# PHYDA Dataset Analysis

Note: These are based on personal analysis and observation and may at some points fail to represent the dataset accurately.

## Links

[NOAA Page](https://www.ncei.noaa.gov/access/paleo-search/study/24230)

[File Location](https://zenodo.org/record/1198817)

Three NetCDF files available for direct download, each ~4 GB and covering a different time span. One covers annual averages beginning April to next year March. The second covers the austral growing season from December to February (DJF). The third covers averages during the boreal growing season from June to August (JJA).

On an external file hosting site, there are additional, presumably more detailed files available for download accompanied by index variable files. These include PDSI (Palmer Drought Severity Index), SPEI (Standardized Precipitation-Evapotranspiration Index), and TAS (Surface Air Temperature) for the three periods of focus (annual, DJF, and JJA). However, I was not able to analyze these as each file is ~22 GB, which my computer does not have sufficient storage space for.

[Associated Publication](https://www.nature.com/articles/sdata201886)

Contains information on methodology as to how the reconstructions were made, but not much on the dataset itself.

[Code Location](https://github.com/njsteiger/PHYDA-v1)

Code used to generate data and perform reconstructions publicly available on GitHub. The creator used MATLAB to generate everything, and all scripts are written in MATLAB code. The files needed to run the code are also available [here](http://clifford.ldeo.columbia.edu/nsteiger/hydro_input/).

## Dataset Information: Annual (April-March)

### Dimensions

Five total dimensions:

* **Time**
  + Name: time
  + Units: Years CE
  + Size: 2000
  + Range: [1, 2000]
  + Step: 1 year CE
* **Latitude**
  + Name: lat
  + Units: Degrees North
  + Size: 96
  + Range: [-90, 90]
  + Step: 1.89473685 degrees North, or, 180/95
* **Longitude**
  + Name: lon
  + Units: Degrees East
  + Size: 144
  + Range: [0, 357.5]
  + Step: 2.5 degrees East
* **Fractional Years**
  + Name: tmon
  + Units: Years CE + Year fraction according to month
  + Size: 24000, or, 2000 x 12
  + Range: [1, 2009.1667]
  + Step: 0.083333 years, or, 1/12
* **Percentile**
  + Name: prctl
  + Units: None, index with 0, 1, and 2 to correspond to the 5th, 50th, and 95th percentiles respectively
  + Size: 3
  + Range: [0, 2]
  + Not actually sure what this dimension indicates, and the values are not printable.
  + Possibilities include that it indicates “percentile”, though I am not sure what relevance this has in the dataset. Prctl() is also a common linux command for thread processes, but as with the first possibility I noted, I am not sure what relevance this has.
    - Upon further analysis, this dimension does in fact, indicate the percentile.

### Variables

There are 67 total variables with data, each with their own dimensions and dimension dependencies (e.g., does it depend on latitude, time, month… etc). However, the ones of interest are most likely the ones related to PDSI. Note that these PDSI values are for land only.

* **PDSI Reconstruction Mean**
  + Name: pdsi\_mn
  + Units: Standardized units of relative dry and wet
  + Size: 2000 x 96 x 144
  + Range: [-12.2195, 10.2998]
  + Dependencies: Year (time), latitude (lat), longitude (lon)
* **PDSI for One Standard Deviation of Ensemble**
  + Name: pdsi\_sg
  + Units: Standardized units of relative dry and wet
  + Size: 2000 x 96 x 144
  + Range: [0.3817, 7.0365]
  + Dependencies: Year (time), latitude (lat), longitude (lon)
* **PDSI Per the 5th, 50th, and 95th Percentiles**
  + Name: pdsi\_pc
  + Units: Standardized units of relative dry and wet
  + Size: 3 x 2000 x 96 x 144
  + Range: [-19.5494, 24.4868]
  + Dependencies: Year (time), latitude (lat), longitude (lon), percentile (prctl)

There are numerous missing values in these PDSI datasets. These following conclusions are based on preliminary experimentation, so they may be corrected later if I discover something that contradicts it. However, generally, it seems as if it is these missing values correspond to location rather than time. Whether values exist or not vary across the range of location indices for a single year, but not across the time spectrum for a single location. Furthermore, the missing value coordinates usually place them in the ocean, which is in accordance with the statement that the data for the PDSI variables are concerned only with land.

