WATER QUALITY

Technology Conference

November 5–9, 2023
Dallas, Texas

Multivariate Fault Detection for Water Reuse: An Ultrafiltration Case Study

(WED12-02)

Taylor Grimm

8 November 2023, 2:00pm CDT

Baylor University

Taylor Grimm
Baylor University

Amos Branch
Carollo Engineers

Kyle Thompson
Carollo Engineers

Andy Salveson
Carollo Engineers

Kate Newhart
U.S. Military Academy

Mandy Hering
Baylor University

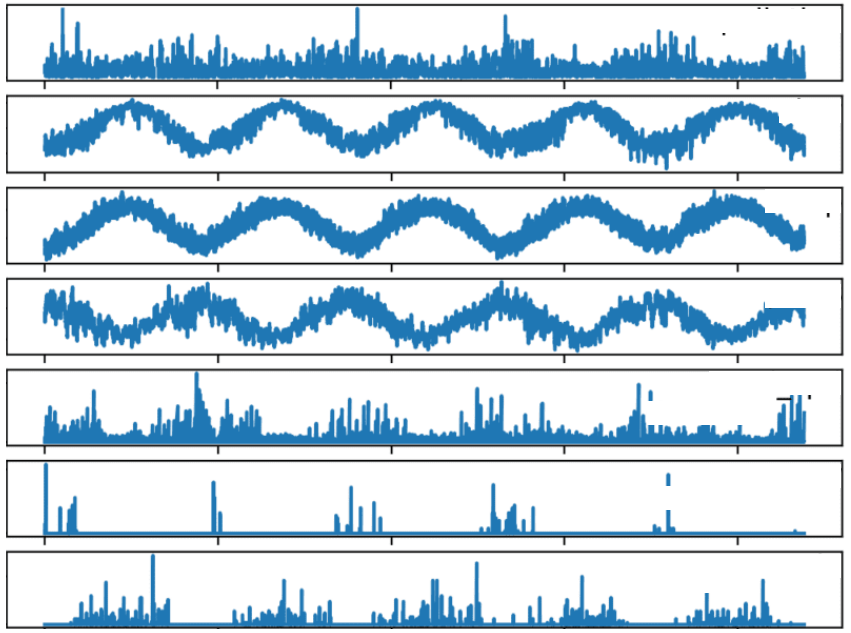
PURE WATER PROJECT
LAS VIRGENES-TRIUNFO
Bringing Our Water Full Circle

Acknowledgement

This work is supported by the National Alliance for Water Innovation (NAWI), funded by the US Department of Energy (DOE), Energy Efficiency and Renewable Energy Office, Advanced Manufacturing Office under Funding Opportunity Announcement DE-FOA-0001905, Project 5.17 Data-Driven Fault Detection and Process Control for Potable Reuse with Reverse Osmosis.

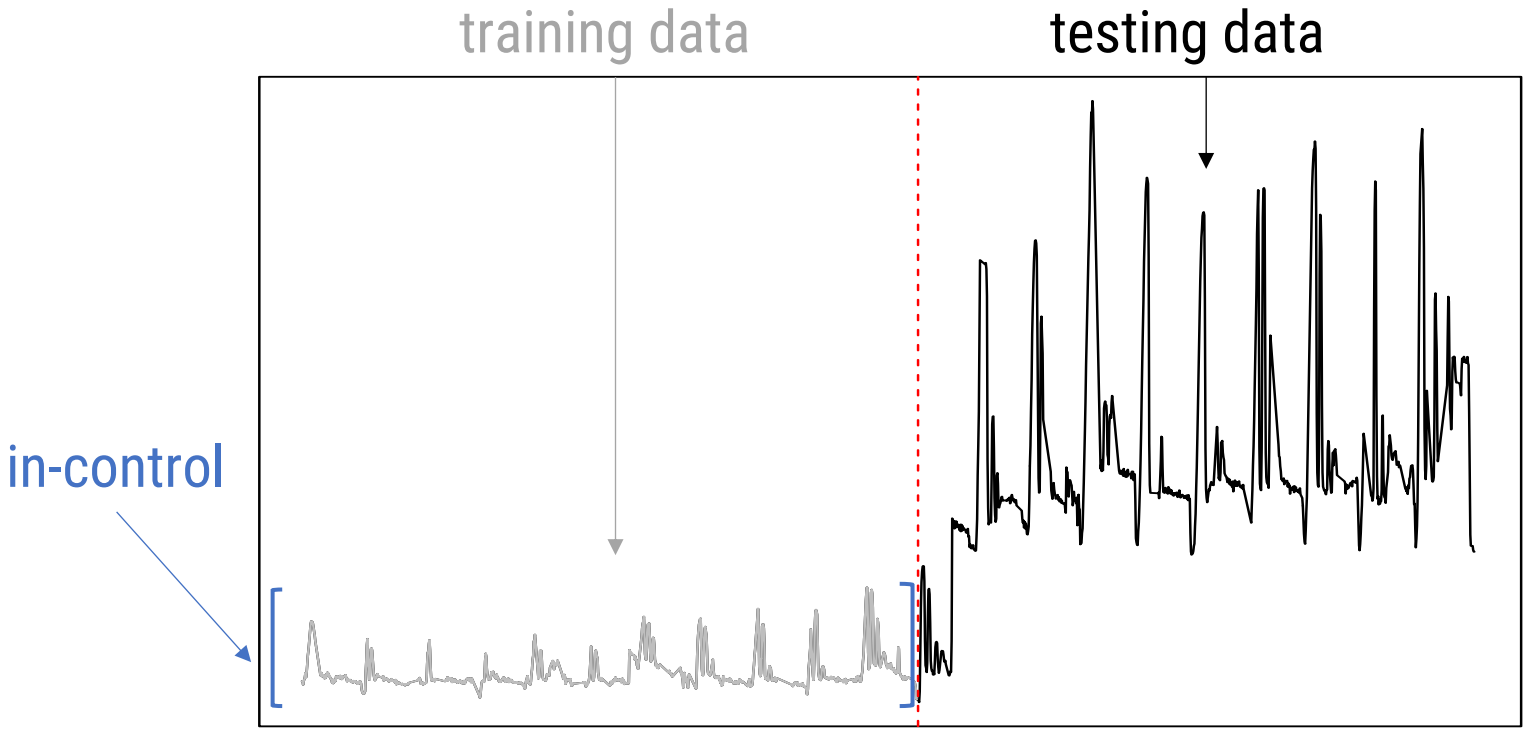
Fault detection

- Important to avoid unnecessary downtime and compromised effluent quality.
- Manually detecting faults is impractical and ineffective, especially in processes with many variables.
- An automated data-driven approach is required for practical application.



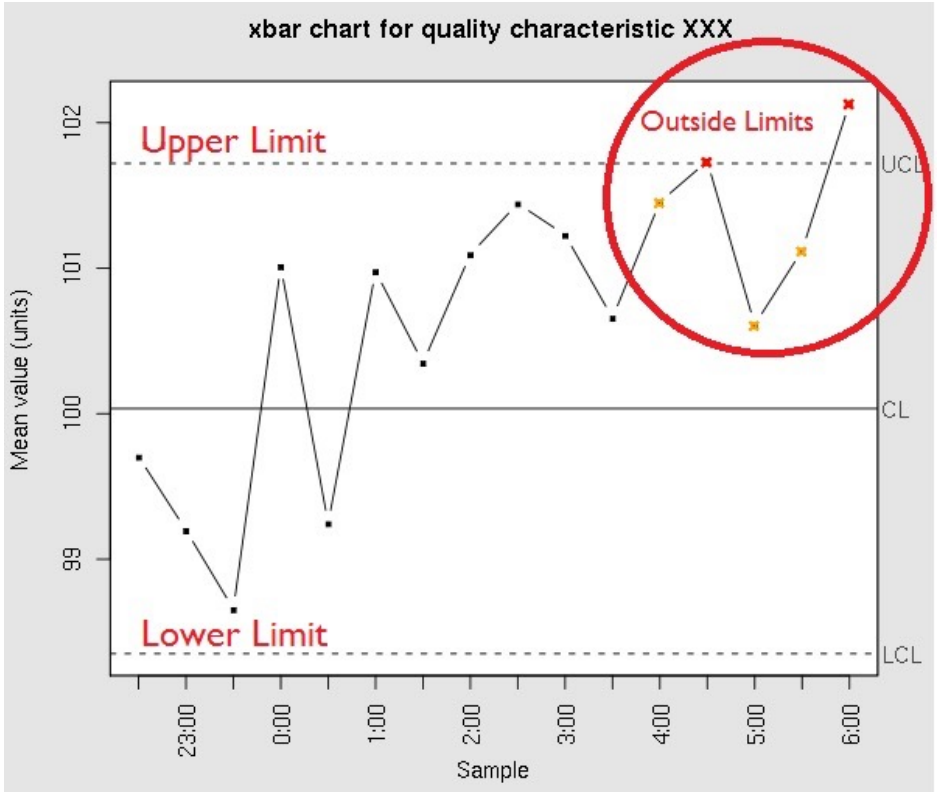
Statistical process monitoring (SPM)

- data-driven approach for real-time fault detection
- compares new *testing* observations to fault-free observations during a previously observed *training* period



Shewhart Control Chart

- Upper and lower limits are set for each process variable.
- An observation is flagged as a fault if it exceeds a control limit.
- Control limits are often set at 3 standard deviations above/below the in-control mean.



