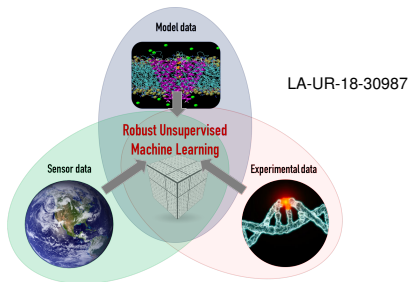


Novel Unsupervised Machine Learning Methods for Extraction of Features Characterizing Datasets and Models

Velimir V. Vesselinov (monty) (vvv@lanl.gov)

Earth and Environmental Sciences Division, Los Alamos National Laboratory, NM, USA



Why unsupervised Machine Learning (ML)?

- **Supervised** ML: requires prior categorization (knowledge) of the processed data

Example: Recognize images of cats and dogs after extensive training; but cannot recognize horses if not trained

Cannot find something that we do not already know

- **Unsupervised** ML: extracts hidden features (signals) in the processed data without any prior information (**exploratory analysis** for **data-driven science**)

Example: Identify features that distinguish images of animals (e.g., cats, dogs, horses, etc.); without prior information or training

Why unsupervised Machine Learning (ML)?

- ▶ **Supervised** ML: requires prior categorization (knowledge) of the processed data

Example: Recognize images of cats and dogs after extensive training; but cannot recognize horses if not trained

Cannot find something that we do not already know

- ▶ **Unsupervised** ML: extracts hidden features (signals) in the processed data without any prior information (**exploratory analysis** for **data-driven science**)

Example: Identify features that distinguish images of animals (e.g., cats, dogs, horses, etc.); without prior information or training

Why unsupervised Machine Learning (ML)?

- ▶ **Supervised** ML: requires prior categorization (knowledge) of the processed data
Example: Recognize images of cats and dogs after extensive training; but cannot recognize horses if not trained
Cannot find something that we do not already know
- ▶ **Unsupervised** ML: extracts hidden features (signals) in the processed data without any prior information (**exploratory analysis** for **data-driven science**)
Example: Identify features that distinguish images of animals (e.g., cats, dogs, horses, etc.); without prior information or training

Why unsupervised Machine Learning (ML)?

- ▶ **Supervised** ML: requires prior categorization (knowledge) of the processed data
Example: Recognize images of cats and dogs after extensive training; but cannot recognize horses if not trained
Cannot find something that we do not already know
- ▶ **Unsupervised** ML: extracts hidden features (signals) in the processed data without any prior information (**exploratory analysis** for **data-driven science**)
Example: Identify features that distinguish images of animals (e.g., cats, dogs, horses, etc.); without prior information or training

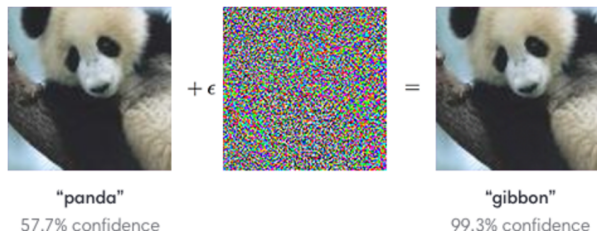
Why unsupervised Machine Learning (ML)?

- ▶ **Supervised** ML: requires prior categorization (knowledge) of the processed data
Example: Recognize images of cats and dogs after extensive training; but cannot recognize horses if not trained
Cannot find something that we do not already know
- ▶ **Unsupervised** ML: extracts hidden features (signals) in the processed data without any prior information (**exploratory analysis** for **data-driven science**)
Example: Identify features that distinguish images of animals (e.g., cats, dogs, horses, etc.); without prior information or training

Why not supervised Machine Learning (ML)

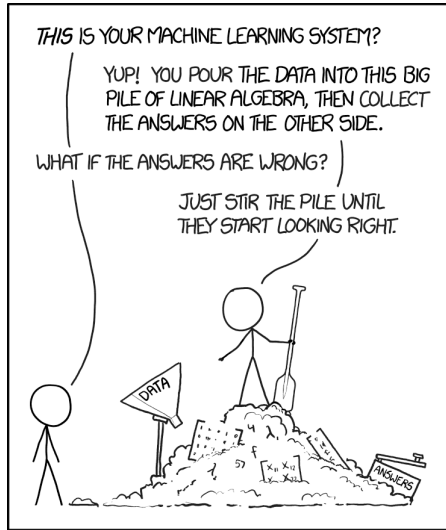
► Supervised ML

- can introduce subjectivity (through the labeling process)
- does not provide insights why horses are different than dogs / cats
- cannot make predictions
- requires huge training (labeled) datasets
- is impacted by “adversarial examples”

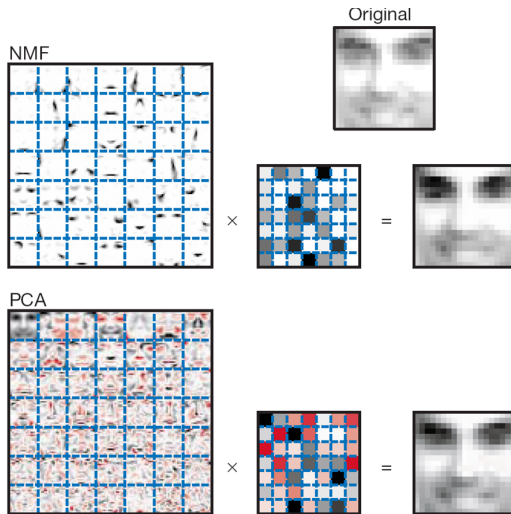


⇒ major limitations of the **supervised** methods
for **data-analytics** and **data-driven science** applications

Supervised Machine Learning (xkcd)



Why nonnegativity?



Nonnegativity constraints provide meaningful and interpretable results (+sparsity)

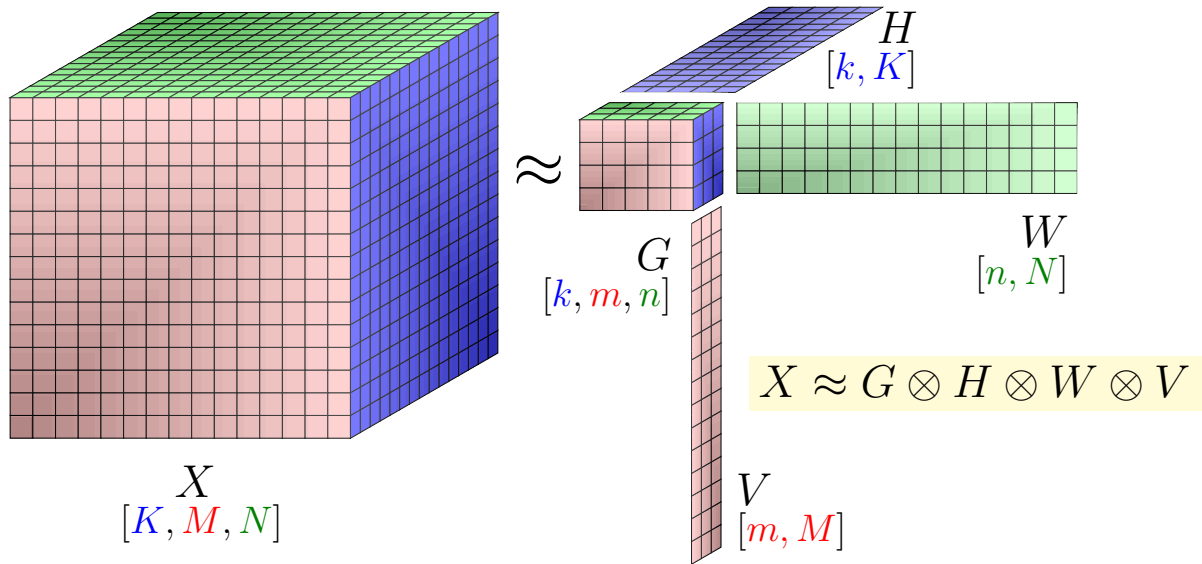
Unsupervised Machine Learning Applications: Exploratory Analysis

- ▶ **Data Analytics:** Identify signals (features) in datasets (latent variables)
 - ▶ Feature extraction (**FE**):
 - ▶ Blind source separation (**BSS**)
 - ▶ Detection of disruptions / anomalies
 - ▶ Image recognition
 - ▶ Discover unknown dependencies and phenomena
 - ▶ Guide development of physics / reduced-order models representing the data
- ▶ **Model Analytics/Diagnostics:** Identify processes (features) in model outputs
 - ▶ Separate processes (inseparable during modeling)
 - ▶ Model reduction (development of reduced-order models)
 - ▶ Identify dependencies between model inputs and outputs
 - ▶ Discover unknown dependencies and phenomena (deterministic/stochastic modeling)
- ▶ **Coupled Data/Model Analytics:**
Simultaneous analyses of data and model outputs (data/model fusion)

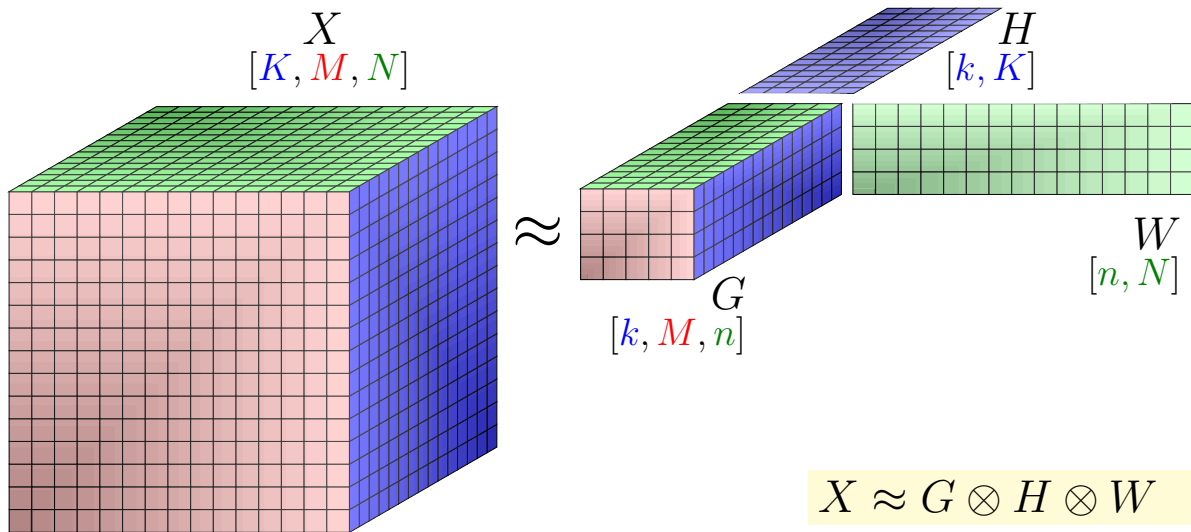
Nonnegative Matrix/Tensor Factorization

- ▶ We have developed a series of novel unsupervised Machine Learning (ML) methods and computational techniques
- ▶ Our methods are based in matrix/tensor factorization coupled with custom k -means clustering and nonnegativity/sparsity constraints:
 - ▶ NMF $_k$: Nonnegative **Matrix** Factorization
 - ▶ NTF $_k$: Nonnegative **Tensor** Factorization
- ▶ NMF $_k$ /NTF $_k$ are capable to efficiently process large datasets (GB/TB's) utilizing GPU's & TPU's (TensorFlow, PyTorch, MXNet)

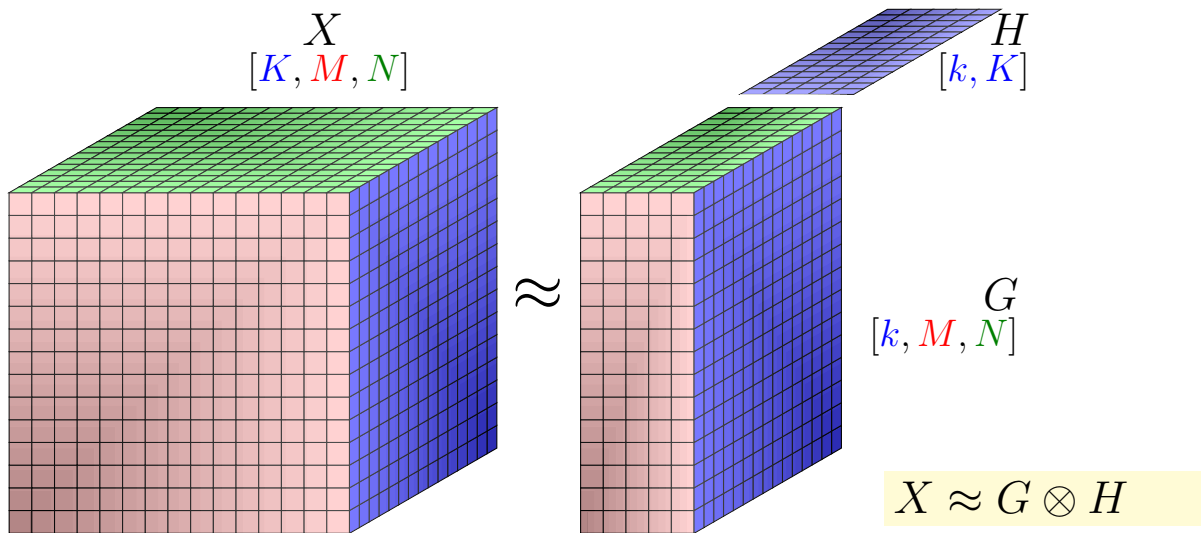
Tucker tensor factorizations (3D case): Tucker-3



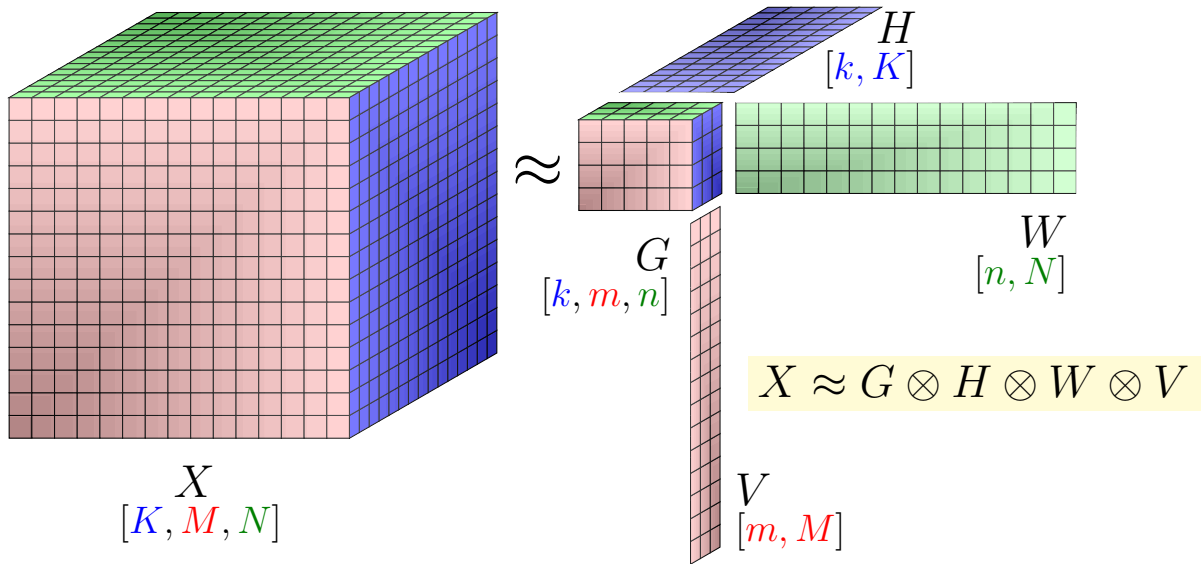
Tucker tensor factorizations (3D case): Tucker-2 (three possible alternatives)



