Will Data Centers Dry Out Texas?

Kendall van Beveren

DTSC 3010 Fall 2025

Abstract

This paper examines how increasing data center development in Texas may interact with long-term trends in statewide water use, drought severity, and reservoir storage. Using annual data from 2000-2024 obtained from the Texas Water Development Board, the dataset was cleaned, merged, and analyzed to assess historical relationships and forecast future conditions. Statistical tests revealed that drought severity significantly predicts reservoir levels, with an OLS model explaining 52.1% of the variance in percent-full values, and that total water use increases during drought, as shown by a significant Pearson correlation ($r = 0.55$, $p = 0.0036$). Three forecasting approaches: vector autoregression, LSTM networks, and recursive random forest model were evaluated, with the random forest model producing the most accurate projections. Forecasts through 2045 suggest that added water demand from data centers could further reduce reservoir storage, particularly during multi-year drought periods. These findings highlight the need for proactive water-resource planning as Texas experiences rapid digital-infrastructure growth.

Within the past decade, Texas has become a hot spot for businesses and data centers have been taking full advantage of the state's prime real estate market. Texas is full of large, inexpensive plots of land, as well as lower power rates and possible incentives offered to businesses by local governments. (Davies, 2025) With that being said, the Texas data center market now has 484 data centers operated by 134 providers. (Baxtel, 2024) One project that has spiked interest in the media, labeled the 'Stargate Project', will contain 875 acres of land in Abilene dedicated to data center development. Stargate was announced by the president of the United States, sharing that the project is a $500 billion dollar investment collaboration between major companies such as: OpenAI, Oracle, MGX, Microsoft, NVIDIA, and SoftBank. However, this isn't the only project of this scale making its way into Texas; various large projects similar to that of Stargate are under construction or in early planning stages for rural Texas.

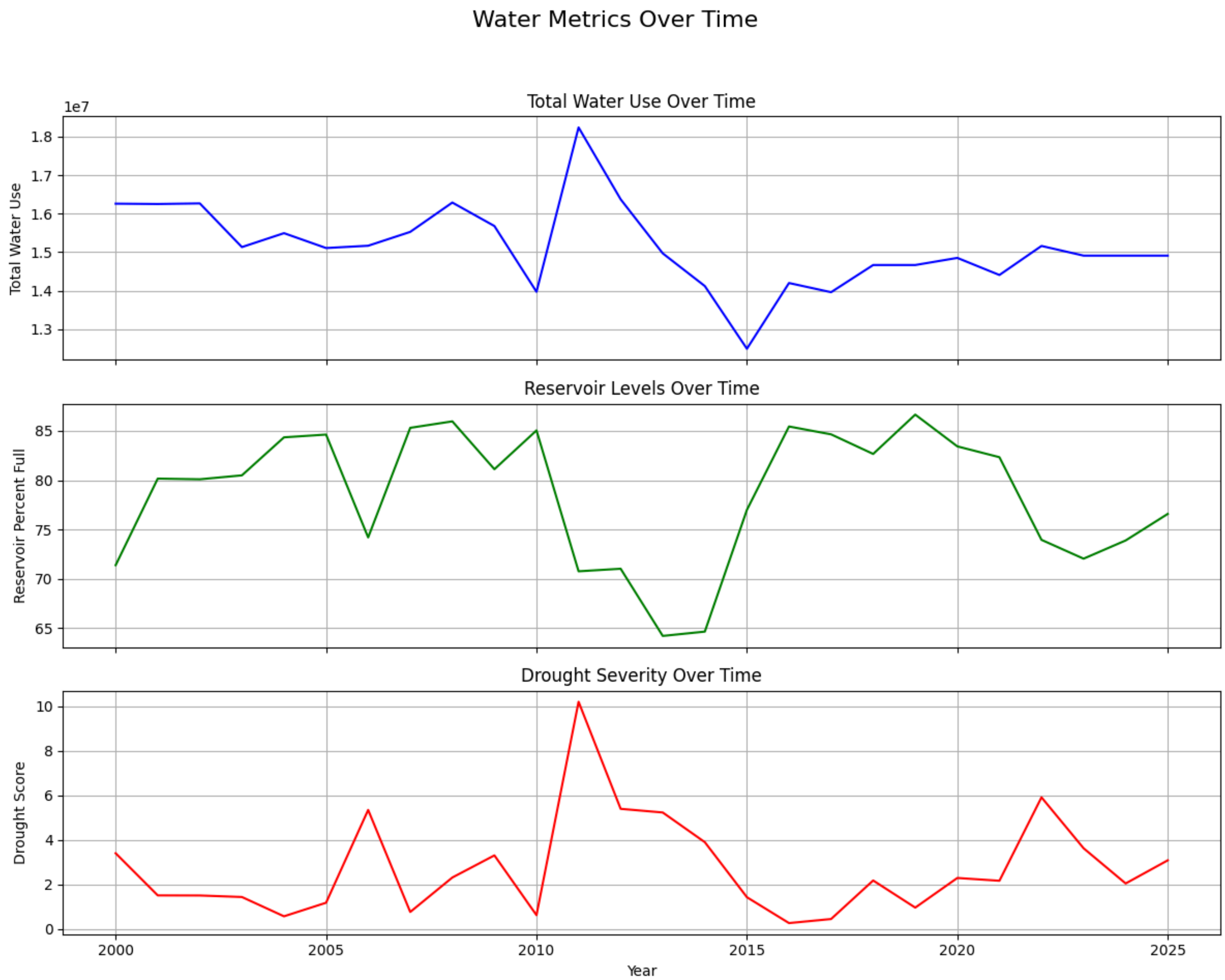
Many discussions regarding the consequences of these centers center around energy consumption, and while that may be a concern, there is a lack of conversation regarding the hidden water consumption within data centers. Texas Public Radio shares, "The average, midsized data center uses 300,000 gallons of water a day, roughly the use of a thousand homes. Larger data centers could use 4.5 million gallons a day, depending on their type of water-cooling system." (Davies, 2025) Water consumption refers to the amount of water pulled from sources without being returned to use, typically warm water from the cooling system discharged by the center, and water withdrawal covers water pulled from sources temporarily. Water usage is categorized into three categories: on-site water use, off-site water use, and water use during manufacturing of equipment. Cooling systems within the facilities take up a majority of the on-site water usage, while off-site contains the amount of water used by plants that supply to power the data centers. The water is sourced from surface water or aquifers, known as 'blue' sources, or recycled water for reuse, known as 'gray' sources (Yañez-Barnuevo, 2025). It's important to understand how exactly these data centers consume water and where it is sourced from in order to understand the possible effects it can have on our residential utilities as well as the environment. For example, areas in Virginia with a large concentration of data centers report, "Despite prioritizing reclaimed water for data center cooling in the past, Loudoun County's data center potable water consumption is now higher, having increased by 250% in the last 4 years, totaling 899 million gallons in 2023." (Trejo-Angeles, 2024) Although Texas has profitable land and energy prices, the climate is not optimal for keeping data centers cool and may not have the resources to keep up with results similar to Loudoun County.

