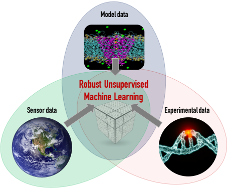
Unsupervised Machine Learning Based on Tensor Factorization

Velimir V. Vesselinov (monty), Daniel O'Malley, Boian S. Alexandrov

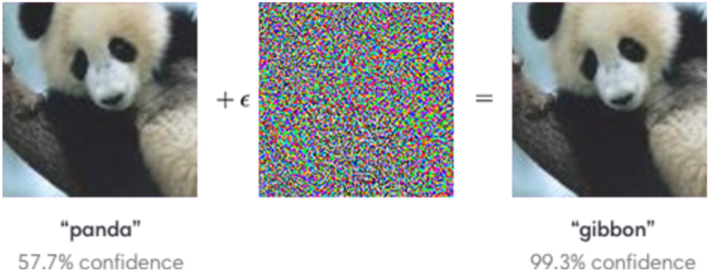
Earth and Environmental Sciences Division — Theoretical Division
Los Alamos National Laboratory, NM 87545, USA



Machine Learning (ML)

- **Supervised** Machine Learning: requires prior categorization of the processed data (can introduce subjectivity)
- **Unsupervised** Machine Learning: discovers hidden features in the processed data without any prior information
- **Deep** Machine Learning: ... coupled supervised and unsupervised techniques

Adversarial Examples



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

- **Data analytics**:
  - Feature extraction (**FE**)
  - Blind source separation (**BSS**)
  - Image recognition
  - Detection of disruptions / anomalies
  - Guide development of physics / reduced-order models representing the data
- **Analyses of model outputs**:
  - Identify dominant processes (features) in the model outputs
  - Guide development of reduced-order models

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
  - identify **the number of robust features** in the data
  - extract **robust features** representing the data
  - extracted features are parts of the data allowing for **intuitive** interpretations
- Selected publications:
  - 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, Water Resources Research, 10.1002/2013WR015037, 2014.
- Machine Learning **Patent**:
  Alexandrov, Vesselinov, Alexandrov, Iliev, Stanev, Source Identification by **Nonnegative Matrix Factorization** Combined with Semi-Supervised Clustering, LANS Ref. No. S133364.000, KS Ref. No. 8472-97415-01, March 2018
- Machine Learning **Copyright Disclosure** :
  Alexandrov, Vesselinov, **Nonnegative Tensor Factorization**, November 2018

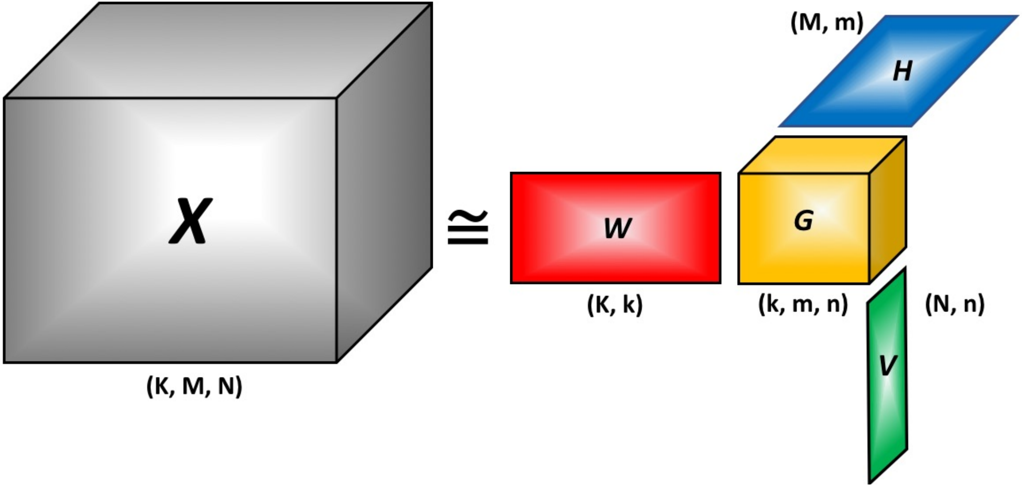
Our Unsupervised Machine Learning Methodology Applications