(https://www.statisticshowto.com/statistical-process-control/)

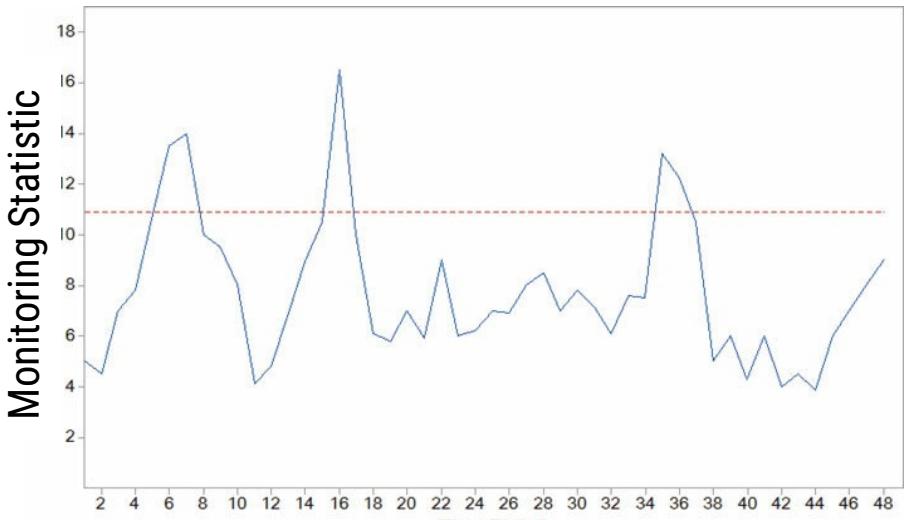
Multivariate SPM (MSPM)

MSPM allows us to

- monitor many variables simultaneously
- detect changes in means and/or covariances
- condense information from many variables into one or two monitoring statistics

Process variables are put into two categories:

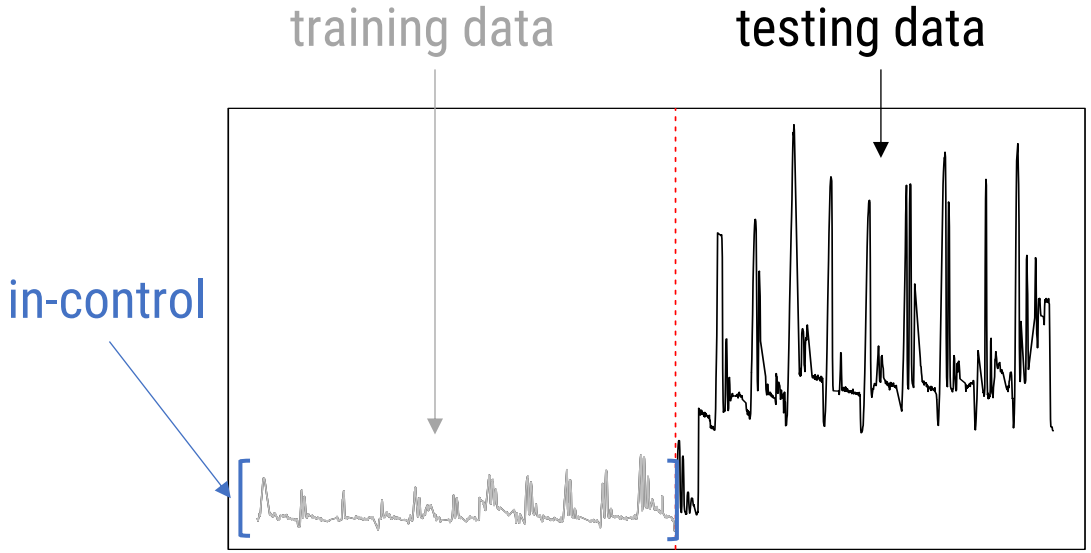
- Monitoring variables: variables we are interested in detecting changes
- Explanatory variables: variables that change and affect monitoring variables

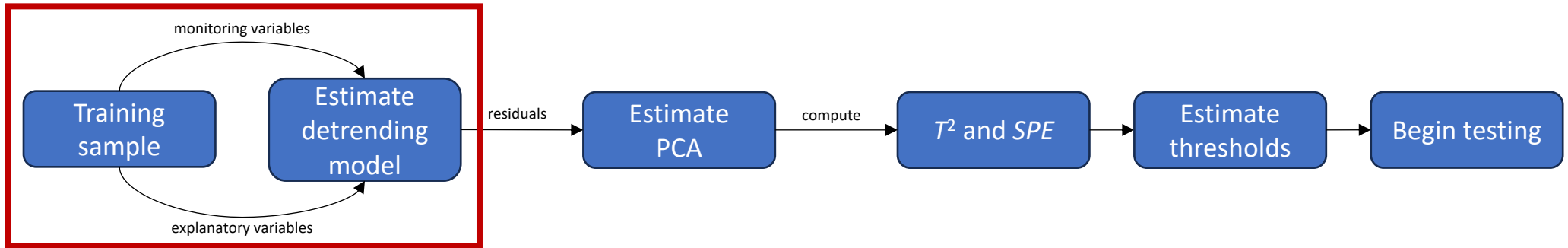


Multivariate SPM (MSPM)

Steps:

1. Determine an in-control training period.
2. Detrend the monitoring variables.
   - Methods to detrend that we tested:
     - No model fit (raw observed data)
     - Adaptive lasso
     - k-nearest neighbors
     - Random forest
     - Extreme gradient boosting
   - Fit a model to remove the expected variability in each monitoring variable
     - Data → model → fitted values
     - Residuals = observed values — fitted values

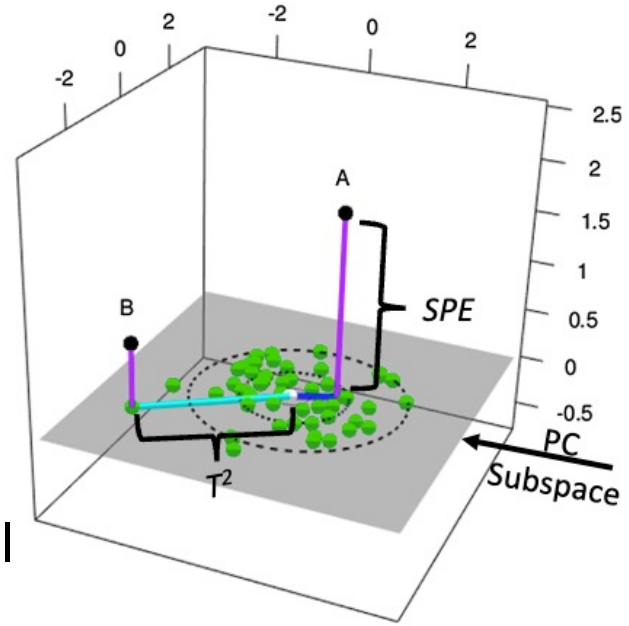




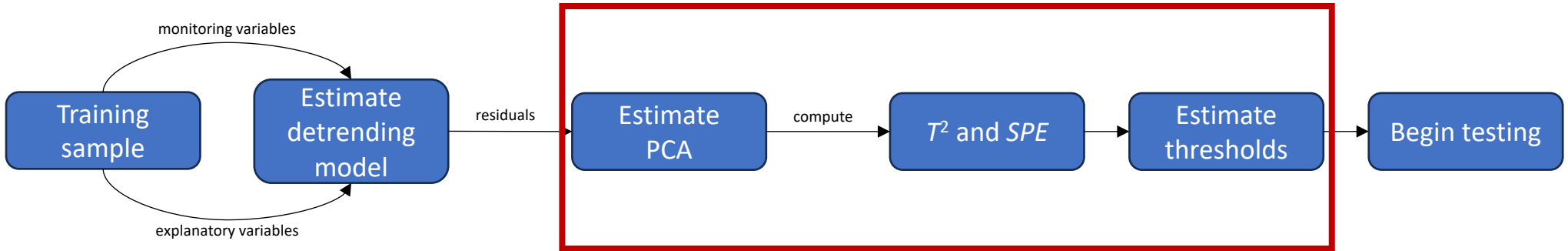
Multivariate SPM (MSPM)

Steps:

3. Apply principal component analysis (PCA) to the residuals from detrending.
    - Method to reduce the dimension of a dataset
4. Compute monitoring statistics.
    - T^2: measures variability within the PCA model
    - Squared prediction error (*SPE*): measures variability outside of the PCA model
5. Determine thresholds for T^2 and *SPE* from the training data
    - Use an upper quantile of computed T^2 and *SPE* values as the threshold to classify future observations as normal or abnormal (fault)

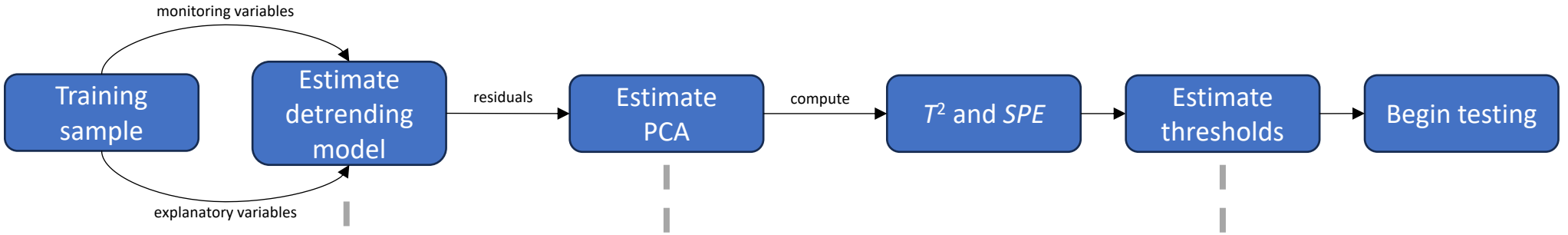


A set of projected observations and two original observations (**A**, **B**) (Kazor et al., 2016).