The objective of this report is to predict the future of Texas water metrics and the possible impact large amounts of data centers could have on the state's resources. The first step included obtaining data covering water use, drought severity, and reservoir levels within Texas and using those values to predict 20 years into the future using a recursive random forest model. After the base line prediction, an additional dataset containing water consumption of various AI models was used to provide an average amount of water used by a quantity of data centers within a year. Thus, predicting water trends that represent the possible strain Texas water resources may face with an increased water consumption from the growth of data centers. Although these data centers can bring jobs and profitability to rural areas, Texas is a very hot and dry state with limited water sources combined with a rapidly growing population threatening the reliability of the state's water supply. I collected data containing water use from three sources: ground water, surface water, and reused water, along with recorded drought index and reservoir levels (percent full) from the Texas Water Development Board. The majority of the data sets were ready to download, however, I had to individually download the recorded water use for each year and combine the files into its own dataset. Preprocessing included finding a standard date format between the three data sets in order to efficiently analyze trend data. The largest interval was yearly, resulting in computing an average for each year within the data sets that had more frequent recordings and placing those values in their own Year column. Data from the year 2000 and onward was used to ensure clean comparisons and reduce gaps in data, along with creating a drought score index (0-5) to create numerical values that represent drought severity categories. Ground water, surface water, and reuse consumption were combined into one category: Total Water Use, before combining drought index, reservoir levels and finally total water use into one csv file. The file then underwent a forward-fill then back-fill method to fill gaps in the data set.

