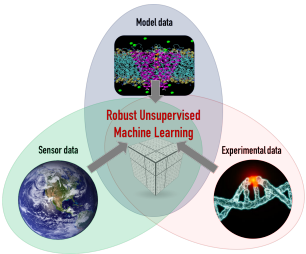
Unsupervised Machine Learning Based on Non-negative Tensor Factorization for Analysis of Filed Data and Simulation Outputs

Velimir V. Vesselinov (monty)
Maruti Mudunuru, Satish Karra, Boian S. Alexandrov, Daniel O'Malley
Los Alamos National Laboratory, NM, USA



Why unsupervised Machine Learning (ML)?

- **Supervised** ML: requires prior categorization (knowledge) of the processed data **random forest**, **neural networks**, **active**, **reinforcement** learning, ...

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

- **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 **random forest**, **neural networks**, **active**, **reinforcement** learning, ...

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

- **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 **random forest**, **neural networks**, **active**, **reinforcement** learning, ...

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

- **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 **random forest**, **neural networks**, **active**, **reinforcement** learning, ...

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

- **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 **random forest**, **neural networks**, **active**, **reinforcement** learning, ...

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

- **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 **random forest**, **neural networks**, **active**, **reinforcement** learning, ...

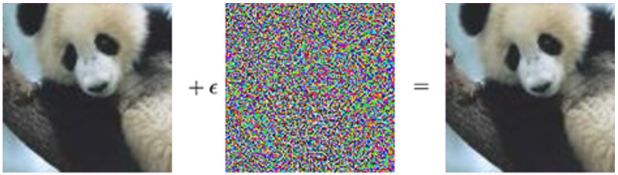
  Example: Recognize images of cats and dogs after extensive training; but cannot recognize horses if not trained

- **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

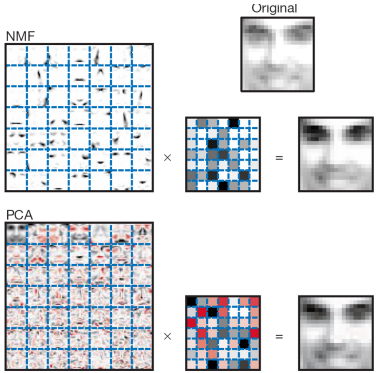
Supervised vs. Unsupervised

Supervised



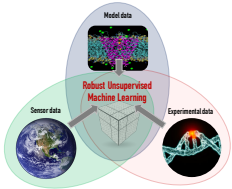
An adversarial input, overlaid on a typical image, can cause a classifier to miscategorize a panda as a gibbon.

Unsupervised

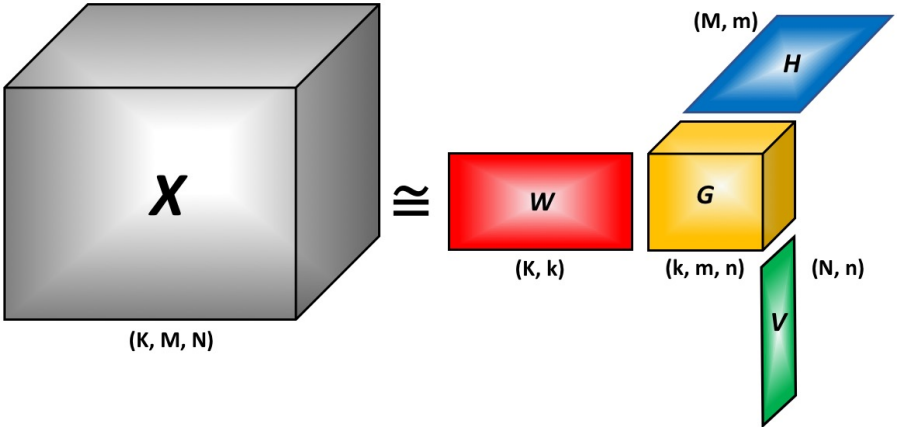


Our Unsupervised Machine Learning Methodology

- We have developed a series of novel Unsupervised Machine Learning methods based on Nonnegative Factorization (Matrix and Tensor) + custom clustering (NMFk / NTFk) that
  - identify **the number of features** (related to the tensor rank) in the data
  - extract **robust features** representing the data
  - extracted features are parts of the data allowing for **intuitive** interpretations
  - applicable to a wide-range of real-world problems (not limited to hydrology)



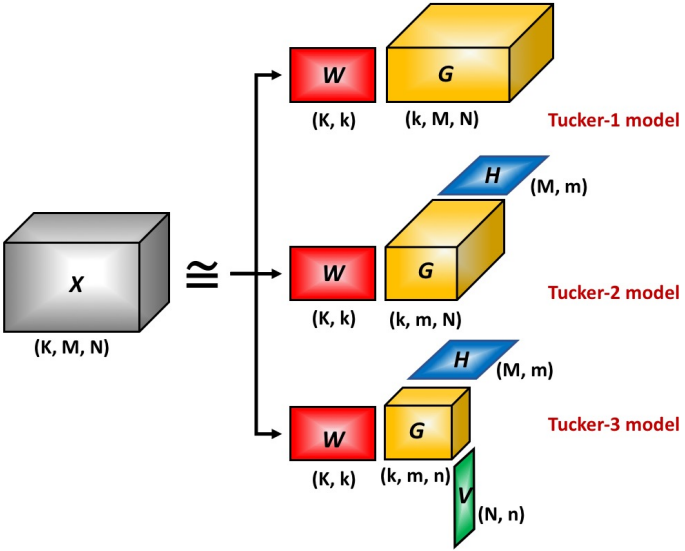
Nonnegative Tucker Tensor factorization (3D Tucker-3)



$X = G \times_k W \times_m H \times_n V$ **Constraints**: G, W, H, V elements ≥ 0

$(K \times M \times N) \rightarrow (k \times m \times n)$ where $K > k$, $M > m$, $N > n$

Nonnegative Tucker Tensor factorization (3D alternatives)



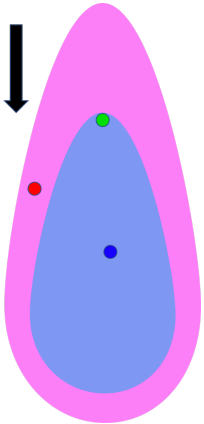
Nonnegative Tensor Factorization (NTFk) Analyses

- **Field Data**: Groundwater contaminant sources ♣ (paper submitted)
- **Lab Data**: X-ray Spectroscopy (paper submitted)
- **Lab Data**: Fluorescence Spectroscopy
- **Model Data**: Reactive mixing $A + B \rightarrow C$ ♣ (paper submitted)
- **Model Data**: Phase separation of co-polymers (paper in preparation)
- **Model Data**: Molecular dynamics of lipids (paper in preparation)
- **Model Data**: Climate model of Europe