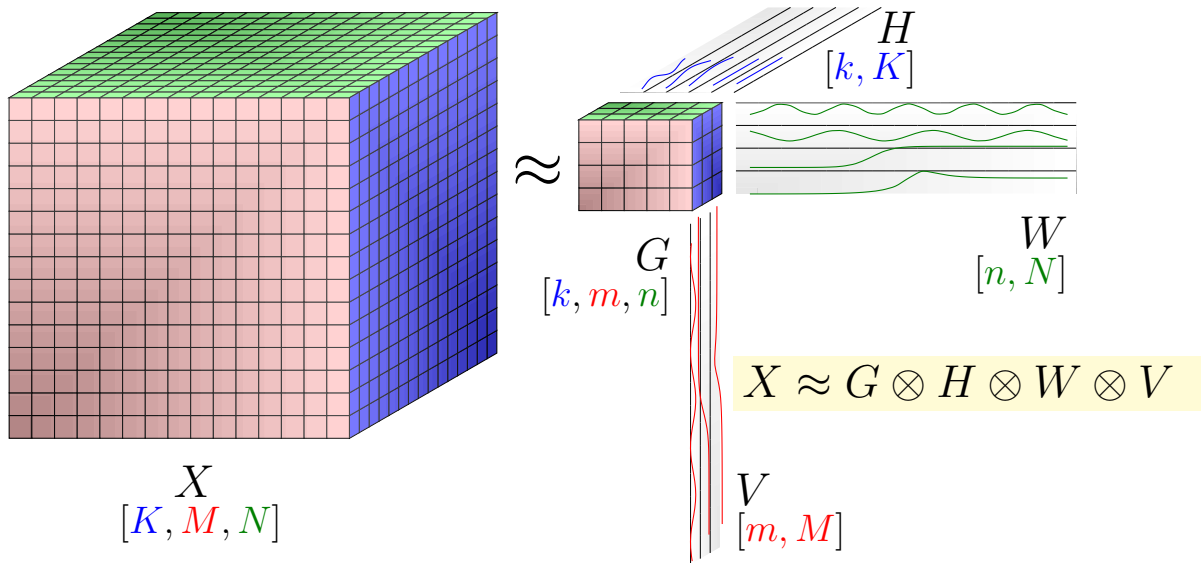
Tucker tensor factorizations (3D case): Tucker-1 (three possible alternatives)



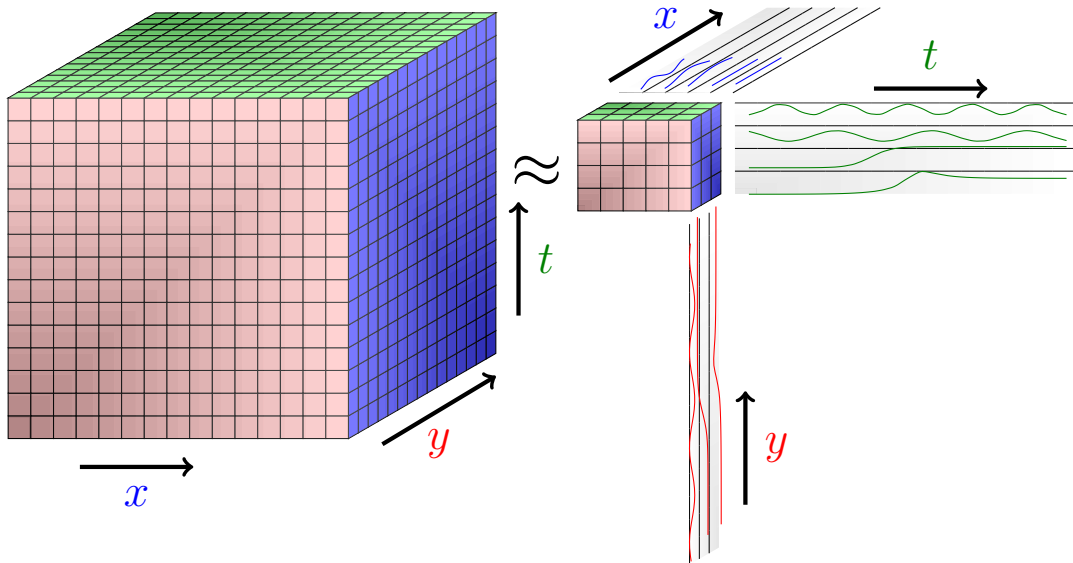
Tucker tensor factorizations (3D case): Tucker-3



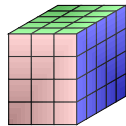
Tucker Tensor Decomposition: Example



Tucker Tensor Decomposition: Example



Tucker Tensor Decomposition: Example



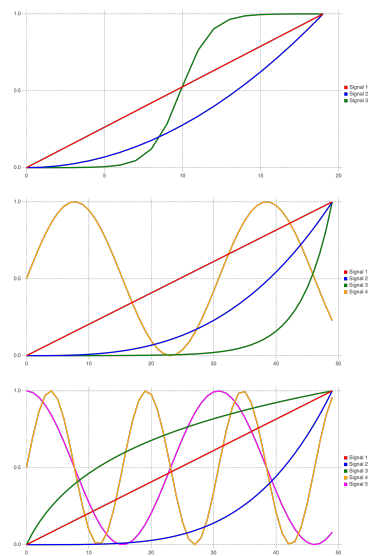
(12 elements)

$$V = \begin{bmatrix} t \\ t^2 \\ \tanh(t - 10) + 1 \end{bmatrix}$$

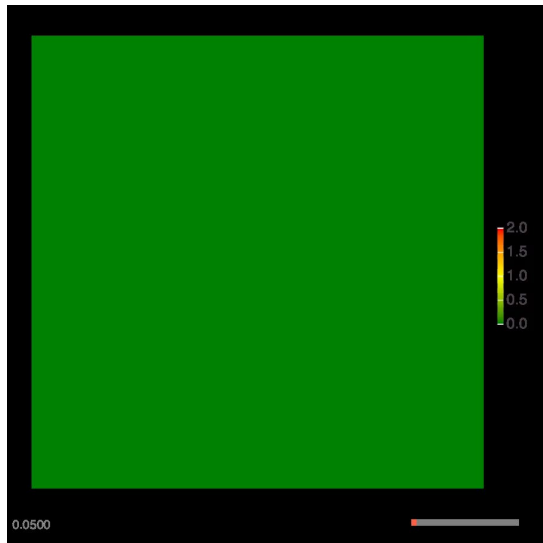
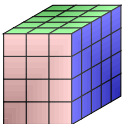
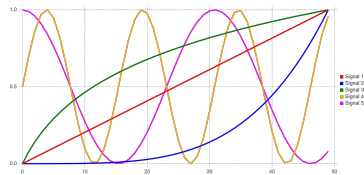
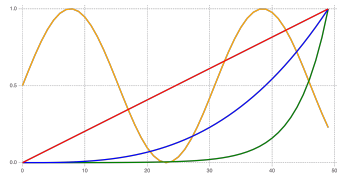
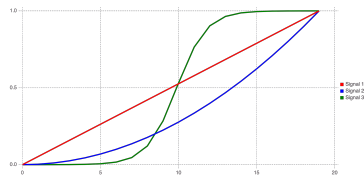
$$W = \begin{bmatrix} x \\ x^3 \\ e^x \\ \sin(x) + 1 \end{bmatrix}$$

$$H = \begin{bmatrix} y \\ y^4 \\ \ln(y) \\ \sin(2y) + 1 \\ \cos(y) + 1 \end{bmatrix}$$

$$\begin{aligned} X &= G \otimes H \otimes W \otimes V \\ &= xyt + xy^4t + xt \ln(y) + \\ &\quad xt(\sin(2y) + 1) + x^3yt + \\ &\quad xt(\cos(y) + 1) + yte^x + \\ &\quad yt(\sin(x) + 1) + \\ &\quad xyt^2 + xy(1 + \tanh(t - 10)) + \\ &\quad t^2e^x(\sin(2y) + 1) + \\ &\quad (1 + \tanh(t - 10))(\sin(x) + 1) \\ &\quad (\cos(y) + 1) \end{aligned}$$



Tucker Tensor Decomposition: Example



Unsup ML

oooooooooooo●oooo

Studies

ooooo

Mixing

ooooo

Climate EU

oooooo o

Climate US

oooooo

Geochem

oooooooo

Geysers

oooooo

Quantum

o

Polymer

o

Summary

oo

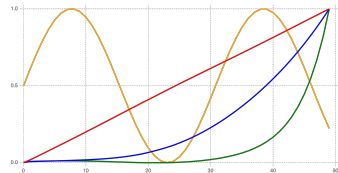
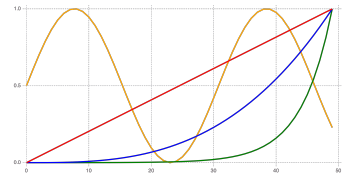
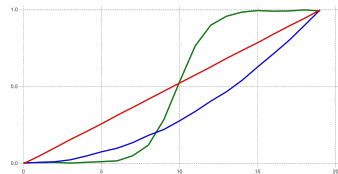
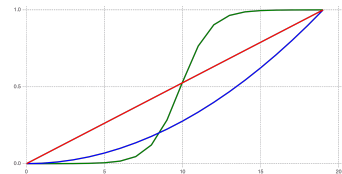
UV

o

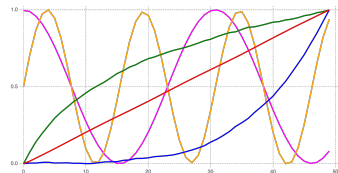
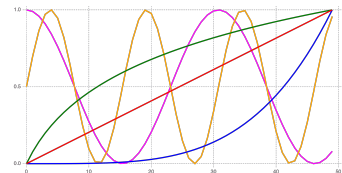
LANSCÉ

oooooooooooo

Tucker Tensor Decomposition: Example



← **Truth vs Predictions** →



Unsup ML

Studies

Mixing

Climate EU

Climate US

Geochem

Geysers

Quantum

Polymer

Summary

UV

LANSCE

○○○○○○○○○○○○○●○○○

○○○○○

○○○○○

○○○○○ ○

○○○○○○

○○○○○○○○○

○○○○○○○

○

○

○○

○

○○○○○○○○○○○

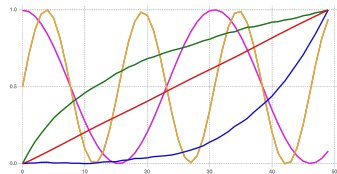
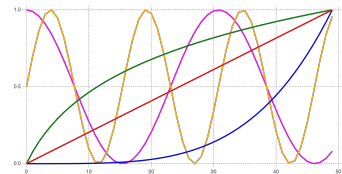
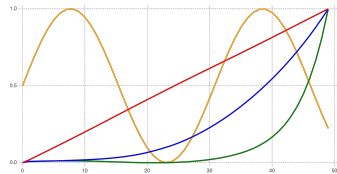
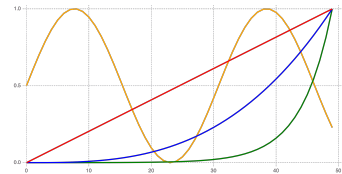
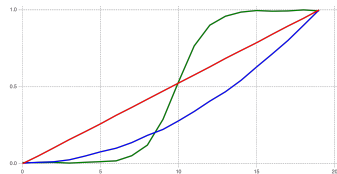
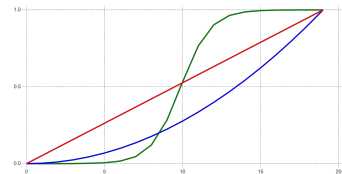
Tucker Tensor Decomposition: Example

← **Truth** vs **Predictions** →

$$V = \begin{bmatrix} t \\ t^2 \\ \tanh(t - 10) + 1 \end{bmatrix}$$

$$W = \begin{bmatrix} x \\ x^3 \\ e^x \\ \sin(x) + 1 \end{bmatrix}$$

$$H = \begin{bmatrix} y \\ y^4 \\ \ln(y) \\ \sin(2y) + 1 \\ \cos(y) + 1 \end{bmatrix}$$



Unsup ML

Studies

Mixing

Climate EU

Climate US

Geochem

Geysers

Quantum

Polymer

Summary

UV

LANSCE

○○○○○○○○○○○○●○○○

○○○○○

○○○○○

○○○○○ ○

○○○○○

○○○○○○○○○

○○○○○

○

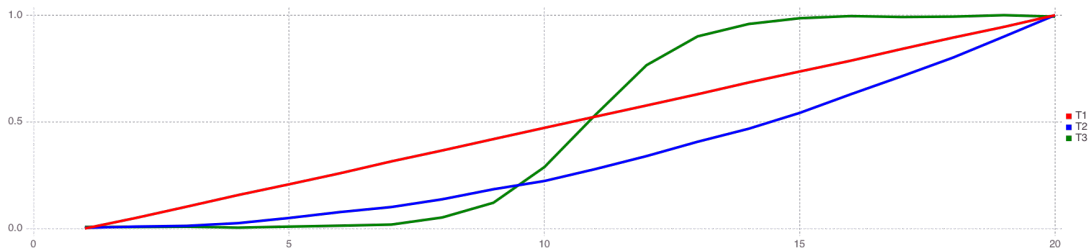
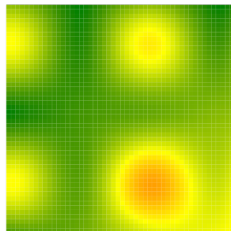
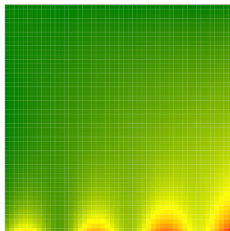
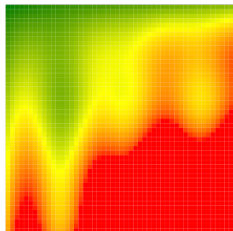
○

○○

○

○○○○○○○○○○

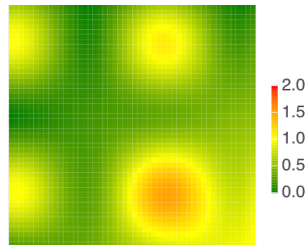
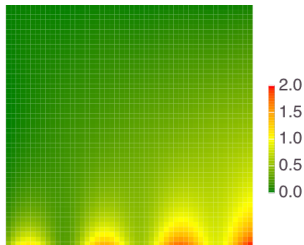
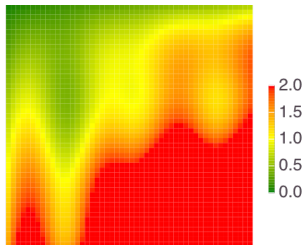
Tucker Tensor Decomposition: Example



Unsup ML Studies Mixing Climate EU Climate US Geochem Geysers Quantum Polymer Summary UV LANSCE

oooooooooooooooo●oo oooooo oooooo oooooo o oooooo oooooo o o oo o ooooooooooooo

Tucker Tensor Decomposition: Example



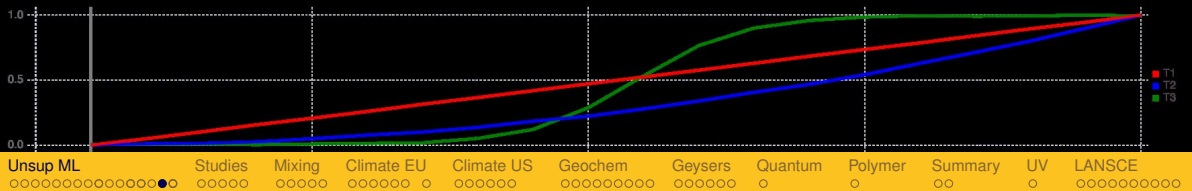
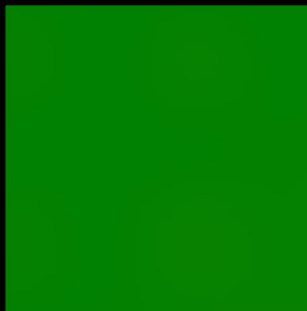
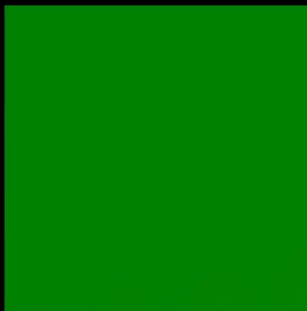
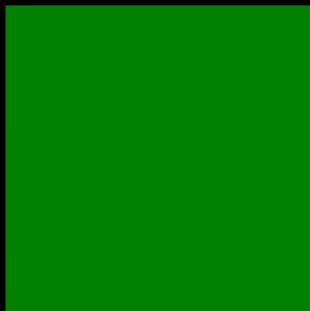
$$M = G \otimes H \otimes W$$

$$M_1 = xy + xy^4 + x \log(y + 1) + \\ x(\sin(2y) + 1) + ye^x \\ x(\cos(y) + 1) + x^3y + \\ y(\sin(x) + 1)$$

$$M_2 = xy + \\ e^x(\cos(z) + 1)$$

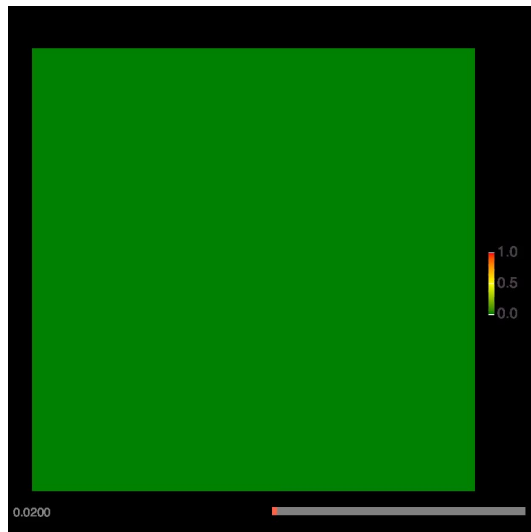
$$M_3 = xy + \\ (\sin(x) + 1)(\cos(y) + 1)$$