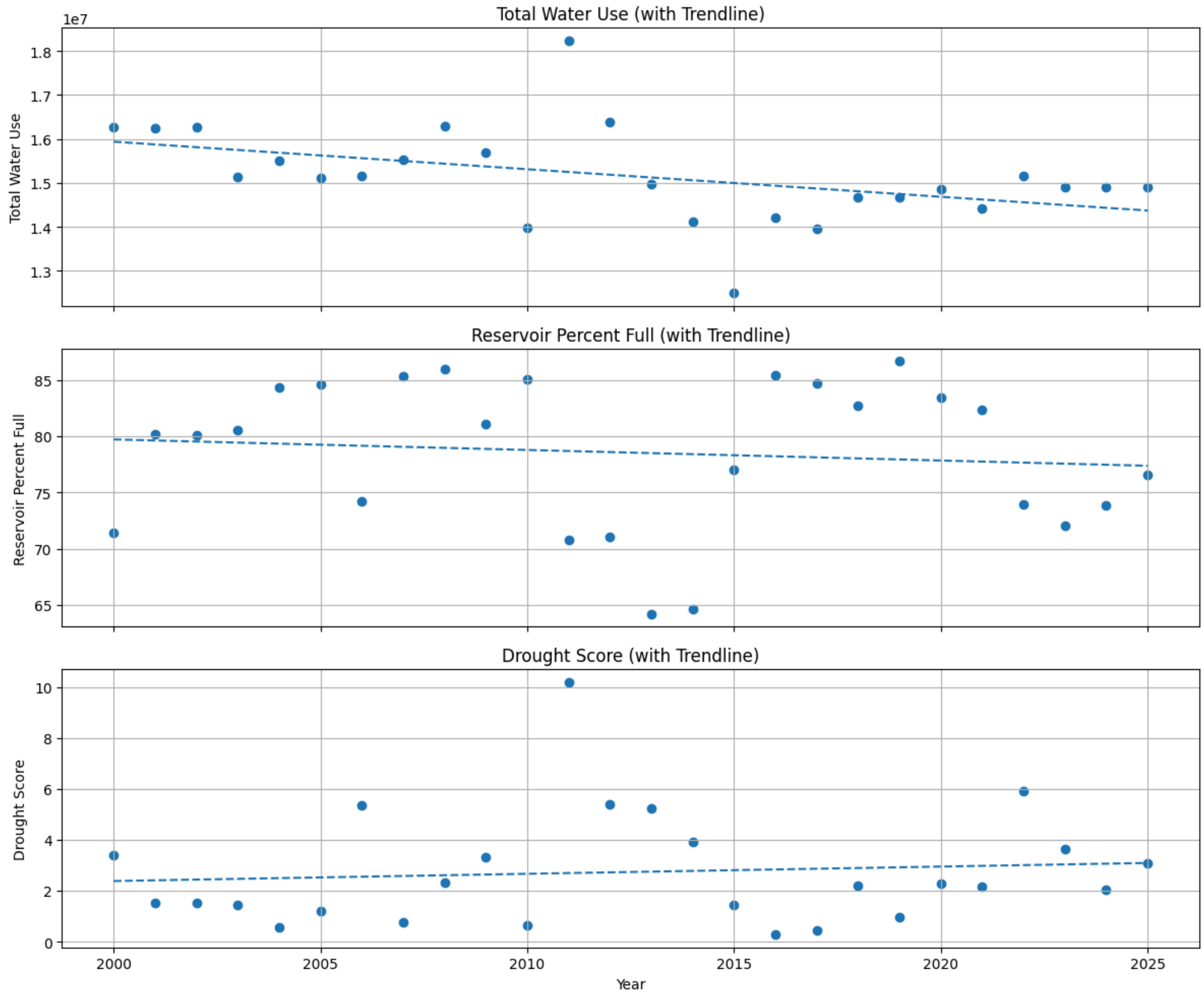
Outside of the water metric data, another dataset contained water consumed per query (mL) for a variety of modern Artificial Intelligence models used in data centers. This data was found in a dashboard that allows users to select different resource consumption options and various models to present data in multiple visualizations, along with allowing users to download the data used for the dashboard. The mean water consumed on-site was averaged between all models, due to lack of knowledge about specific models that will be implemented within Texas data centers, and instead created an estimate of water used per query for the average AI data center. Assuming a large data center runs around 100 million queries a day, the water per query value was then multiplied by 3.65e10 to represent 36.5 billion queries a year. This value was then divided by 1.233e9 to convert the milliliters into the standard acre-feet, and finally multiplied by 500 to represent the data centers operating in Texas.

Exploratory Data Analysis

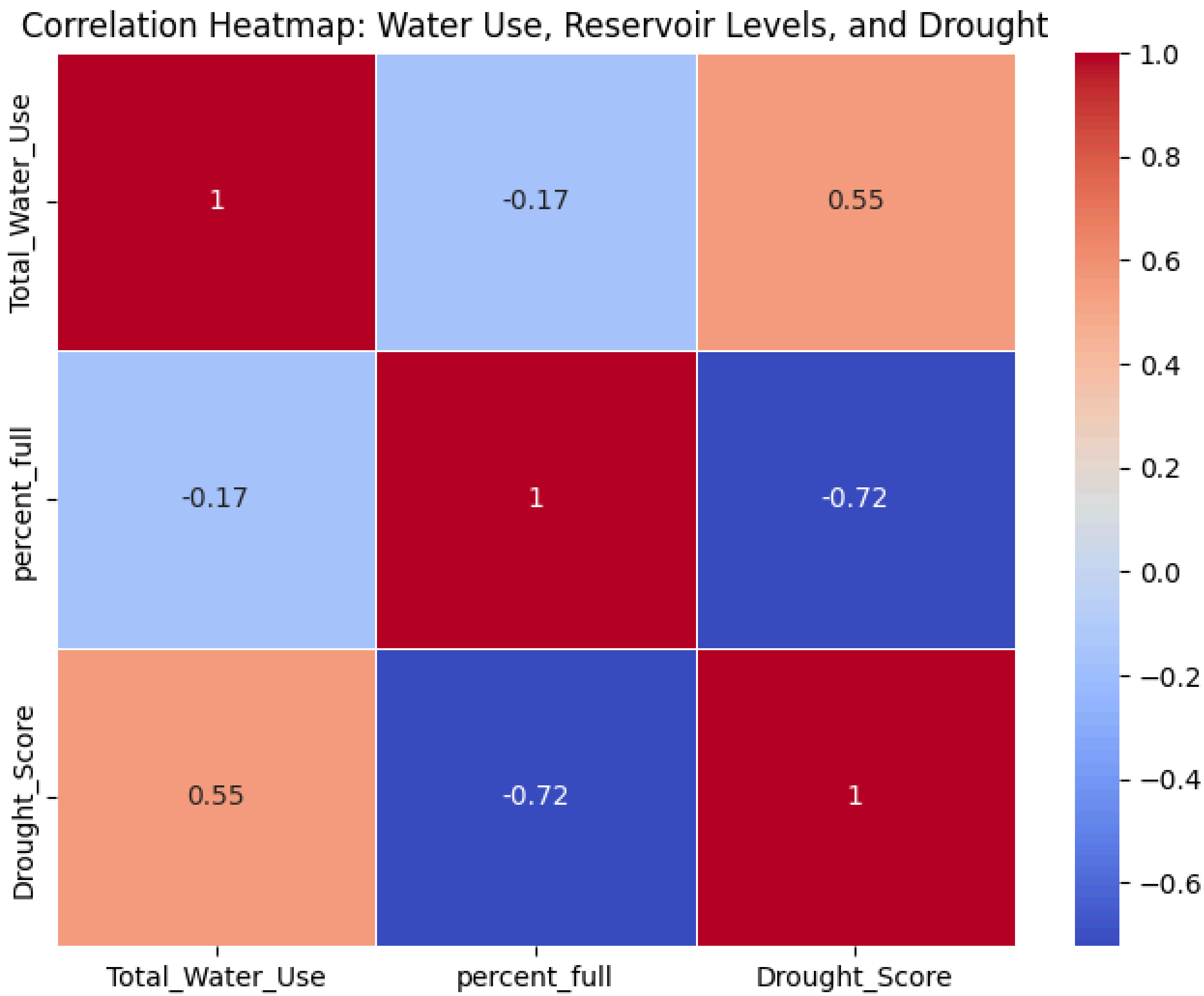
The Analysis began by creating a linear graph of the recorded water metrics to evaluate trends and interactions between total water use, reservoir level percentage, and drought index.