- **NMFk**: matrix-based (two-dimensional) algorithms (well-tested; widely used)
  - Extract barometric and pumping effects in pressure data
  - Identify and predict processes for optimal control of the LANSCE particle accelerator
  - Characterize materials using X-ray
  - Analyze model predictions of molecular dynamics trajectories
  - Characterize influenza epidemics
  - Extract image features using Quantum Computing (**D-Wave**)
  - Identify cancer signatures in human genomes (**30+** papers in Nature/Science/Cell)
- **NTFk**: tensor-based (high-dimensional) algorithms (actively developed at the moment)
- Here, we present several **NTFk** applications

Matrix/Tensor Factorization (NTF) challenges

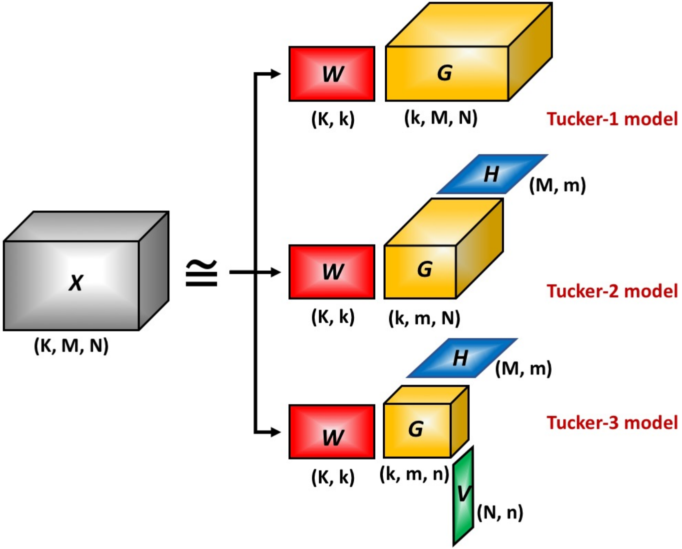
- identifying the number of unknown features (groundwater types) **K** (in **NMF**k, resolved using custom clustering based on the Frobenius norm and cluster Silhouettes; identification under **NTF**k is much more challenging ...)
- solving the constraint optimization problem to estimate matrix/tensor elements
- dealing with large high-dimensional datasets (high-performance computing)
- ...

Tucker-3 tensor factorization (3D case)

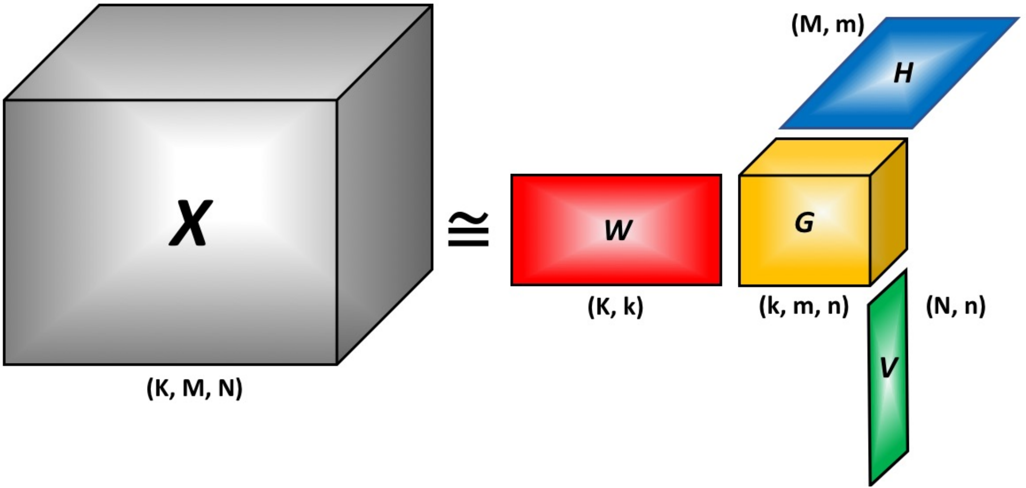


Factorizing all 3 dimensions $(K \to k,\ M \to m,\ N \to n)$

Tucker tensor factorizations (3D case)