Tucker Tensor Decomposition: Example

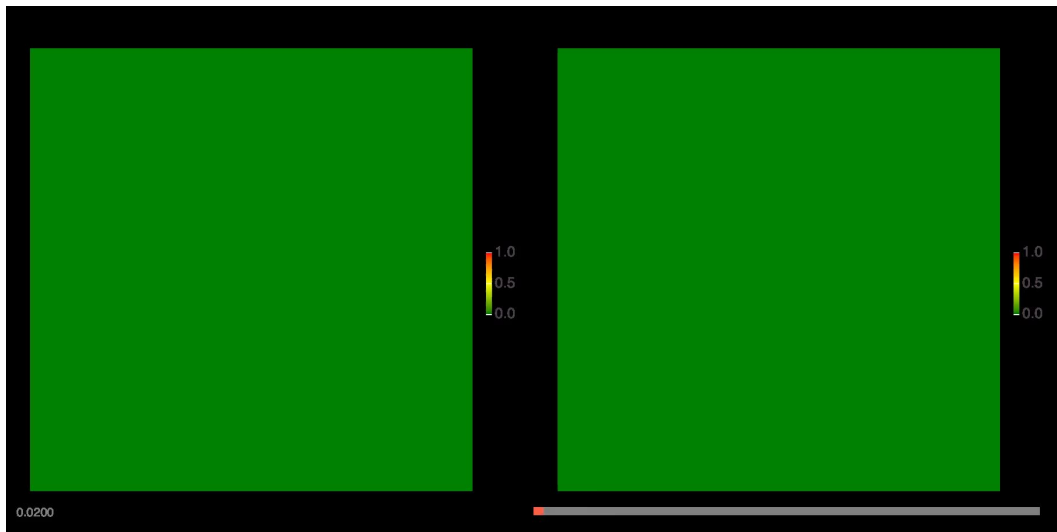


Tucker Tensor Decomposition: Example II

- ▶ $(50 \times 50 \times 50)$ tensor
- ▶ 50 columns in x
- ▶ 50 rows in y
- ▶ 50 time frames
- ▶ 'ones' swimming in a sea of 'zeros'



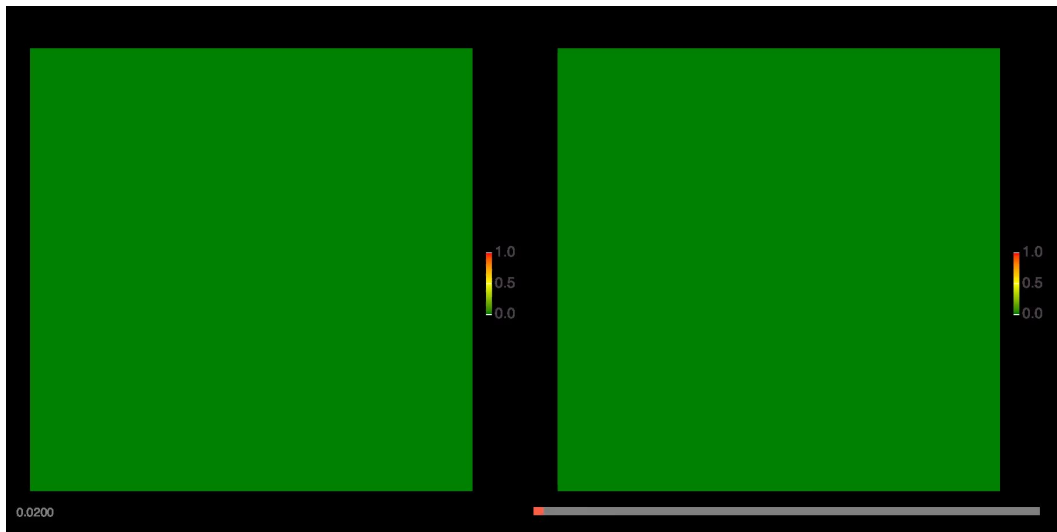
Tucker Tensor Decomposition: Example II



Factorizing all **3** dimensions $(50 \times 50 \times 50) \rightarrow (6 \times 44 \times 48)$

Unsup ML	Studies	Mixing	Climate EU	Climate US	Geochem	Geysers	Quantum	Polymer	Summary	UV	LANSCE
oooooooooooooooo●	ooooo	ooooo	oooooo o	oooooo	oooooooo	oooooo	o	o	oo	o	oooooooooooo

Tucker Tensor Decomposition: Example II



6 groups of swimmers (x); 44 lanes occupied (y); 48 time frames (first/last empty)

Unsup ML	Studies	Mixing	Climate EU	Climate US	Geochem	Geysers	Quantum	Polymer	Summary	UV	LANSCE
oooooooooooooooo●	ooooo	ooooo	oooooo o	oooooo	oooooooo	oooooo	o	o	oo	o	oooooooooooo

► Identifying the number of unknown features:

- resolved using custom k -means clustering and sparsity constraints on the core tensor
- number of features identified based on the reconstruction quality (e.g., Frobenius norm) and cluster Silhouettes

► Solving a non-unique optimization problem:

- addressed through multistarts, regularization and nonnegativity constraints
- applying diverse optimization techniques (Multiplicative/Alternating Least Squares algorithms, NLOpt, Ipopt, Gurobi, MOSEK, GLPK, Clp, Cbc, ...)
- accounting for the physics

► Processing Big Data:

- GPU's / TPU's / Distributed computing
- Account for data sparsity and structure
- Nonnegative Tensor Trains

► Dealing with Noisy Data:

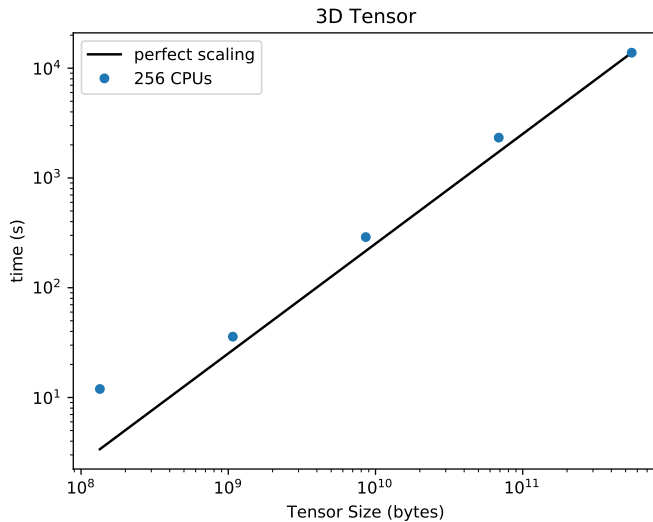
- Random noise impacts accuracy but its accountable
- Systematic noise is identified as separate signals

4GB Tensor (1000 × 1000 × 1000)

Framework	Execution time (seconds)
MATLAB	2634
NumPy	881
MXNet	644
PyTorch	121
TensorFlow	119
Julia	109



NMF_k / NTF_k Scalability



Unsup ML
oooooooooooooooo

Studies
oo●oo

Mixing
ooooo

Climate EU
oooooo o

Climate US
oooooo

Geochem
oooooooo

Geysers
oooooo

Quantum
o

Polymer
o

Summary
oo

UV
o

LANSCE
oooooooooooo

► Field Data:

- Characterization of groundwater contaminant sources
- US Climate data
- Geothermal data
- Seismic data

► Lab Data:

- X-ray Spectroscopy
- UV Fluorescence Spectroscopy
- Microbial population analyses

► Operational Data:

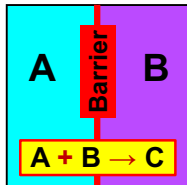
- LANSCE: Los Alamos Neutron Accelerator
- Hydrocarbon (oil/gas) production

► Model Data:

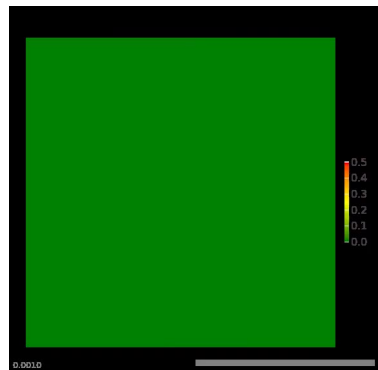
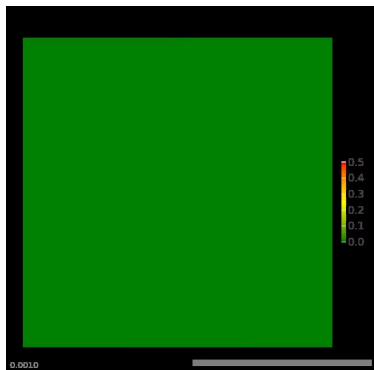
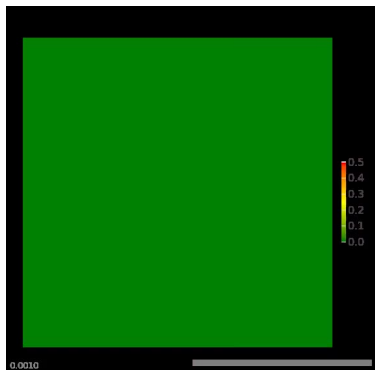
- Reactive mixing $A + B \rightarrow C$
- Phase separation of co-polymers
- Molecular Dynamics of proteins
- EU Climate modeling (Helmholtz Institute, Germany)

- ▶ Stanev, Vesselinov, Kusne, Antoszewski, Takeuchi, Alexandrov, Unsupervised Phase Mapping of X-ray Diffraction Data by Nonnegative Matrix Factorization Integrated with Custom Clustering, **Nature Computational Materials**, 2018.
- ▶ Vesselinov, Munuduru, Karra, O'Malley, Alexandrov, Unsupervised Machine Learning Based on Non-Negative Tensor Factorization for Analyzing Reactive-Mixing, **Journal of Computational Physics**, (in review), 2018.
- ▶ Vesselinov, O'Malley, Alexandrov, Nonnegative Tensor Factorization for Contaminant Source Identification, **Journal of Contaminant Hydrology**, (n review), 2018.
- ▶ O'Malley, Vesselinov, Alexandrov, Alexandrov, Nonnegative/binary matrix factorization with a D-Wave quantum annealer, **PLOS ONE**, (accepted), 2018.
- ▶ Vesselinov, O'Malley, Alexandrov, Contaminant source identification using semi-supervised machine learning, **Journal of Contaminant Hydrology**, 10.1016/j.jconhyd.2017.11.002, 2017.
- ▶ Alexandrov, Vesselinov, Blind source separation for groundwater level analysis based on nonnegative matrix factorization, **WRR**, 10.1002/2013WR015037, 2014.

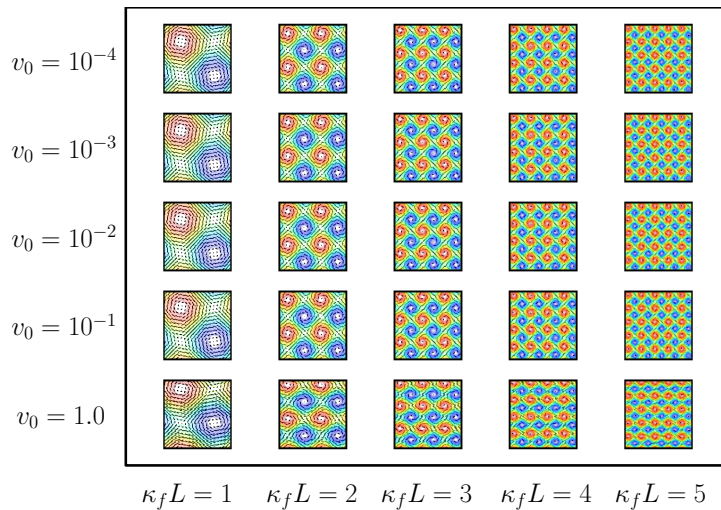
Reactive mixing



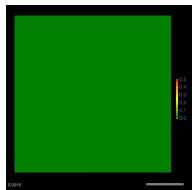
- ▶ > 2000 simulations of C concentrations in time/space with varying model inputs representing reactive mixing (5 input model parameters)
- ▶ NTF_k identifies physics processes impacting C concentrations and their relationship to model inputs



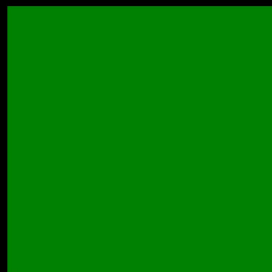
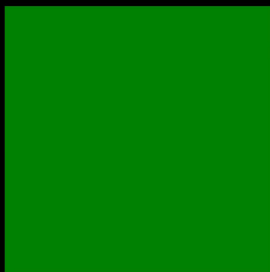
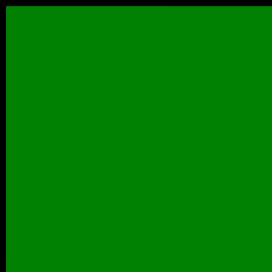
Reactive mixing: Flow fields



Reactive mixing: NTF_k results



- ▶ NTF_k extracts the dominant time/space features (**processes** / **vortices**) and compresses the model outputs
- ▶ Compression: $> 200\text{GB} \rightarrow \sim 70\text{MB}$ (ratio ~ 3000)
Here, $(1000 \times 81 \times 81) \rightarrow (3 \times 12 \times 13)$ (*time* $\times$ *rows* $\times$ *columns*)

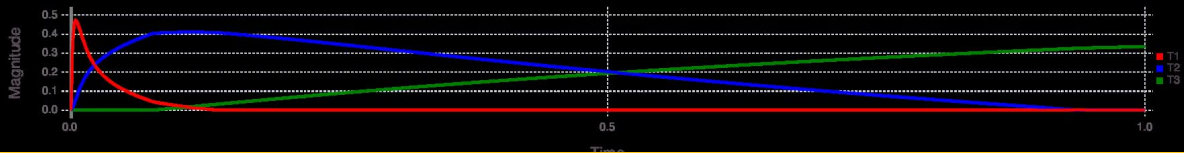
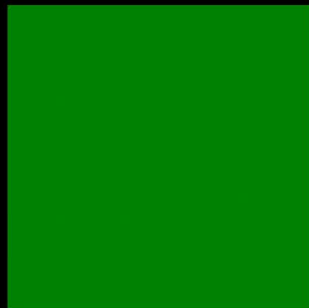
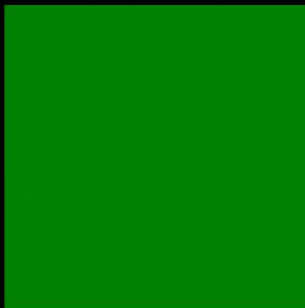
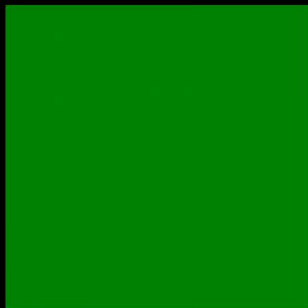