Total water use fluctuates between 13 and 17 million acre-feet and has a moderate variability throughout the 25 year period. Although there is a spike around 2010-2011 where the data reaches its highest point, followed by its local minimum around 2015, the data stabilizes around 14-15 million acre-feet after 2016. Reservoir levels have a high volatility and are sensitive to drought and water use. There are recurring cycles of consistent refilling after a steep drop in levels. The drought score has some short-term clustering, where high drought scores occur in multiple year clusters as seen in 2011-2013. The three graphs show negative correlation at a notable peak in 2010 where drought severity and water use spike resulting in a lowered reservoir level, followed by a lowered drought severity and water use restoring reservoir levels over the next five years.



This scatterplot with a trendline shows a slight decrease in water use over time, as well as a subtle increase in drought severity over the 25 year period.



The heatmap created from the water metric data demonstrates a positive correlation between drought severity and water use, while displaying the strong negative correlation between reservoir level percentage and drought severity.

The statistical tests conducted on the water metric data included an OLS regression and Pearson correlation. The OLS regression test was used to clarify the relationship between drought severity and reservoir levels, and more specifically if drought index can help predict reservoir levels.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           percent_full   R-squared:                       0.521
Model:                            OLS   Adj. R-squared:                  0.501
Method:                 Least Squares   F-statistic:                     26.14
Date:                Fri, 05 Dec 2025   Prob (F-statistic):           3.12e-05
Time:                        06:49:20   Log-Likelihood:                -76.167
No. Observations:                  26   AIC:                             156.3
Df Residuals:                      24   BIC:                             158.9
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const            84.3928      1.472     57.332      0.000      81.355      87.431
Drought_Score    -2.1425      0.419     -5.113      0.000      -3.007      -1.278
==============================================================================
Omnibus:                        3.190   Durbin-Watson:                   0.790
Prob(Omnibus):                  0.203   Jarque-Bera (JB):                2.130
Skew:                          -0.698   Prob(JB):                        0.345
Kurtosis:                       3.130   Cond. No.                         5.88
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

The model revealed a moderately strong relationship between the two variables since the drought score explains 52.1% of the variation in reservoir percent full. Each unit increase in drought severity is associated with an average 2.14% decrease in reservoir levels and has a highly significant effect due to the p-score being significantly less than 0.05. However, the Durbin-Watson value of 0.79 implies a positive autocorrelation.

The Pearson correlation between drought severity and total water use revealed a significant moderate positive relationship from the results r = 0.55 and p = 0.0036. This means as drought severity increases, total water use tends to increase as well, while the p-score indicates the correlation is statistically significant.

