HydroScale: Forecasting Water Efficiency for Data Centers

Fall 2024 MIDS Capstone - Marlon Fu, Austin Ho, Nora Povejsil, Suhas Prasad, Derek Yao

Our Team

Marlon Fu

Derek Yao

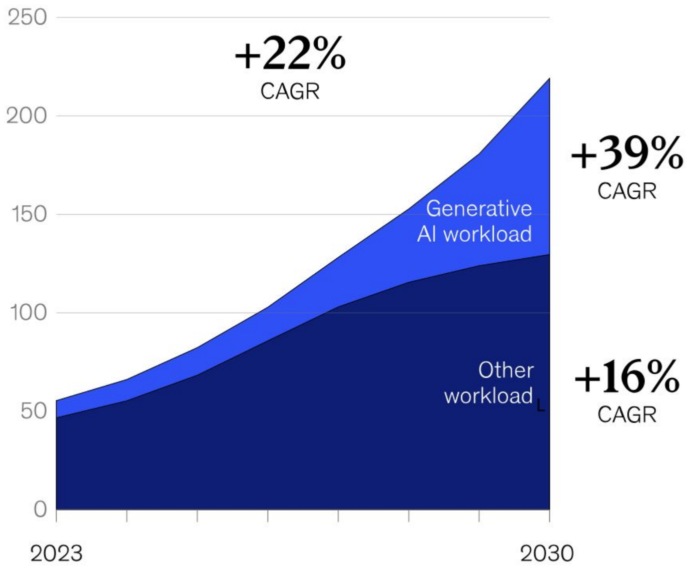
Nora Povejsil

Suhas Prasad

Austin Ho

Behind Every AI is a Data Center

Estimated global data center capacity demand,[1] gigawatts



CNBC

Data centers powering artificial intelligence could use more electricity than entire cities

Data center campuses power artificial intelligence and cloud computing; The campuses could grow so large that finding enough power and...

Yahoo Finance

Meta Expands AI Infrastructure with $10 Billion Louisiana Data Center

Meta

The Economic Times

Microsoft deal signals booming demand from data centres to power AI

US utilities are securing lucrative power supply deals with data center operators amid surging AI-driven demand.

McKinsey & Company, AI power: Expanding data center capacity to meet growing demand

Water: The Unseen Resource

300,000 Gal.

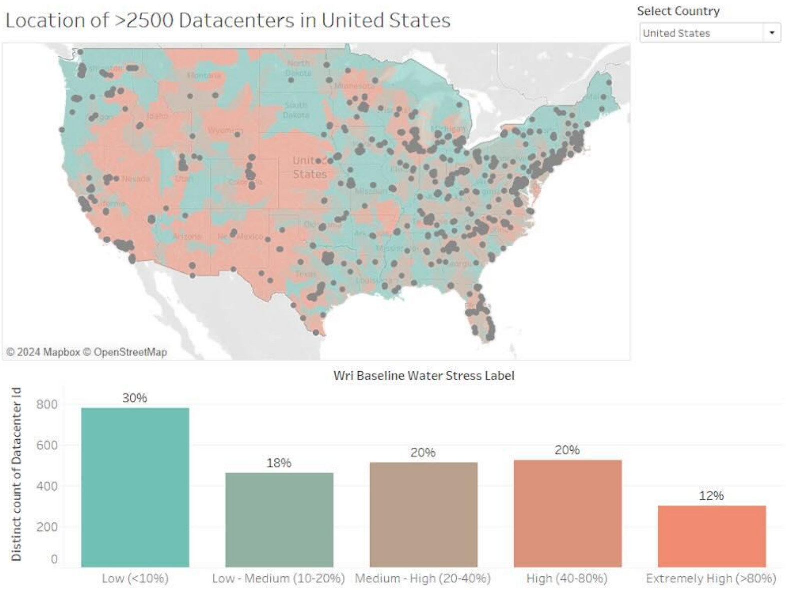
Average Daily Water Withdrawn

52% of DCs

Using Medium to High Stress Watersheds

96 of 204

Water Basins Facing Shortages in 50 Years



Ana Pinheiro Privette, AI's Challenging Waters
Understanding Water Availability | U.S. Geological Survey
Planet Tracker, AI needs to reduce its water dependency

Why Act Now?

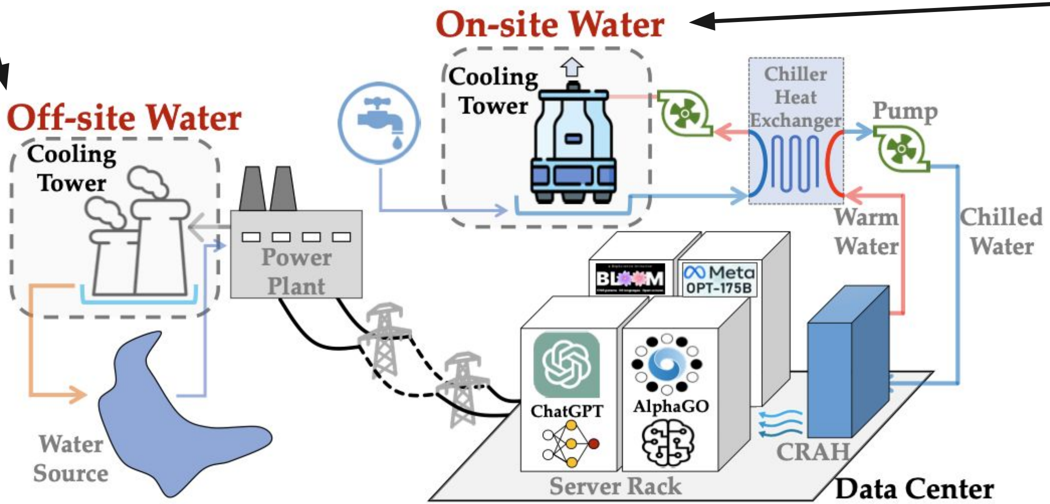
*“A large focus for data center efficiency has been on minimizing energy use […]. **The need to minimize water consumption has received considerably less attention**.”*

- Ana Pinheiro Privette, University of Illinois Urbana-Champaign

Two Types of DC Water Consumption

Water used **indirectly** in generating electricity for the data center.

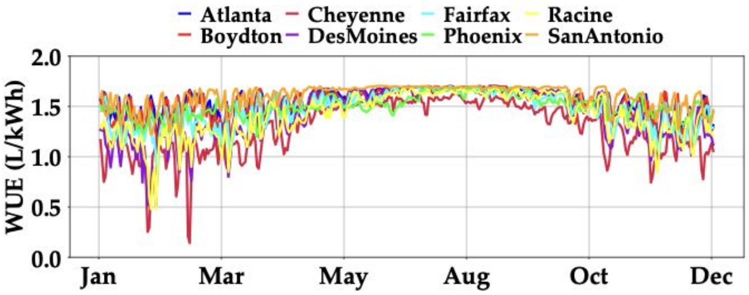
Water used **directly** for server cooling at the data center.

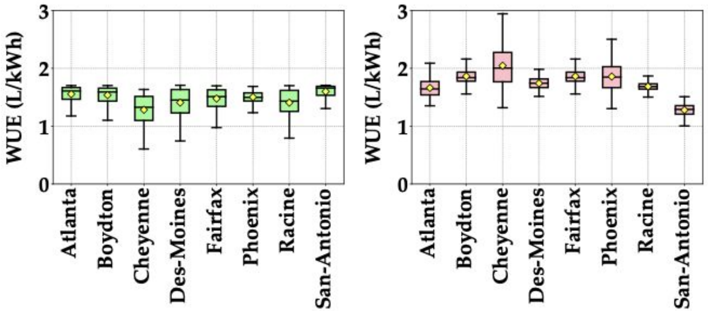


Pengfei Li, Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models

Measuring Efficiency Through
Water Usage Effectiveness (WUE)

WUE [L/kWh] is the industry standard metric for water efficiency, with an ideal value of 0.





*"By exploiting **spatial-temporal diversity** of water efficiency, we can **dynamically schedule** AI model training and inference to cut the water footprint."*

- Pengfei Li et al., University of California - Riverside

Pranjol Sen Gupta et al., A Dataset for Research on Water Sustainability

Our Mission

HydroScale empowers data centers to optimize water efficiency with **72-hour, nationwide water usage effectiveness (WUE) forecasts**, providing spatial and temporal insights for **both on-site and off-site** water.