Advection

Dispersion

Diffusion

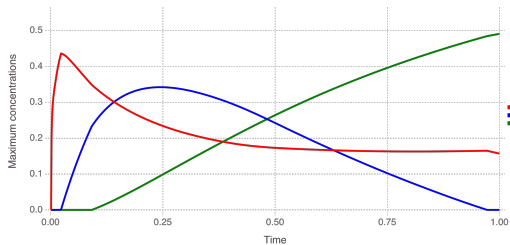
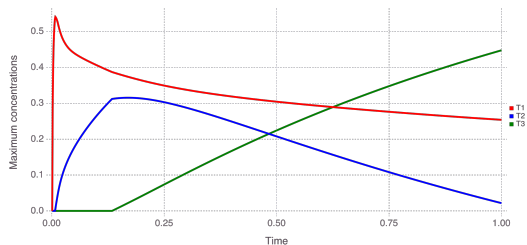
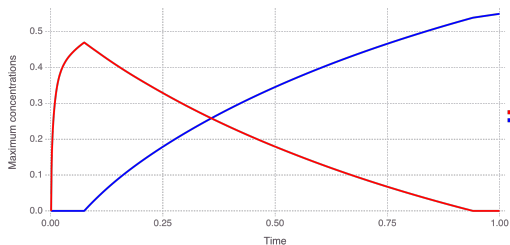
Reactive mixing: NTF_k results



Unsup ML Studies Mixing Climate EU Climate US Geochem Geysers Quantum Polymer Summary UV LANSCE

oooooooooooooooooooo oooooo ooo●o oooooo o oooooo oooooooooo oooooo o o oo o oooooooooooooo

Reactive mixing: NTF_k results ($\kappa_f L$ impacts)



Unsup ML

oooooooooooooooo

Studies

ooooo

Mixing

oooo●

Climate EU

oooooo o

Climate US

oooooo

Geochem

ooooooooo

Geysers

oooooo

Quantum

o

Polymer

o

Summary

oo

UV

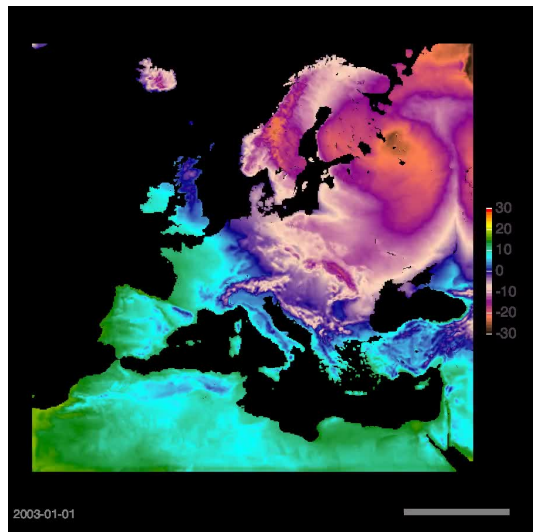
o

LANSCE

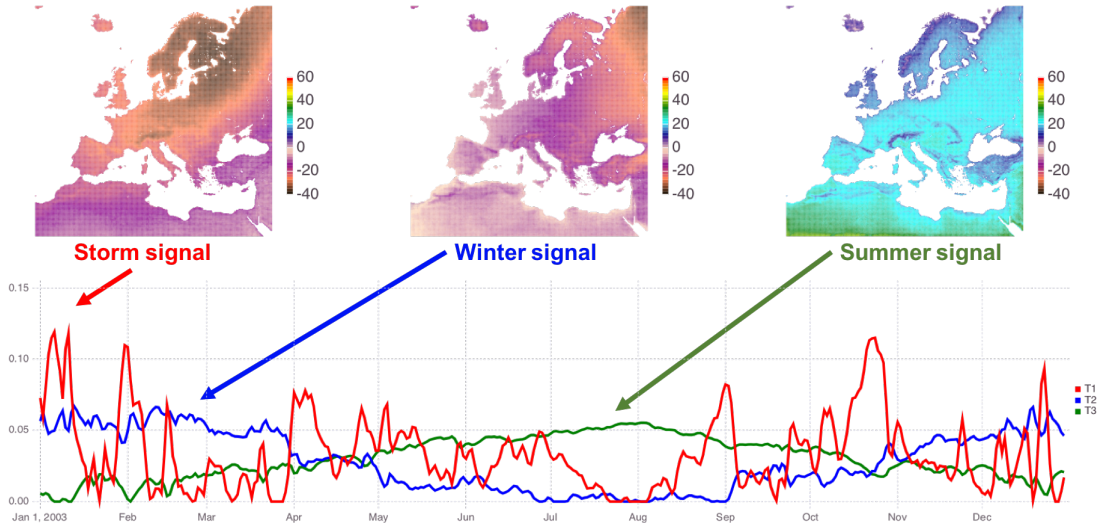
oooooooooooo

Climate model of Europe: air temperature

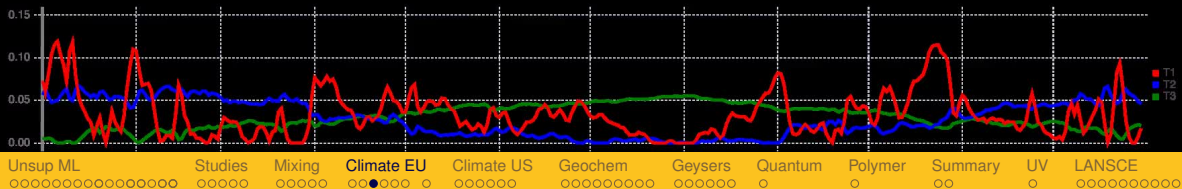
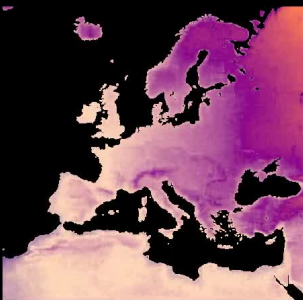
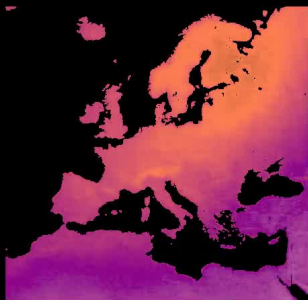
- ▶ fluctuations in the air temperature [$^{\circ}\text{C}$]
- ▶ $(424 \times 412 \times 365)$
(*columns* $\times$ *rows* $\times$ *days*)
- ▶ **NTF_k** extracts spatial and temporal footprints of dominant features:
 - ▶ storm signal
 - ▶ winter seasonal signal
 - ▶ summer seasonal signal
 - ▶
- ▶ Data compression: $\sim 4\text{GB} \rightarrow \sim 0.5\text{MB}$
Compression ratio: ~ 8000



Climate model of Europe: air temperature fluctuations represented by 3 signals

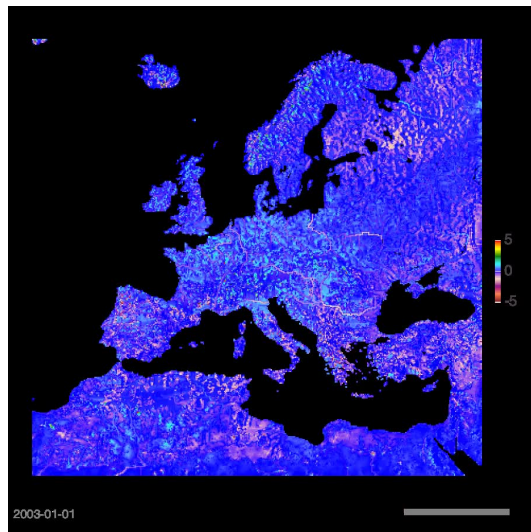


Climate model of Europe: air temperature fluctuations represented by 3 signals



Climate model of Europe: water-table depth

- ▶ fluctuations in the water-table depth [m]
- ▶ $(424 \times 412 \times 365)$
(*columns* $\times$ *rows* $\times$ *days*)
- ▶ **NTF_k** extracts spatial and temporal footprints of dominant signals
 - ▶ spring snowmelt signal
 - ▶ summer rainfall signal
 - ▶ seasonal signal
 - ▶



Climate model of Europe: water-table fluctuations represented by 3 signals



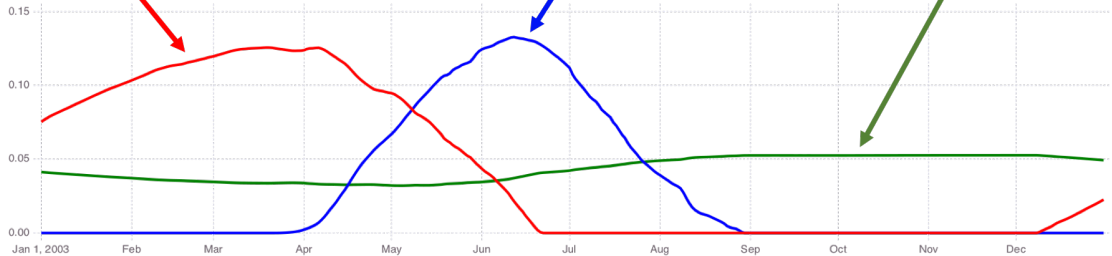
Spring snowmelt



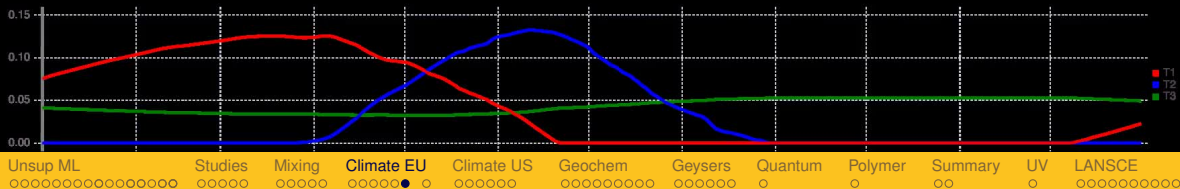
Summer rainfalls



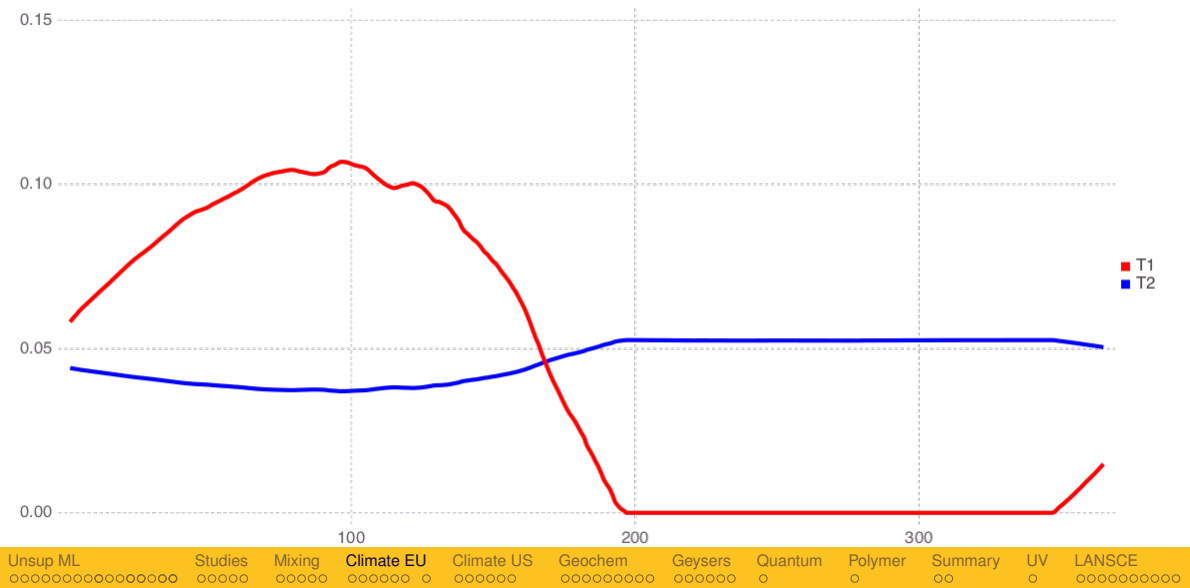
Seasonality



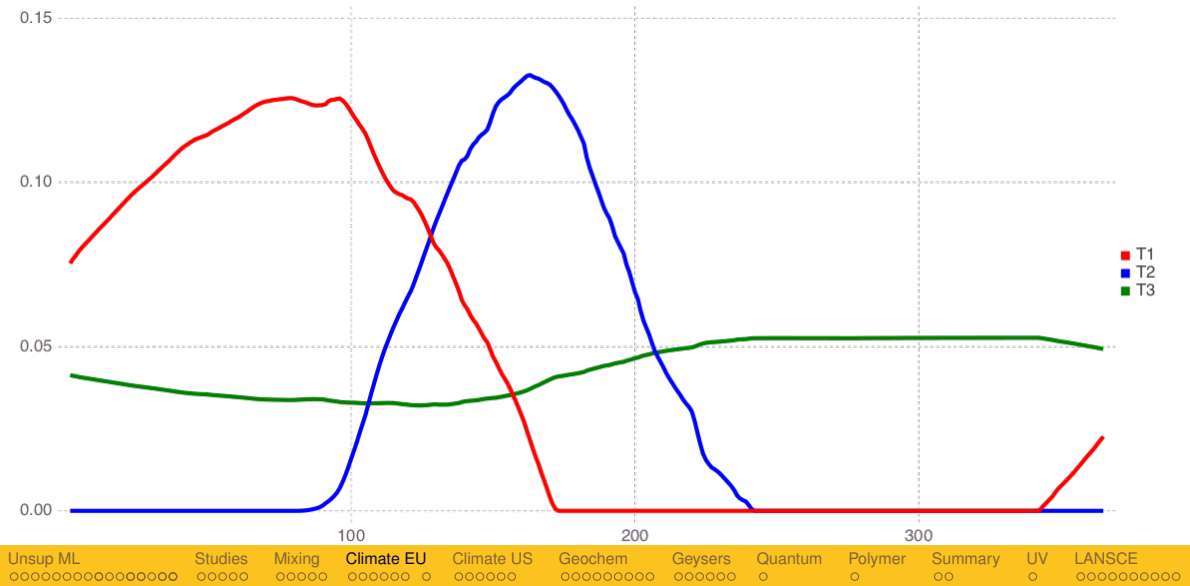
Climate model of Europe: water-table fluctuations represented by 3 signals



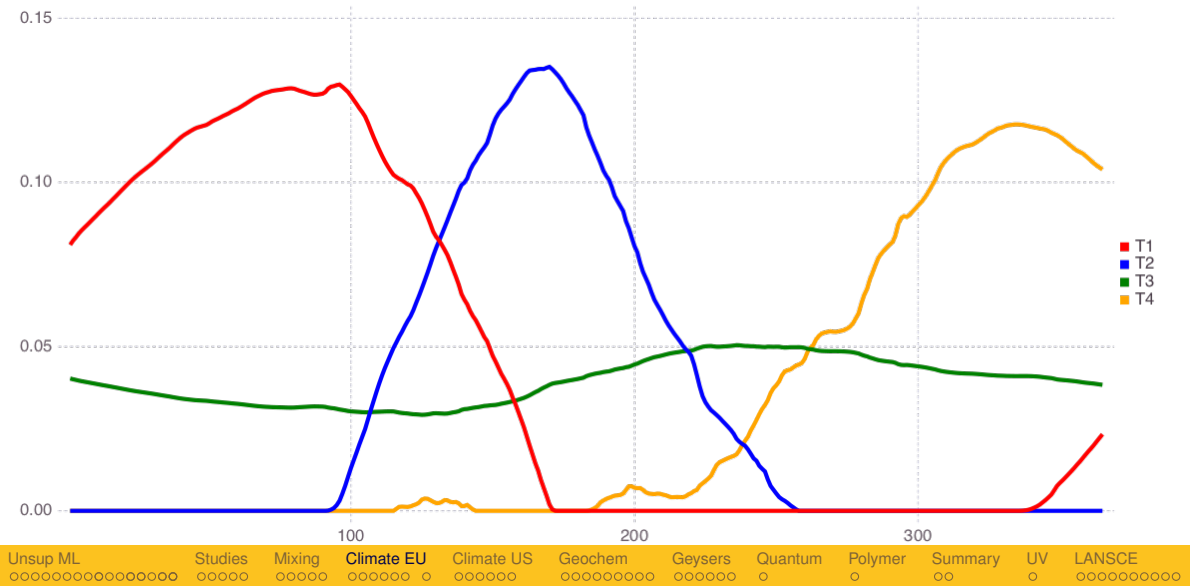
Climate model of Europe: Water-table fluctuations represented by 2 signals