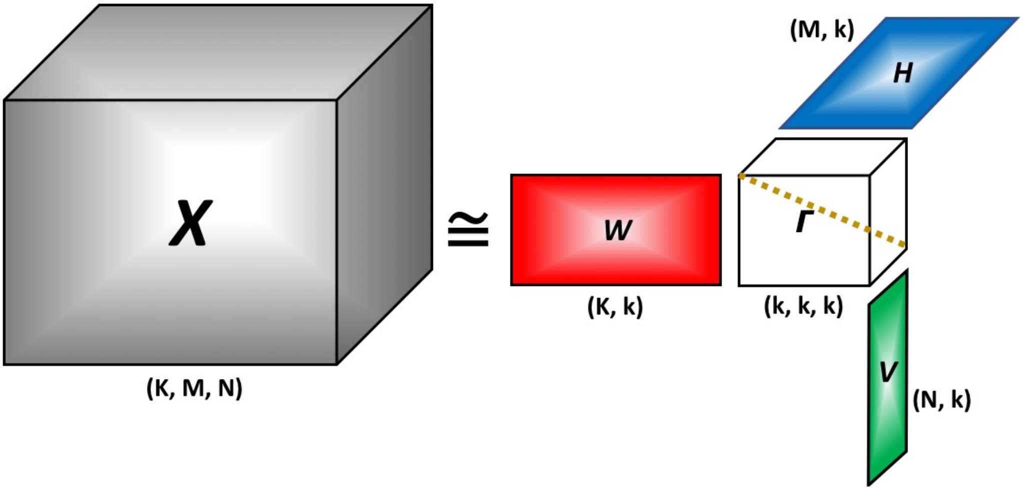


Tucker-3 tensor factorization (3D case)



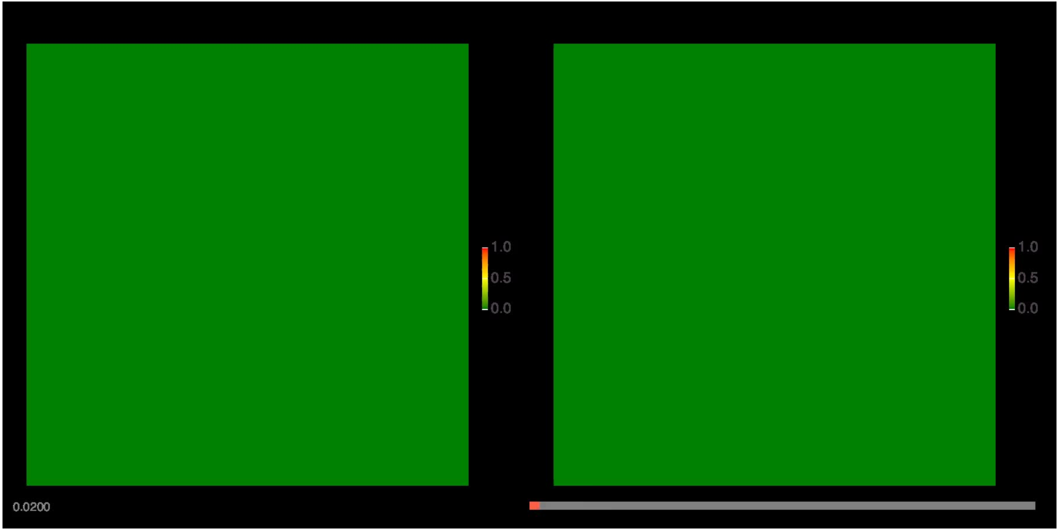
Factorizing all 3 dimensions ($K \to k$, $M \to m$, $N \to n$)

Candecomp/Parafac (CP) tensor factorization (3D case)