Proactiveness

Forecasted WUE guide DCs in adapting their operations to **avoid water inefficient times and locations**.

Visibility

Spatial aspect provides visibility into how water efficiency is **related to local water scarcity** impacts.

Benchmarking

Forecasts enable DCs to compare anticipated water usage against **regulatory and ESG benchmarks**.

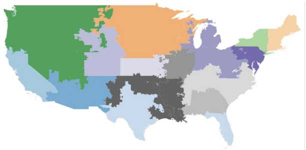
Data **Sources**

Off-site

- Time series WUE simulated from energy generation fuel mix data
  - Time: **5 yrs, hourly**
  - Space: **19 eGRID regions**
- eGRID bounding shapefiles

Independent Statistics and Analysis
U.S. Energy Information Administration

EPA



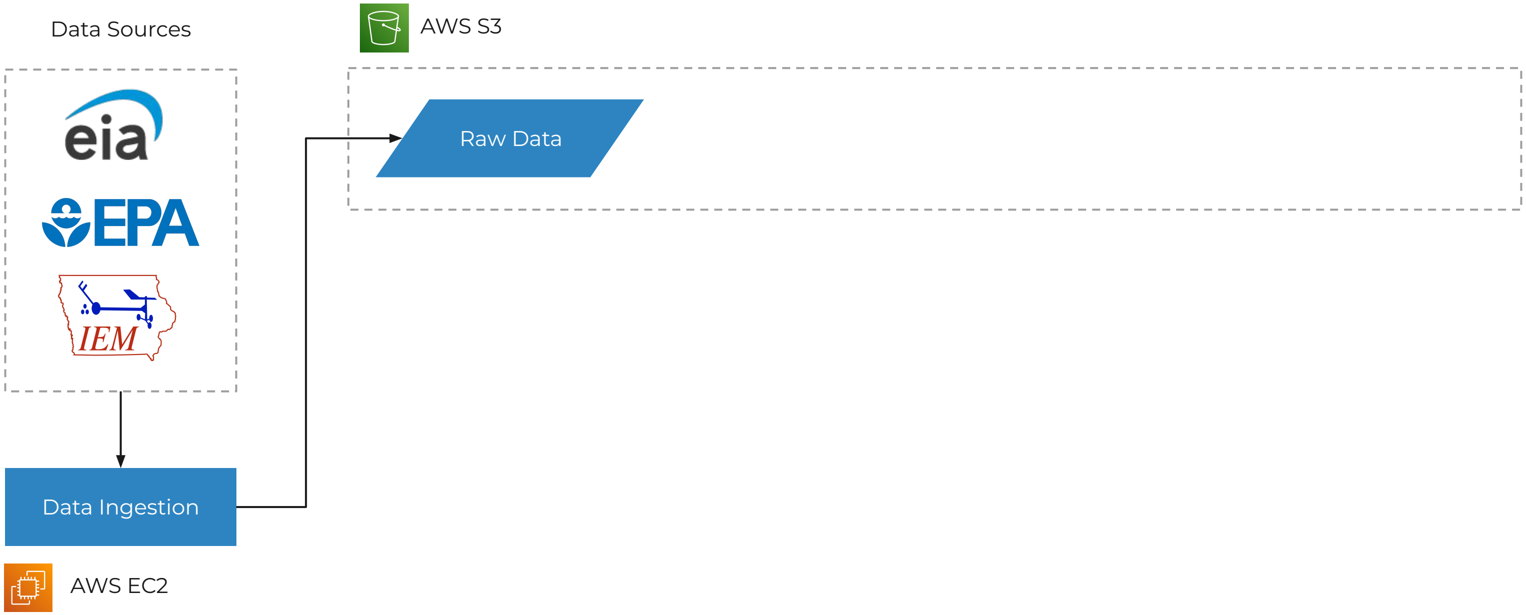
On-site

- Time series WUE simulated from historical weather data
  - Time: **5 yrs, hourly**
  - Space: **1153 weather stations**
- Station geospatial metadata

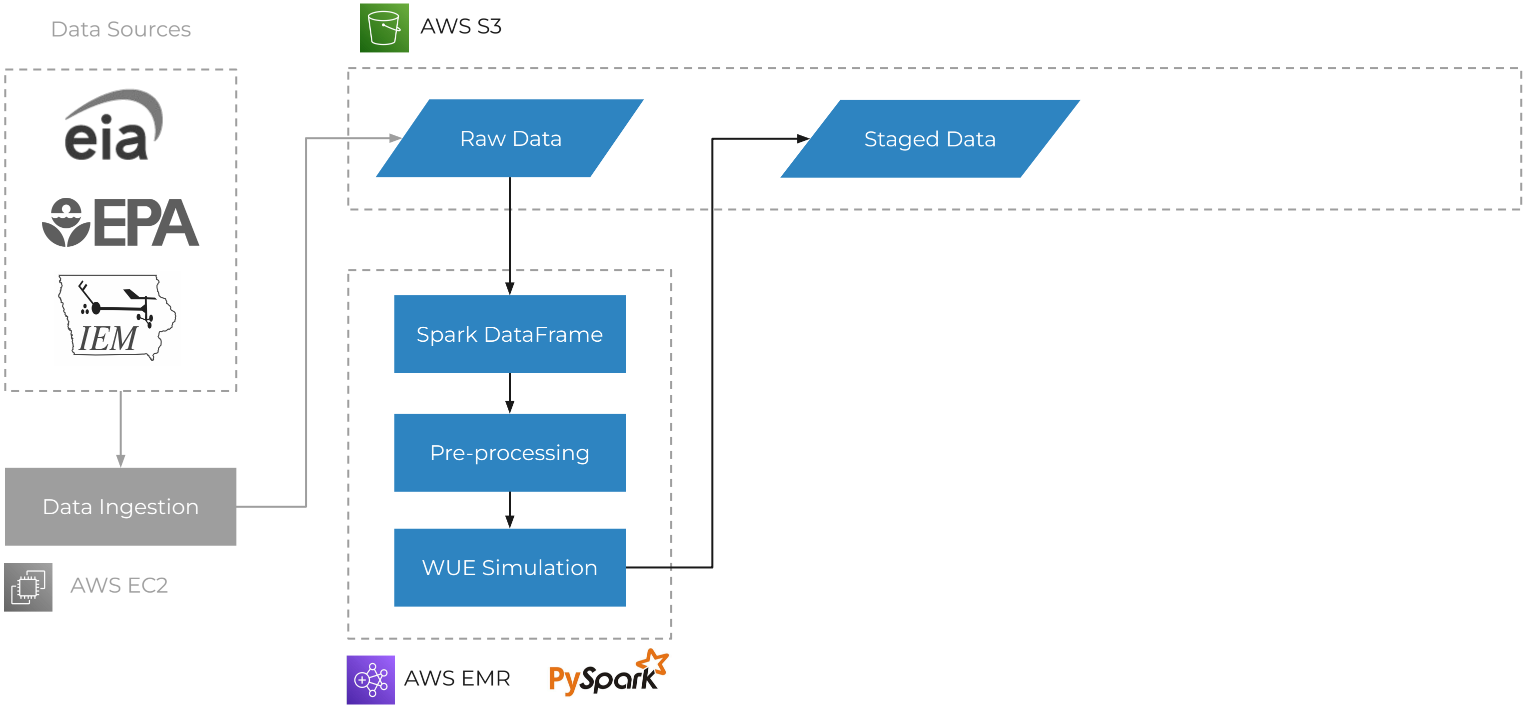
IEM

We follow the methodology in *A Dataset for Research on Water Sustainability* by Pranjol Sen Gupta et al. for simulating WUE based on our available data.