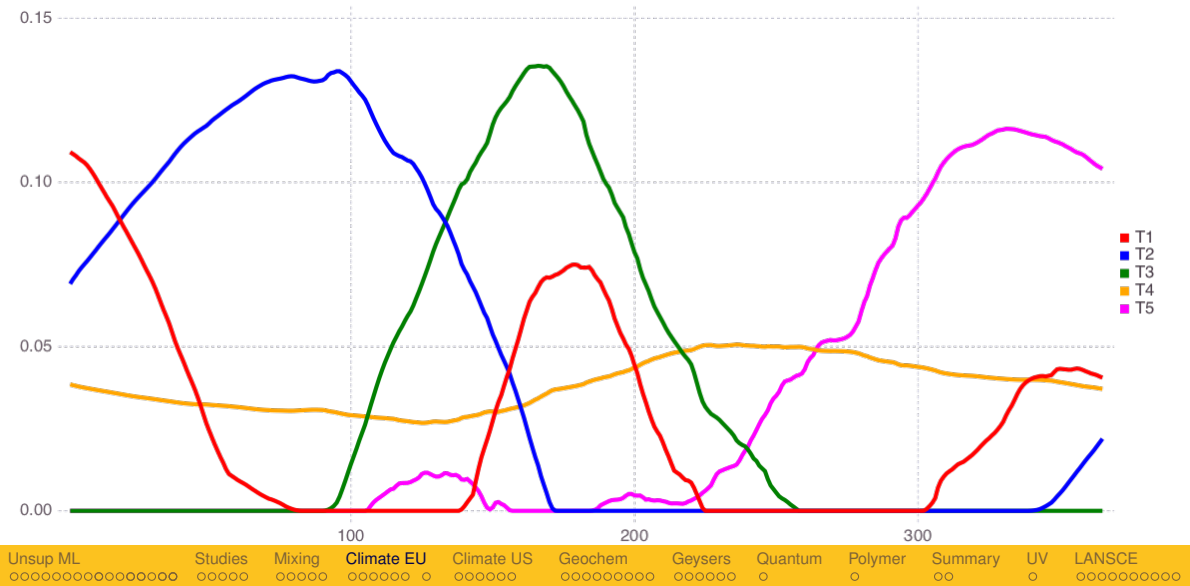
Climate model of Europe: Water-table fluctuations represented by 3 signals



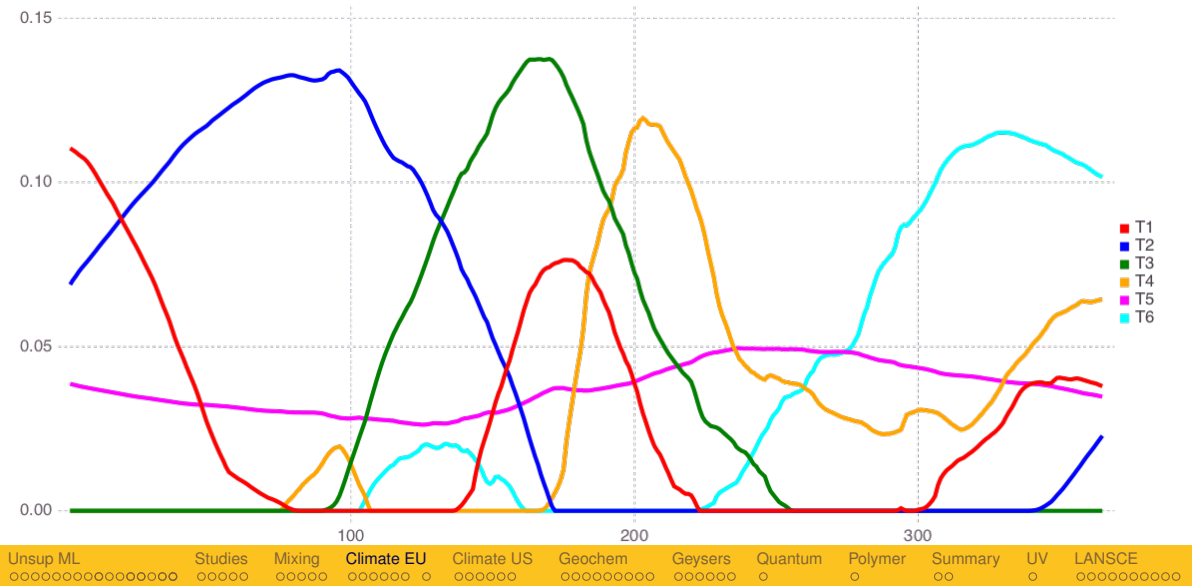
Climate model of Europe: Water-table fluctuations represented by 4 signals



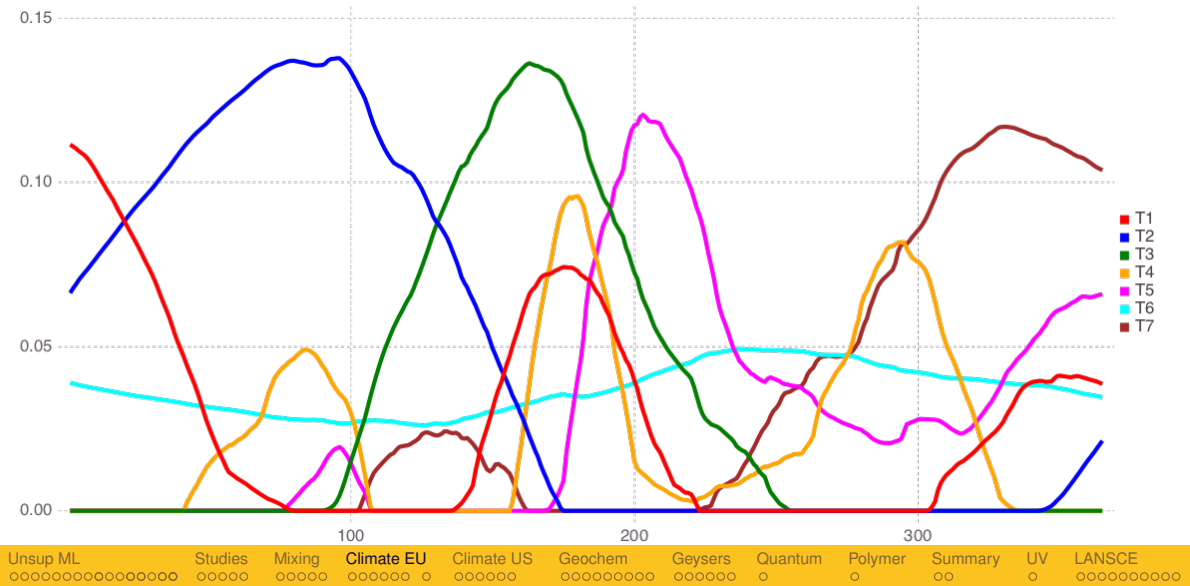
Climate model of Europe: Water-table fluctuations represented by 5 signals



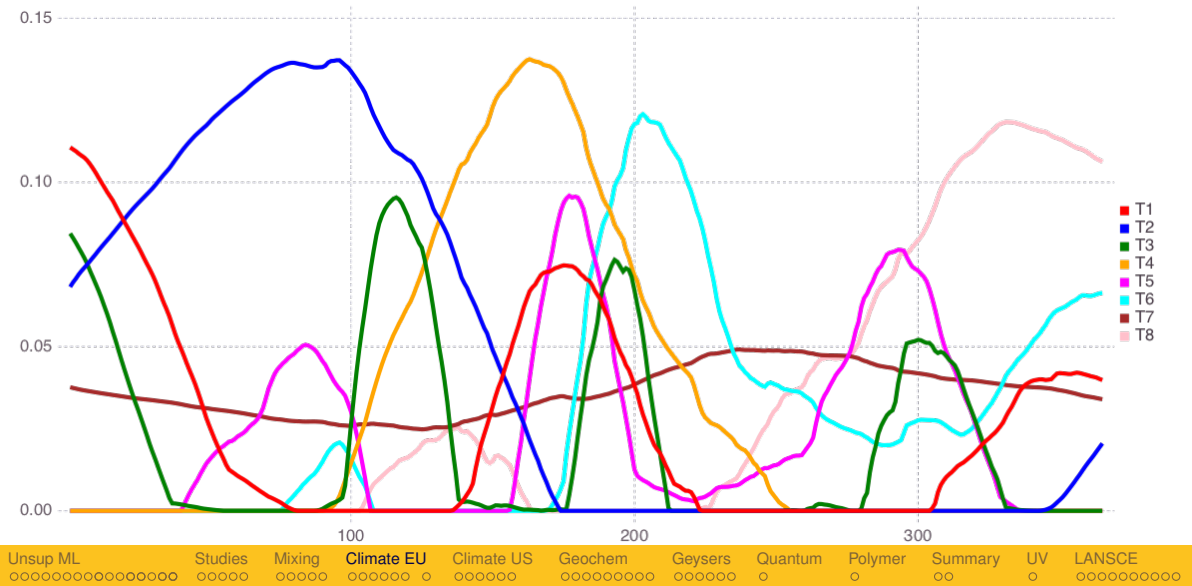
Climate model of Europe: Water-table fluctuations represented by 6 signals



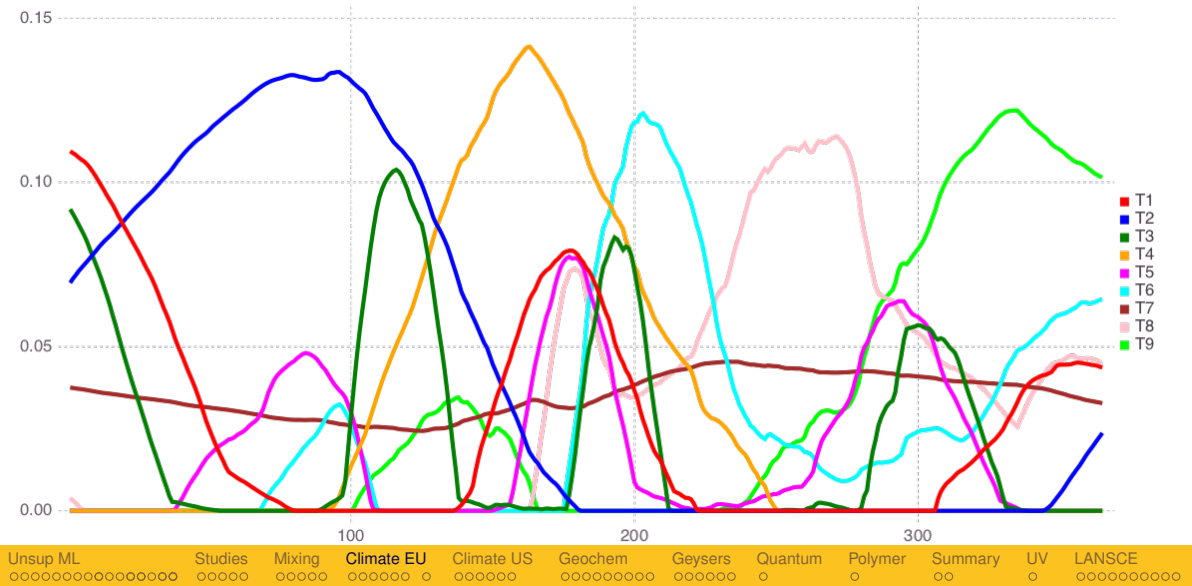
Climate model of Europe: Water-table fluctuations represented by 7 signals



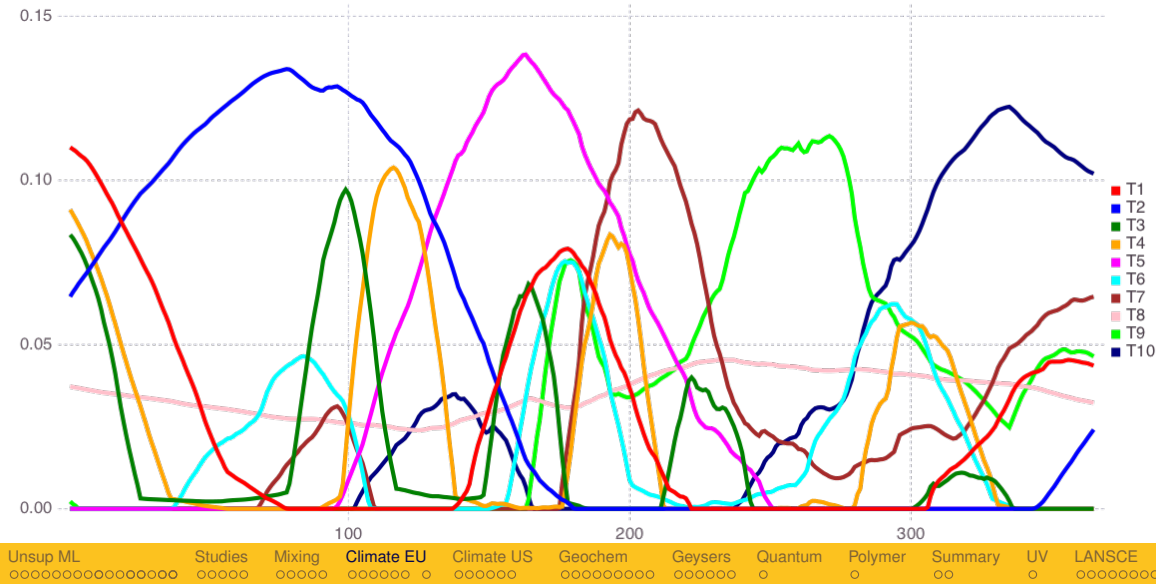
Climate model of Europe: Water-table fluctuations represented by 8 signals



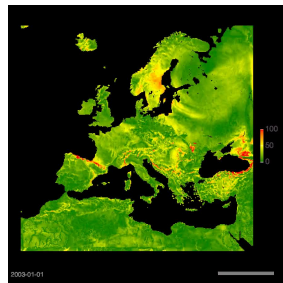
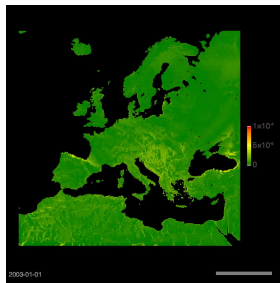
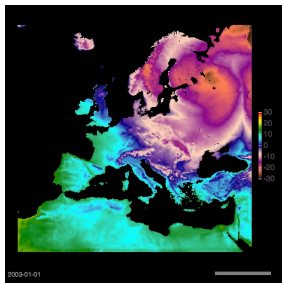
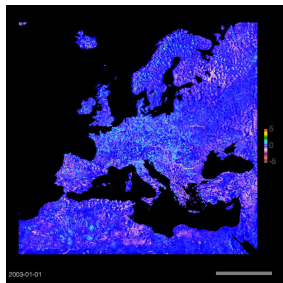
Climate model of Europe: Water-table fluctuations represented by 9 signals



Climate model of Europe: Water-table fluctuations represented by 10 signals



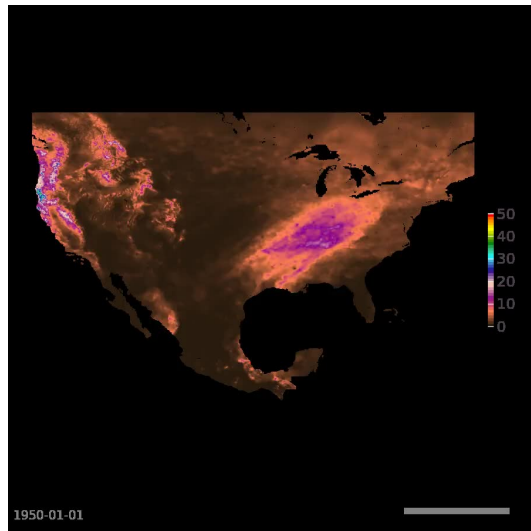
Climate model of Europe: analyze all model outputs (>40) simultaneously



