of a star is to calculate the star's luminosity, as it's not provided in the NASA dataset. To do this, we utilized factors that were found in the dataset, including stellar distance, the star's apparent magnitude, and stellar type. We used three formulas suggested in Tom E. Morris' paper to calculate the luminosity for all the stars in the dataset. From there we used the luminosity to calculate the inner and outer bounds of each star's CHZ. The dataset includes the semi-major axes for all the planets but not the semi-minor axes. However, we needed both the semi-major and semi-minor axes to ensure that the planet stays within the CHZ throughout its entire orbit. We took advantage of the formula for the eccentricity of an ellipse — $'semi\_minor' = ('semi\_major') * \sqrt{(1-('orb\_eccen')^2)}$, where 'semi_minor' is the semi-minor axis, 'semi_major' is the semi-major axis, and 'orb_eccen' is the orbital eccentricity—and were able to calculate the semi-minor axes for planets that had both orbital eccentricity and semi-major axes values. Finally, we checked if the planet's semi-major and semi-minor axes lie within the calculated host star's CHZ (Figure 1).

Upon testing our code, we realized that some exoplanets that we know are located in the CHZ, such as K2-18b, were not showing up as part of the output. Thus, we decided to check the accuracy of Morris' method by comparing the outputs of our code with the findings from the Planetary Habitability Laboratory (PHL)—a list of 55 potentially habitable exoplanets derived by measuring their similarity to Earth—to see if there were any common planets(11). For the planets that appeared on the PHL list but not on ours, we looked at the ranges and average of the differences between their CHZ bounds and axes (Figure 2).

```python
import pandas as pd
import numpy as np

df = pd.read_csv('TestingExoplanetSources.csv')

for row in df.index:
    # First step is to grab the stellar distance, magnitude, and stellar type values from the data set
    st_dist = df.loc[row, 'st_dist']
    st_optmag = df.loc[row, 'st_optmag']
    absolute_magnitude = st_optmag-5*(np.log10(st_dist/10))
    if df.loc[row, 'st_teff'] >= 2400 and df.loc[row, 'st_teff'] <=3700:
        BC = -2.0
    elif df.loc[row, 'st_teff'] >= 3700 and df.loc[row, 'st_teff'] <=5200:
        BC = -0.8
    elif df.loc[row, 'st_teff'] >= 5200 and df.loc[row, 'st_teff'] <=6000:
        BC = -0.4
    elif df.loc[row, 'st_teff'] >= 6000 and df.loc[row, 'st_teff'] <=7500:
        BC = -0.15
    elif df.loc[row, 'st_teff'] >= 7500 and df.loc[row, 'st_teff'] <=10000:
        BC = -0.3
    elif df.loc[row, 'st_teff'] >= 10000 and df.loc[row, 'st_teff'] <=30000:
        BC = -2.0
    bolometric_magnitude = absolute_magnitude + BC
    # Second step is to use the values collected and calculate the absolute luminosity of the host star
    absolute_luminosity = 10**((bolometric_magnitude-4.72)/-2.5)
    # Third step is using the calculated absolute luminosity to calculate the respective inner and outer bound of the CHZ
    inner_bound = np.sqrt(absolute_luminosity/1.1)
    outer_bound = np.sqrt(absolute_luminosity/0.53)
    # Grabs the semi major axis and eccentricity values from data set to calculate semi minor axis
    max_axis = df.loc[row, 'pl_orbsmax']
    eccentricity = df.loc[row, 'pl_orbeccen']
    min_axis = max_axis*(np.sqrt(1-(eccentricity**2)))
    #Last it checks if the semi minor and semi major axis are within the inner and outer CHZ bounds
    # and prints out a list of exoplanets from the data set fitting the habitability criteria
    if (min_axis > inner_bound and max_axis < outer_bound):
        print (df.loc[row, 'pl_name'])
```