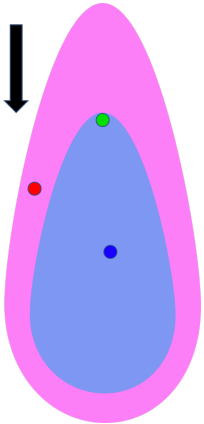
Our Recent ML Publications

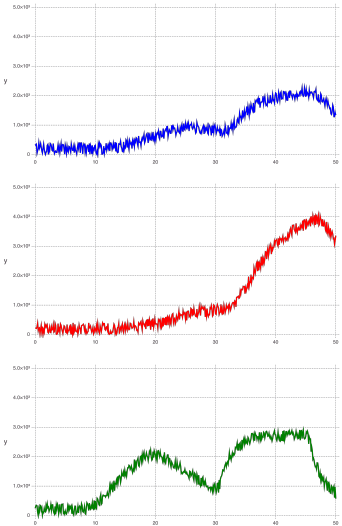
- Stanev, Vesselinov, Kusne, Antoszewski, Takeuchi, Alexandrov, Unsupervised Phase Mapping of X-ray Diffraction Data by Nonnegative Matrix Factorization Integrated with Custom Clustering, **Nature Computational Materials**, (submitted), 2018.
- Vesselinov, Munuduru, Karra, O'Maley, Alexandrov, Unsupervised Machine Learning Based on Non-Negative Tensor Factorization for Analyzing Reactive-Mixing, **Journal of Computational Physics**, (submitted), 2018.
- Vesselinov, O'Malley, Alexandrov, Nonnegative Tensor Factorization for Contaminant Source Identification, **Journal of Contaminant Hydrology**, (submitted), 2018.
- O'Malley, Vesselinov, Alexandrov, Alexandrov, Nonnegative/binary matrix factorization with a D-Wave quantum annealer, **PLOS ONE**, (submitted), 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.

Geochemistry: ML for extracting contaminant plumes

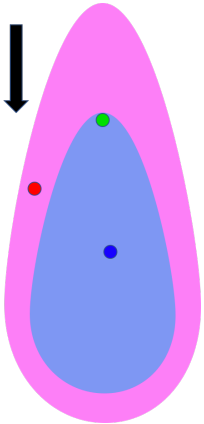


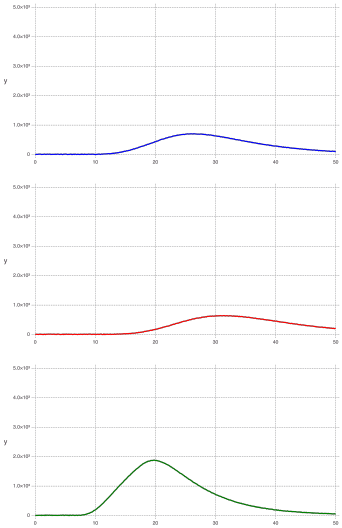
Geochemistry: ML for extracting contaminant plumes

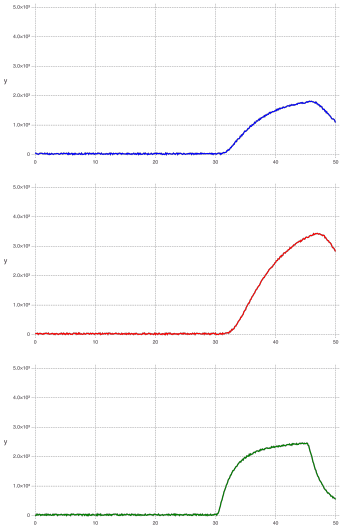




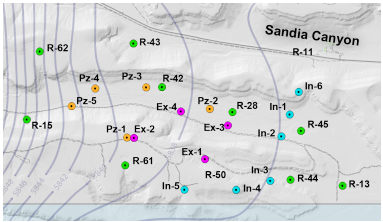
Geochemistry: ML for extracting contaminant plumes



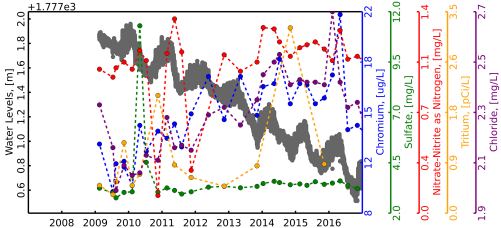




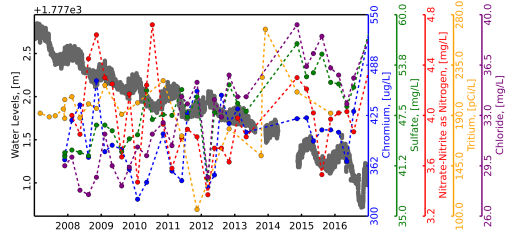
Geochemistry: LANL hydrogeochemical dataset