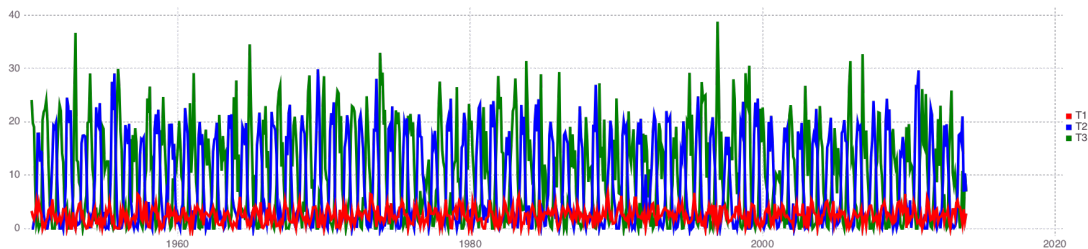
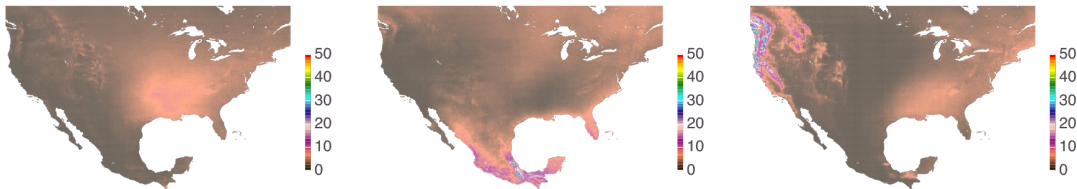
- Find interconnections among model outputs
- Evaluate impacts of different model setups
- Find dominant processes impacting model predictions
(e.g., climate impacts on groundwater resources, impacts of subsurface processes on atmospheric conditions)

US Climate data: precipitation

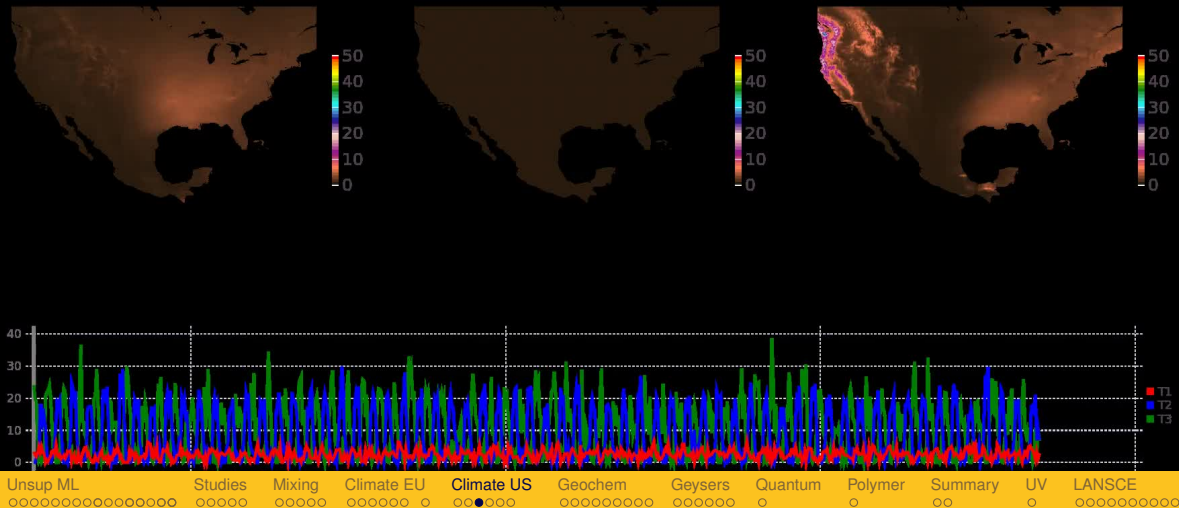
- ▶ NOAA data set of observed monthly averaged precipitation with spatial resolution of 6km ($1/16^\circ$)
- ▶ fluctuations in monthly averaged precipitation [inches]
- ▶ $(928 \times 614 \times 768)$
(*columns* $\times$ *rows* $\times$ *months*)



US precipitation: NTF_k results

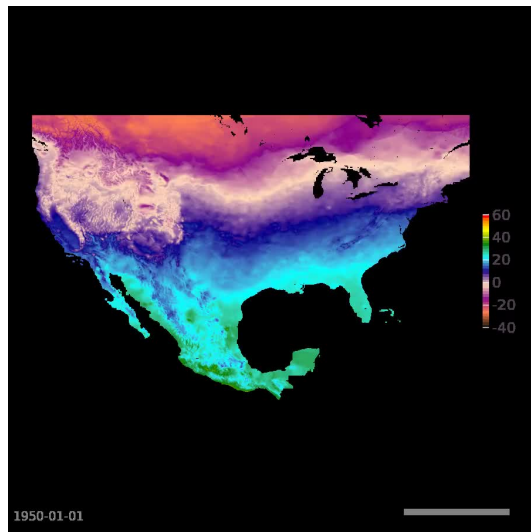


US precipitation: NTF_k results

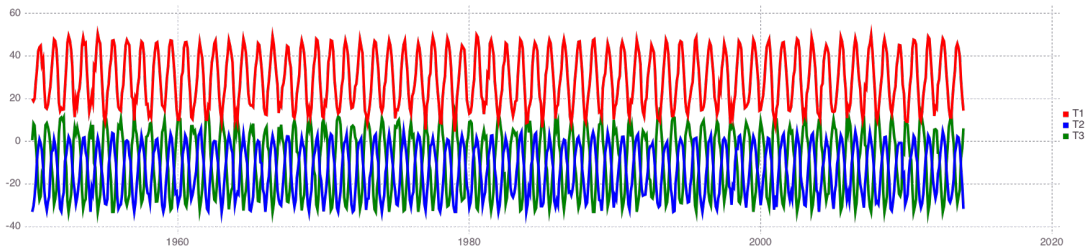
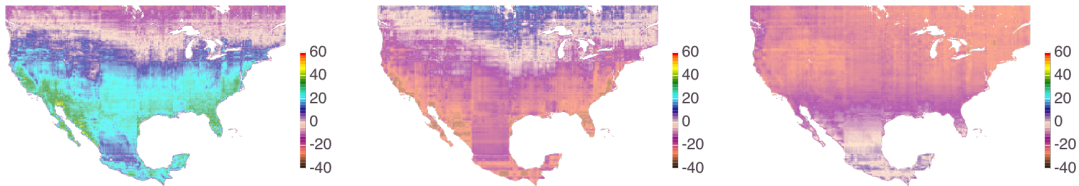


US Climate data: temperature

- ▶ NOAA data set of observed maximum monthly temperatures with spatial resolution of 6km (1/16°)
- ▶ fluctuations in maximum monthly temperatures [$^{\circ}\text{C}$]
- ▶ $(928 \times 614 \times 768)$
(*columns* $\times$ *rows* $\times$ *months*)

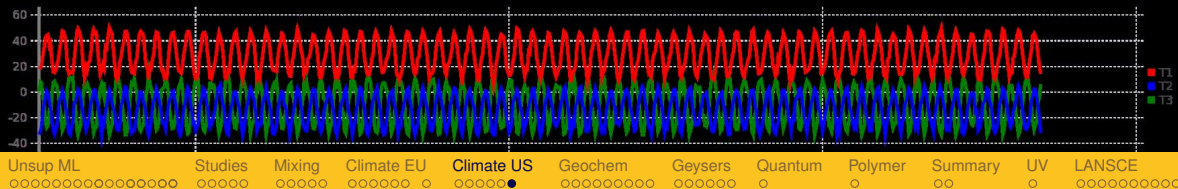


US maximum monthly temperatures: NTF_k results

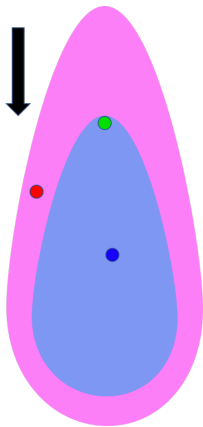


Unsup ML	Studies	Mixing	Climate EU	Climate US	Geochem	Geysers	Quantum	Polymer	Summary	UV	LANSCÉ
oooooooooooooooooooo	ooooo	ooooo	oooooo o	oooo●o	oooooooooo	ooooooo	o	o	oo	o	oooooooooooo

US maximum monthly temperatures: NTF_k results



ML for extracting contaminant plumes



Unsup ML
oooooooooooooooo

Studies
ooooo

Mixing
ooooo

Climate EU
oooooo o

Climate US
oooooo

Geochem
●oooooooo

Geysers
oooooo

Quantum
o

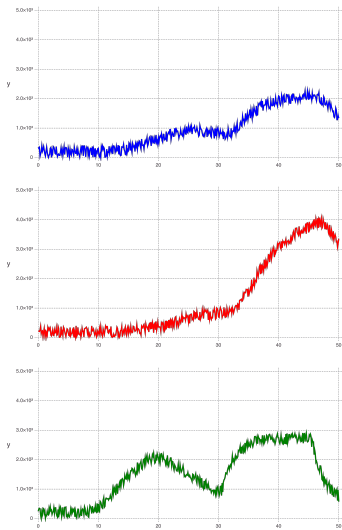
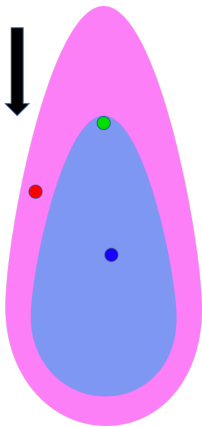
Polymer
o

Summary
oo

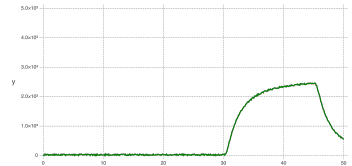
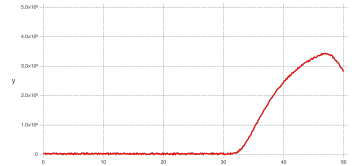
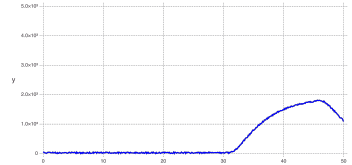
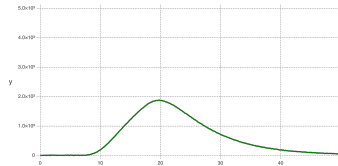
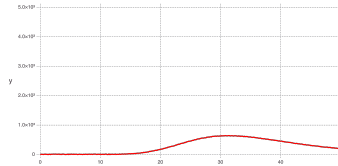
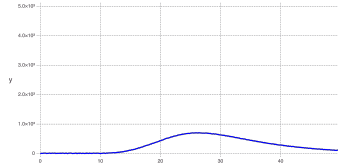
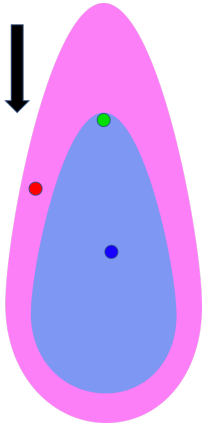
UV
o

LANSCE
oooooooooooo

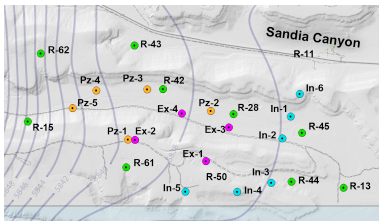
ML for extracting contaminant plumes



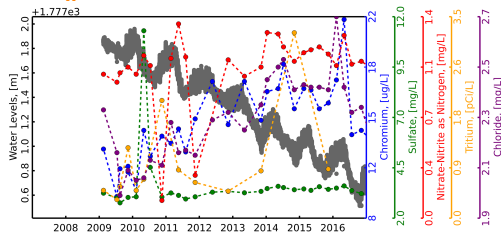
ML for extracting contaminant plumes



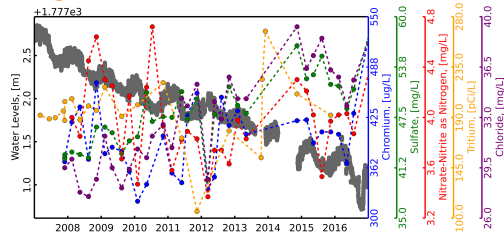
Geochemistry: LANL hydrogeochemical dataset



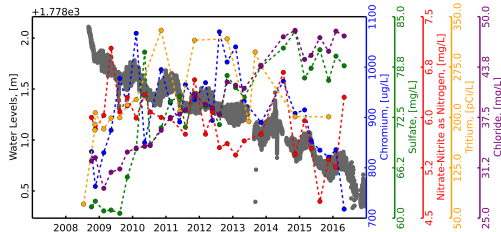
R-44#1



R-28

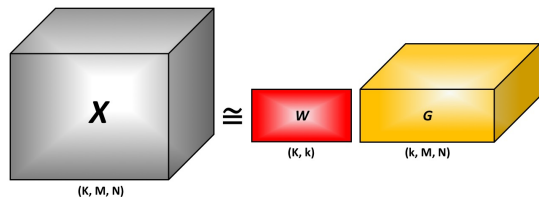


R-42



$(18 \times 8 \times 12)$ tensor (*wells* $\times$ *species* $\times$ *years*)

Geochemistry: Nonnegative Tensor Factorization based on Tucker-1 decomposition



- ▶ X : data tensor
- ▶ W : source (groundwater type) matrix (**unknown**)
- ▶ G : mixing tensor (**unknown**)

- ▶ M : number of observation points (wells)
- ▶ N : number of observation times (e.g., 2001, 2002, ..., 2017)
- ▶ K : number of geochemical species observed (e.g., Cr^{6+} , SO_4^{2+} , NO_3^- , etc.)
- ▶ k : number of **unknown** groundwater types mixed at each well

- ▶ **Constraints:**
all tensor/matrix elements ≥ 0

$$\sum_{i=1}^k G_{i,j,t} = 1 \quad \forall j, t$$

NTF_k analysis estimated 7 groundwater types

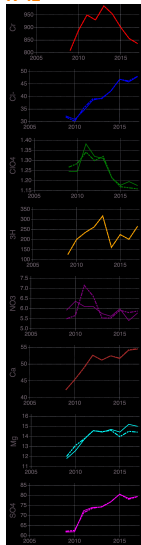
Sources	<i>Cr</i> (<i>μg/L</i>)	<i>Cl</i> ⁻ (<i>mg/L</i>)	<i>ClO</i> ₄ (<i>μg/L</i>)	³ <i>H</i> (<i>pCi/L</i>)	<i>NO</i> ₃ (<i>mg/L</i>)	<i>Ca</i> (<i>mg/L</i>)	<i>Mg</i> (<i>mg/L</i>)	<i>SO</i> ₄ (<i>mg/L</i>)
S1	2970.00	63.00	0.00	0.00	14.00	73.00	25.00	170.00
S5	21.00	51.00	0.00	950.00	2.40	67.00	15.00	50.00
S6	1.50	64.00	0.00	0.00	2.80	51.00	10.00	68.00
S2	0.79	0.35	14.00	0.00	0.50	5.30	1.70	0.60
S4	0.50	0.14	0.00	0.00	10.00	21.00	5.00	10.00
S3 (B)	0.25	3.60	0.00	0.00	0.01	41.00	11.00	0.06
S7 (B)	0.10	0.03	0.00	0.00	0.01	0.40	0.80	0.90