Fig. 1. Python code calculating the CHZ of host stars and determining if orbiting planets stay in the CHZ

```python
    # checks if exoplanet appears on PHL list of habitable exoplanets
    is_habitable = df.loc[row, 'Source 1']
    # Case 1: within bounds
    if is_habitable == 'yes' and (min_axis > inner_bound and max_axis < outer_bound):
        print (df.loc[row, 'pl_name'] + ' --> Inside bounds')
    # Case 2: outside inner bound
    elif is_habitable == 'yes' and (min_axis < inner_bound):
        difference = inner_bound - min_axis
        total_difference = total_difference + difference
        print (df.loc[row, 'pl_name'] + ' --> Outside bounds by: ' + str(difference))
    # Case 3: outside outer bound
    elif is_habitable == 'yes' and (max_axis > outer_bound):
        difference = max_axis - outer_bound
        total_difference = total_difference + difference
        print (df.loc[row, 'pl_name'] + ' --> Outside bounds by: ' + str(difference))
# For cases 2 and 3, the difference between the axis and CHZ bound is taken into account. Finally
# the average difference is calculated by dividing total difference by the 31 planets on PHL list
average_difference = total_difference/31
print (average_difference)
```

Fig. 2. Python code checking our CHZ calculations against the PHL list and calculating error differences

K2-18b Ruleset. Our second approach consisted of creating an algorithm to find exoplanets with similar attributes to that of K2-18b. Since K2-18b was discovered to have atmospheric water vapor and is a potential candidate for habitability, we developed an array of rules to identify similar exoplanets.

To do so, we took multiple factors into account, including spectral type, density, orbital eccentricity and period, discovery method, and status. We first filtered the dataset using spectral type; since the spectral type of a star largely impacts the habitability of an exoplanet, we strictly looked into exoplanets orbiting M-type stars, the same star type that K2-18b orbits.

We then filtered the dataset and only kept exoplanets with a density between 4 and 7 grams per cubic centimeter g/cm^3, which ensured that only Earth-like rocky planets are being considered. All gas giants were excluded from habitability considerations because, as far as we know, there is no evidence of life on gaseous planets.

One problem we encountered was the large number of missing values in the planet density column. There are only 542 planets with density values, translating to 3,655 blanks, out of the 4,197 exoplanets in the dataset. Thus, we turned to estimating the density by using other variables, specifically

mass and radius. For planets that did not have a density value but had mass and radius values, we roughly estimated the density by dividing the planet's mass by its volume (calculated as the volume of a sphere). We wrote a Python program to automate this process and were able to estimate the densities for an additional 242 exoplanets (Figure 3). Despite this, the large majority of the dataset remained unused due to the aforementioned limitation.

```
import pandas as pd
import numpy as np

url = 'https://raw.githubusercontent.com/allench36/NASA-Caltech-Exoplanet-Archive/master/updatedexoplanet.csv'

exo = pd.read_csv(url)

# create new dataframe for planets with jupiter radii value, mass value, and no density value
new_exo_mass = exo.loc[(exo['pl_dens'].isnull()) & (exo['pl_radj'].notnull()) & (exo['pl_massj'].notnull())]
# filter original dataframe for planets with density value
exo = exo.loc[exo['pl_dens'].notnull()]

# calculate densities for child dataframe
new_exo_mass['pl_dens'] = (new_exo_mass['pl_massj'] * (1.898 * (10 ** 27) * 1000)) \
/ (4 * np.pi * ((new_exo_mass['pl_radj'] * 43441 * 160934) ** 3) / 3)

# append everything into 'total_dens'
total_dens = exo.append(new_exo_mass, ignore_index=True)
```

Fig. 3. Python code calculating the densities of planets which have mass and radius values

Next, we accounted for orbital eccentricity, the extent to which an object's orbit deviates from a perfect circle. In context, an eccentricity of 0 represents a perfectly circular orbit, an eccentricity of 1 represents a parabolic escape orbit, and an eccentricity between 0 and 1 represents an elliptical orbit. Since K2-18b's orbital eccentricity is around 0.2, we omitted planets with an orbital eccentricity greater than that. A higher eccentricity indicates that the planet is more likely to stray out of the CHZ. We also accounted for the orbital period, which is the amount of time it takes for the planet to make a complete orbit around its host star. K2-18b has an orbital period of 32.94 Earth days, so we used a range of 20.94 to 44.94, or +-12 Earth days.