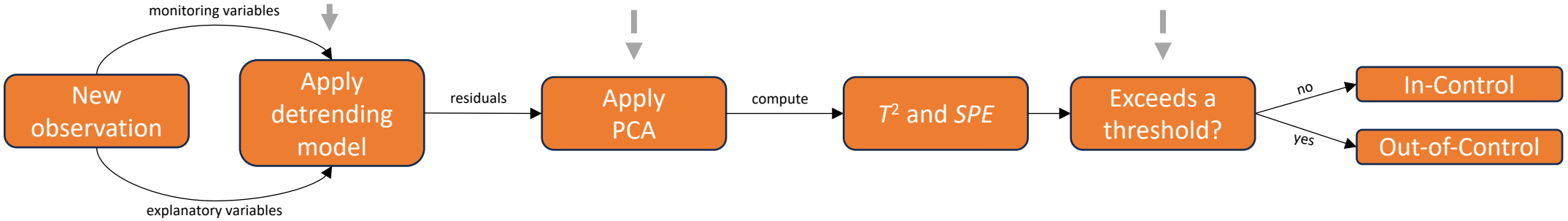


Multivariate SPM (MSPM) Process

Training

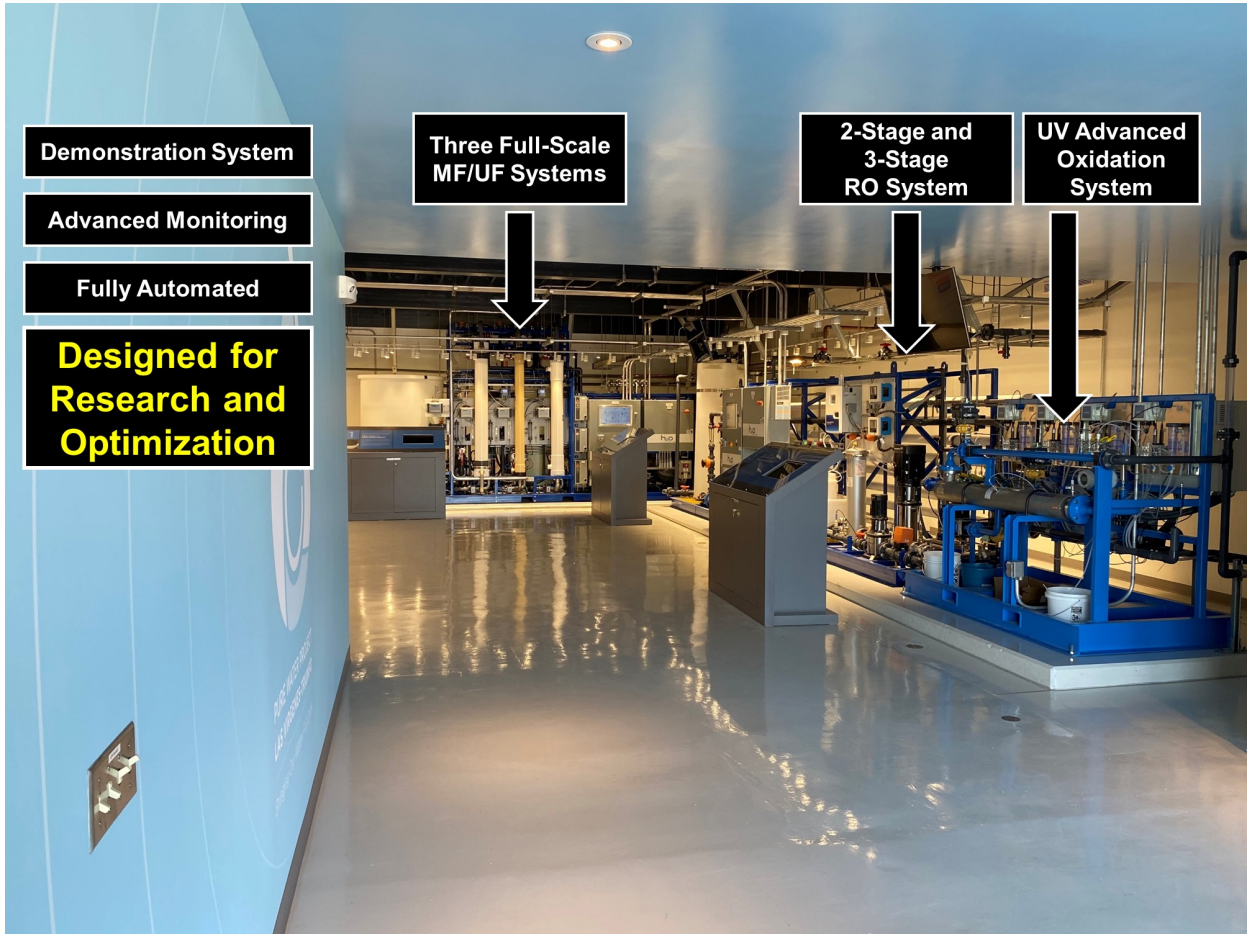


Testing



Case Study

- Pure Water Demo Facility
- Constructed by the Las Virgenes Municipal Water District – Triunfo Water & Sanitation District Joint Powers Authority
- UF + RO + UV/AOP System
- 100 gpm
- Operational since 2020
- Calabasas, CA



Case Study Data

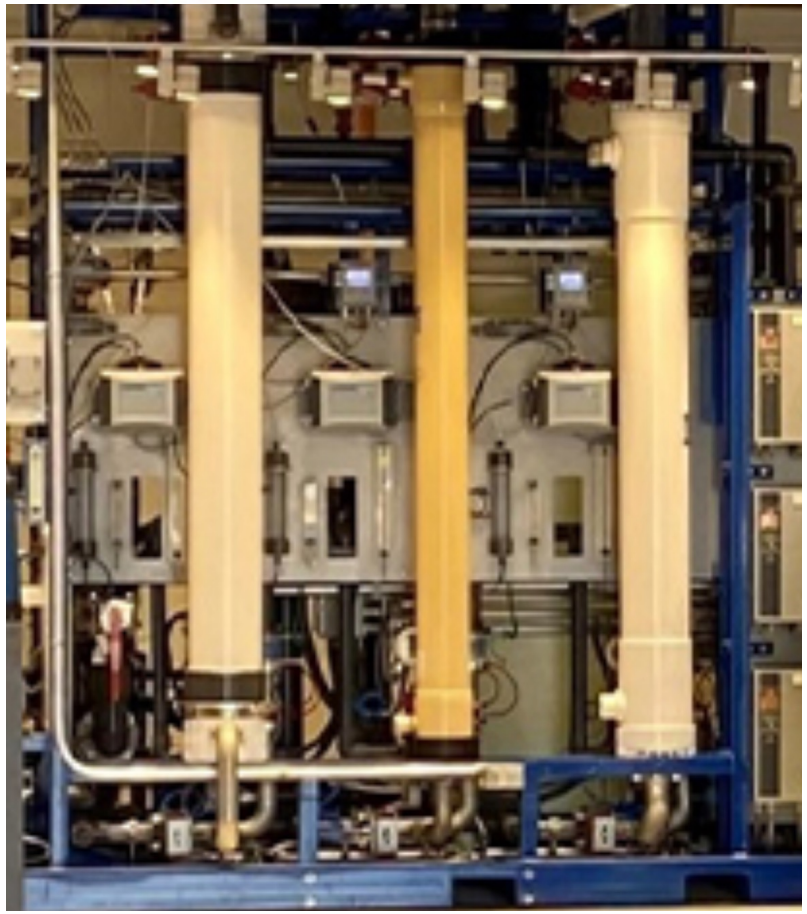
PURE WATER PROJECT
LAS VIRGENES-TRIUNFO
Bringing Our Water Full Circle

Two IC training periods:

1. April 2, 2021 - April 13, 2021
2. April 27, 2022 - May 11, 2022

Training periods were selected by finding periods of time that are visually

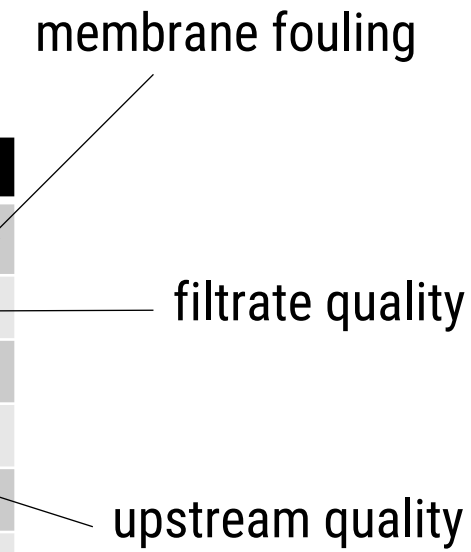
- stable for several days (no obvious faults)
- shortly followed by an obvious change in at least one monitoring variable



Case Study Data

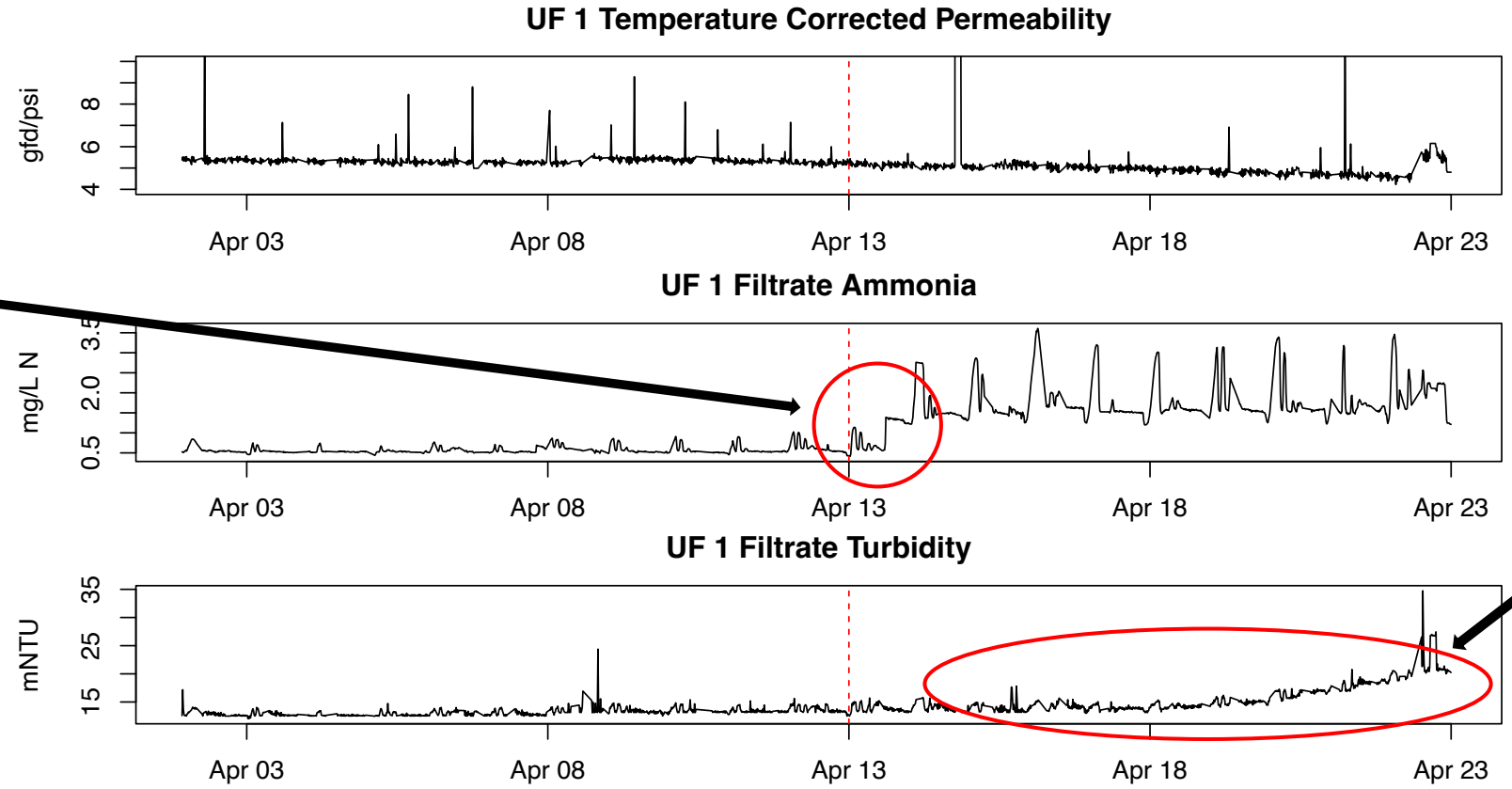
15 minute averages of process variables:

Explanatory Variables	Monitoring Variables
UF Filtrate pH	UF (1, 2, 3) Temperature-Corrected Permeability
UF Filtrate ORP	UF (1, 2, 3) Filtrate Turbidity
UF Filtrate Total Chlorine	UF Filtrate Ammonia
UF Feed Turbidity	
UF Backwash Flow	
UF Feed Temperature	
UF (1, 2, 3) TMP	
UF (1, 2, 3) Flux	
UF (1, 2, 3) Temperature-Corrected TMP	
UF (1, 2, 3) Permeability	
RO Feed Conductivity	
RO Feed TOC	



Short-term Case Study #1

Training: April 2nd, 2021 – April 13th, 2021



Upward shift

Slow drift

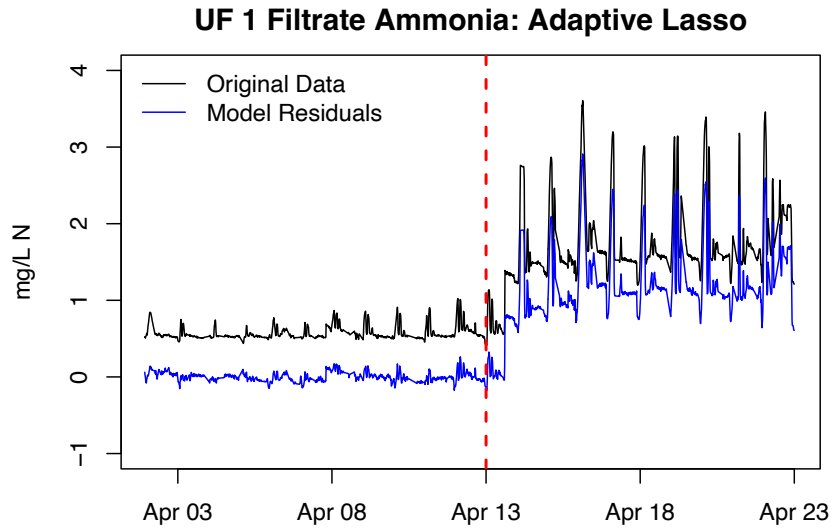
Detrending Plots

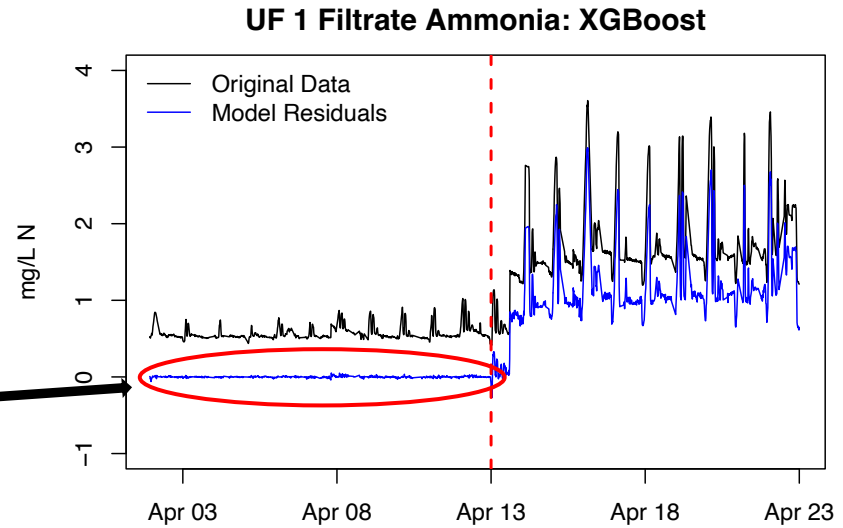
The red line separates training (left) from testing (right).

- No model predicts the upward shift.
- RF and XGBoost may be overfit.

Overfitting: model fits well to training data, predicts poorly on testing data.

- Often a problem with complex machine learning models.

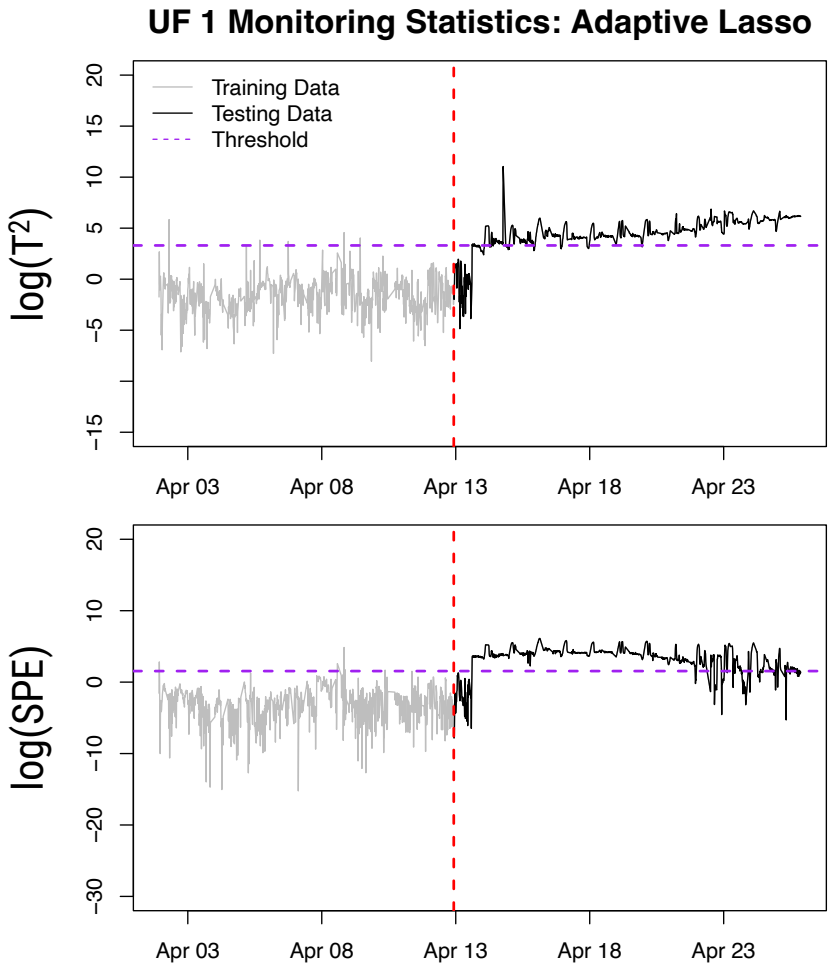




Overfit?

Monitoring Statistic Plots

- All models are able to detect a change in the process shortly after testing begins.
- Results from RF and XGBoost provide stronger evidence of a fault but could be due to overfitting.



MSPM with PCA

To avoid false alarms, adjustments to the PCA model are sometimes required to account for dependence in the data or natural process changes over time.

- **Dynamic PCA:** incorporates lags of process variables in the model to account for dependence in the observations.
- **Adaptive PCA:** updates the model regularly with in-control (IC) observations to adapt to natural long-term process changes.
- **Adaptive-Dynamic PCA (AD-PCA):** uses both modifications.

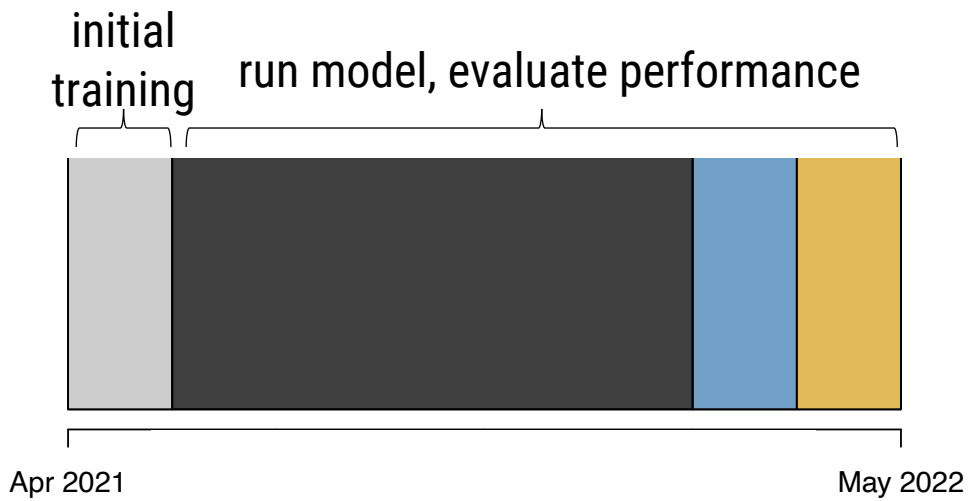
Long-term Case Study

Studies usually evaluate methods on short-term case studies.