## Validation

The dataset was validated using the Pearson correlation coefficient and the continuous ranked probability skill score (CRPSS), though greater emphasis was placed on CRPSS. For the SPEI and TAS reconstructions, skill and correlation metrics were computed against the [CRU TS3](https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.3711?casa_token=hHk3qnQ7iRsAAAAA:i9KbJYxfevMgB8p-RzaGmVnHEerXLKdNWuomoaL8mv_MRagcVWEzhbabWgf5YggszkngwEqNIRGQaQ) surface dataset using the [Berkeley Earth](https://static.berkeleyearth.org/papers/Methods-GIGS-1-103.pdf) reconstruction methodology for the 1901-2000 time interval. As for the PDSI reconstructions, skill and correlation metrics over the same 1901-2000 time interval were computed with regard to the observational data noted in a [1900-2008 PDSI study](https://www.researchgate.net/publication/251426941_Characteristics_and_trends_in_various_forms_of_the_Palmer_Drought_Severity_Index_during_1900-2008).

## Analysis

### Tabora

Assuming that like the LMR, the data is organized such that each coordinate point is a cross point representing a 1.895 latitude by longitude area, Tabora’s coordinates of -6N, 32E is covered in PHYDA’s - 6.6316N and 32.5E coordinate point.

#### Correlations

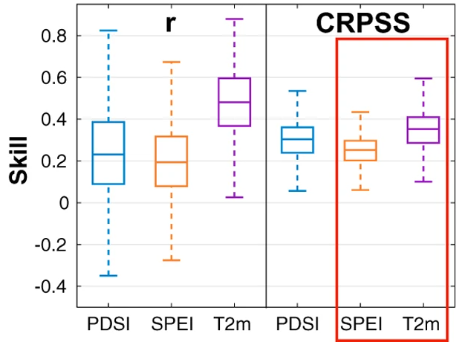
As with previous correlation analyses, the Pearson correlation coefficient was used.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **LMR mean PDSI** | **LMR mean PRATE** | **PHYDA 1 std PDSI** | **PHYDA 5th Percentile PDSI** | **PHYDA 50th Percentile PDSI** | **PHYDA 95th Percentile PDSI** |
| **PHYDA mean PDSI** | 0.22362 | 0.33287 | 0.06449 | 0.99770 | 0.99996 | 0.99512 |

|  |  |  |
| --- | --- | --- |
| **LMR MCrun Index** | **PHYDA mean PDSI Correlation with LMR PDSI** | **PHYDA mean PDSI Correlation with LMR PRATE** |
| **0** | -0.21889 | 0.1343 |
| **1** | 0.2513 | 0.27971 |
| **2** | 0.01305 | 0.15509 |
| **3** | 0.30268 | 0.30354 |
| **4** | 0.12653 | 0.18728 |
| **5** | 0.17627 | 0.26315 |
| **6** | 0.22142 | 0.21906 |
| **7** | 0.12998 | 0.09093 |
| **8** | 0.24787 | 0.31624 |
| **9** | 0.13493 | 0.18797 |
| **10** | 0.04061 | 0.18128 |
| **11** | 0.07992 | 0.01201 |
| **12** | 0.04053 | 0.01687 |
| **13** | 0.09494 | 0.21161 |
| **14** | -0.03231 | 0.1132 |
| **15** | 0.30398 | 0.31241 |
| **16** | 0.12363 | 0.26975 |
| **17** | 0.17943 | 0.14902 |
| **18** | 0.08323 | 0.239 |
| **19** | -0.07471 | 0.05912 |
| **Mean** | 0.11122 | 0.1851 |

There is a low amount of correlation between any of the PHYDA and LMR datasets. In fact, some of the correlation coefficients are even negative, indicating that the data tends to move in opposite directions. This is interesting, considering how the LMR V2 dataset is in fact included as a reconstruction source. When viewing the list of proxies used, we can see that proxy data is very sparse across the continent of Africa, and within East Africa, mostly coral records are used. According to the dataset’s corresponding paper when discussing verification, the correlation is much higher in areas that have a larger density of proxies, such as the United States and Western Europe, when evaluated with regards to the various Drought Atlases from 1500-2000. Relatedly, the majority of the proxy records are concentrated within the same time period from 1500-2000. However, it should be noted that this section says nothing about how the PHYDA dataset correlates with the LMR dataset.