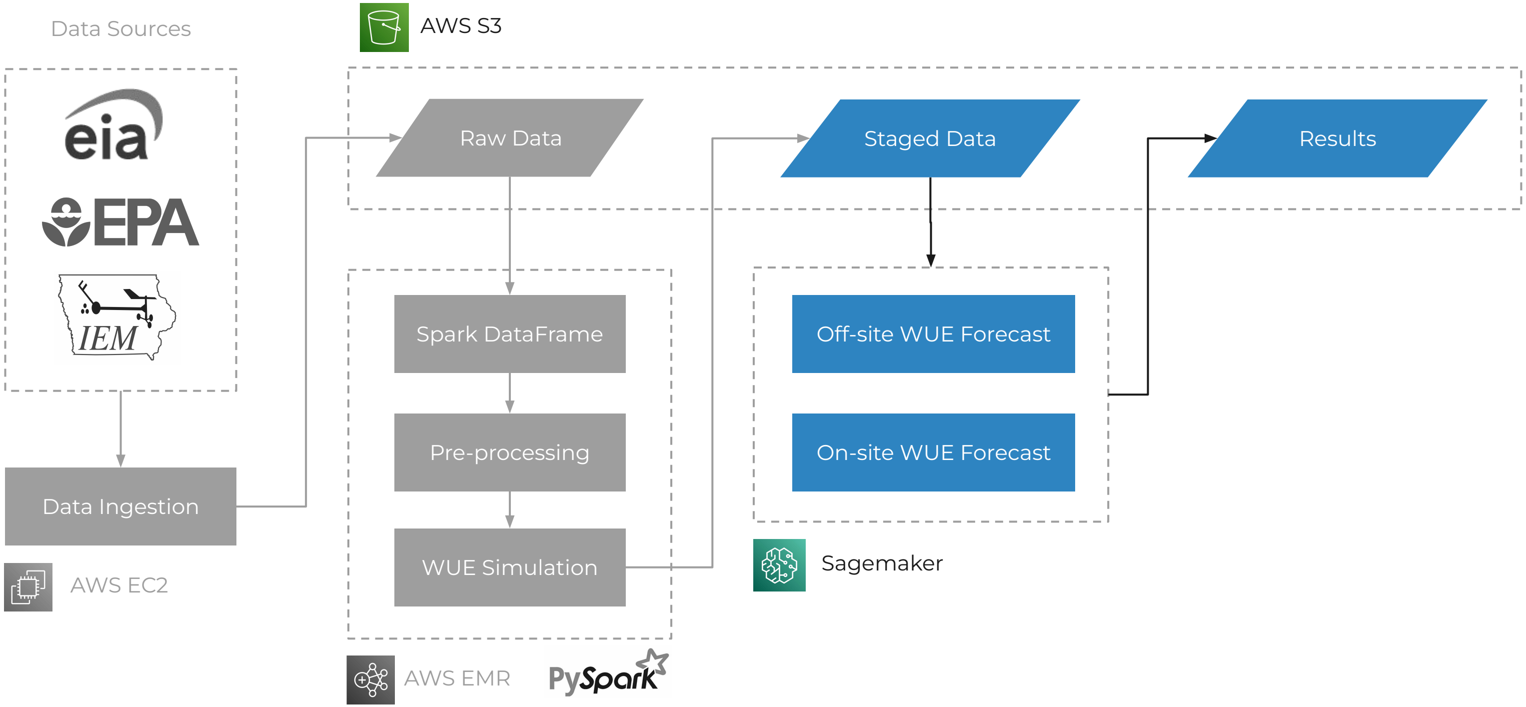
Data Pipeline **Overview**



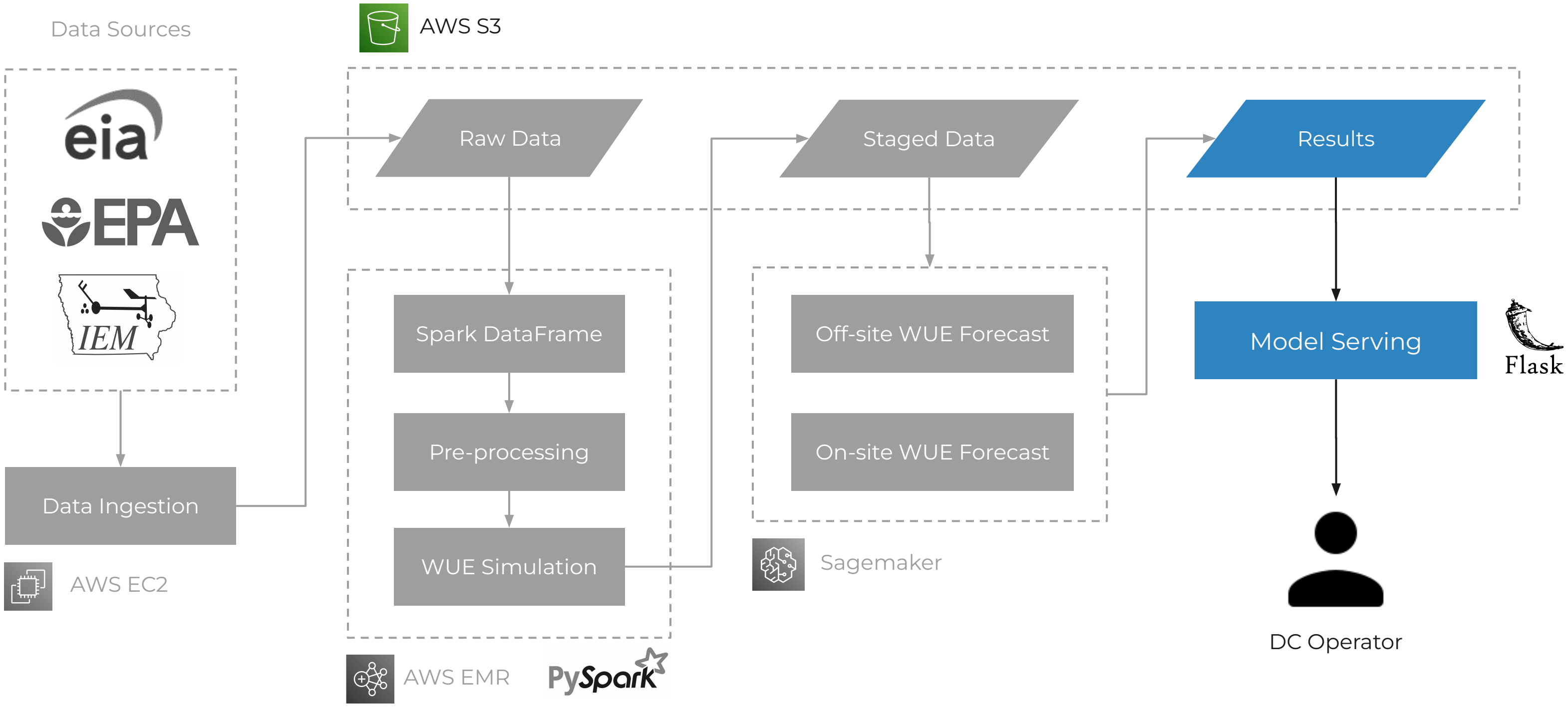
Data Pipeline Overview



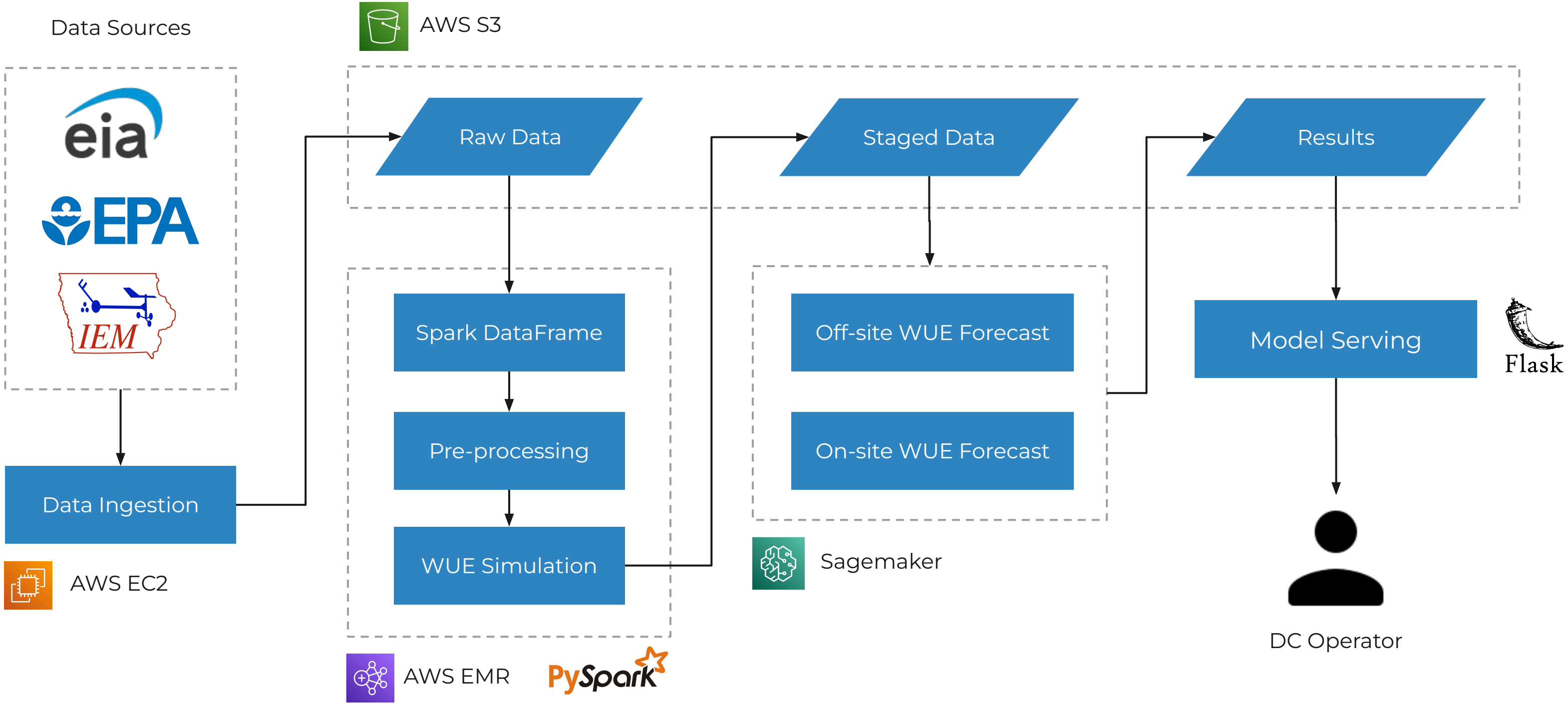
Data Pipeline Overview



Data Pipeline Overview



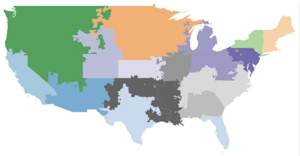
Data Pipeline Overview



Off-Site WUE Forecasting

Off-site WUE Forecasting Methods

5 yr. Hourly Data
19 eGRID Regions



Train/Val/Test Split

Train (~3yrs)*

Backtest (72 hours)

Forecast (72 hours)

Models

SARIMA

LSTM

TimesFM*

MAE, RSME, and MAPE Analysis

Write Chosen Forecast to S3

*TimesFM accepts a maximum training window of 21 days