- Satisfactory short-term performance does not guarantee adequate long-term performance.

We assess long-term monitoring performance through an experiment:

- Train the model on the first in-control training period (April 2-13 2021)
- Evaluate the performance on a:
  - **long, unspecified period** (April 14 2021 – April 26 2022)
  - known in-control period (April 27 2022 – May 11 2022)
  - known out-of-control period (May 12 2021 – May 25 2022)



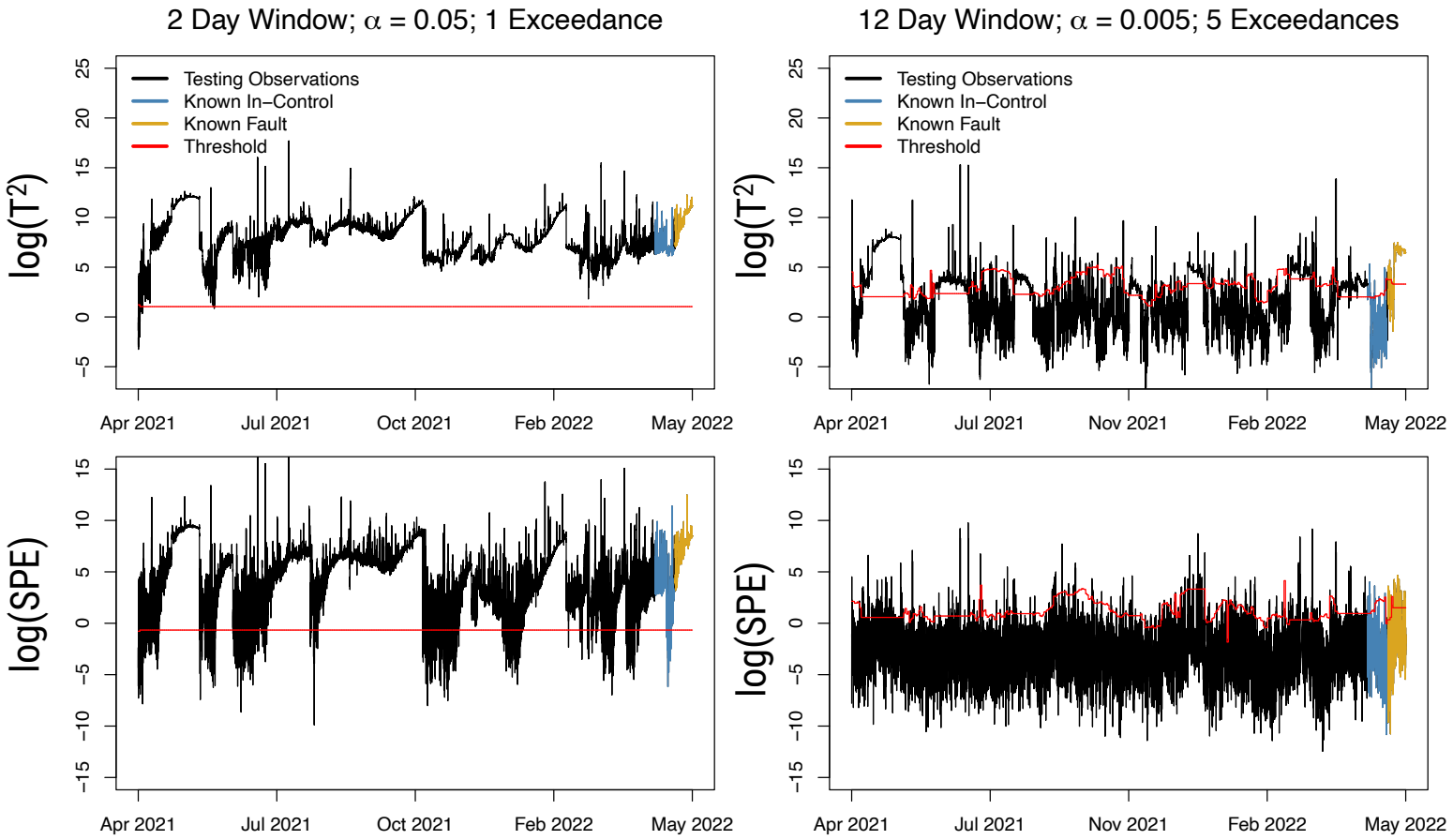
We compare

- MSPM with AD-PCA and adaptive lasso detrending
- Adaptive univariate control charts

MSPM with AD-PCA

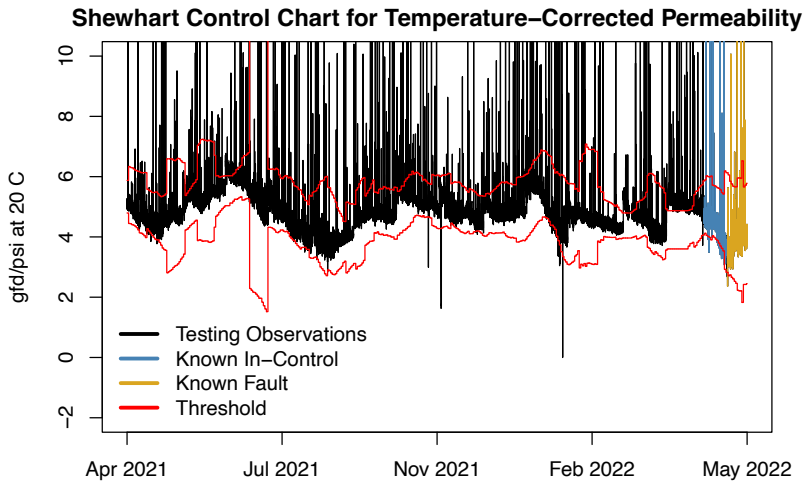
- The following are investigated:
  - Window size for retraining
  - Number of exceedances prior to issuing an alarm
  - Level of the threshold to issue an alarm

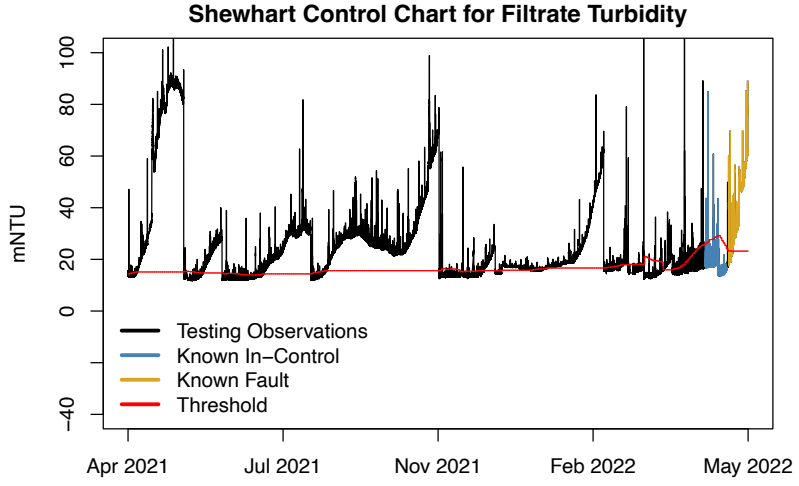
- Adjustments are required to achieve satisfactory long-term fault detection performance.

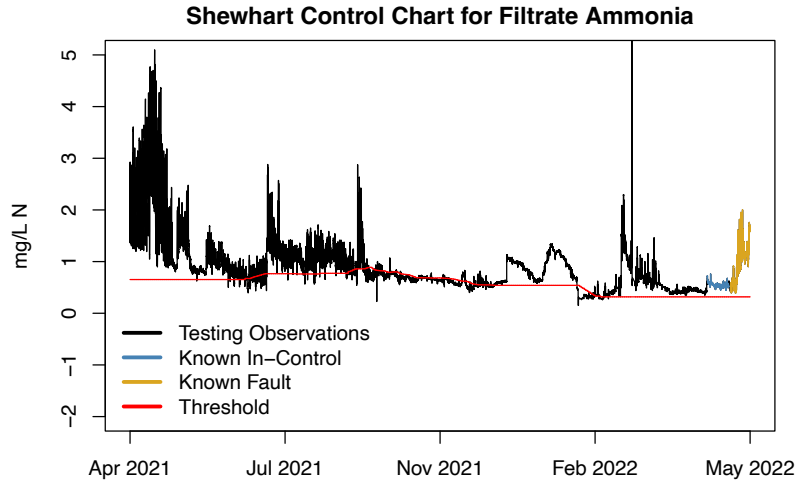


Univariate Shewhart Control Charts

- These charts are more complex than what is commonly used in practice.
- Charts for turbidity and ammonia rarely update.
- Large periods of time are flagged for turbidity and ammonia.







Long-term Case Study

Percent of observations labeled as OC for a 12-day training window when α = 0.005, and 5 exceedances in a row are required to trigger an alarm.

	Testing (n = 36014)	Known IC (n = 1419)	Known Fault (n =1238)
Shewhart			
Temp. Corrected Permeability	9.8%	5.2%	1.5%
Filtrate Turbidity	64.0%	1.6%	88.5%
Filtrate Ammonia	74.8%	100.0%	100.0%
AD-PCA			
T^2	36%	16.6%	75.8%
SPE	2.1%	1.8%	9.9%

Conclusion

- Adjustments must be made for long-term application of fault detection methods.
- Increasing the number of exceedances before raising an alarm results in delayed detection of faults.
  - Not increasing this number causes too many alarms, and the model fails to update.
- Multivariate methods become increasingly useful as the number of monitoring variables increases.
- AD-PCA performs as expected but requires some tuning for long-term monitoring.
- Univariate Shewhart charts are sensitive to outliers and perform poorly for some variables.