Finally, we filtered for discovery method and status. We only considered exoplanets discovered via "transit," because this allows us to directly observe their atmospheric composition and determine if water vapor is present. Lastly, we only considered exoplanets that had unquestioned confirmation statuses in published literature (Figure 4).

```
# create child with 'pl_orbper' value
total_y_orbper = total_dens.loc[total_dens['pl_orbper'].notnull()]
# create child with no 'pl_orbper' and 'pl_ratdor' but with 'pl_orbsmax'
total_n_orbper = total_dens.loc[(total_dens['pl_orbper'].isnull())
& (total_dens['pl_ratdor'].isnull()) & (total_dens['pl_orbsmax'].notnull())]

# calculate 'pl_orbper' for planets without it
total_n_orbper['pl_orbper'] = np.sqrt(total_n_orbper['pl_orbsmax']**3) * 365

# append all child dataframes into 'total_orbper'
total_orbper = total_y_orbper.append(total_n_orbper, ignore_index=True)

# filter out planets without 'pl_orbeccen' value
total = total_orbper.loc[total_orbper['pl_orbeccen'].notnull()]

# filter by matching K2-18b (M-type host star, density between 4 and 7 g/cm^3, no controversy
# discovered via transit, eccentricity 0.2 or less, orbital period ~10 or +20)
total = total.loc[(total['st_spstr'].str.startswith('M', na=False)) & (total['pl_dens'] >= 4) & (total['pl_dens'] <= 7)
& (total['pl_controvflag'] == 0) & (total['pl_discmethod'].str.contains('Transit', na=False)) & (total['pl_orbeccen'] <= 0.2 )
& (total['pl_orbper'] >= 22.94) & (total['pl_orbper'] <= 52.94)]
```

Fig. 4. Python code used to filter the dataset using a K2-18b-matching ruleset

Results

Of the 4,201 confirmed exoplanets in NASA's dataset, 1,263 had the attributes necessary to calculate respective CHZs. We found that 59 exoplanets stayed in their respective CHZs according to our calculations. As mentioned earlier, we also tested our code on the PHL list of potentially habitable planets to get a sense of how our calculations compared with others'. First, we ran our code on the entire NASA dataset, finding the number of planets with axes below the inner bound, the number of planets with axes within the inner and outer bound, and the number of planets with axes above the outer bound (Figure 5). Next, we used the same method but on the PHL list (Figure 6). Finally, we looked at the ranges and averages of the differences between the axes and CHZ bounds for planets that were outside of their star's calculated bounds (Figure 7).

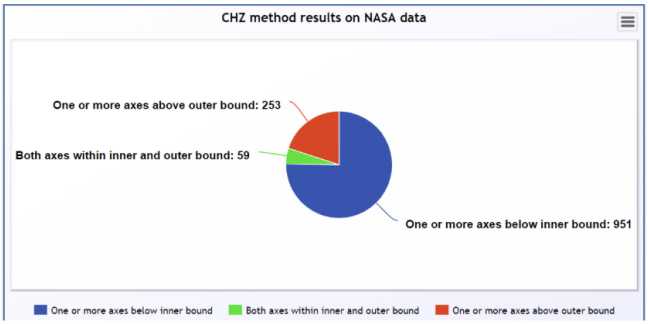

Fig. 5. The NASA dataset has 59 planets within their respective CHZ boundaries, 951 planets with one or more axes below the inner bound, and 253 planets with one or more axes above the outer bound

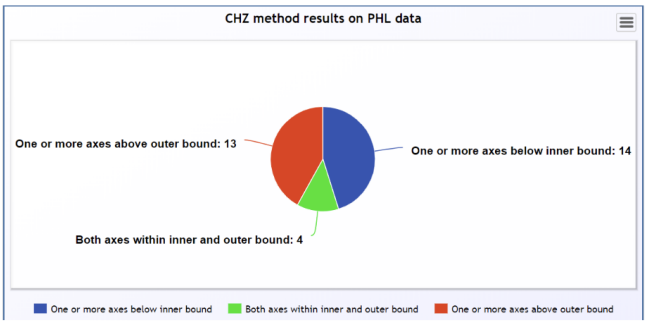

Fig. 6. The PHL list has 4 planets within their respective CHZ boundaries, 14 planets with one or more axes below the inner bound, and 13 planets with one or more axes above the outer bound.