Off-site Modeling

SARIMA

- ➔ **Auto-ARIMA models with seasonal components**
- ➔ 19 time series from unique eGRID regions
- ➔ 72 hr forecast
- ➔ 72 hr backtest

LSTM

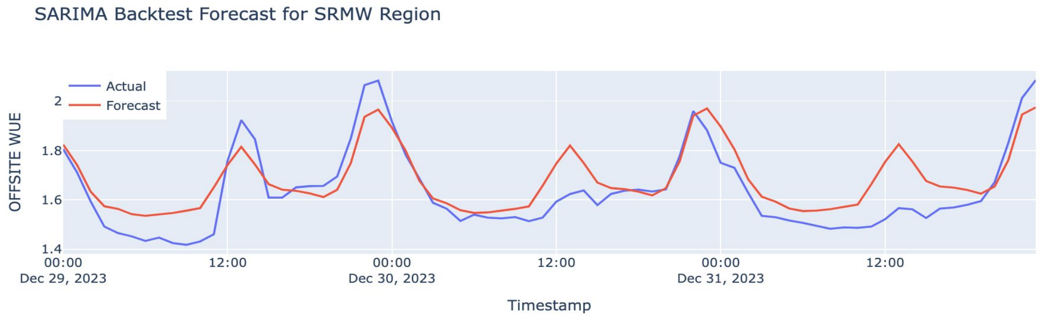
- ➔ **Sequential RNN**
  - **adam optimizer**
  - **relu activation**
  - **mse loss**
- ➔ 19 time series from unique eGRID regions
- ➔ 72 hr forecast
- ➔ 72 hr backtest

TimesFM

- ➔ **Google foundational model**
- ➔ 19 time series from unique eGRID regions
- ➔ 72 hr forecast
- ➔ 72 hr backtest
- ➔ **Can only use 21 days of training data**

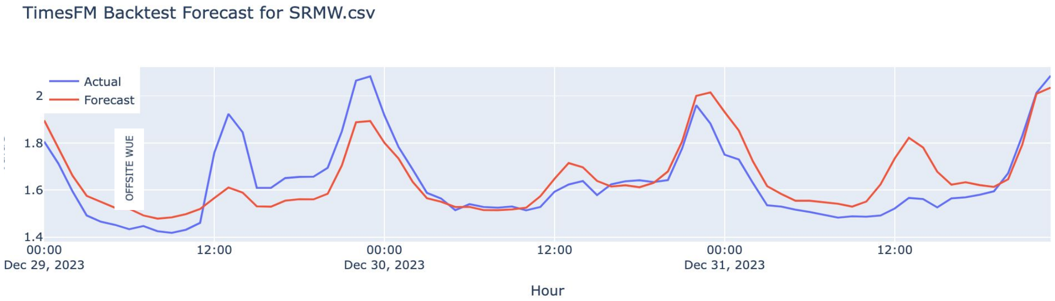
Model Evaluation (Off-site)