NTF_k estimated concentrations at various wells

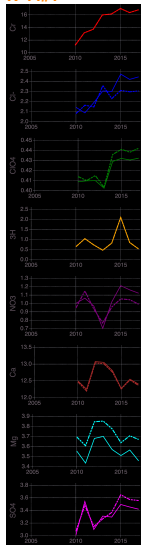
R-28



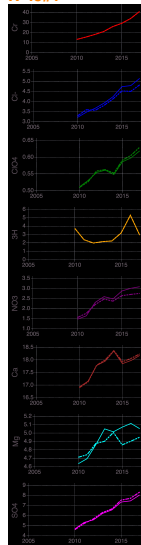
R-42



R-44#1



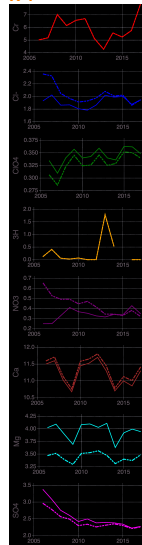
R-45#1



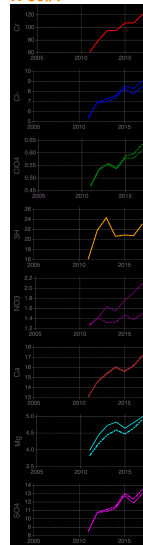
R-11



R-1



R-50#1



Unsup ML

oooooooooooooooo

Studies

ooooo

Mixing

ooooo

Climate EU

oooooo o

Climate US

oooooo

Geochem

ooooo●ooo

Geysers

oooooo

Quantum

o

Polymer

o

Summary

oo

UV

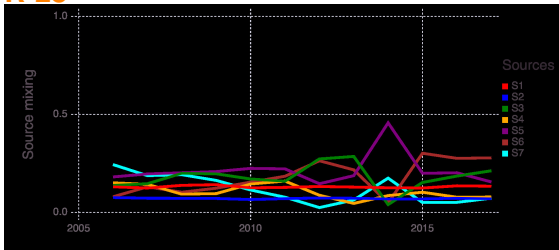
o

LANSCE

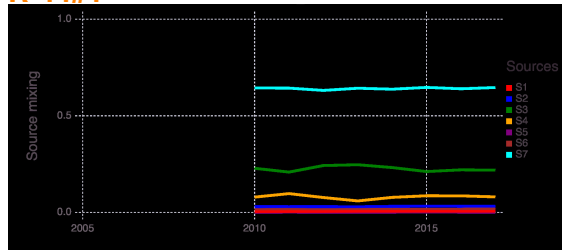
oooooooooooo

NTF_k estimated time-dependent mixing of 7 groundwater types at various wells

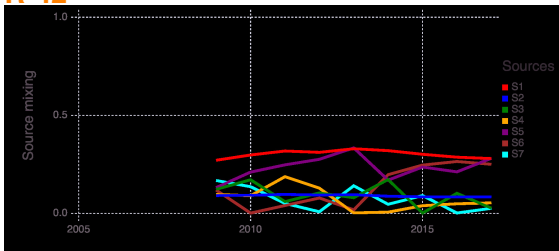
R-28



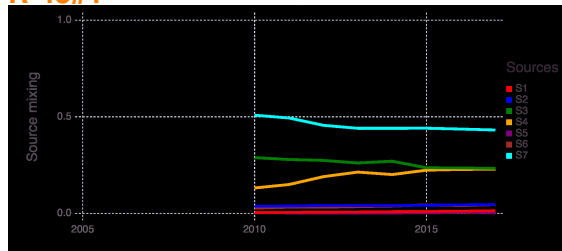
R-44#1



R-42

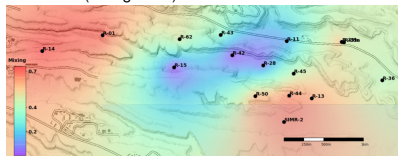


R-45#1

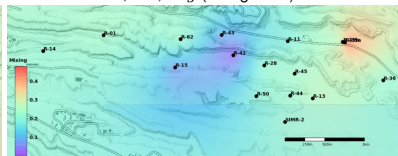


NTF_k identified sources (groundwater types) Jan-Dec 2016

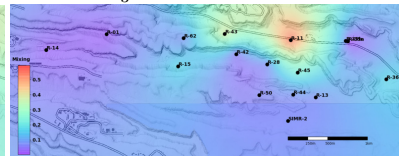
Source 7: (background)



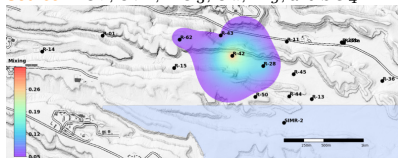
Source 3: Cl^- , Ca , Mg (background)



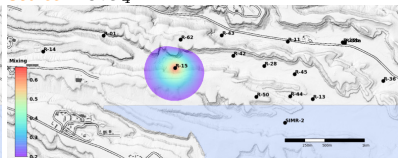
Source 4: NO_3



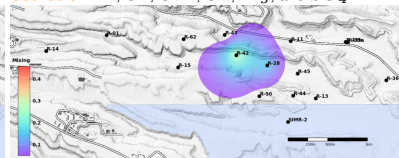
Source 1: Cr , Cl^- , NO_3 , Ca , Mg , and SO_4



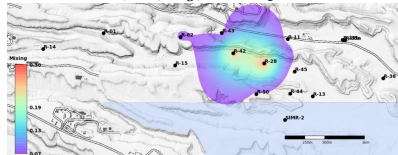
Source 2: ClO_4



Source 5: 3H , Cr , Cl^- , Ca , Mg , and SO_4

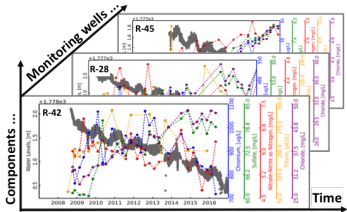


Source 6: Cl^- , Ca , Mg , and SO_4



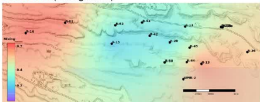
NTF_k analysis of LANL hydrogeochemical datasets

January 2005 – December 2005

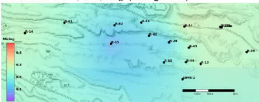


(18 × 8 × 12) tensor
(wells × species × years)

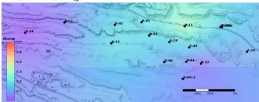
Source 7: (background)



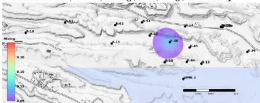
Source 3: Cl⁻, Ca, Mg (background)



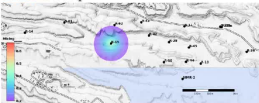
Source 4: NO₃



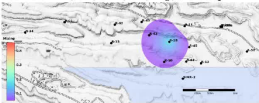
Source 1: Cr, Cl⁻, NO₃, Ca, Mg, and SO₄



Source 2: ClO₄



Source 5: ³H, Cr, Cl⁻, Ca, Mg, and SO₄

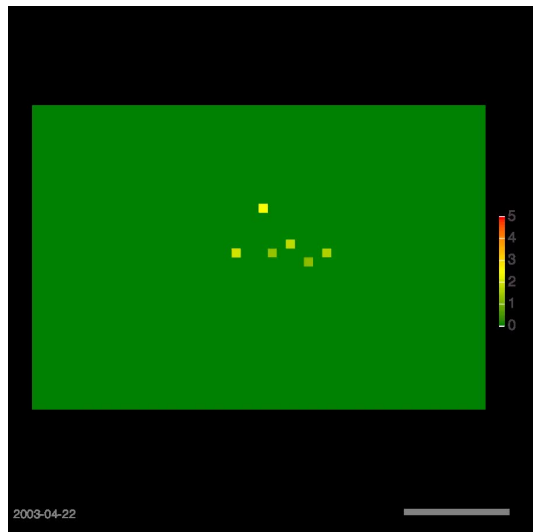


Source 6: Cl⁻, Ca, Mg, and SO₄

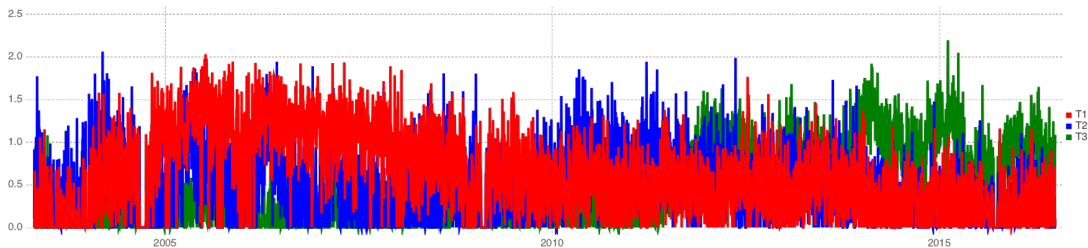
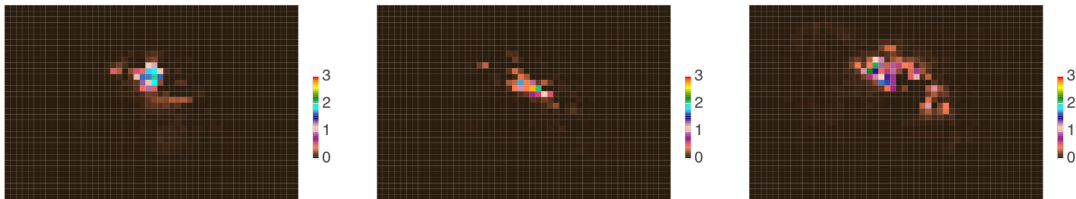


Geysers Geothermal Field Seismic Events

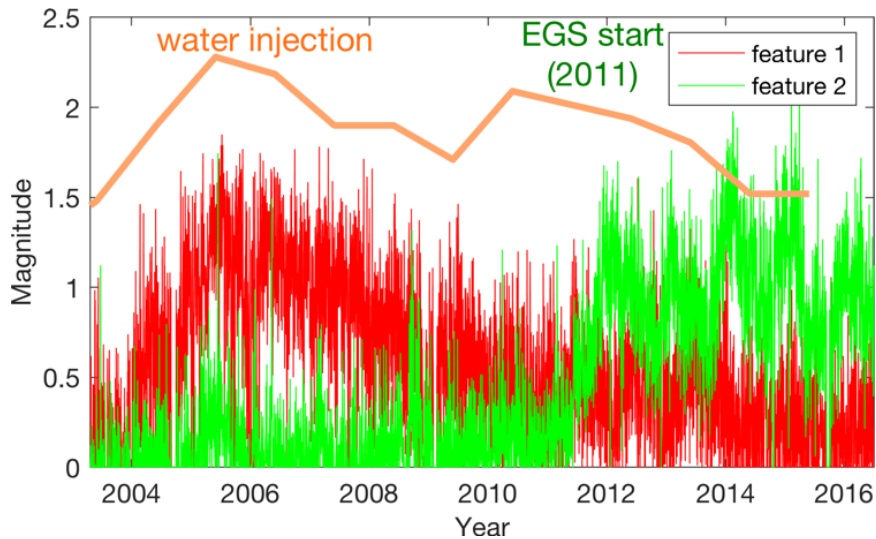
- ▶ 470,263 seismic events have been identified between 2003 and 2016
- ▶ spatial and temporal properties converted to a tensor
- ▶ $(50 \times 34 \times 4818)$
(*columns* $\times$ *rows* $\times$ *days*)
- ▶ **NTF_k** extracts spatial and temporal footprints of dominant features:
 - ▶ EGS injection starts on November 6th, 2011
 - ▶



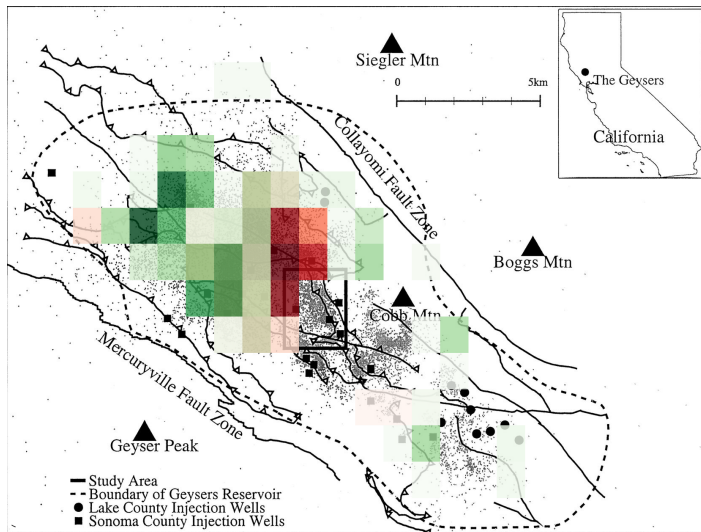
Seismic Events: NTF_k results



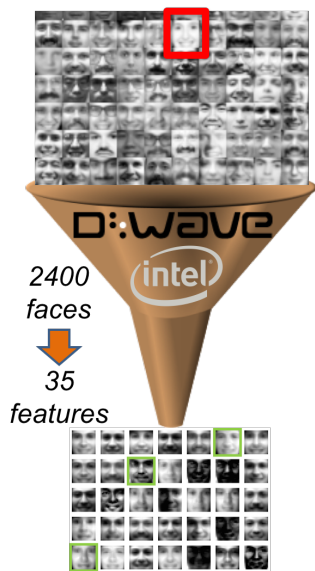
Seismic Events: NTF_k results



Seismic Events: NTF_k results



Quantum Machine Learning using D-Wave



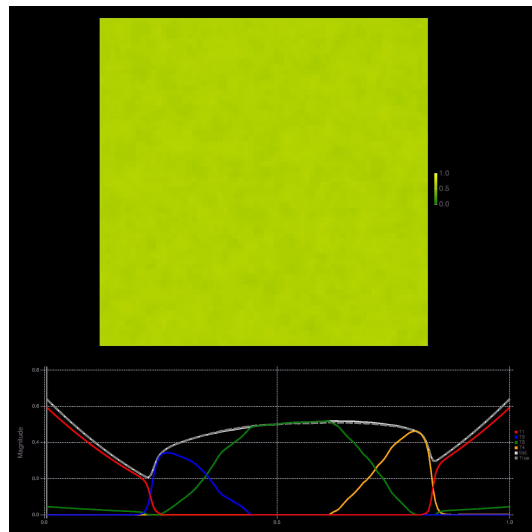
$$\text{Face Image} \approx \text{Basis Image} = \text{Basis Image}_1 + \text{Basis Image}_2 + \text{Basis Image}_3$$

The diagram illustrates the Non-negative Binary Matrix Factorization (NBMF) process. It shows a target face image (highlighted with a red border) being approximated (indicated by the ≈ symbol) by a single basis image (blurred). This is then equated (indicated by the = symbol) to the sum of three basis images (each highlighted with a green border), demonstrating how the target image is decomposed into its constituent features.