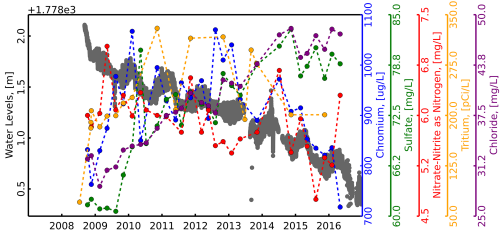
R-44#1



R-28

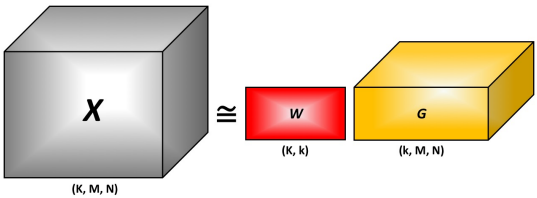


R-42



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

Geochemistry: Nonnegative Tensor Factorization based on Tucker-1 decomposition



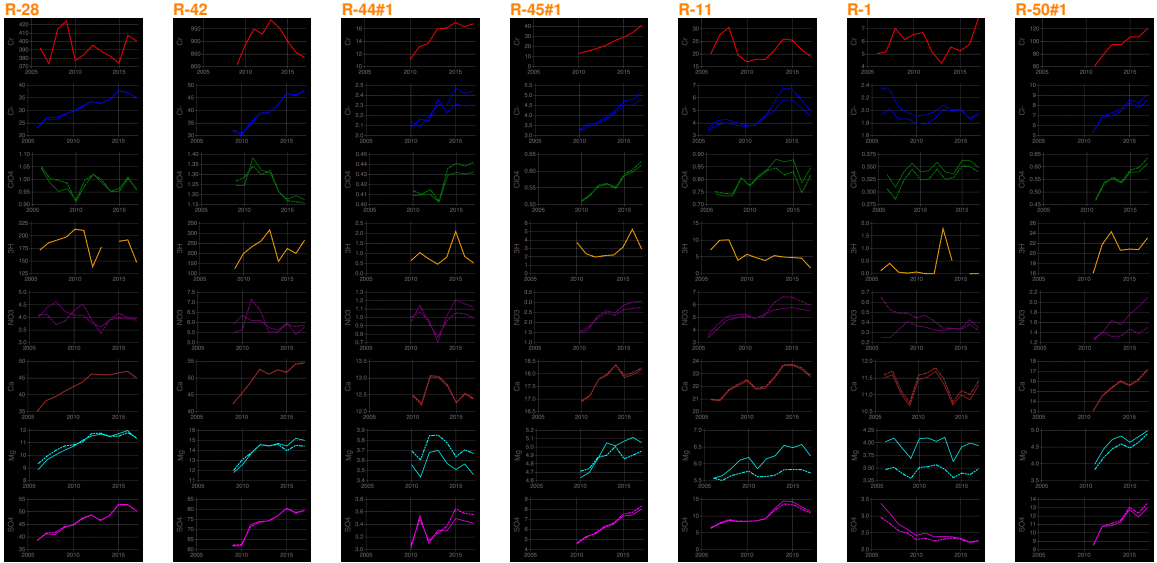
- **X**: data tensor
- **W**: source (groundwater type) matrix (**unknown**)
- **G**: mixing tensor (**unknown**)

- **M**: number of observation points (wells; 18)
- **N**: number of observation times (12: 2005, 2006, ..., 2017)
- **K**: number of geochemical species observed (8: Cr^{6+}, SO_4^{2+}, NO_3^-, etc.)
- **k**: number of **unknown** groundwater types mixed at each well
- **Constraints**:
  all tensor/matrix elements ≥ 0
  $$\sum_{i=1}^{k} G_{i,j,t} = 1 \; \forall j, t$$

NTFk analysis estimated 7 groundwater types

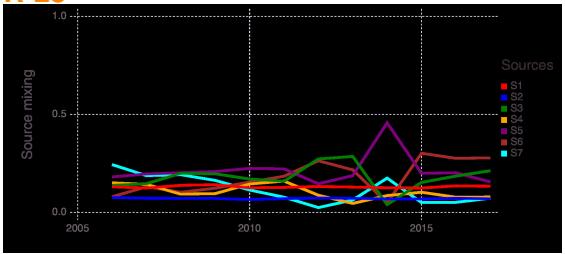
Sources	Cr $(\mu g/L)$	Cl^- (mg/L)	ClO_4 $(\mu g/L)$	^3H (pCi/L)	NO_3 (mg/L)	Ca (mg/L)	Mg (mg/L)	SO_4 (mg/L)
S1	**2970.00**	**63.00**	0.00	0.00	**14.00**	**73.00**	**25.00**	**170.00**
S5	**21.00**	**51.00**	0.00	**950.00**	2.40	**67.00**	15.00	**50.00**
S6	1.50	**64.00**	0.00	0.00	2.80	51.00	10.00	**68.00**
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

NTFk estimated concentrations at various wells

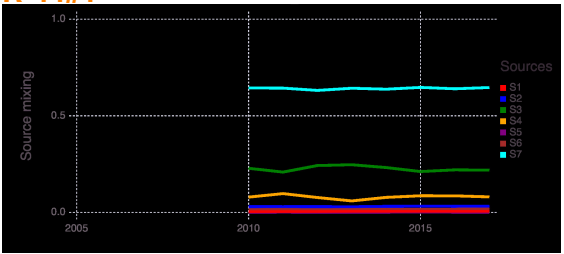


NTFk estimated time-dependent mixing of 7 groundwater types at various wells

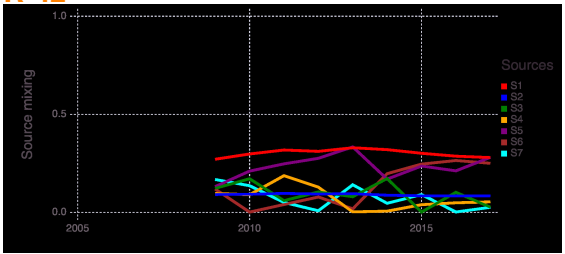
R-28



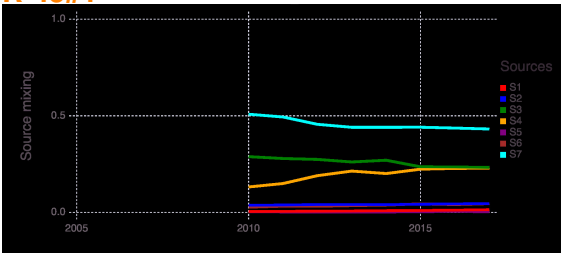
R-44#1



R-42

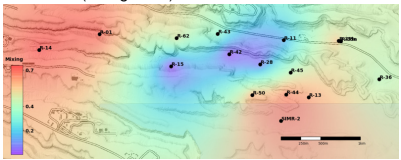


R-45#1

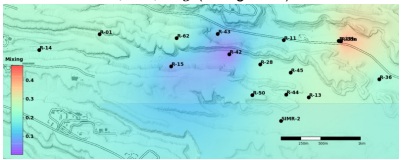


NTFk identified sources (groundwater types) Jan-Dec 2016

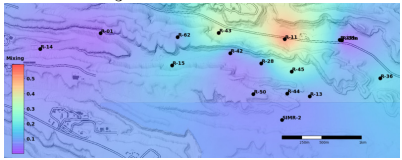
Source 7: (background)



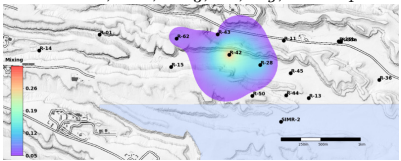
Source 3: Cl^-, Ca, Mg (background)



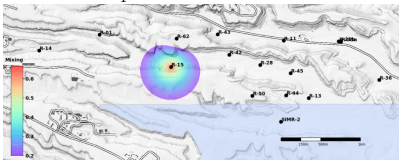
Source 4: NO_3



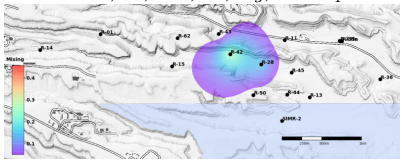
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