SARIMA

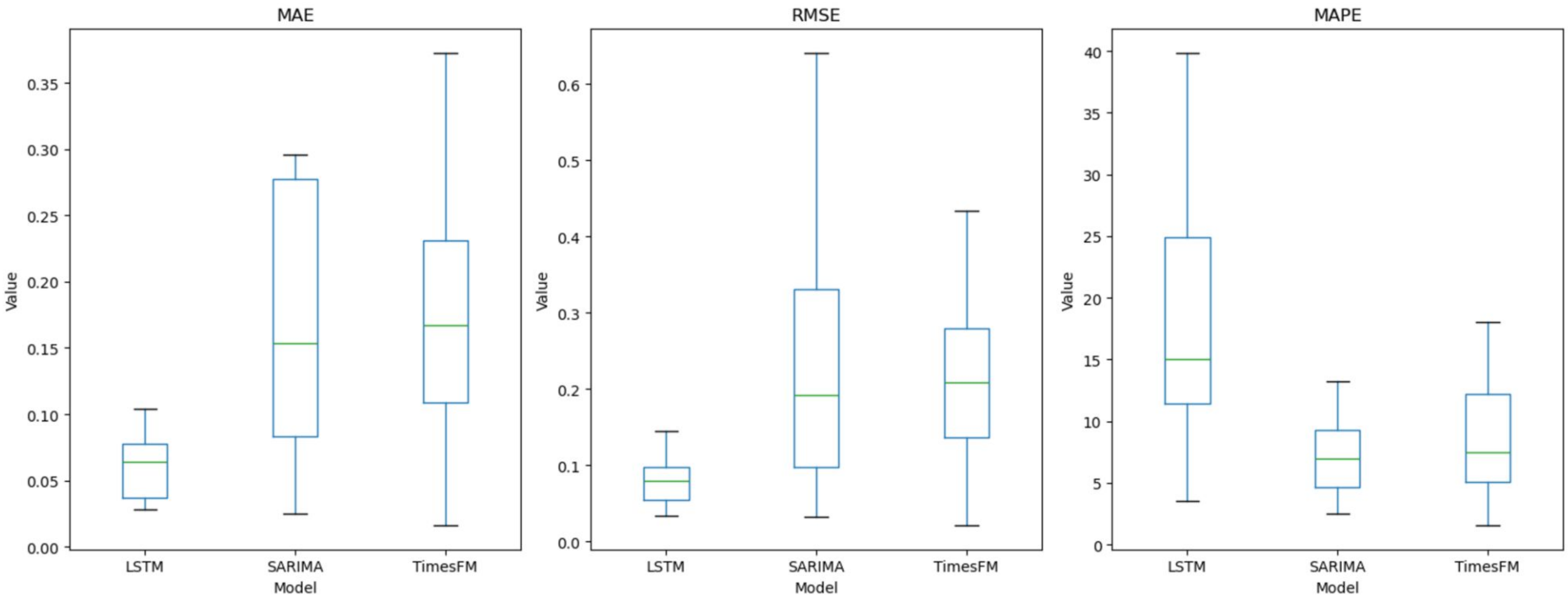


LSTM

TimesFM

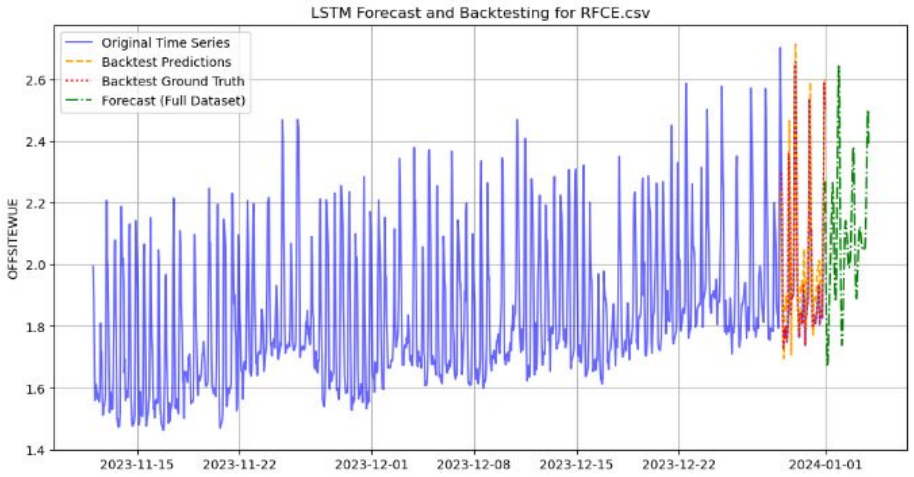


Model **Evaluation (Off-site)**

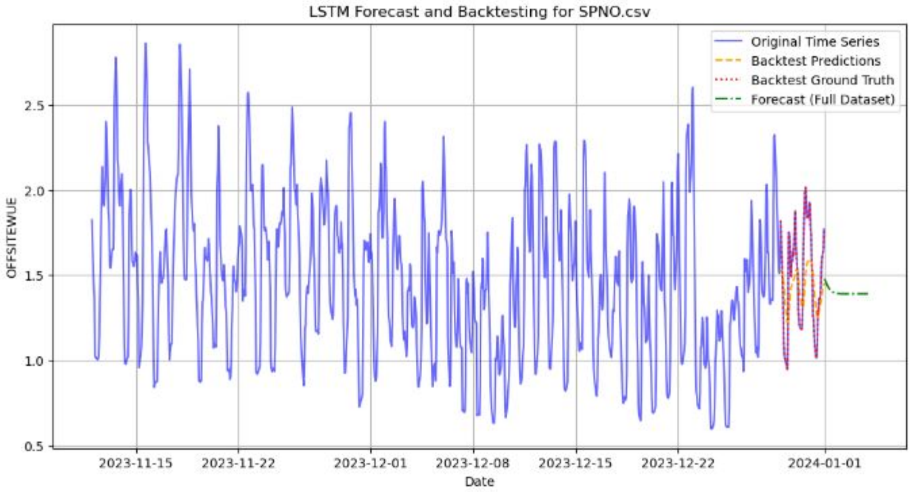


Comparing model performance metrics for model types across all 19 eGRID regions (57 models total)

LSTM Evaluation (Off-site)

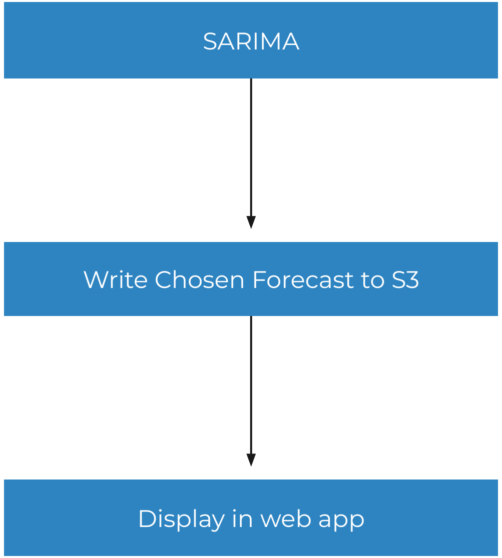


Performing well



Flatlining

Model Selection (Off-site)



Similar/better performance on RMSE, MAE, MAPE metrics

Ability to expand to multivariate time series forecasting

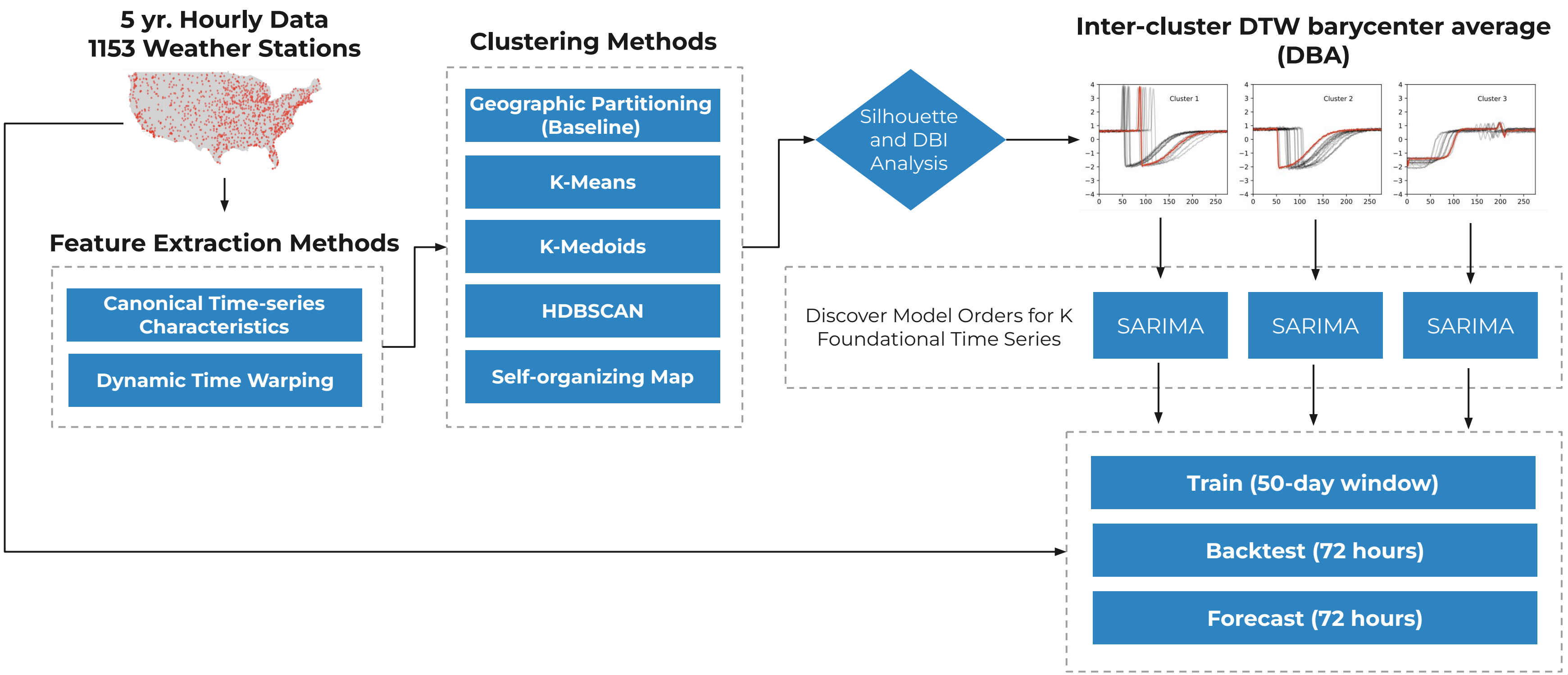
Simpler and more reliable model can provide interpretability to end users