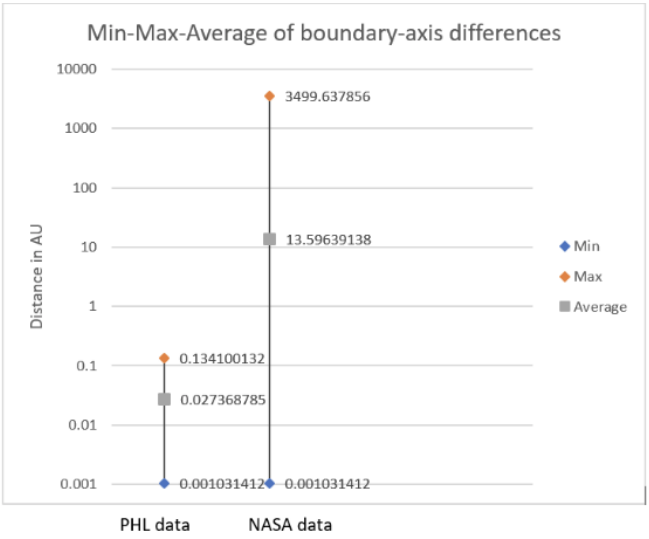

Fig. 7. Shown above is the average and lower-upper range of deviation between the semi-major and semi-minor axes and CHZ boundaries for the PHL and NASA data

For the K2-18b ruleset method, we found that only K2-3d satisfies our conditions. K2-3d, discovered in 2015, is located 147 light-years away and was discovered by NASA's Kepler K2 Mission. It is a rocky super Earth and completes an orbit around the M-type red dwarf K2-3 every 45 days.
We plan to expand our ruleset to try and identify more planets similar to K2-18b. By doing this, we can correlate the two sets of data and determine the correlation coefficient between them.

Discussion

Our initial plan to probe habitability was to determine if a planet could sustain an atmosphere. To do so, we would have to calculate the escape velocity of gas molecules on that planet. However, the escape velocity is a function of the planet's temperature, which was minimally represented in the dataset. We sought to derive planets' temperatures using other formulas, but most of them required knowledge of the planet's bond albedo, or the fraction of power in the total electromagnetic radiation incident on an object that is scattered back into space. Calculating a planet's bond albedo necessitated a level of understanding that we were not able to accumulate in a short two months time, so we undertook an alternative approach: calculating planets' respective CHZs and identifying planets with similar attributes to that of K2-18b, a potentially habitable exoplanet with a known atmosphere, accounting for factors including spectral type, density, semi-major and semi-minor axes, luminosity, orbital eccentricity and period.
Initially, there were 2 possibilities as to why our code was not recognizing K2-18b as a habitable-zone exoplanet: the calculations done by Morris were incorrect, or there was some explanatory variable not accounted for by Morris' work. After running our code on the list of potentially habitable planets from PHL, we found an average deviation between the axes and CHZ boundaries of 0.02 AU. Analyzing the average and range of deviation, we can conclude that the error isn't significant, indicating that some explanatory variable is not being accounted for by Morris.
K2-3d, the planet that satisfied the K2-18b ruleset, is larger than Earth but smaller than most gas giants, which is the reason for its classification as a super Earth(12). It has a density of 5.62 g/cm^3, so K2-3d is likely to be rocky. Considering its density and mass, K2-3d may possess the right conditions to support life. However, some concerns for the habitability of K2-3d include the fact that its semi-major axis is relatively close to its host star, meaning that the K2-3d may be too hot to support life. In the future, we plan to expand our ruleset to try and identify more planets similar to K2-18b.
Despite our findings, much of our analysis could be subject to error because many planets were discovered via transit, so many values in the dataset were estimated, not directly observed. And although 59 planets were found in the CHZ, the formula we used is merely one way to model the CHZ and not guaranteed to be entirely accurate. In addition, our K2-18b ruleset may be flawed because although K2-18b is able to support an atmosphere, the extent to which it's habitable is not conclusive(7).
For these reasons, there is still much work to be done regarding exoplanetary habitability. In this study, we found 59 planets located in their respective CHZ, though more analysis is necessary to draw any conclusions. In addition, planets orbiting M-type stars may not be able to sustain life. For one, planets orbiting M-type stars have a high chance of becoming tidally locked; second, due to the star's low radiation, planets must orbit fairly close to the star to have the appropriate temperatures to support life; lastly, many M-type stars are flare stars, meaning that their inconsistent brightness could make life on orbiting planets very difficult(13).