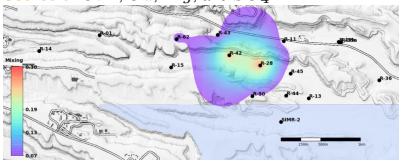
Source 2: ClO_4



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

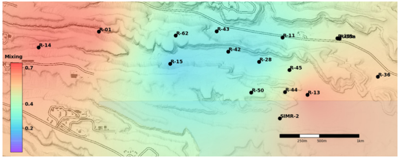


Source 6: Cl^-, Ca, Mg, and SO_4

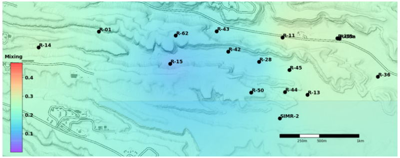


NTFk identified sources (groundwater types) Jan-Dec 2005

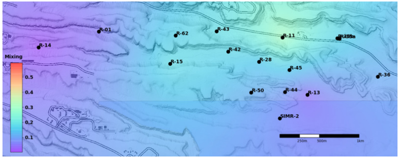
Source 7: (background)



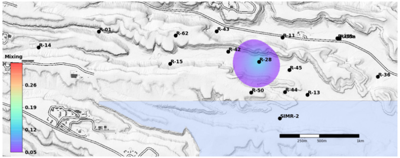
Source 3: Cl^-, Ca, Mg (background)



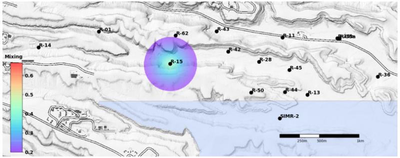
Source 4: NO_3



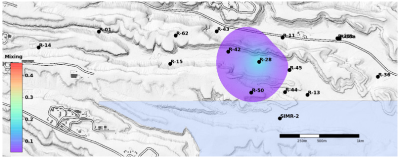
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



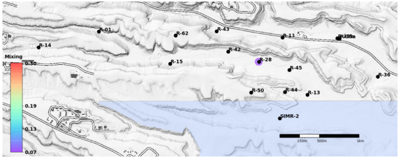
Source 2: ClO_4



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

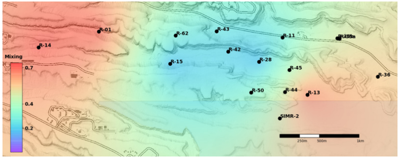


Source 6: Cl^-, Ca, Mg, and SO_4

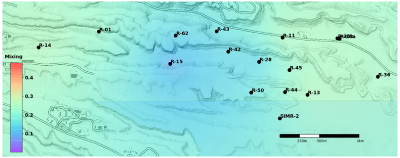


NTFk identified sources (groundwater types) Jan-Dec 2006

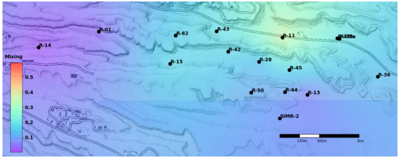
Source 7: (background)



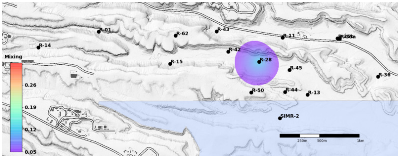
Source 3: Cl^-, Ca, Mg (background)



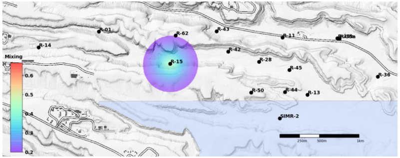
Source 4: NO_3



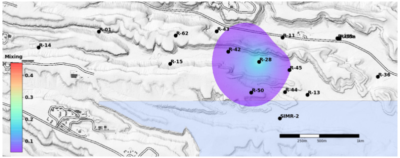
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



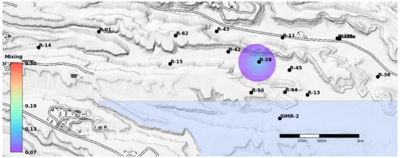
Source 2: ClO_4



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

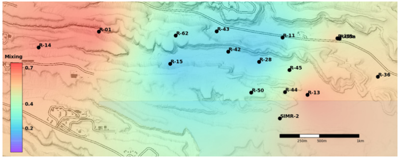


Source 6: Cl^-, Ca, Mg, and SO_4

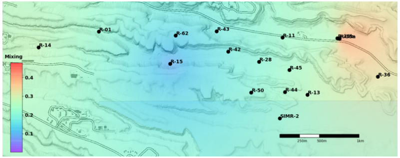


NTFk identified sources (groundwater types) Jan-Dec 2007

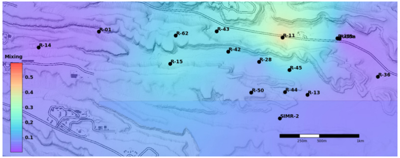
Source 7: (background)



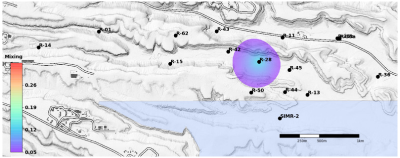
Source 3: Cl^-, Ca, Mg (background)



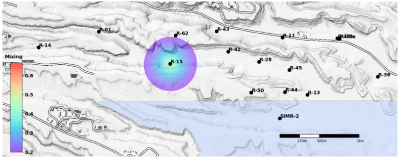
Source 4: NO_3



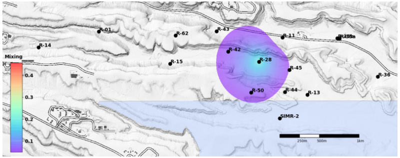
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



Source 2: ClO_4



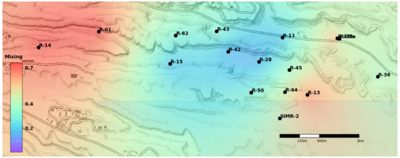
Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4



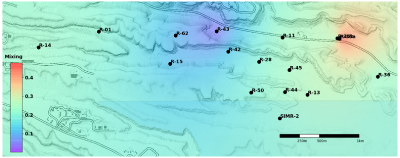
Source 6: Cl^-, Ca, Mg, and SO_4

NTFk identified sources (groundwater types) Jan-Dec 2008

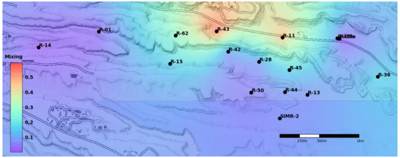
Source 7: (background)



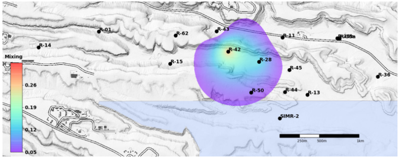
Source 3: Cl^-, Ca, Mg (background)