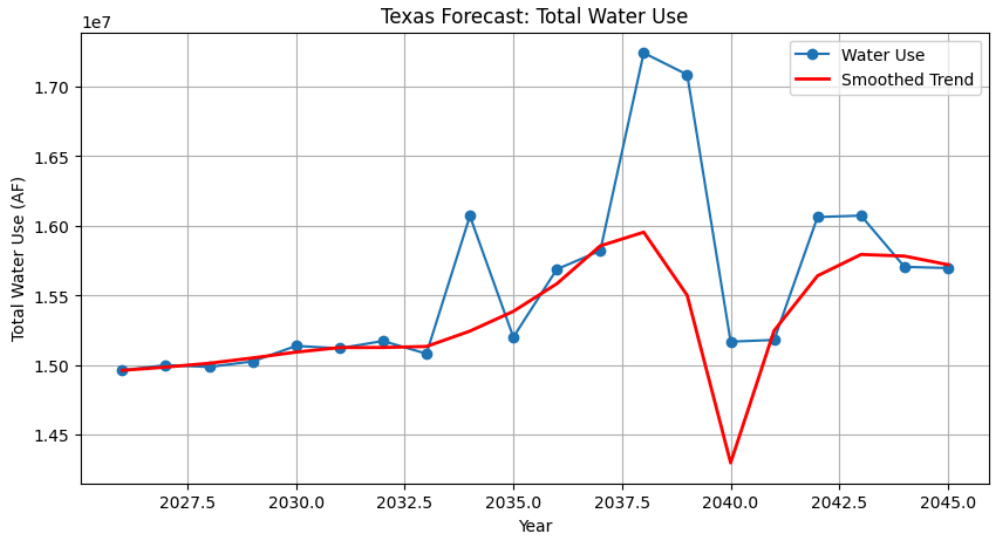
Models

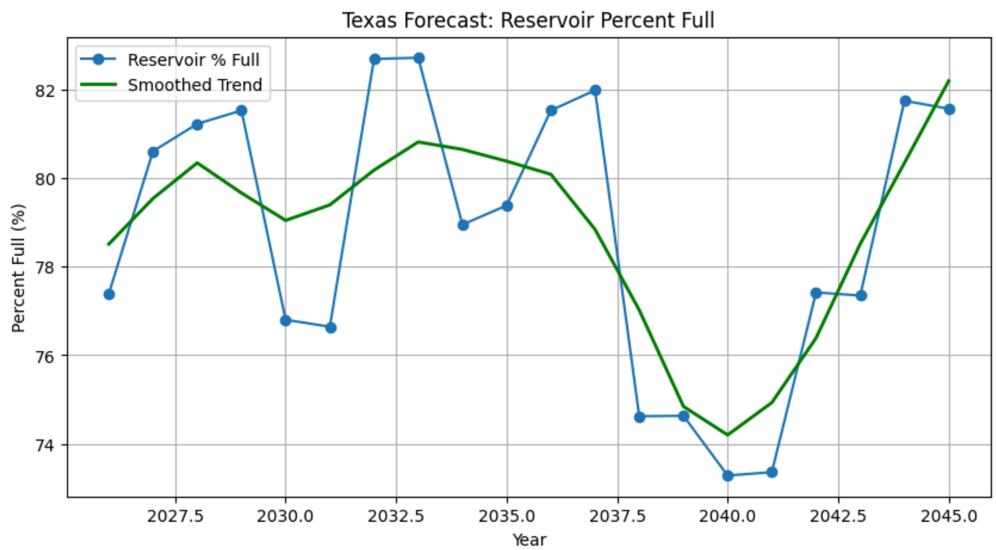
Three separate machine learning models were then tested with the water metric data including: vector autoregression, long short-term memory (a type of recurrent neural network), and a recursive random forest model. To determine the best model for the data, I calculated and compared the mean absolute error, root mean squared error, and mean absolute percentage error for the predicted water use and reservoir levels using the different models. The tables below represent the error metrics for water use and reservoir levels in that order.

MODEL	MAE	RMSE	MAPE
VAR	3256980	3256980	21.569
RF	1308070	1664558	9.074
LSTM	1913856	2094095	12.665

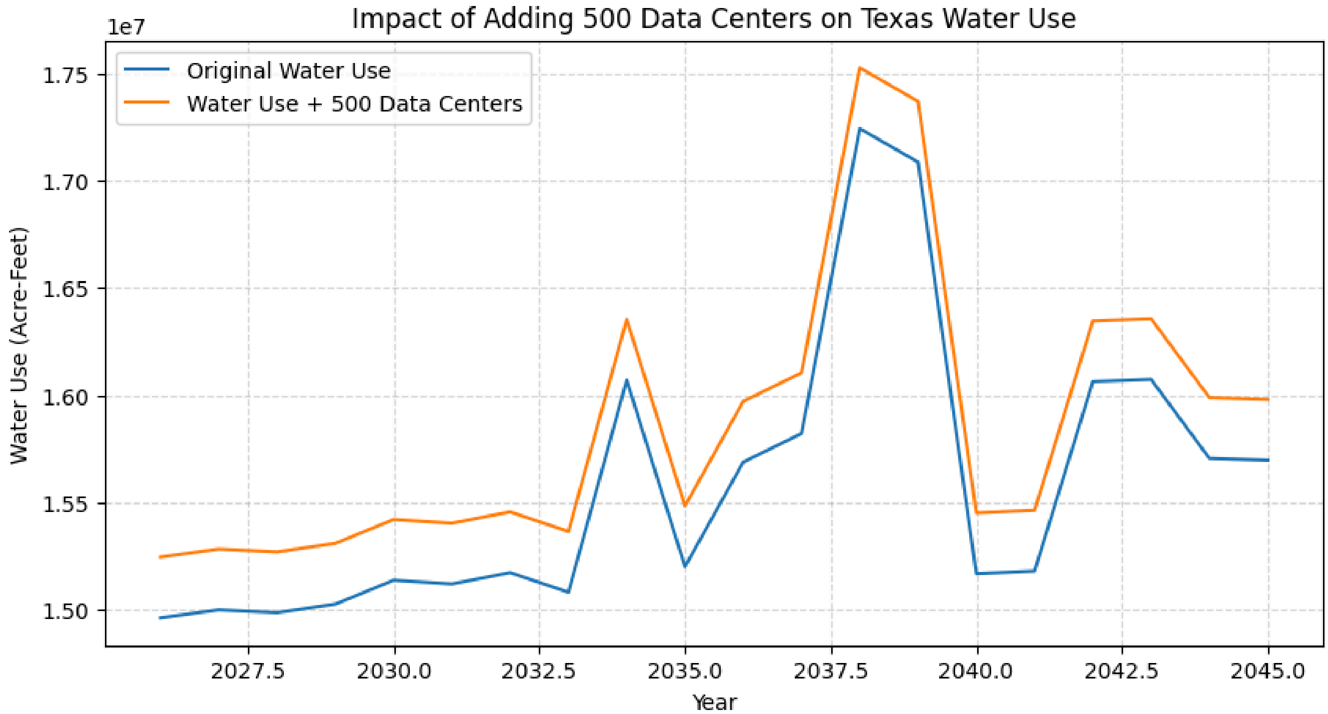
MODEL	MAE	RMSE	MAPE
VAR	21.769	21.769	31.23
RF	7.375	8.176	9.90
LSTM	11.795	13.916	14.693