Fast to compute if we expand to live forecasting

On-Site WUE Forecasting

Images Credits: Aerco Systems and Meta

On-site WUE Forecasting Methods



Simplifying with Time Series Clustering

Scalable Forecasts

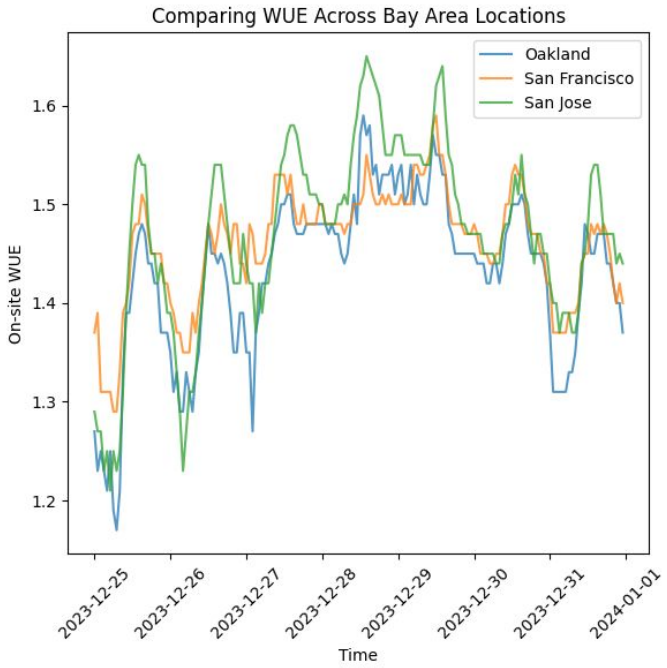
- Training and fine-tuning individual models for each of the **1000+ 4 year hourly locations doesn't scale**

Reduced Redundancy

- Clustering **similar patterns** and characteristics to **derive foundational time series**
- Eliminates the need for individual models

Interpretability

- Geographically visualize locations with similar WUE profiles

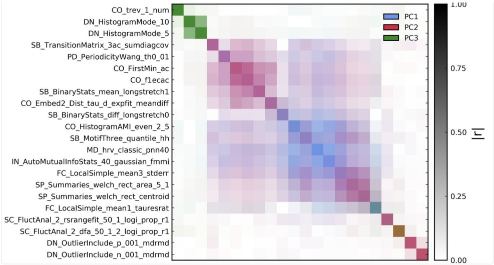


Operationalizing Clusters

Feature-Based Clustering

Uses statistical and mathematical properties of time series

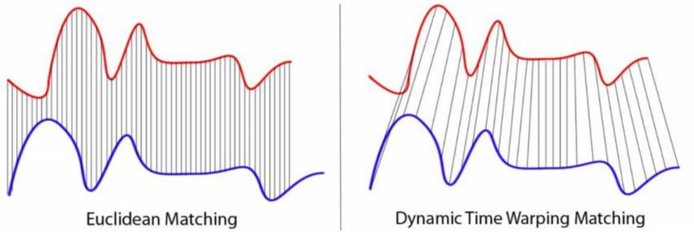
- **Catch 22** (22 canonical statistical features capturing key time-series behaviors)



Shape-Based Clustering

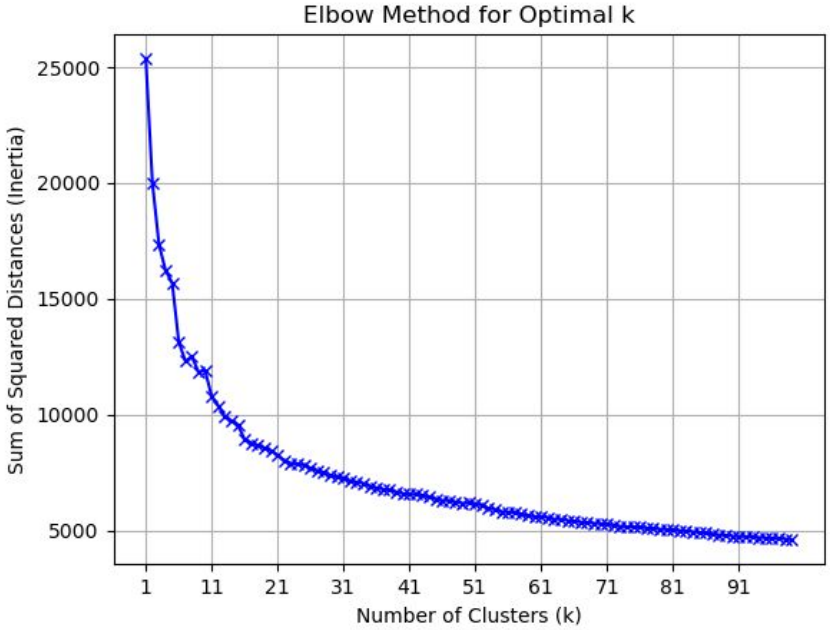
Directly compares the shapes of time series using pairwise similarity measures

- **Dynamic Time Warping** (time adjusted Euclidean distance)

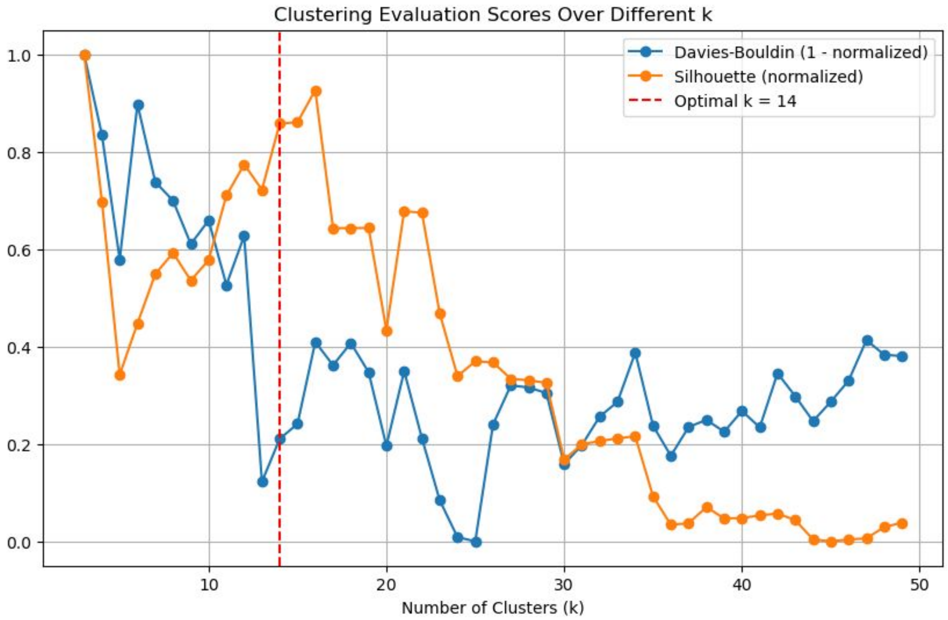


Cluster **Optimization**

*Catch 22 K-Means



The "elbow point" where the improvement flattens is considered the optimal number of clusters



The optimal k is where the silhouette score is highest, indicating well-defined and well-separated clusters **AND** where the davies-bouldin index is lowest indicating compact and well-separated clusters

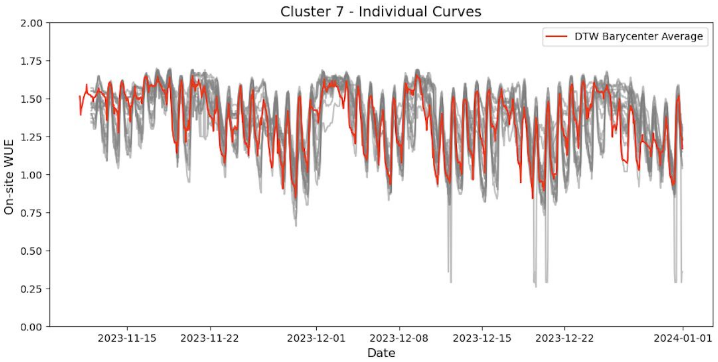
Time Series Clustering Candidates

Model	DTW Features			Catch22 Features		
	Optimal Num. Clusters	Avg. Silhouette	Avg. DBI	Optimal Num. Clusters	Avg. Silhouette	Avg. DBI
Geographic Partitioning	9	0.054	2.620	9	-0.037	3.975
K-Means	**9**	**0.236**	**1.500**	**14**	**0.462**	**0.862**
K-Medoids	**10**	**0.226**	**1.988**	4	0.114	2.760
HDBSCAN	7	-0.146	2.157	3	0.350	2.091
SOM	7	0.104	1.110	27	0.094	1.943