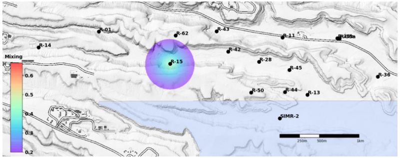
Source 4: NO_3



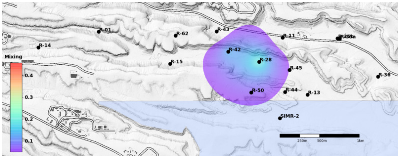
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



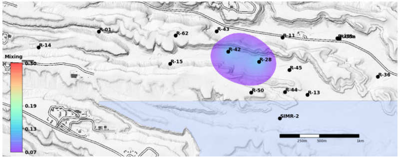
Source 2: ClO_4



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

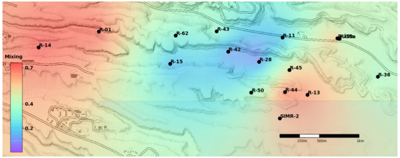


Source 6: Cl^-, Ca, Mg, and SO_4

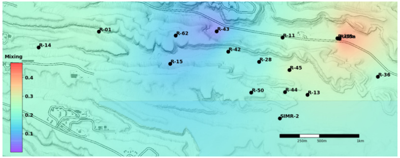


NTFk identified sources (groundwater types) Jan-Dec 2009

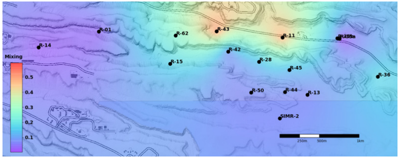
Source 7: (background)



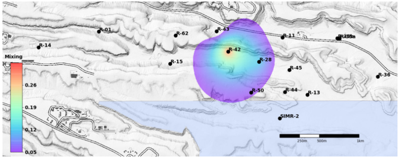
Source 3: Cl^-, Ca, Mg (background)



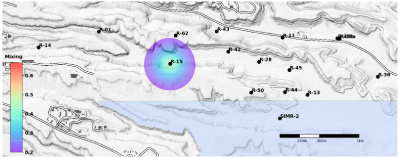
Source 4: NO_3



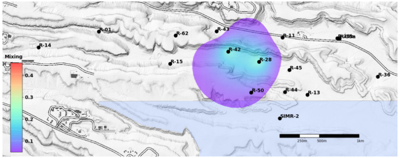
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



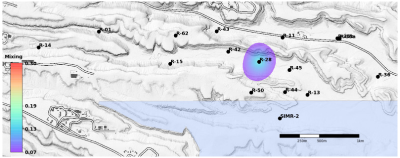
Source 2: ClO_4



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

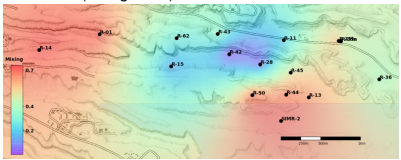


Source 6: Cl^-, Ca, Mg, and SO_4

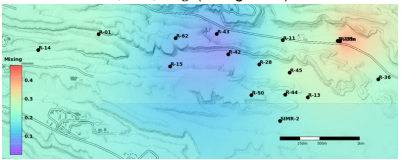


NTFk identified sources (groundwater types) Jan-Dec 2010

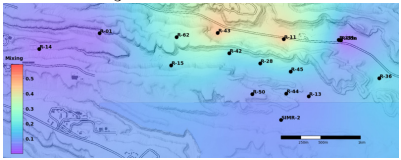
Source 7: (background)



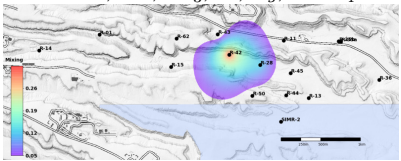
Source 3: Cl^-, Ca, Mg (background)



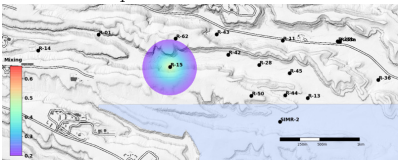
Source 4: NO_3



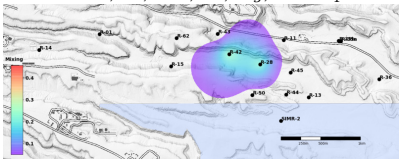
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



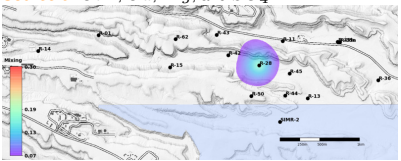
Source 2: ClO_4



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

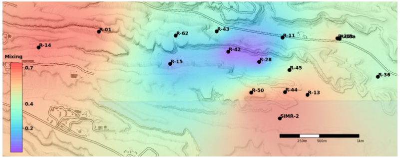


Source 6: Cl^-, Ca, Mg, and SO_4

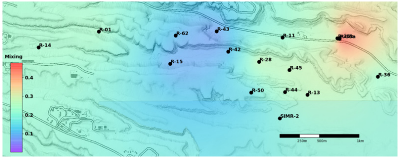


NTFk identified sources (groundwater types) Jan-Dec 2011

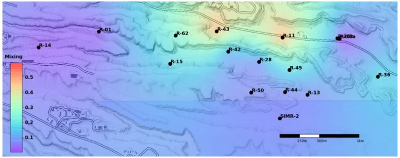
Source 7: (background)



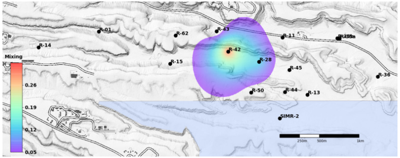
Source 3: Cl^{-}, Ca, Mg (background)



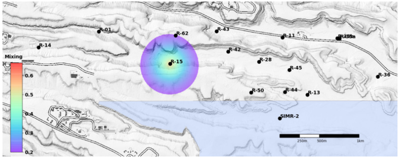
Source 4: NO_3



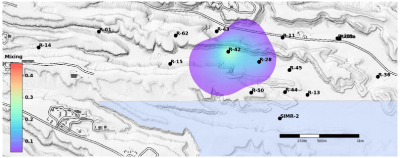
Source 1: Cr, Cl^{-}, NO_3, Ca, Mg, and SO_4



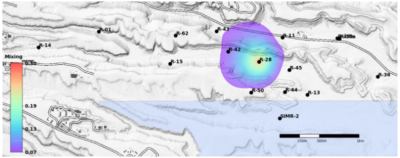
Source 2: ClO_4



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

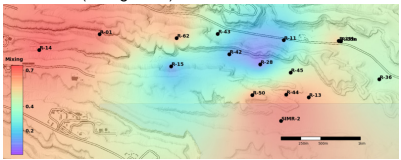


Source 6: Cl^{-}, Ca, Mg, and SO_4

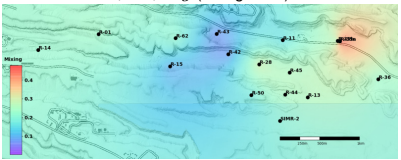


NTFk identified sources (groundwater types) Jan-Dec 2012

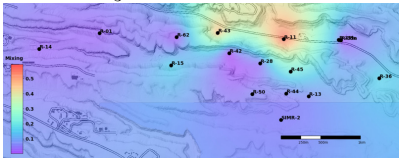
Source 7: (background)