References

Hotelling, H. (1947) “Multivariate quality control" In Eisenhart, C., Hastay, M. W., and Wallis, W., editors, Techniques of Statistical Analysis, pages 111–184. McGraw-Hill, New York.

Jackson, J. E. and Mudholkar, G. S. (1979) “Control procedures for residuals associated with principal component analysis," Technometrics, 21(3):341–349.

Kazor, K., Holloway, R., Cath, T., and Hering, A. S. (2016) “Comparison of linear and nonlinear dimension reduction techniques for automated process monitoring of a decentralized wastewater treatment facility,” Stochastic Environmental Research and Risk Assessment, 30: 1527-1544.

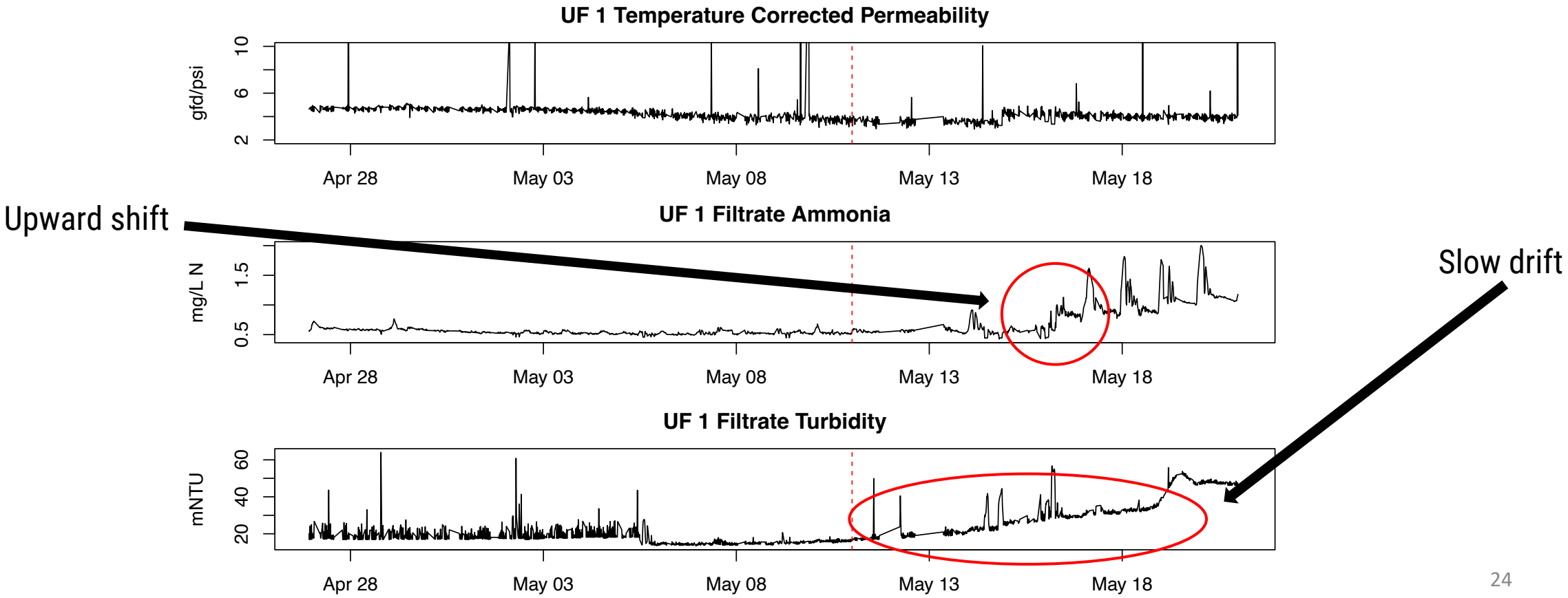
Klanderman, M., Newhart, K.B., Cath. T.Y., Hering, A.S. (2020) “Fault isolation for a complex decentralized wastewater treatment facility,” Journal of the Royal Statistical Society, Series C., 69, 931-951.

Klanderman, M. C., Newhart, K. B., Cath, T. Y., and Hering, A. S. (2020) “Case studies in real-time fault isolation in a decentralized wastewater treatment facility,” Journal of Water Process Engineering, 38: 101556.

Ku, W., Storer, R. H., and Georgakis, C. (1995). “Disturbance detection and isolation by dynamic principal component analysis,” Chemometrics and Intelligent Laboratory Systems, 30(1):179–196.

Short-term Case Study #2

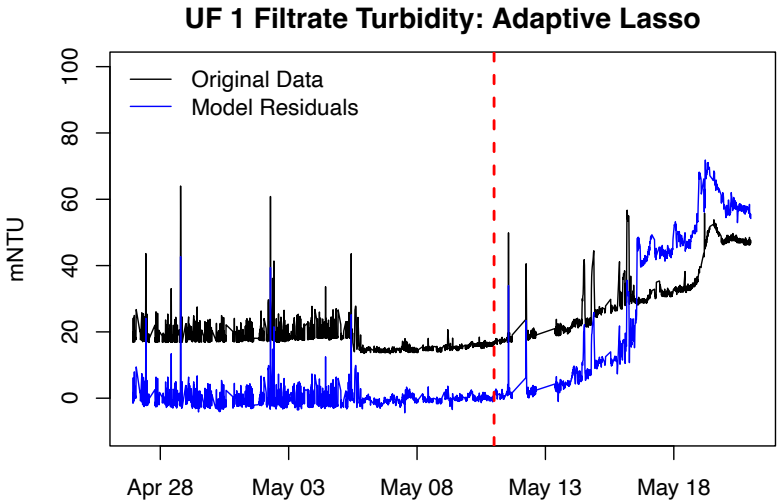
Training: April 27, 2022 - May 11, 2022

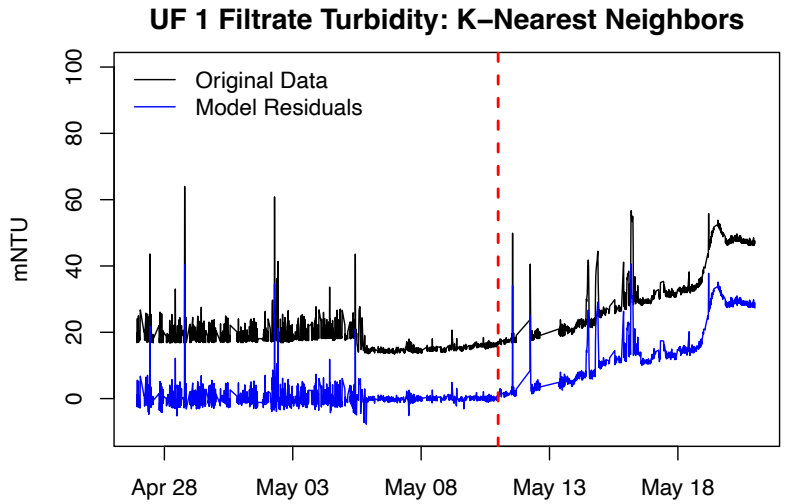


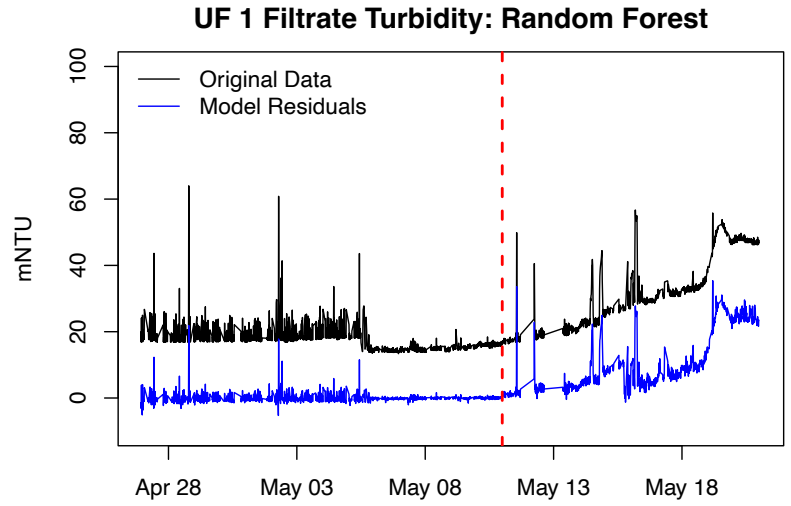
Detrending Plots

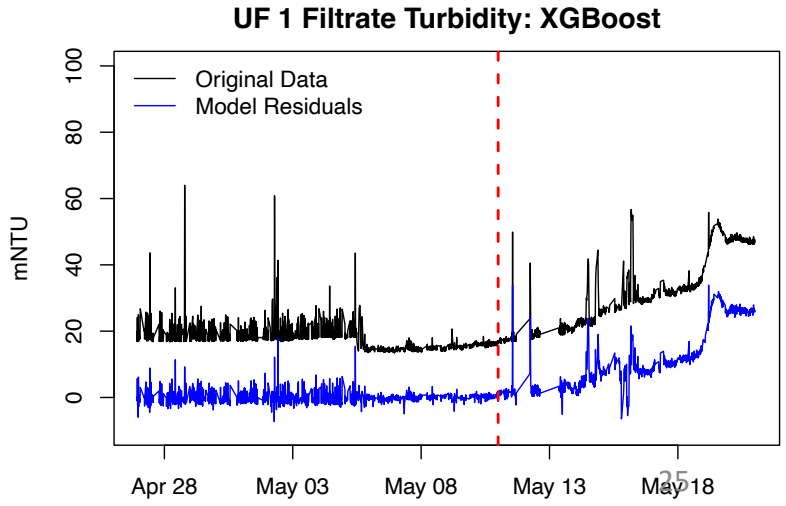
The red line separates training (left) from testing (right).

- No models predict the upward drift.
- Adaptive lasso begins to severely underpredict before May 18th.