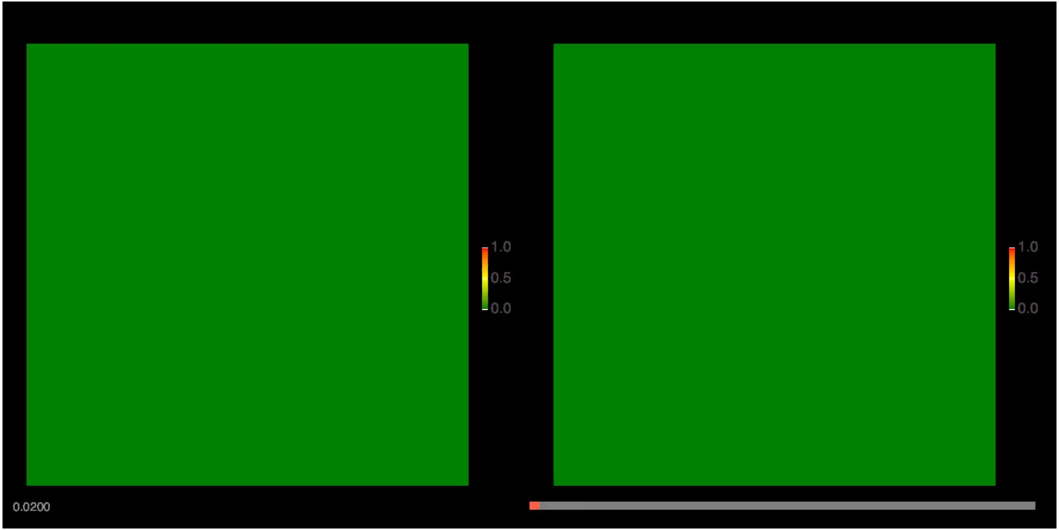
Factorizing all 3 dimensions ($K \to k$, $M \to k$, $N \to k$)

Tucker-3 tensor example

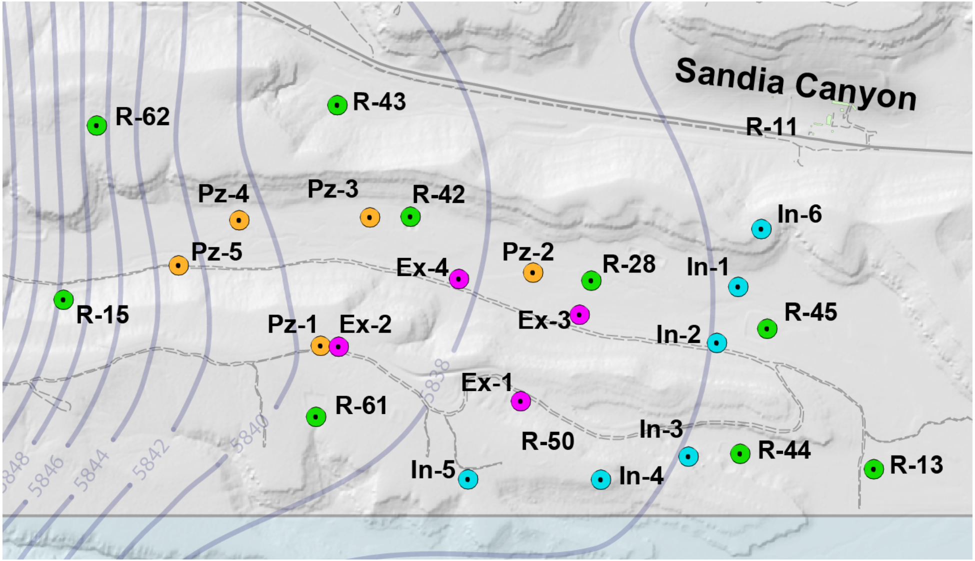


Factorizing all **3** dimensions $(50 \to 6, 50 \to 44, 50 \to 48)$

Tucker-3 tensor example

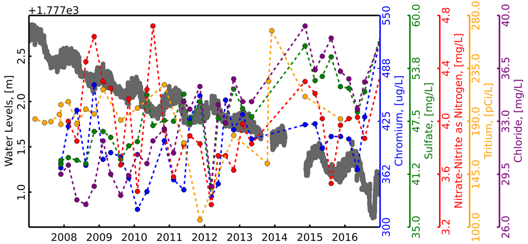


6 groups of swimmers (x); 44 lanes occupied (y); 48 time frames (first/last empty)

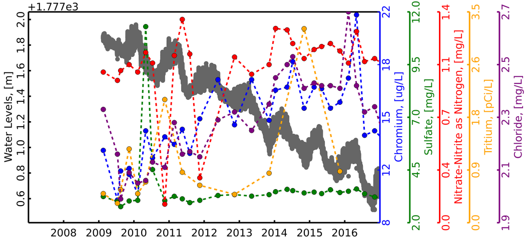

Sandia Canyon
R-62
R-43
R-11
Pz-4
Pz-3
R-42
In-6
Pz-5
Ex-4
Pz-2
R-28
In-1
R-15
Pz-1
Ex-2
Ex-3
R-45
In-2
Ex-1
R-61
R-50
In-3
R-44
R-13
In-5
In-4
5848
5846
5844
5842
5840
5838