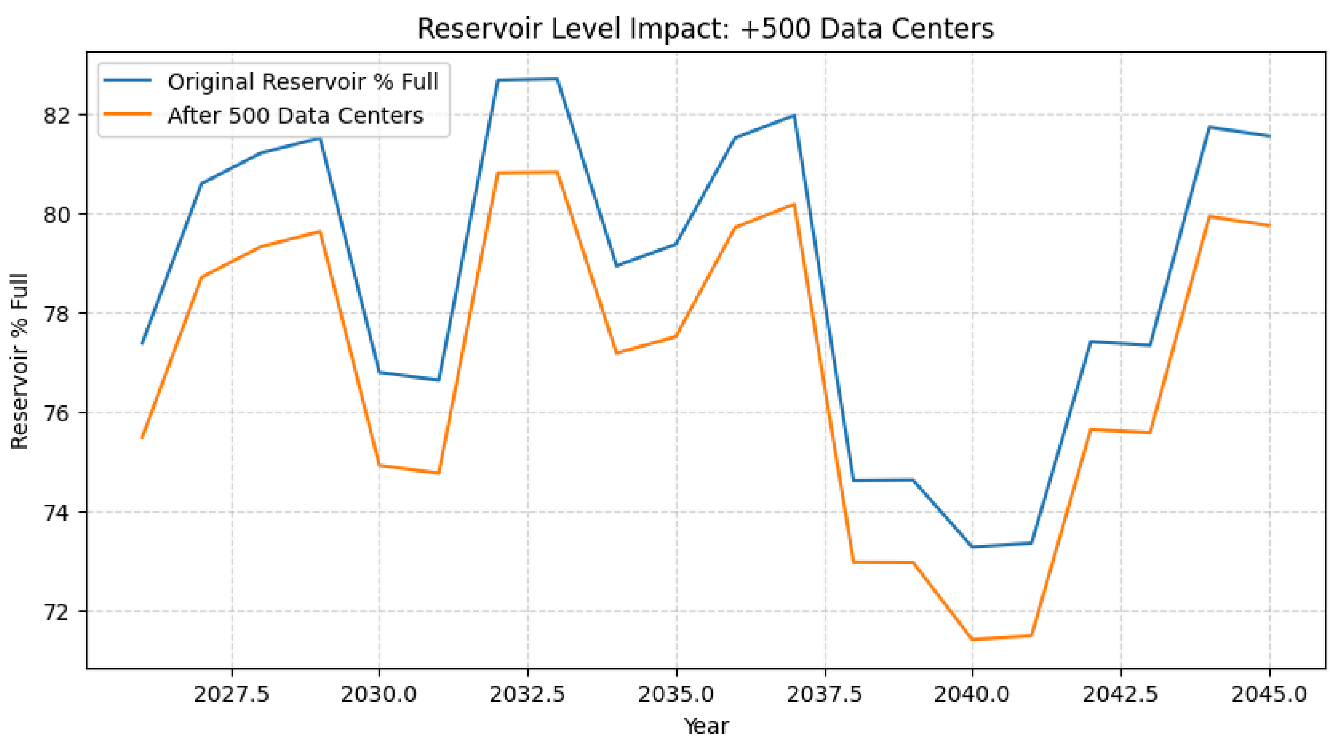
The error metrics show that random forest would be the best model option for my data and goals. Implementing random forest utilized the recorded data to predict 20 years into the future and estimate water use and reservoir levels up to the year 2045. VAR was not a good candidate for this data and presented the highest values of the three, and although LSTM has fairly good scores, the run time and retracing slowed the code down significantly and required a lower amount of epochs thus reducing the accuracy of the prediction.