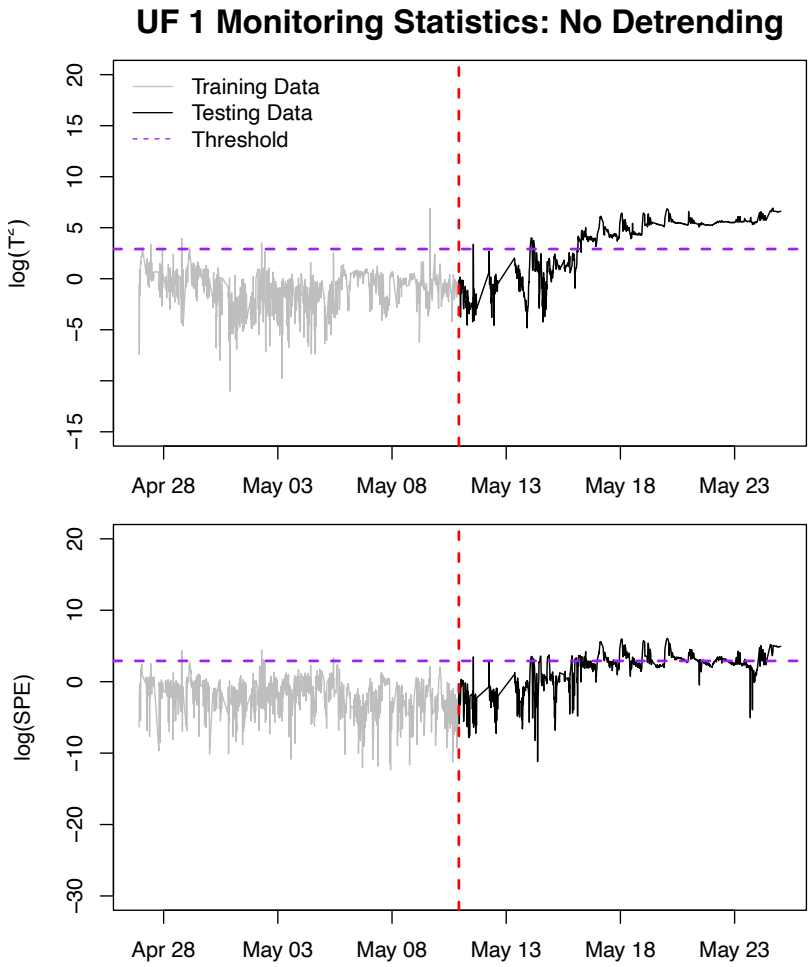


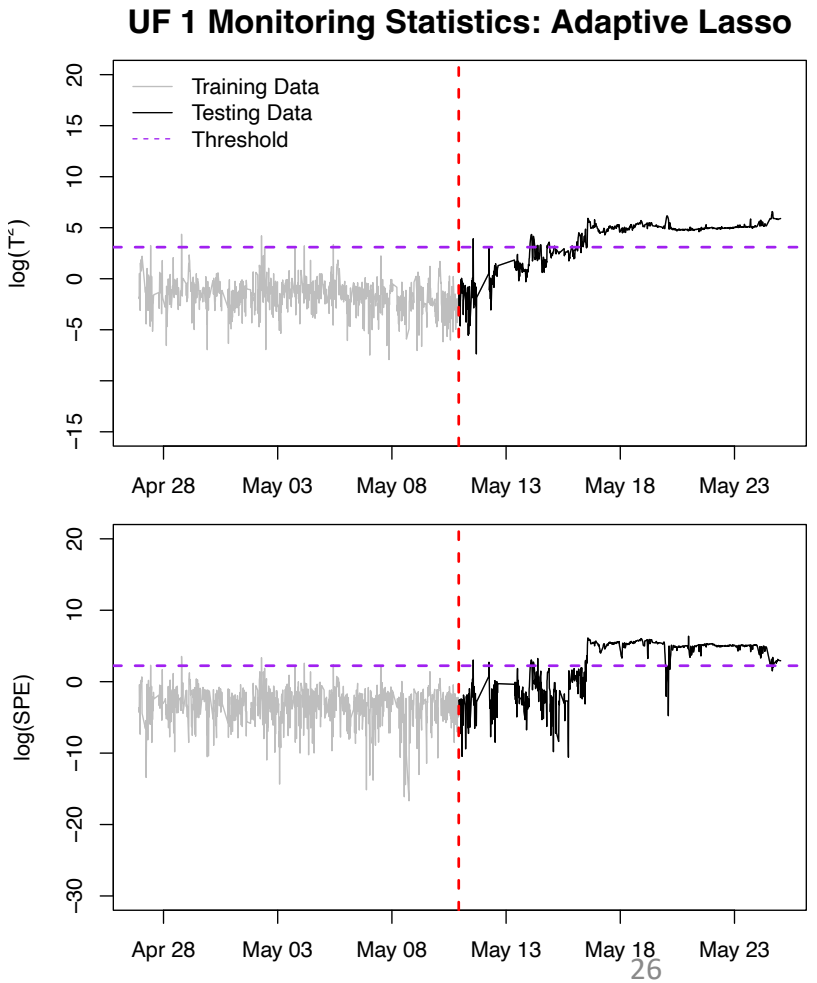




Monitoring Statistic Plots

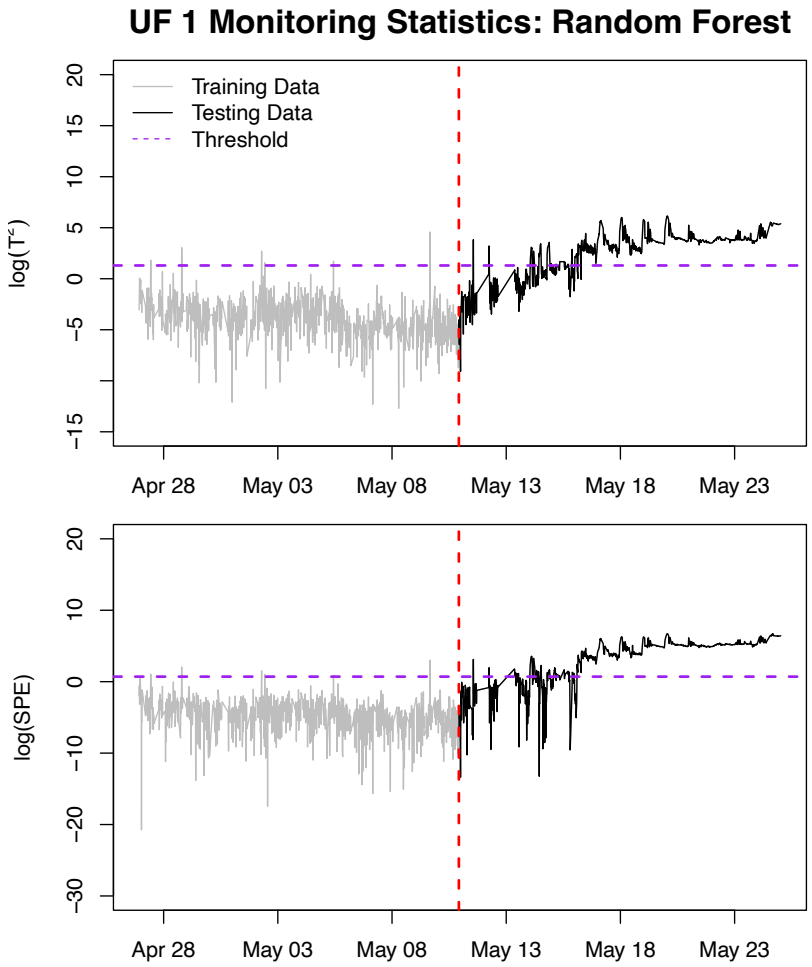
- Both plots show a fault.
- Adaptive lasso provides stronger evidence
  - *SPE* consistently exceeds the threshold once the fault is found.

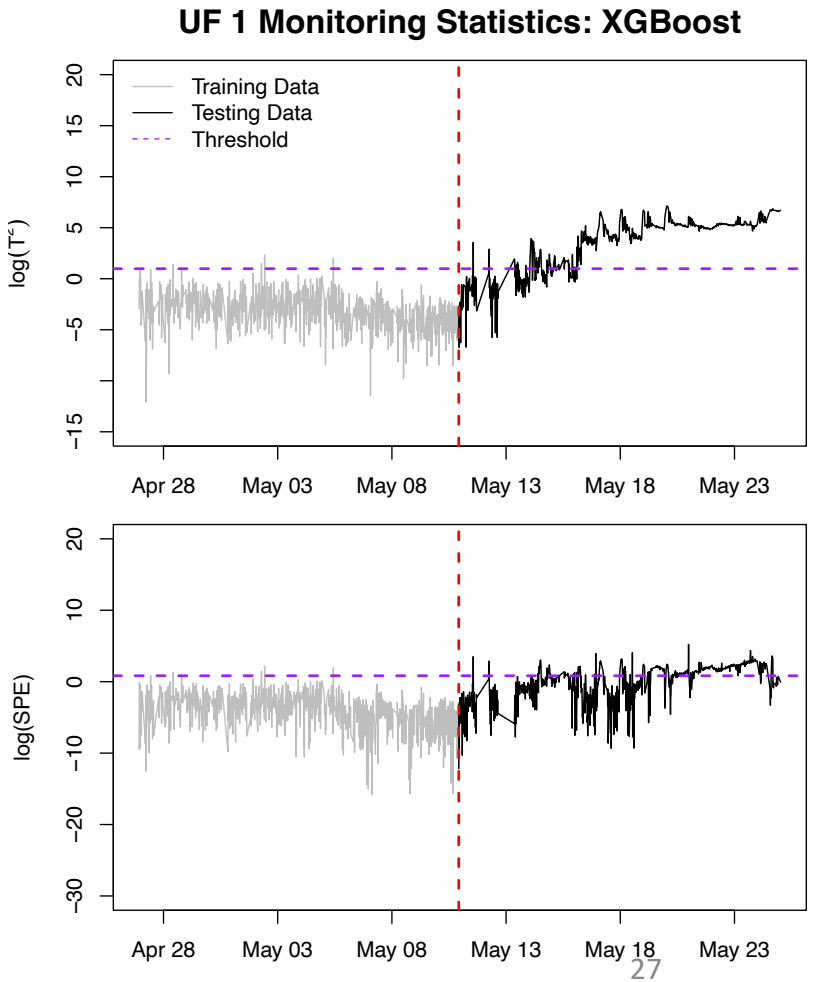




Monitoring Statistic Plots

- Both models identify a fault.
- T^2 values are larger for XGBoost, but *SPE* values are smaller than RF.
  - With XGBoost detrending, the PCA model better explains the variability.