LANL hydrogeochemical datasets

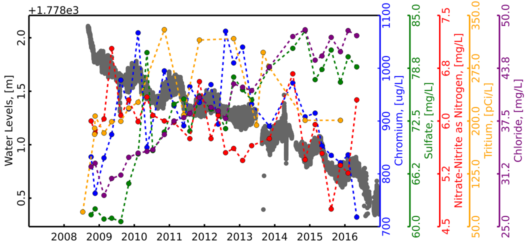
R-28



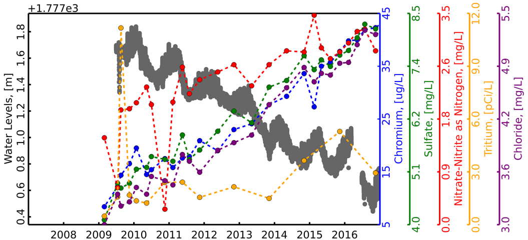
R-44#1



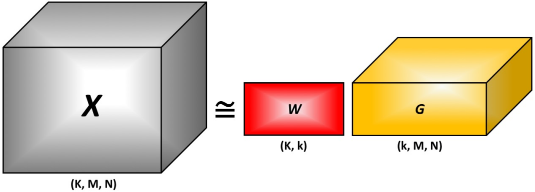
R-42



R-45#1



Geochemistry: Nonnegative Tensor Factorization based on Tucker-1 decomposition



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

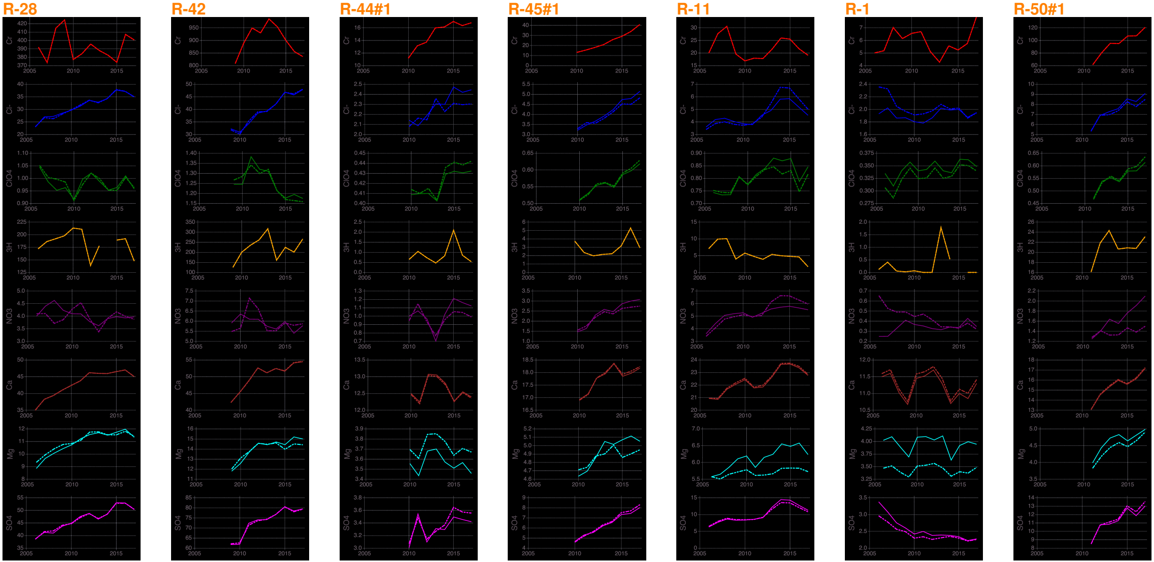
- **M**: number of observation points (wells)
- **N**: number of observation times (e.g, 2001, 2002, ..., 2017)
- **K**: number of geochemical species observed (e.g., Cr^{6+}, SO_4^{2+}, NO_3^-, etc.)
- **k**: number of **unknown** groundwater types mixed at each well
- **Constraints**:
  all tensor/matrix elements ≥ 0