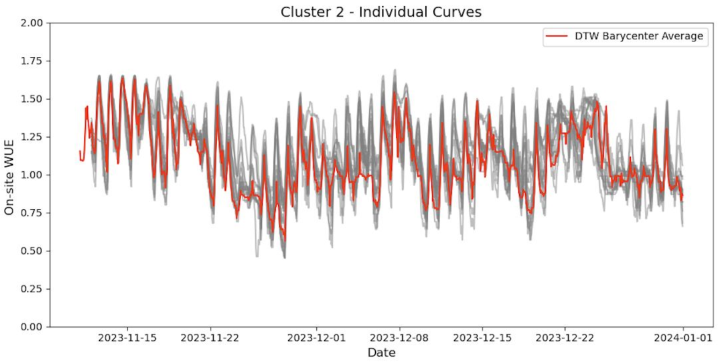
Deriving Model Orders from DBA

Dynamic Time Warping Barycenter Averaging (DBA) is a time series analysis method that uses Dynamic Time Warping (DTW) to create representative sequences for data categories.

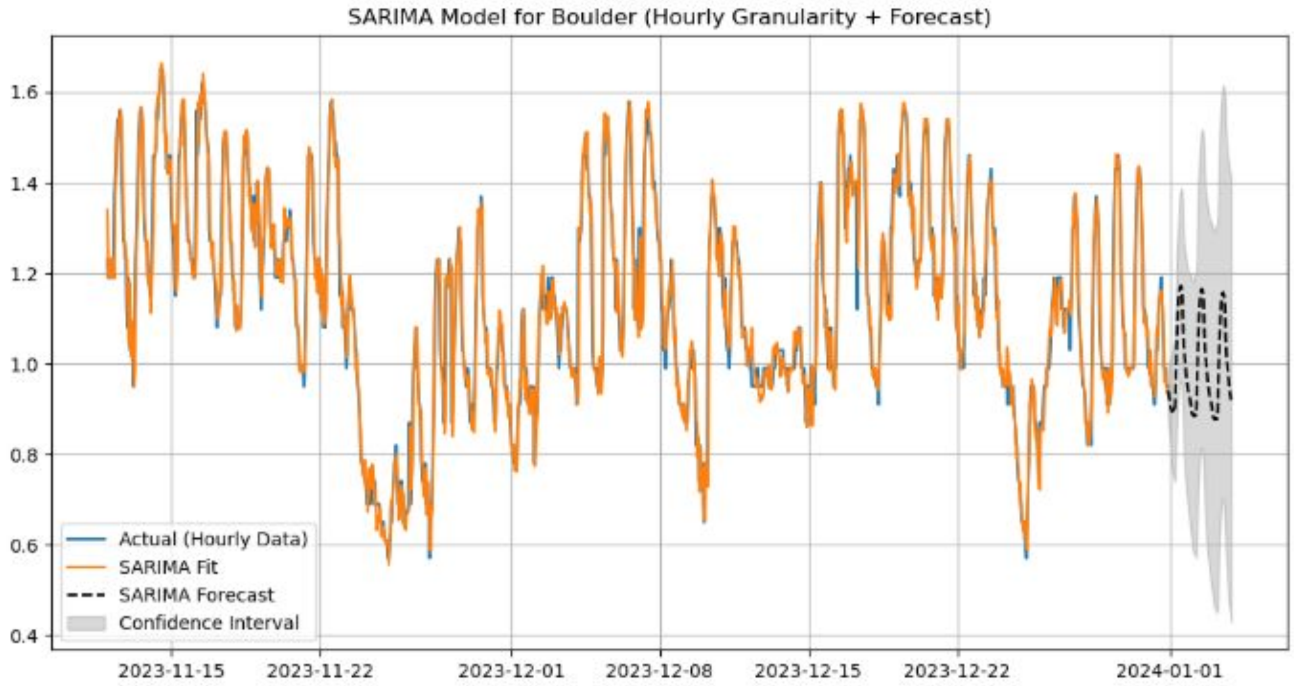
ARIMA (2, 1, 2), (1, 0, 2, 24)



ARIMA (2, 1, 3), (2, 0, 2, 24)



Model **Evaluation (On-site)**



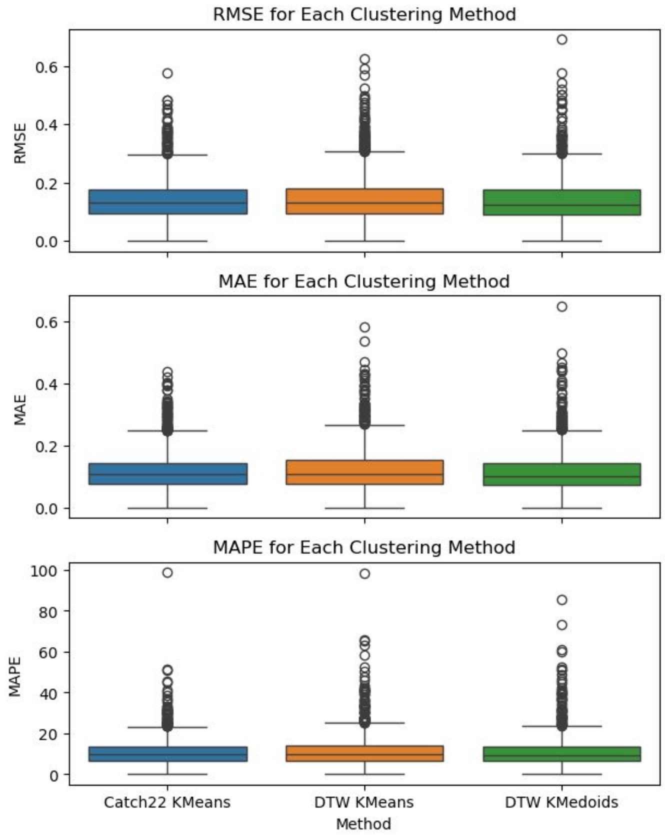
Model Evaluation for Boulder (Hourly Granularity):

- MAE: 0.0299
- RMSE: 0.0375
- MAPE: 2.75%

Forecast Performance by Clustering Method

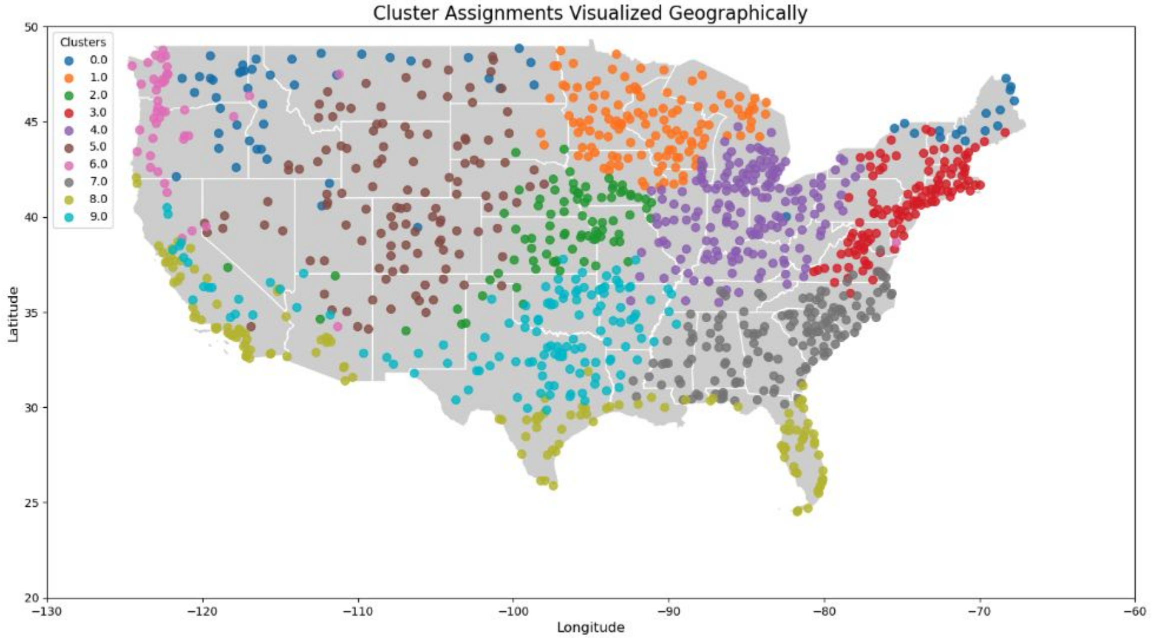
Metrics Shown for Mean Absolute Percent Error