Conclusion

Using tools like Python and Microsoft Excel, we were able to determine that 59 exoplanets are found in the CHZ, and one planet—K2-3d—has a high degree of similarity to K2-18b. However, much more research is required before we can definitively say that any one of these 60 planets is able to sustain life. Although we cannot draw any conclusions, our research shows that there is great potential for life to exist elsewhere in the universe. In the future, we plan to analyze the 59 habitable-zone exoplanets in more detail in an effort to uncover more information about their surface conditions. We also plan to focus our resources towards the study of K2-3d, a rocky super Earth that can potentially possess habitable conditions. As more powerful telescopes are created, we will continue our search for extraterrestrial life, perhaps with more and more success.
All leading authors contributed to the methods to find exoplanets with potential to contain life and the finding of K2-3d as one such. All authors contributed to the writing and formatting of this paper.

ACKNOWLEDGEMENTS
We would like to thank Professor Robert Downing for his valuable aid and support throughout our research. In addition, we would like to thank NASA and the California Institute of Technology for access to the exoplanet data set. The authors gratefully acknowledge the Aspiring Scholars Directed Research Program and sponsors for providing us with the opportunity to conduct this research.

Bibliography

1. Aleksander Wolszczan and Dail A Frail. A planetary system around the millisecond pulsar psr1257+ 12. *Nature*, 355(6356):145–147, 1992.
2. Michel Mayor, Christophe Lovis, and Nuno C Santos. Doppler spectroscopy as a path to the detection of earth-like planets. *Nature*, 513(7518):328–335, 2014.
3. David G Koch, William J Borucki, Larry Webster, Edward W Dunham, Jon M Jenkins, John Marriott, and Harold J Reitsema. Kepler: a space mission to detect earth-class exoplanets. In *Space telescopes and instruments V*, volume 3356, pages 599–607. International Society for Optics and Photonics, 1998.
4. Giovanna Tinetti, Jonathan Tennyson, Caitlin A Griffith, and Ingo Waldmann. Water in exoplanets. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1968):2749–2764, 2012.
5. Frances Westall and André Brack. The importance of water for life. *Space Science Reviews*, 214(2):50, 2018.
6. James F Kasting, Daniel P Whitmire, and Ray T Reynolds. Habitable zones around main sequence stars. *Icarus*, 101(1):108–128, 1993.
7. Björn Benneke, Ian Wong, Caroline Piaulet, Heather A Knutson, Joshua Lothringer, Caroline V Morley, Ian JM Crossfield, Peter Gao, Thomas P Greene, Courtney Dressing, et al. Water vapor and clouds on the habitable-zone sub-neptune exoplanet k2-18b. *The Astrophysical Journal Letters*, 887(1):L14, 2019.
8. Angelos Tsiaras, Ingo P Waldmann, Giovanna Tinetti, Jonathan Tennyson, and Sergey N Yurchenko. Water vapour in the atmosphere of the habitable-zone eight-earth-mass planet k2-18 b. *Nature Astronomy*, 3(12):1086–1091, 2019.
9. Nikku Madhusudhan, Matthew C Nixon, Luis Welbanks, Anjali AA Piette, and Richard A Booth. The interior and atmosphere of the habitable-zone exoplanet k2-18b. *The Astrophysical Journal Letters*, 891(1):L7, 2020.

10. Tom E. Morris. Calculating the habitable zone, Oct 2010.
11. The habitable exoplanets catalog.
12. Sara Seager, M Kuchner, CA Hier-Majumder, and Burkhard Militzer. Mass-radius relationships for solid exoplanets. *The Astrophysical Journal*, 669(2):1279, 2007.
13. Krisztián Vida, Zs Kővári, András Pál, Katalin Oláh, and Levente Kriskovics. Frequent flaring in the trappist-1 system—unsuited for life? *The Astrophysical Journal*, 841(2):124, 2017.