- ▶ Performed unsupervised feature extraction using Non-negative Binary Matrix Factorization (NBMF)
- ▶ Coupling quantum (**D-Wave**) and classical computing
- ▶ Analyzed the faces from the original NMF paper published in Nature (Lee & Seung, 1999)

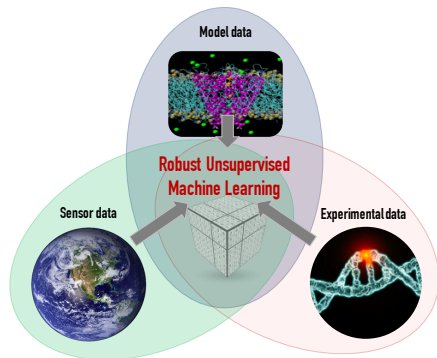
Polymer-chain folding

- ▶ polymer transitions between different states
- ▶ **NTF_k** extracts phase-transition stages
- ▶ $(201 \times 64 \times 64 \times 3) \rightarrow (5 \times 12 \times 12 \times 1)$
(*state* $\times$ *rows* $\times$ *columns* $\times$ *phases*)



Summary

- ▶ Developed **novel** unsupervised ML methods and computational tools based on Nonnegative Factorization (Matrices/Tensors)
- ▶ Our ML methods have been used to solve various real-world problems (brought breakthrough discoveries related to human cancer research)
- ▶ Our ML work already contributes to program development at LANL in mission-critical areas
- ▶ Our goal is to extend the ML methods and tools to solve big ($>$ terabyte-scale) high-dimensional (> 3) data



- ▶ $\text{NMF}_k + \text{ShiftNMF}_k + \text{GreenNMF}_k$
- ▶ NTF_k
- ▶ NBMF: Quantum machine learning using **D-Wave**

- ▶ MADS: Model-Analyses & Decision Support

open-source, version-controlled, high-performance computational framework

<http://mads.lanl.gov>

<http://madsjulia.github.io/Mads.jl>

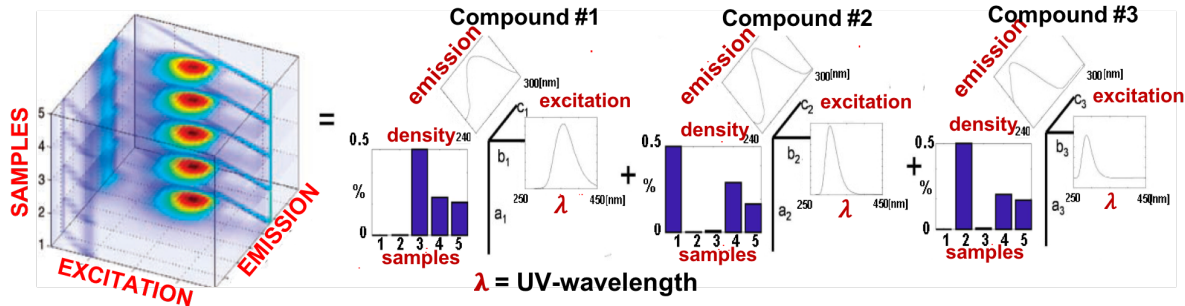
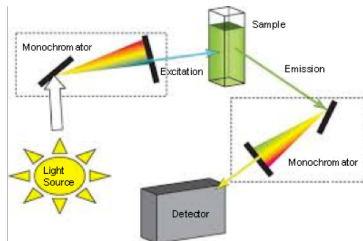


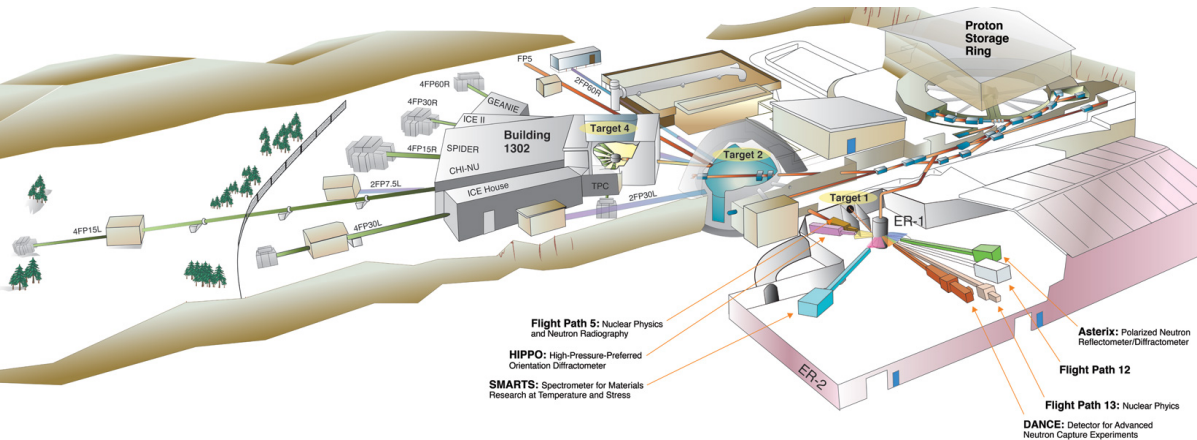
- ▶ Blind Source Separation examples:

http://madsjulia.github.io/Mads.jl/Examples/blind_source_separation

Fluorescence Spectroscopy

- ▶ 5 cuvettes (samples) with unknown mixtures of 3 unknown compounds
- ▶ Emissions and excitations measured ($5 \times 250 \times 200$ tensor)
- ▶ **NTF_k** identifies the mixing ratios of the compounds and their characteristic spectra





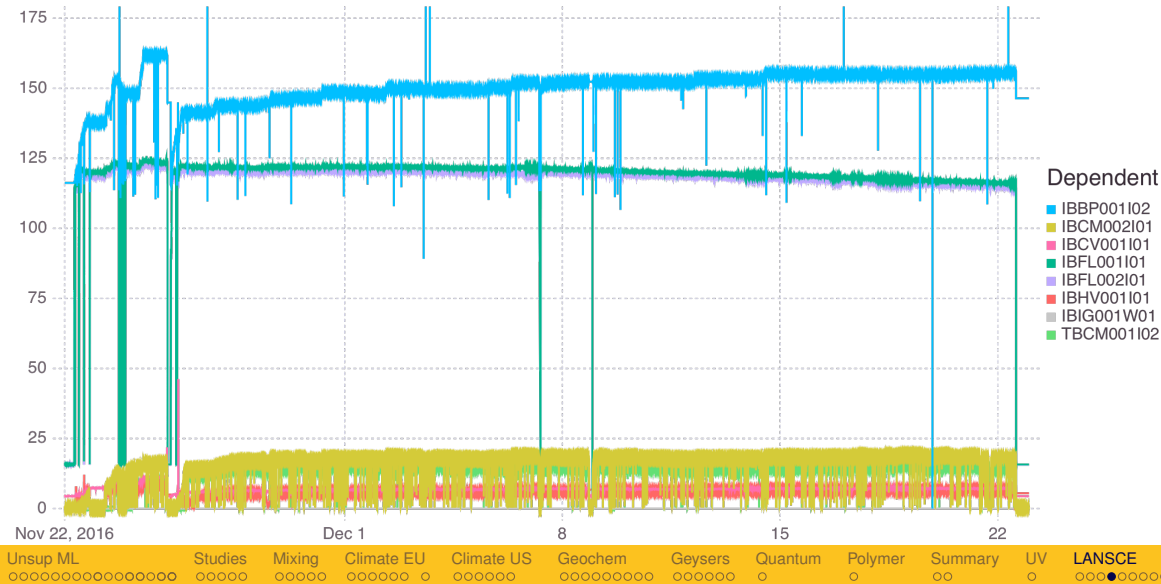
- ▶ Numerous independent variables ... knobs controlling the accelerator performance
15 provided / analyzed
- ▶ Numerous dependent variables ... observations representing the accelerator performance
8 provided / analyzed
- ▶ $312,326 * 23 = 7,183,498 \approx 100$ MB (per month)
Full monthly dataset ≈ 2 GB; For 30 years ≈ 1 TB;
- ▶ Goal **#1**: ML a reduced-order model to simulate the performance (physics model does not exist and cannot be built)
- ▶ Goal **#2**: Use ML to control the knobs to optimize the performance

LANSCÉ: Independent variables = 15

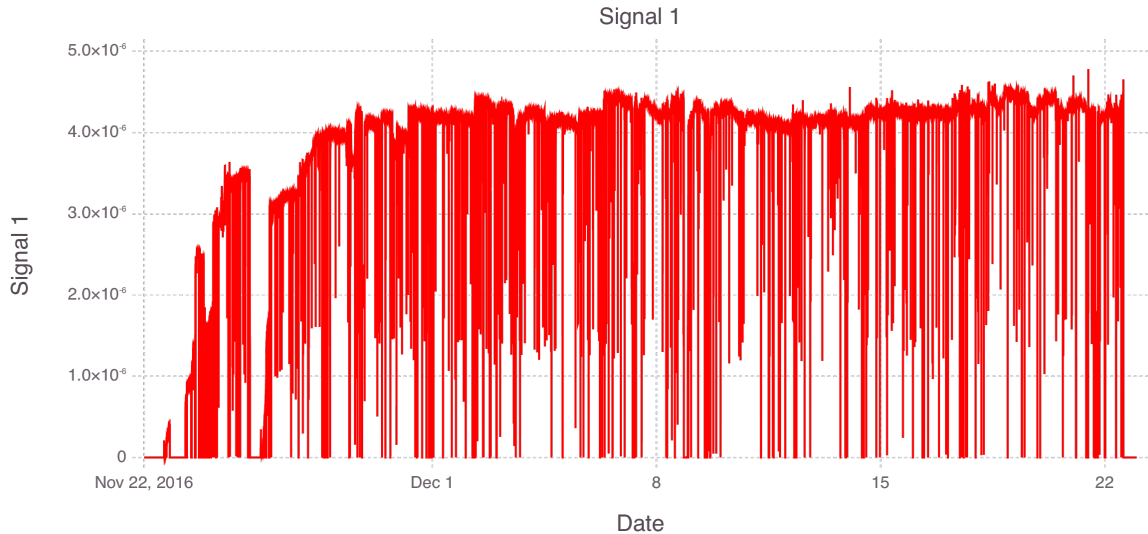


Unsup ML	Studies	Mixing	Climate EU	Climate US	Geochem	Geysers	Quantum	Polymer	Summary	UV	LANSCE
oooooooooooooooooooo	oooooo	oooooo	oooooooo o	oooooooo	oooooooooooo	oooooooo	o	o	oo	o	oo●oooooooo

LANSCE: Dependent variables = 8

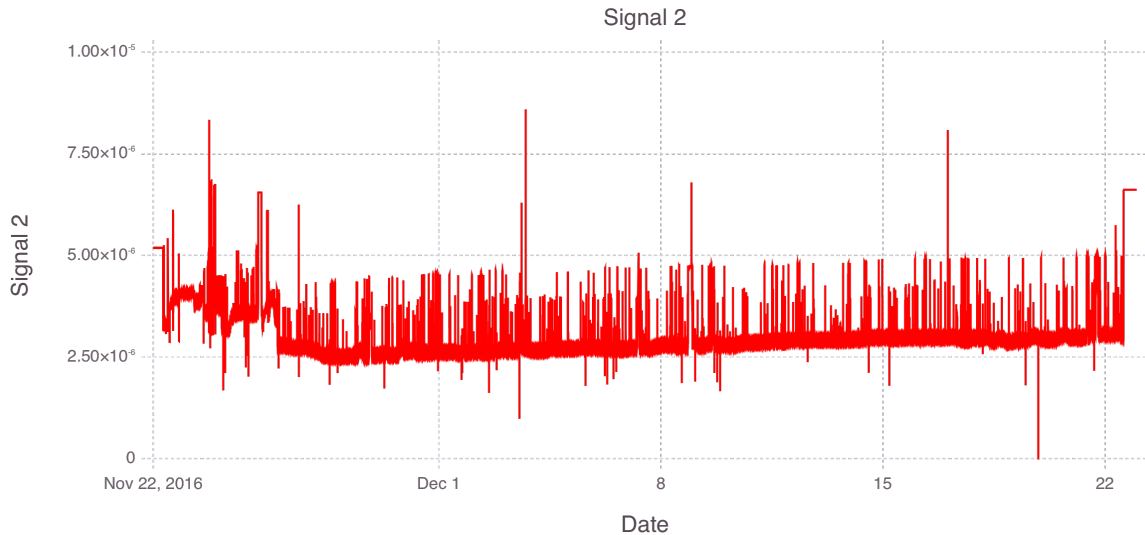


LANSCE: NMF_k estimated signal #1 based on the dependent variables

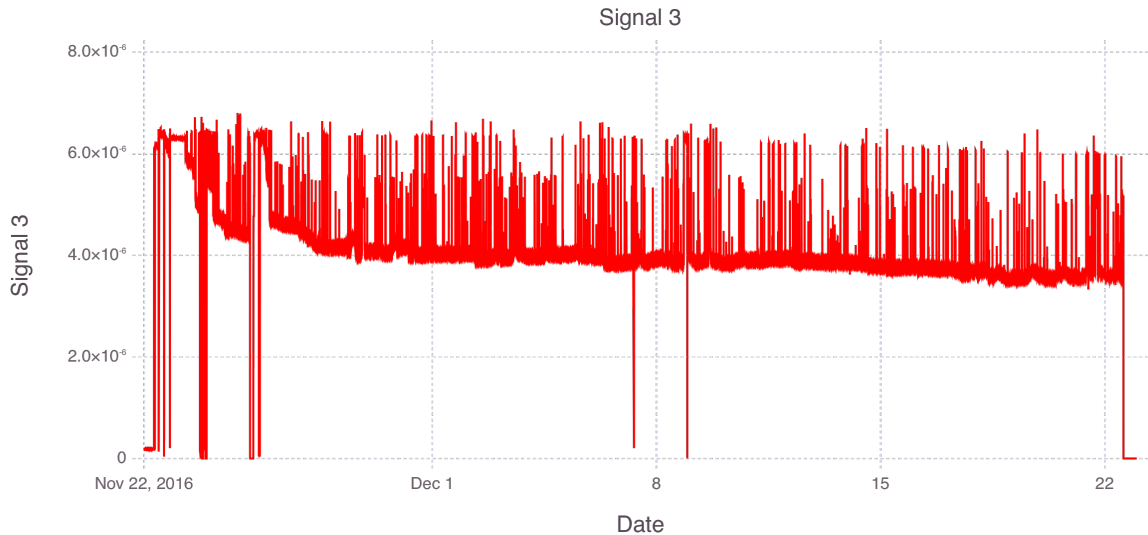


Unsup ML	Studies	Mixing	Climate EU	Climate US	Geochem	Geysers	Quantum	Polymer	Summary	UV	LANSCE
oooooooooooooooo	ooooo	ooooo	oooooo o	oooooo	ooooooooo	oooooo	o	o	oo	o	oooo●ooooo

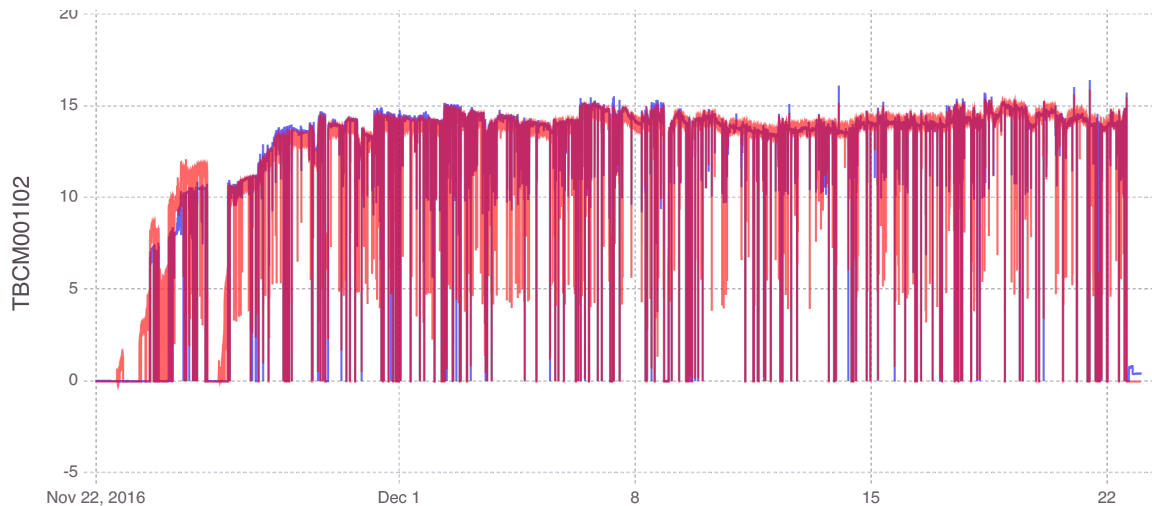
LANSCE: NMF_k estimated signals #2 based on the dependent variables



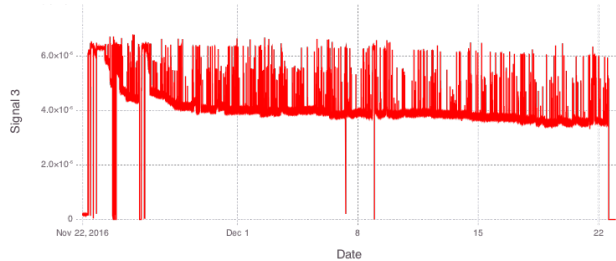
LANSCE: NMF_k Estimated signal #3 based on the dependent variables



LANSCe: NMF_k reconstruction of one of the dependent variables



LANSCE: NMF_k estimated signal #3 vs one of the independent variables



Unsup ML
oooooooooooooooo

Studies
ooooo

Mixing
ooooo

Climate EU
oooooo o

Climate US
ooooooo

Geochem
ooooooooo

Geysers
ooooooo

Quantum
o

Polymer
o

Summary
oo

UV
o

LANSCE
ooooooooo●o

LANSCE: Reduced-order (SVR) Model