The box-whisker plot below shows that for the PHYDA PDSI data, 50% of the Pearson correlation coefficients fall between roughly 0.1 – 0.4, with a median of approximately 0.25. This corresponds well to data gathered from Tabora, though the two datasets evaluated against (Drought Atlas and LMR) are different. The correlation between the mean PHYDA PDSI data and the LMR PDSI data is 0.22362 – only slightly below 0.25. Interestingly, the mean of the PHYDA PDSI data correlated with each MCrun index from the LMR PDSI dataset is lower than both, at 0.1851. The highest correlation coefficient is not significantly higher than 0.25, at 0.31624, and the lowest is 0.01201. This potentially suggests that for Tabora, at the very least, the mean ensemble results are generally more robust than any individual reconstruction. This possibility can be drawn from the correlations between the PHYDA PDSI data and the LMR precipitation data as well. The correlation of the ensemble mean data is 0.33287, while the mean of correlation coefficients with all the different LMR reconstructions is 0.11122 with a maximum of 0.30398 and minimum of -0.21889.



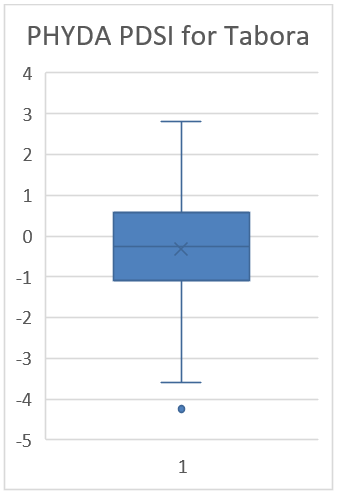
Box and whisker plot for the April-March dataset showing the Pearson correlation coefficient (r) and CRPSS score.

They also state in their usage notes that based on Drought Atlas reconstruction methodology, variance in the ensemble mean will decrease as the proxy data decreases and rely more heavily on the prior. Using machine learning terminology, this just means that as the amount of data available decreases, the more the dataset relies on educated guesses. If you’re interested in the technical details, a breakdown is in the box below. No definitions for the terms are given in the paper, so this may be misinterpreted, though I do think it makes sense.

|  |
| --- |
| In machine learning, “variance” is a measure of how much a variable changes according to the data. For example, take the image below, where the figure on the right has high variance and the one on the left as low variance. High variance generally means that the variable is fit more to the observed data. However, when the variance is too high, it means that the variable ignores the general trends and instead follows the exact path of the data. This is a problem in machine learning, which aims to make predictions based on general trends, though it may be less relevant in climatology when investigating anomalous climate.    Meanwhile, the “prior” used here may be referring to the priors used in Bayesian methodology, where a posterior prediction, or final prediction, is made based two factors – a prior and likelihood distribution. The prior is essentially an estimate as to what the final distribution will be, while the likelihood is actual observed data used to adjust the prior. For example, take flipping a coin. We know beforehand that coins have a 50% chance of landing on either heads or tails, so we make this our prior estimation. However, we then run a few tests and see that 80% of the time, the coin lands on heads. This raises the potential that the coin may be unfair, but the data we gathered was highly limited and hardly represents the entirety of the possible probability distribution, though we still need to take the test results into account. We therefore make this our likelihood distribution and combine it with the prior to produce a posterior distribution that acts as the final predictor. More weight is placed on the likelihood as the amount of observed data increases, and the effect of the prior can actually become negligible if a sufficient amount is given. |

In either case, the creator advises that the dataset should only be used after checking the proxy and spacial verification maps to determine whether or not the dataset will actually be useful.