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

Machine Learning analyses estimated 7 groundwater types

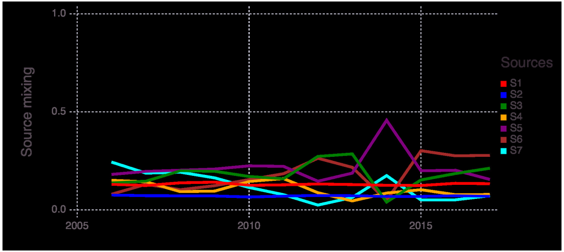
Sources	Cr $(\mu g/L)$	Cl^{-} (mg/L)	ClO_4 $(\mu g/L)$	^{3}H (pCi/L)	NO_3 (mg/L)	Ca (mg/L)	Mg (mg/L)	SO_4 (mg/L)
S1	**2970.00**	**63.00**	0.00	0.00	**14.00**	**73.00**	**25.00**	**170.00**
S5	**21.00**	**51.00**	0.00	**950.00**	2.40	**67.00**	15.00	**50.00**
S6	1.50	**64.00**	0.00	0.00	2.80	51.00	10.00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

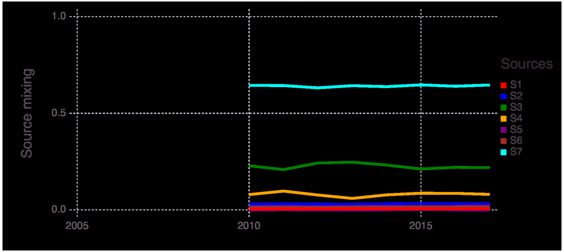


NTFk estimated time-dependent mixing of 7 groundwater types at various wells

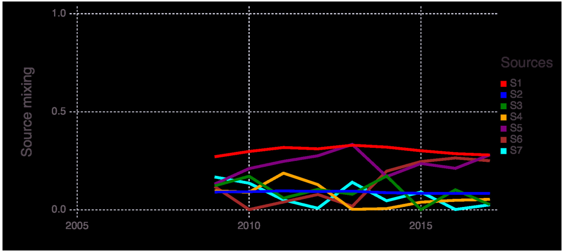
R-28



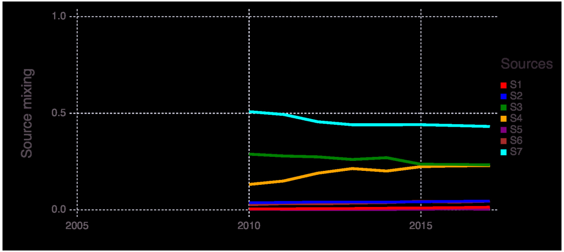
R-44#1



R-42

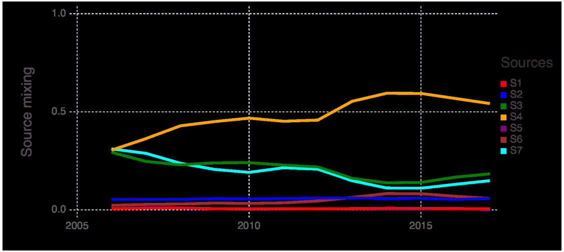


R-45#1

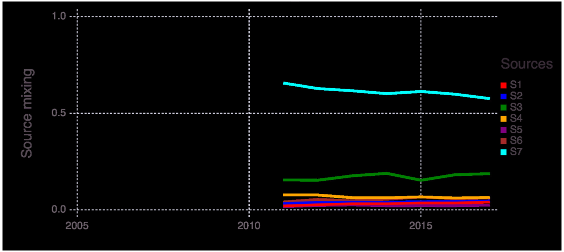


NTFk estimated time-dependent mixing of 7 groundwater types at various wells

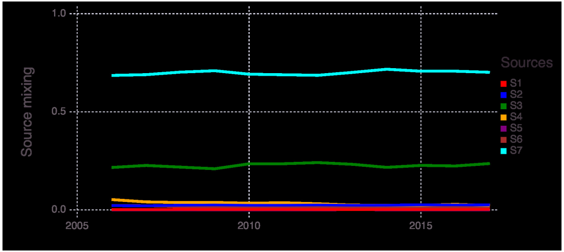
R-11



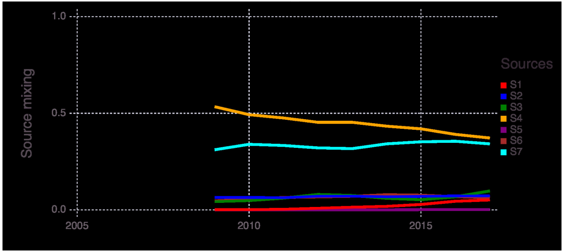
R-50#1



R-1