	Catch22 K-Means	DTW K-Means	DTW K-Medoids
Mean	**10.99**	11.66	11.14
Standard Deviation	**6.92**	8.21	7.89
25%	6.82	6.72	**6.54**
50%	9.70	9.86	**9.30**
75%	13.47	14.17	**13.47**



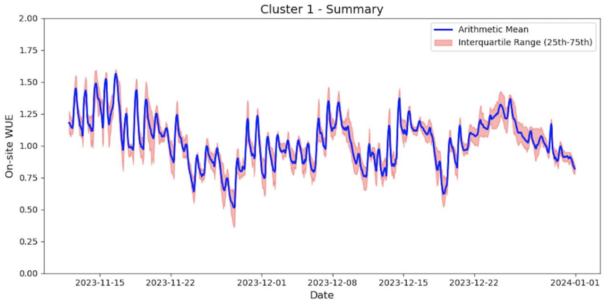
Time Series Clustering Results

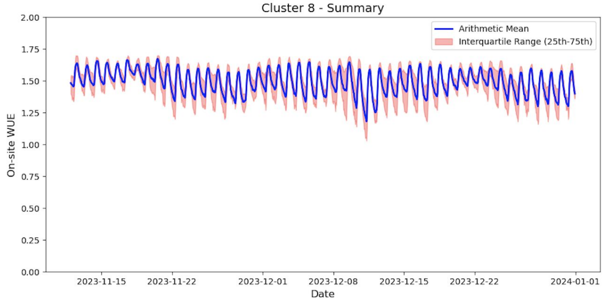
Geographic Distribution of Time Series Clusters



*Shown are example clusters from DTW K-Medoids

Cluster Examples





Key Discoveries

Off-site WUE Forecasting

Though LSTM and TimesFM are impressive, SARIMA models are dependable and deterministic

On-site WUE Forecasting

We use feature-based clustering to create **10** representative cluster-level SARIMA models and use these to forecast city-level WUE data for **1153** locations

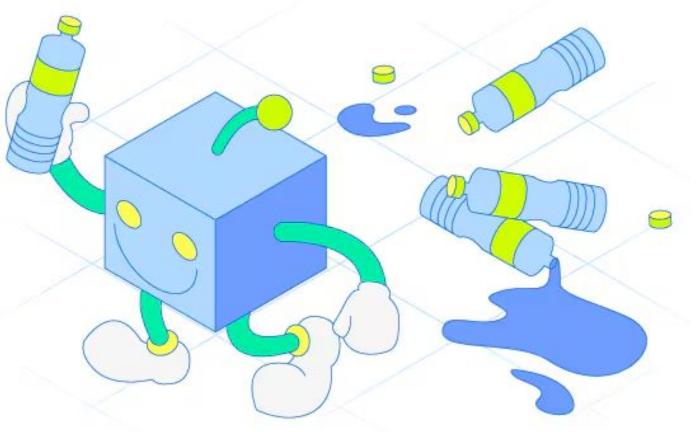


Image Credits: The Washington Post

Ethical Considerations

Greenwashing

"large companies often present their sustainability practices in a positive light [...] to **give the appearance of greater sustainability**" (McCauley).

Overconsumption

Unintentionally incentivizing data center construction, **root problem of resource overconsumption** in the tech world.

Local Impact

Awareness of water scarcity and water consumption by data centers (secretive industry). Information for communities to negotiate or fight back.

Images Credits: 100DaysofRealFood.com, ITU News, Associated Press

Future Direction

Real-time Inference

Consider live forecasting by streaming data from our data sources, or batch forecasting by periodically updating our data and models..

Tailored Forecasts

Refine on-site WUE simulation to **accommodate varying operational scenarios**. Bolster off–site WUE methods by **investigating additional data sources** and multivariate forecasts.

Outreach

Leverage insights to **increase public awareness** about the impact that data centers have on communities' water supplies.

HydroScale

Facilitating Water Stewardship in a Digital Future

Marlon Fu - marlonfu@berkeley.edu
Derek Yao - derekyao672@berkeley.edu
Nora Povejsil - npovejsil@berkeley.edu
Suhas Prasad - pseuhas@berkeley.edu
Austin Ho - aicho@berkeley.edu

Appendix

Acknowledgements

We extend our sincerest gratitude to Shaolei Ren, Pengfei Li, and their team at the University of California—Riverside, whose input, methodology, and data we used to design our own processes.

Additionally, we express our appreciation for various industry experts, including Ivy Chan and her team at Equinix, as well as Jim Gao at Phaidra. Their domain insights and direct feedback were instrumental in deepening our understanding of the data center space.

We also deeply thank our capstone instructors, Joyce Shen and Morgan Ames, for their continued guidance and mentorship throughout this project.

AWS Infra Specs

- EC2
    - Instance: c5.12xlarge (x10)
- Elastic MapReduce (EMR Studio)
    - Number of Spark drivers: 1
    - Size of driver: 4 vCPUs, 16 GB memory
    - Driver disk details: Standard, 20 GB disk
    - Number of Spark executors: 5
    - Size of executor: 4 vCPUs, 8 GB memory
    - Executor disk details: Standard, 20 GB disk
- Sagemaker (Shared Workspace)
    - Instance: ml.m5.2xlarge
    - Image: SageMaker Distribution 2.1.0
    - Storage (GB): 12

References

Beveridge, Nathaniel R., "Deep Learning for Weather Clustering and Forecasting" (2021). Theses and Dissertations. 5082. https://scholar.afit.edu/etd/5082

Eure, J. (2024, April 24). *The world's AI generators: Rethinking water usage in data centers to build a more sustainable future*. Lenovo StoryHub. https://news.lenovo.com/data-centers-worlds-ai-generators-water-usage/

Privette, A. P. (2024, October 11). *Ai's challenging Waters*. Grainger Engineering Office of Marketing and Communications. https://cee.illinois.edu/news/AIs-Challenging-Waters

Gupta, P. S., Hossen, M. R., Li, P., Ren, S., & Islam, M. A. (2024). A dataset for research on Water Sustainability. *The 15th ACM International Conference on Future and Sustainable Energy Systems*, 442–446. https://doi.org/10.1145/3632775.3661962

Li, P., Yang, J., Islam, M. A., & Ren, S. (2023). Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2304.03271

References

Lubba, C. H., Sethi, S. S., Knaute, P., Schultz, S. R., Fulcher, B. D., & Jones, N. S. (2019). catch22: CAnonical Time-series CHaracteristics. Data Mining and Knowledge Discovery, 33(6), 1821-1852.

Peer, R. A., Grubert, E., & Sanders, K. T. (2019). A regional assessment of the water embedded in the US electricity system. Environmental Research Letters, 14(8), 084014. https://doi.org/10.1088/1748-9326/ab2daa

Srivathsan, B., Sorel, M., & Sachdeva, P. (2024, October 29). *AI power: Expanding Data Center capacity to meet growing demand*. McKinsey & Company. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/ai-power-expanding-data-center-capacity-to-meet-growing-demand

Tingström, C., & Åkerblom Svensson, J. (2023). Forecasting With Feature-Based Time Series Clustering (Dissertation). Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:hj:diva-61877