The PDSI standard deviation and percentile data can give us an idea of the uncertainty for the PDSI with regards to this region. The mean standard deviation is 2.101, and the standard deviation of the standard deviation data is a low 0.063873. Considering how the PDSI values for Tabora fall within a range of -4.2458 to 2.8157, this is not insignificant. The box-whisker plot below for the mean PDSI reconstruction shows that 50% of the values lie between roughly -1 – 0.7.



To evaluate against another dataset, the mean PDSI and precipitation data for both the PHYDA and LMR datasets with regards to Tabora were correlated with the results from the 20th Century Reanalysis dataset. All data was limited to the range 1836-2000 and resulted in the table of correlation coefficients below. Note that the correlation coefficients between the PHYDA and LMR datasets differ from the ones presented previously due to the more limited temporal range.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Reanalysis Annual Precipitation** | **PHYDA Mean PDSI** | **LMR Mean PDSI** | **LMR Mean Precipitation** |
| **Reanalysis Annual Precipitation** | --- | 0.12127 | 0.09685 | 0.13841 |
| **PHYDA Mean PDSI** | 0.12127 | --- | 0.25140 | 0.44162 |
| **LMR Mean PDSI** | 0.09685 | 0.25140 | --- | 0.79139 |
| **LMR Mean Precipitation** | 0.13841 | 0.44162 | 0.79139 | --- |

The highest overall correlation is between the LMR precipitation and PDSI data, which is to be expected considering how they were derived using the same proxies and records. Interestingly, the correlation between the LMR and PHYDA datasets are higher, with 0.25140 and 0.44162 as opposed to 0.22362 and 0.33287 respectively when limited to 1836-2000 rather than when spanning 1700-2000. As for the Tabora Reanalysis precipitation dataset, however, correlations with the Tabora LMR and PHYDA datasets result in the lowest overall correlation coefficients. The maximum is 0.13704 when correlated with the LMR mean precipitation data, and the lowest with the LMR mean PDSI data at 0.09622.