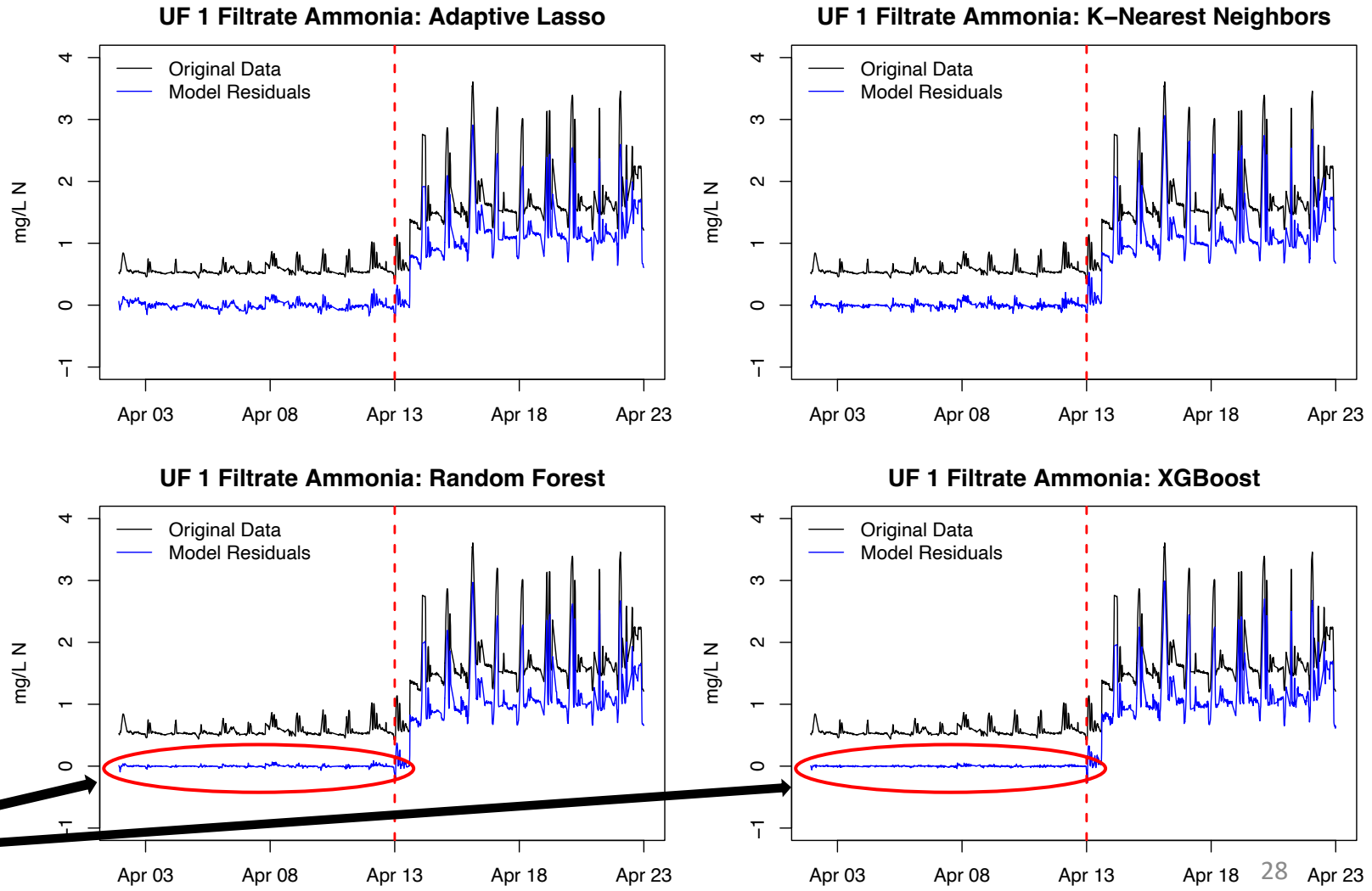
Detrending Plots

The red line separates training (left) from testing (right).

- No model predicts the upward shift.
- RF and XGBoost may be overfit.

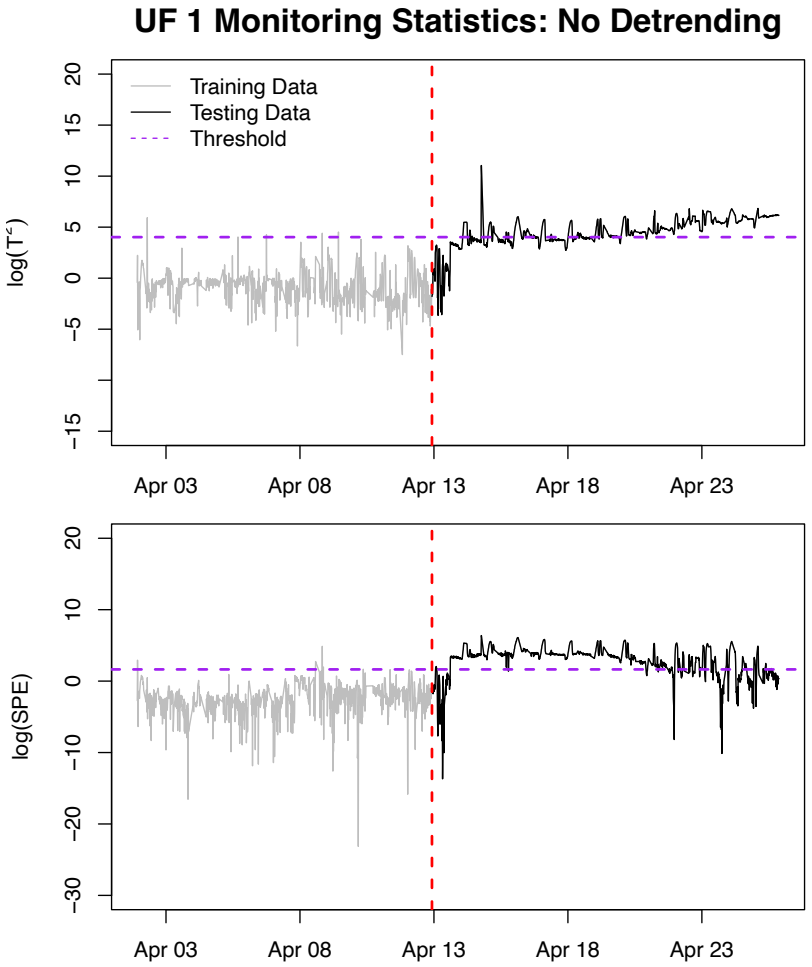
Overfitting: model fits well to training data, predicts poorly on testing data.

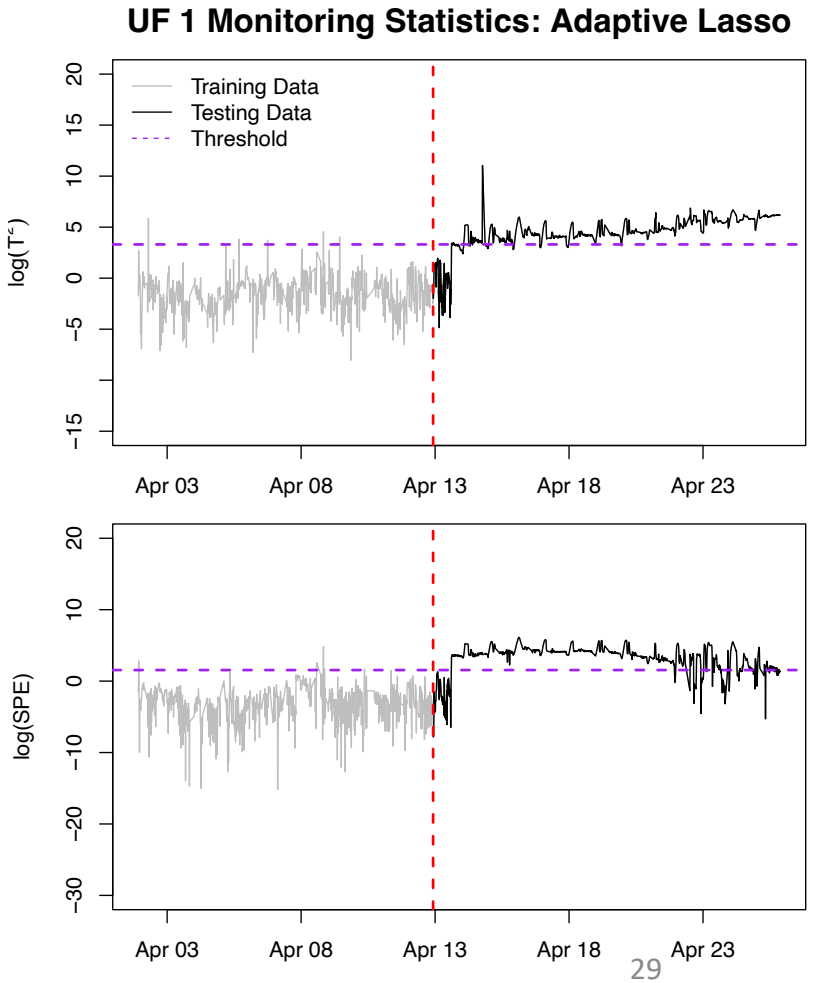
- Often a problem with complex machine learning models.



Monitoring Statistic Plots

- Even without detrending, both plots show a fault shortly after testing begins.
- Adaptive lasso values are slightly larger
  - More consistently above the T^2 threshold
  - Detrending helps





Monitoring Statistic Plots

- Both models detect a fault shortly after testing begins.
- T^2 and *SPE* are much larger than before.
  - Stronger evidence of a fault
  - Could be caused by overfitting