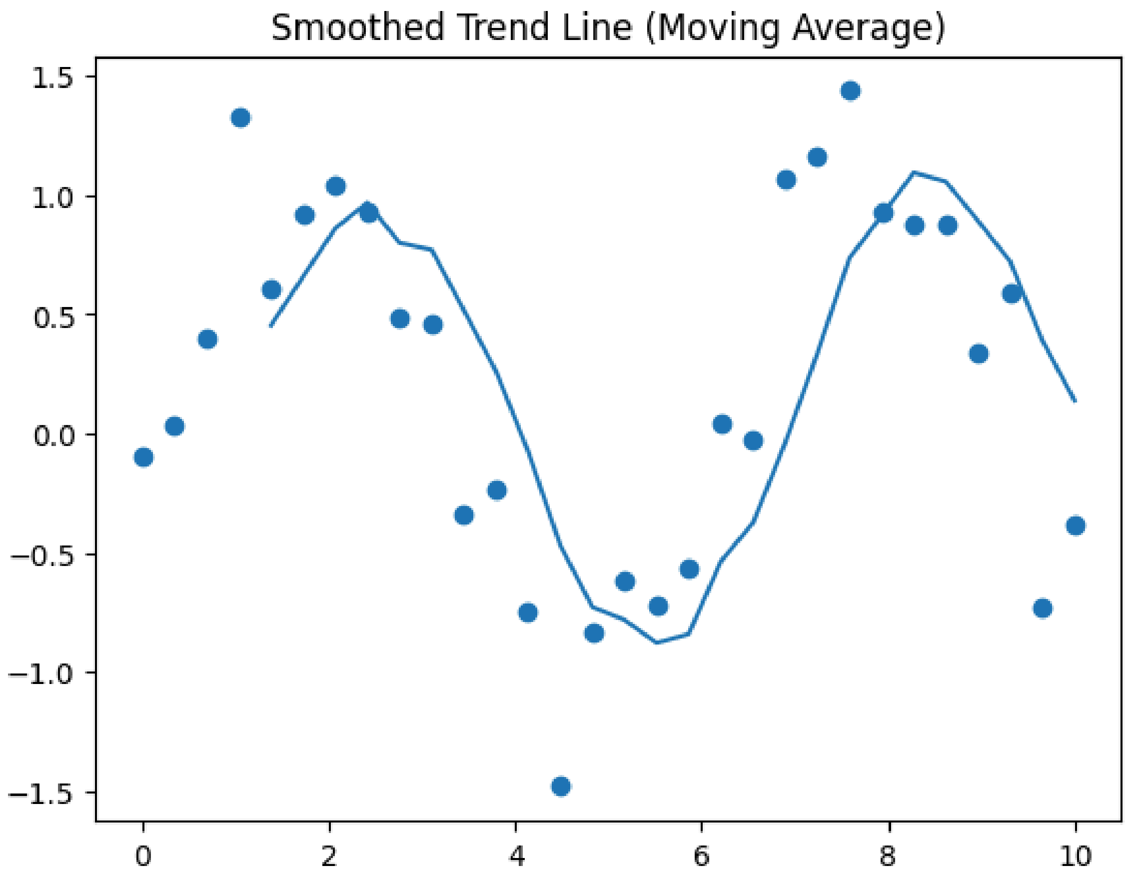


The predicted graphs have a similar correlation to the original data graphs, however there is an outlier in the total water use graph around 2040 that does not match the previous patterns between water use and reservoir data. This data would serve as the baseline for the future predictions and analysis for data center impact on Texas water resources.





Taking the baseline prediction data and implementing the estimated values for data center water consumption, we are able to envision the strain that data centers could place onto future water trends within Texas, especially around the year 2040 where the reservoir levels reach a concerning new minimum.



Last, I computed a moving average with a trend line to predict and document possible reservoir capacity that would be lost due to the extra strain of data center water consumption, on top of Texan's yearly water use.

Overall, data centers will consume more water within Texas as they continue to grow and expand throughout the state. Reservoir levels show decline corresponding to heightened periods of water use and severe drought, as well as showing a gradual increase in total water use over time. Water resources are already strained in periods of severe drought and heightened water use, resulting in dangerously low reservoir levels with the addition of rapidly growing data centers consuming large amounts of water. Water infrastructure and conservation must become a higher priority in order to preserve the Texas water supply and ensure safe reservoir levels.

Summary of Learning

During my time conducting this project, I have learned that planning is absolutely crucial. It is very common for mishaps and issues that would cause major setbacks without proper time allotted for each step. I have also learned that preprocessing data is crucial when conducting higher levels of analyzation and implementing machine learning models. For example, I didn't think too much about the units of measure for water, but simply converting the different datasets containing water information took up a very large portion of my preprocessing work as well as impacting my results for the entire project, such as displaying acre-feet rather than a more common unit of measure. One interesting finding was the correlation between drought and water use. I did not expect water use to increase during periods of severe drought, in fact I expected the opposite but was able to use the results to think of reasons why they are correlated, such as a possible need to pump more ground water, or an increase in agricultural water demand to compensate for the dry weather. I also tried my best to attempt every model or statistical test myself based on online resources before searching into more specific problems or errors I encountered, and I believe it allowed me to become more comfortable with the methodologies and improve problem-solving skills with python as I continued to work on the project.

References

Abera, L. (2025). *The Growing Role of Artificial Intelligence in Tomorrow's Urban Hydrological Infrastructure Liya Abera*.