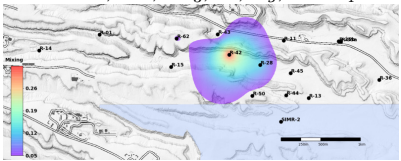
Source 3: Cl^-, Ca, Mg (background)



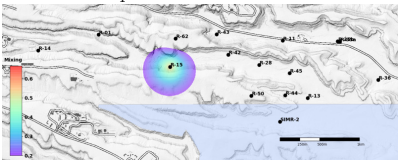
Source 4: NO_3



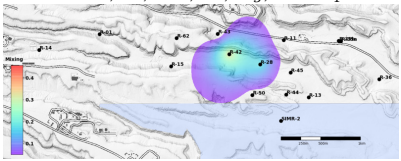
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



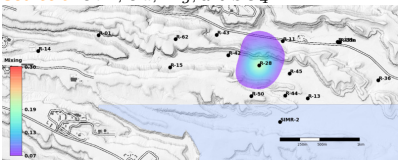
Source 2: ClO_4



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

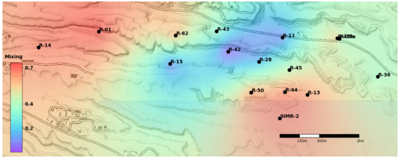


Source 6: Cl^-, Ca, Mg, and SO_4

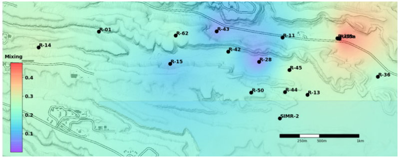


NTFk identified sources (groundwater types) Jan-Dec 2013

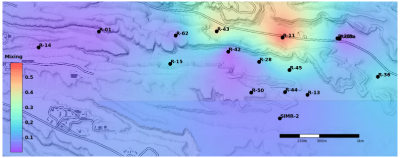
Source 7: (background)



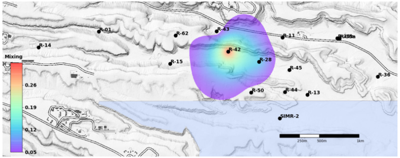
Source 3: Cl^-, Ca, Mg (background)



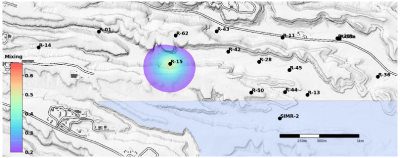
Source 4: NO_3



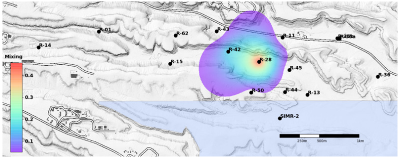
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



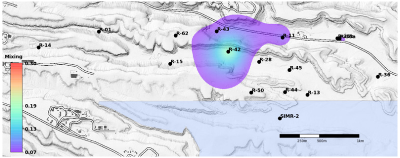
Source 2: ClO_4



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

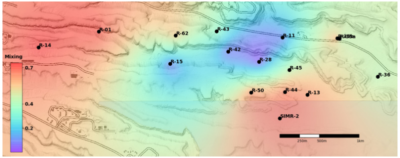


Source 6: Cl^-, Ca, Mg, and SO_4

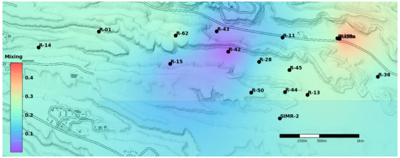


NTFk identified sources (groundwater types) Jan-Dec 2014

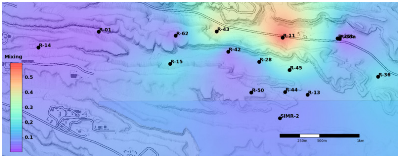
Source 7: (background)



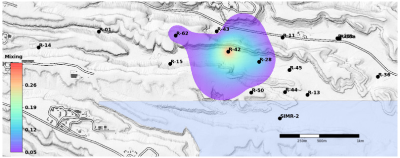
Source 3: Cl^-, Ca, Mg (background)



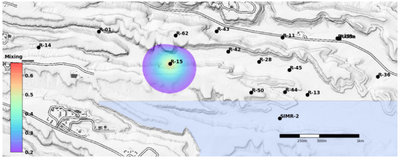
Source 4: NO_3



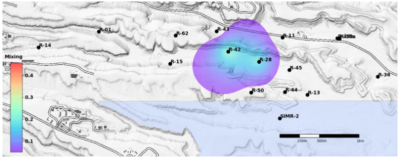
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



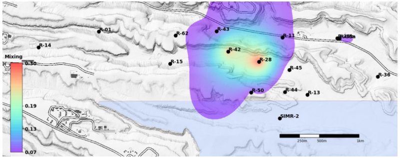
Source 2: ClO_4



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

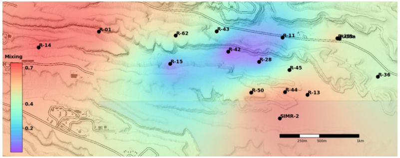


Source 6: Cl^-, Ca, Mg, and SO_4

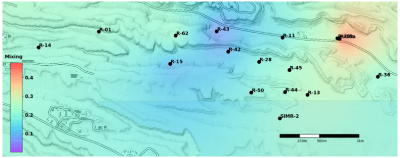


NTFk identified sources (groundwater types) Jan-Dec 2015

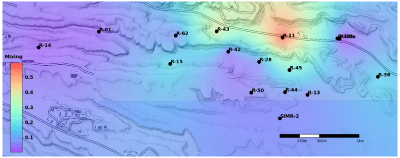
Source 7: (background)



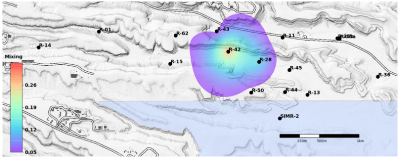
Source 3: Cl^-, Ca, Mg (background)



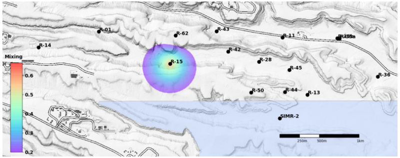
Source 4: NO_3



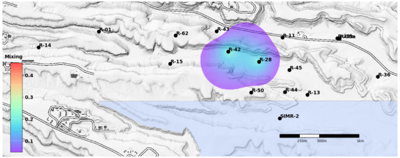
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