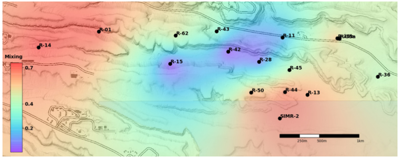


R-43#1

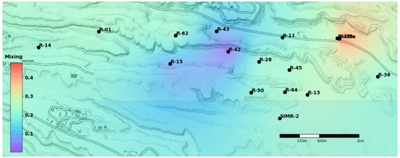


NTFk identified sources (groundwater types) Jan-Dec 2016

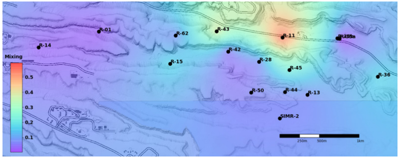
Source 7: (background)



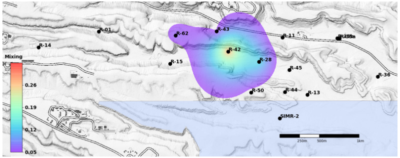
Source 3: Cl^-, Ca, Mg (background)



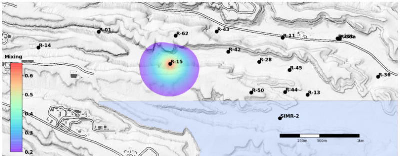
Source 4: NO_3



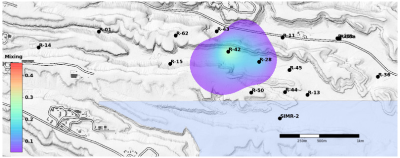
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



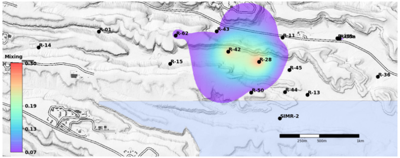
Source 2: ClO_4



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

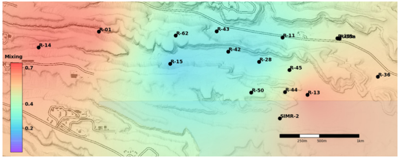


Source 6: Cl^-, Ca, Mg, and SO_4

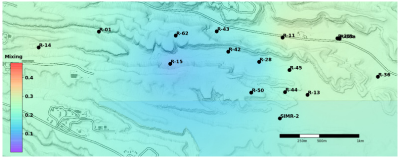


NTFk identified sources (groundwater types) Jan-Dec 2005

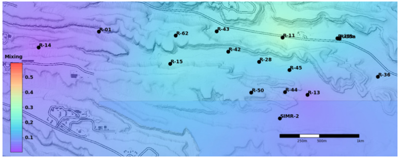
Source 7: (background)



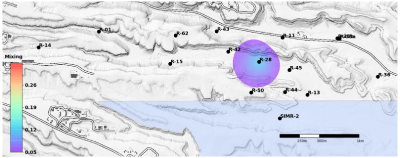
Source 3: Cl^-, Ca, Mg (background)



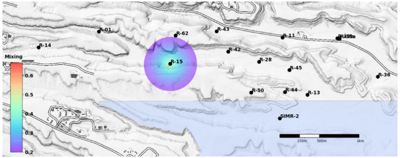
Source 4: NO_3



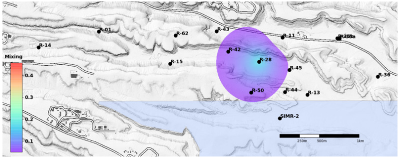
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