Atkinson, L., & Miniard, D. (2025). *POWER STRUGGLES: TEXANS' THOUGHTS ON THE AI DATA CENTER BOOM AND ITS ENVIRONMENTAL IMPACTS POWER STRUGGLES: TEXANS' THOUGHTS ON THE AI DATA CENTER BOOM AND ITS ENVIRONMENTAL IMPACTS 2 BACKGROUND*. https://mediaengagement.org/wp-content/uploads/2025/07/Texans-Thoughts-on-the-AI-Data-Center-Boom-and-Its-Environmental-Impacts.pdf

Baxtel. (2024). *Texas Data Centers & Colocation*. @Baxtel. https://baxtel.com/data-center/texas

Davies, D. M. (2025, August 8). *Texas Matters: The future of AI and water in Texas*. TPR. https://www.tpr.org/podcast/texas-matters/2025-08-08/texas-matters-the-future-of-ai-and-texas-water

Elsworth, C., Huang, K., Patterson, D., Schneider, I., Sedivy, R., Goodman, S., Townsend, B., Ranganathan, P., Dean, J., Vahdat, A., Gomes, B., & Manyika, J. (2025). *Measuring the environmental impact of delivering AI at Google Scale*. https://arxiv.org/pdf/2508.15734

Jegham, N. (2025). *Power BI Report*. Powerbi.com. https://app.powerbi.com/view?r=eyJrIjoiZjVmOTI0MmMtY2U2Mi00ZTE2LTk2MGYtY2ZjNDMzODZkMjlmIiwidCI6IjQyNmQyYThkLTljY2QtNDI1NS04OTNkLTA2ODZhMzJjMTY4ZCIsImMiOjF9

Li, P., Yang, J., Islam, M., & Ren, S. (2025). *Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models*. https://arxiv.org/pdf/2304.03271

Martinez, A. (2025, September 25). *Texas planners don't know how much water data centers will need*. The Texas Tribune. https://www.texastribune.org/2025/09/25/texas-data-center-water-use/

McBride, S., Ghaffary, S., & Ford, B. (2025, January 24). Stargate's First Data Center Site is Size of Central Park, With At Least 57 Jobs. Yahoo Finance. https://finance.yahoo.com/news/stargate-first-data-center-size-202854022.html

Office of the Texas State Climatologist . (2025). *Water Data For Texas*. Waterdatafortexas.org. https://www.waterdatafortexas.org/drought/drought-monitor?period=2025-09-23&areaType=state&areaName=tx

OpenAI. (2025, January 21). *Announcing The Stargate Project*. Openai.com. https://openai.com/index/announcing-the-stargate-project/

Our contribution to a global environmental standard for AI | Mistral AI. (2025). Mistral.ai. https://mistral.ai/news/our-contribution-to-a-global-environmental-standard-for-ai

Primm, G. (2024, March 12). *Forecasting Data Center Immersion Cooling Technology for the Year Ahead*. Green Revolution Cooling. https://www.grcooling.com/blog/forecasting-data-center-immersion-cooling-technology/

Sen Gupta, P., Hossen, R., Li, P., Ren, S., Atiqul Islam, M., & Rahim Talukder, Z. (2024, May 3). *A Dataset for Research on Water Sustainability* . OSF; National Science Foundation. https://osf.io/8zuhy/metadata/?format=datacite-json

Shumba, N., Tshekiso, O., Li, P., Fanti, G., & Ren, S. (2025). *A Water Efficiency Dataset for African Data Centers*. 453–460. https://doi.org/10.1145/3715335.3735483

Texas Water Development Board. (n.d.). *Water Data For Texas*. Www.waterdatafortexas.org. https://www.waterdatafortexas.org/drought/overview

Texas Water Development Board. (2022). *2022 State Water Plan Water User Group Water Management Strategy Supply - Texas Water Data Hub*. Txwaterdatahub.org. https://txwaterdatahub.org/dataset/2022-state-water-plan-wug-wms-supply

Texas Water Development Board. (2025a). 2026 Region C Regional Water Plan. *Region c Draft 2026 RWP*, *1*. https://www.twdb.texas.gov/waterplanning/rwp/plans/2026/index.asp

Texas Water Development Board. (2025b). *Draft Regional Water Plan Data 2026*. Texasstatewaterplan.org. https://2027.texasstatewaterplan.org/region/C

Texas Water Development Board. (2025c). *Region-C Planning Region Reservoirs*. Waterdatafortexas.org. https://waterdatafortexas.org/reservoirs/region/region-c

Texas Water Development Board. (2025d). *Reservoirs Statewide 1 Year*. Waterdatafortexas.org. https://www.waterdatafortexas.org/reservoirs/statewide-1year.csv

Texas Water Development Board. (2025e). *Texas Water Development Board*. Texas.gov. https://www.twdb.texas.gov/surfacewater/conditions/report/archives/index.asp

Texas Water Development Board. (2025f, August 7). *Historical Water Use Survey Data | Texas Water Development Board*. Www.twdb.texas.gov. https://www.twdb.texas.gov/waterplanning/waterusesurvey/dashboard/index.asp

Wu, Y., Hua, I., & Ding, Y. (2025). Not All Water Consumption Is Equal: A Water Stress Weighted Metric for Sustainable Computing. *ACM SIGEnergy Energy Informatics Review*, *5*(2), 84–90. https://doi.org/10.1145/3757892.3757904

Yañez-Barnuevo, M. (2025, June 25). *Data Centers and Water Consumption | Article | EESI*. Eesi.org. https://www.eesi.org/articles/view/data-centers-and-water-consumption