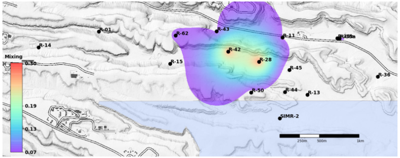
Source 2: ClO_4



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

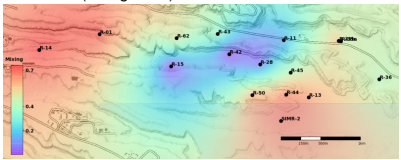


Source 6: Cl^-, Ca, Mg, and SO_4

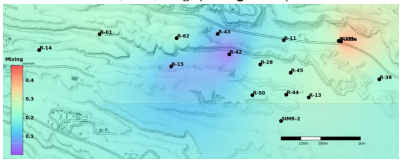


NTFk identified sources (groundwater types) Jan-Dec 2016

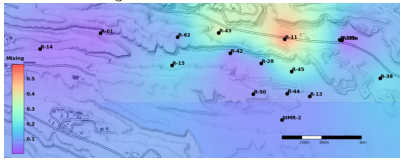
Source 7: (background)



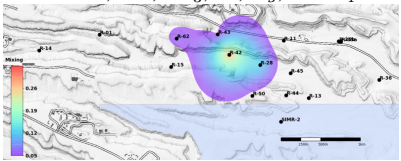
Source 3: Cl^-, Ca, Mg (background)



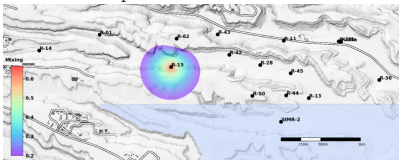
Source 4: NO_3



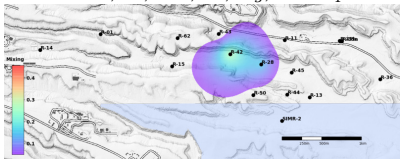
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



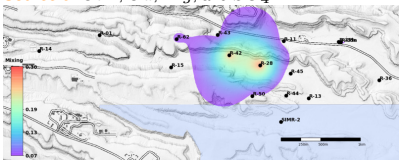
Source 2: ClO_4



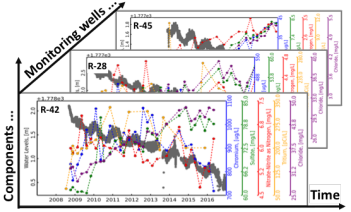
Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4



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

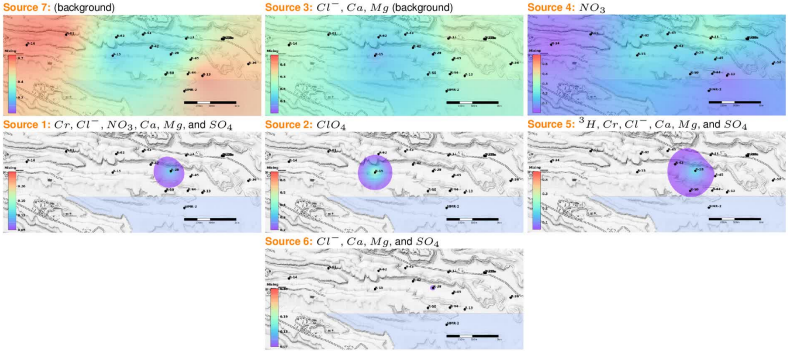


NTFk analysis of LANL hydrogeochemical datasets

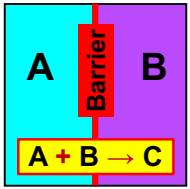


$(18 \times 8 \times 12)$ tensor
$(wells \times species \times years)$

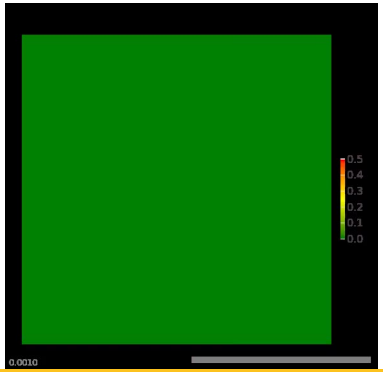
January 2005 – December 2005

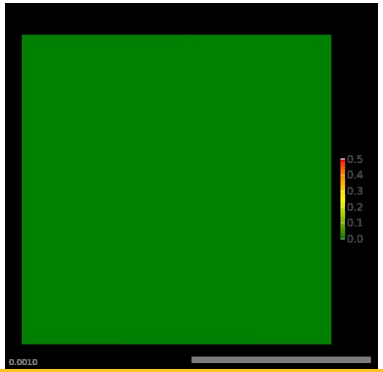


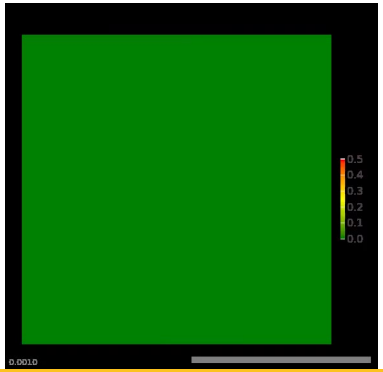
Reactive mixing



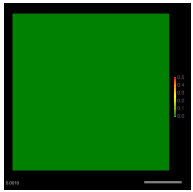
- > 2000 simulations of C concentrations in time/space with varying model inputs representing reactive mixing (5 input model parameters)
- **NTF**k identifies physics processes impacting C concentrations and their relationship to model inputs