### Additional Tests

For the sake of sampling how the number of proxies affected correlation with the LMR datset and checking the range of correlation coefficients, I picked three locations with varying amounts of proxy data. These locations are Santa Fe (US), Bellinzona (Switzerland), and Banda Aceh (Indonesia). Of them, Bellinzona has the highest density of proxies, Santa Fe has a lower density but a larger number in the surrounding area, and Banda Aceh had a single coral record nearby.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Santa Fe** | | **Bellinzona** | | **Banda Aceh** | |
|  | **LMR mean PDSI** | **LMR mean PRATE** | **LMR mean PDSI** | **LMR mean PRATE** | **LMR mean PDSI** | **LMR mean PRATE** |
| **PHYDA mean PDSI** | 0.25443 | 0.18577 | 0.39708 | 0.04644 | 0.48354 | 0.33740 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Santa Fe** | | **Bellinzona** | | **Banda Aceh** | |
| **LMR MCrun Index** | **PHYDA PDSI**  **X**  **LMR PRATE** | **PHYDA PDSI**  **X**  **LMR PDSI** | **PHYDA PDSI**  **X**  **LMR PRATE** | **PHYDA PDSI**  **X**  **LMR PDSI** | **PHYDA PDSI**  **X**  **LMR PRATE** | **PHYDA PDSI**  **X**  **LMR PDSI** |
| **0** | 0.12685 | 0.15095 | 0.25643 | 0.39197 | 0.36477 | 0.46565 |
| **1** | 0.1031 | 0.16921 | -0.02054 | 0.07114 | -0.00809 | 0.00532 |
| **2** | 0.12757 | 0.21363 | 0.07034 | 0.29398 | 0.33826 | 0.4167 |
| **3** | 0.26699 | 0.29261 | 0.23828 | 0.34593 | 0.25779 | 0.25871 |
| **4** | 0.18897 | 0.23342 | 0.009 | 0.24987 | -0.06971 | 0.05539 |
| **5** | 0.0756 | 0.14243 | 0.14782 | 0.2774 | 0.18591 | 0.3684 |
| **6** | 0.25694 | 0.28563 | -0.20959 | 0.03421 | 0.35411 | 0.49258 |
| **7** | 0.17765 | 0.2299 | -0.10905 | 0.10016 | 0.25322 | 0.27386 |
| **8** | 0.13973 | 0.20465 | -0.13349 | 0.16427 | 0.11626 | 0.2739 |
| **9** | 0.1172 | 0.14618 | -0.0361 | 0.32687 | 0.23652 | 0.40445 |
| **10** | 0.17991 | 0.19633 | 0.10721 | 0.34624 | 0.16505 | 0.35058 |
| **11** | 0.03862 | 0.12547 | -0.06026 | 0.14516 | 0.34643 | 0.35739 |
| **12** | 0.19316 | 0.22927 | -0.08483 | 0.24428 | 0.36352 | 0.39044 |
| **13** | 0.20252 | 0.27993 | 0.14383 | 0.40441 | 0.17979 | 0.25944 |
| **14** | 0.09722 | 0.14625 | 0.02074 | 0.31303 | 0.11152 | 0.27321 |
| **15** | 0.10777 | 0.19784 | -0.01724 | 0.1902 | 0.18576 | 0.32122 |
| **16** | 0.19551 | 0.20244 | 0.00525 | 0.13445 | 0.28744 | 0.28637 |
| **17** | 0.1064 | 0.26285 | -0.06176 | 0.15575 | 0.11758 | 0.14148 |
| **18** | 0.1885 | 0.20612 | 0.10111 | 0.30506 | 0.08151 | 0.41455 |
| **19** | 0.0814 | 0.13094 | 0.18118 | 0.30962 | 0.26455 | 0.38908 |
| **Mean** | 0.148581 | 0.202303 | 0.027417 | 0.2402 | 0.20661 | 0.309936 |

As seen in the tables above, the number of proxies does not seem to have affected the correlation between the PHYDA and LMR datasets in a discernable way. There is a wide range of correlation coefficients, with the highest actually belonging to the information gathered for Banda Aceh and the lowest for Bellinzona. Of course, this is also a small sample size and shouldn’t be used as evidence for a general trend. However, it is likely that the results of the LMR dataset and the Drought Atlases vary significantly.

## Advantages

* **Numerous Metrics Available:** Allows you to obtain data based on mean reconstruction value, standard deviation, and different percentiles in an accessible way that does not require you to compute it yourself.
  + Transparency regarding margin of error.
  + There are also numerous other variables in this dataset that may be useful for other studies. For example, El Nino anomalies and Atlantic Multidecadal Oscillation.
* **All Sources Available:** All code as well as proxy and reconstruction data used to generate the dataset is publicly available. This allows others to perform their own reconstructions using the methods presented in the original publication paper.
  + However, it should be noted that computational power required for certain operations is vey significant. The creator suggests an upper limit over 100 GB of RAM required.
  + Similarly, the total size of the input proxy and reconstruction data is very large.

## Disadvantages

* **Low Resolution:** Coordinate boxes are 1.895 latitude by 2.5 longitude.
  + Exacerbates the missing values problem, as not every coordinate box will be completely land or completely sea. Dataset may be less usable for locations close to the shore and small islands.
* **Documentation:** Less documentation in comparison to the LMR dataset with regards to the dataset.
  + **Less Detailed Metadata:** Required more research and analysis to determine the meanings of the variables and dimensions. Data ranges had to be individually found.
  + **PDSI Range:** What is the allowable range of the PDSI? Not explicitly stated in the metadata. How does this affect how the data is interpreted? For example, an actual range of -1 to +1 on an allowable range of -10 to +10 is very different than on an allowable range of -2 to +2.
    - Is there a range that the PDSI is set to? Online resources state the index range to be [-4, 4] though the range was [-40, 40] in the LMR dataset.
* **No Exact Rainfall Measurements:** Although the PDSI can act as a relative substitute