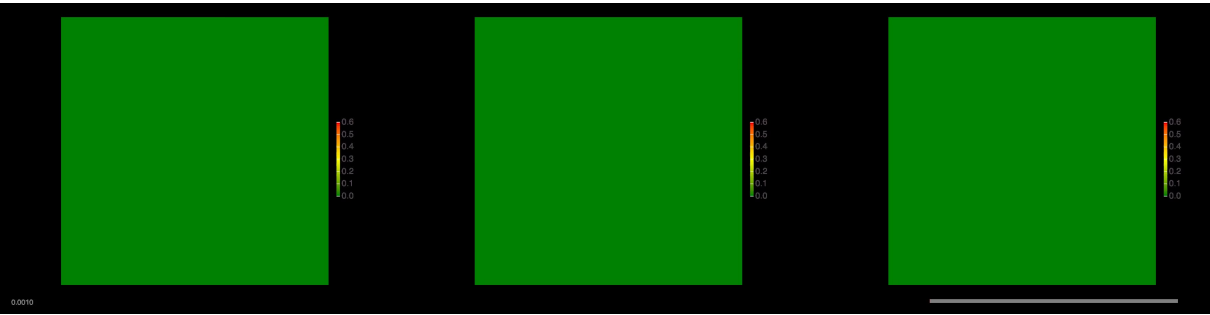




Reactive mixing: NTFk results

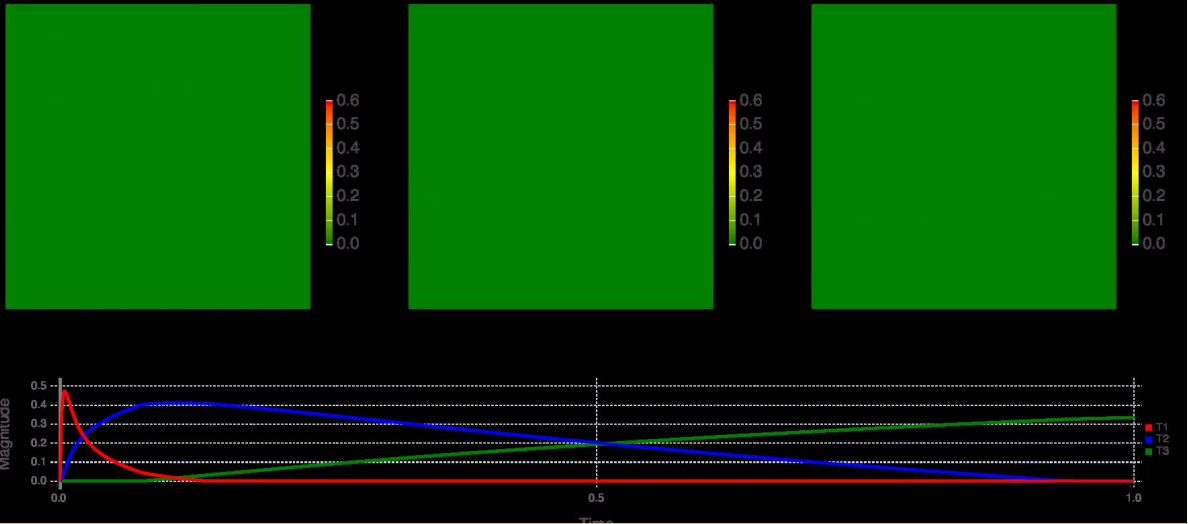


- NTFk extracts the dominant time/space features (**processes** / **vortices**) and compresses the model outputs
- Compression: > 200GB $\rightarrow \sim 70$MB (ratio ~ 3000
  Here, $(1000 \times 81 \times 81) \rightarrow (3 \times 12 \times 13)$ $(time \times rows \times columns)$



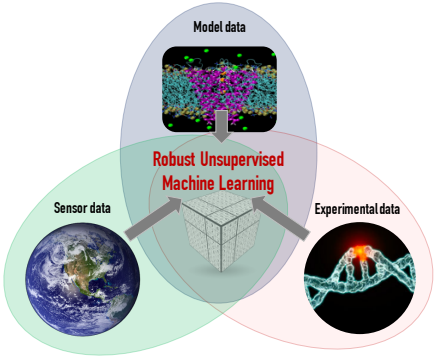
Advection **Dispersion** **Diffusion**

Reactive mixing: NTFk results



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 goal is to extend the ML methods and tools to solve big ($>$ terabyte-scale) high-dimensional (> 5) data



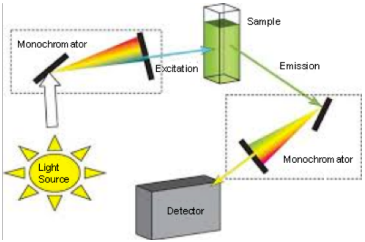
Machine Learning (ML) Algorithms / Codes developed by our team

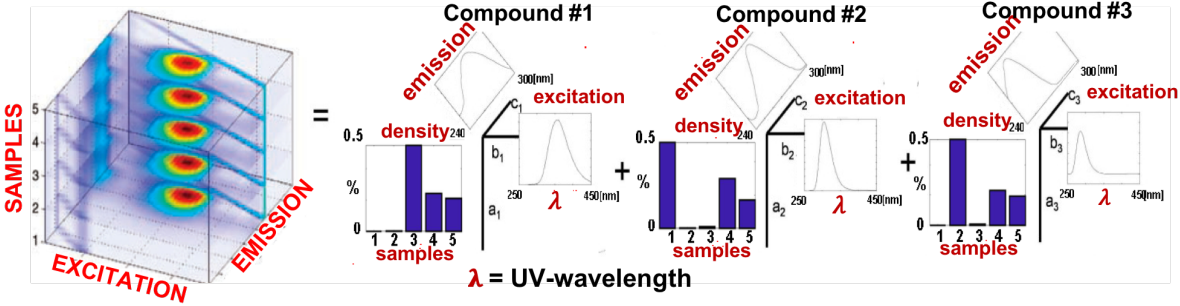
- NMFk + ShiftNMFk + GreenNMFk **(patent)**
- NTFk **(copyright disclosure)**
- NBMF: Quantum machine learning using **D-Wave** **(paper in review)**
- MADS: Model-Analyses & Decision Support
  open-source, version-controlled, high-performance computational framework
  `http://mads.lanl.gov` `http://madsjulia.github.io/Mads.jl`

- Blind Source Separation examples:
  `http://madsjulia.github.io/Mads.jl/Examples/blind_source_separation`

Fluorescence Spectroscopy

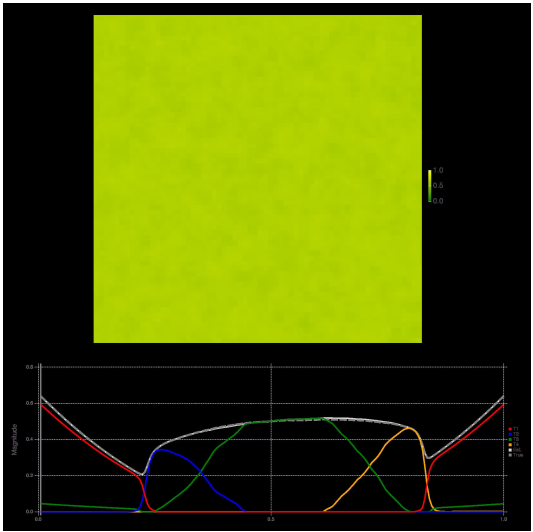
- 5 cuvettes (samples) with unknown mixtures of 3 unknown compounds
- Emissions and excitations measured ($5 \times 250 \times 200$ tensor)
- **NTF**k identifies the mixing ratios of the compounds and their characteristic spectra





Polymer-chain folding

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



« Play/Pause