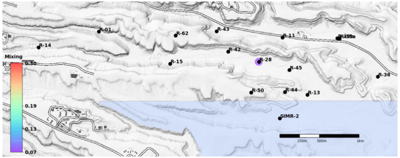
Source 2: ClO_4



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

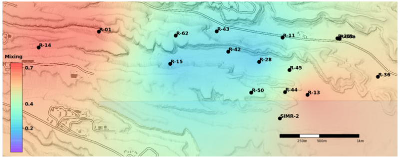


Source 6: Cl^-, Ca, Mg, and SO_4

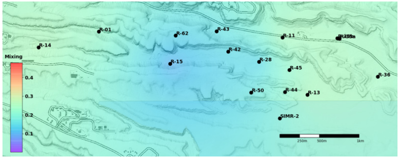


NTFk identified sources (groundwater types) Jan-Dec 2006

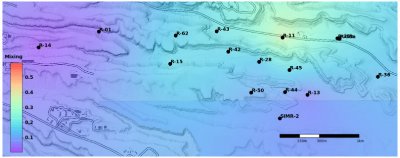
Source 7: (background)



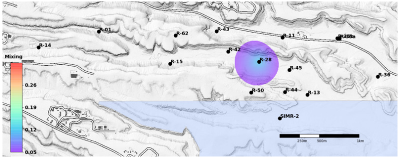
Source 3: Cl^-, Ca, Mg (background)



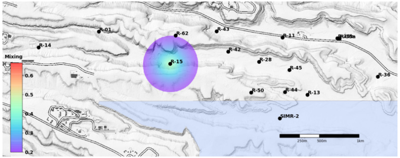
Source 4: NO_3



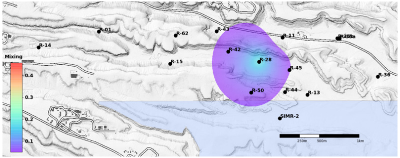
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



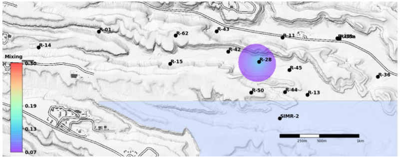
Source 2: ClO_4



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

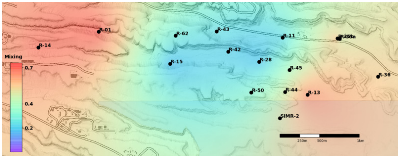


Source 6: Cl^-, Ca, Mg, and SO_4

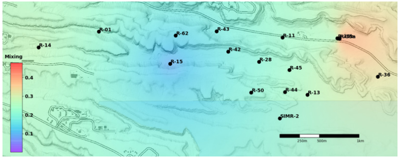


NTFk identified sources (groundwater types) Jan-Dec 2007

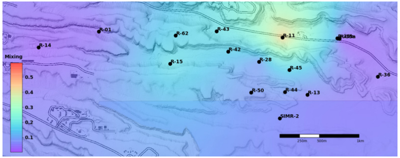
Source 7: (background)



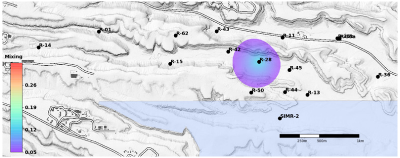
Source 3: Cl^-, Ca, Mg (background)



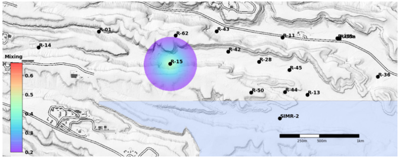
Source 4: NO_3



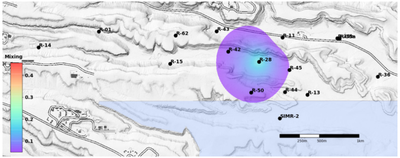
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



Source 2: ClO_4



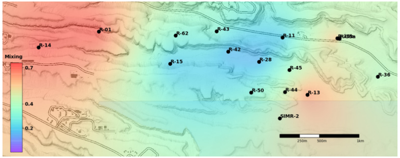
Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4



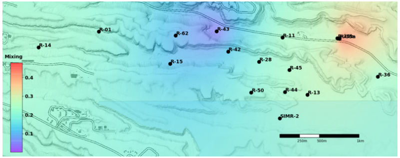
Source 6: Cl^-, Ca, Mg, and SO_4

NTFk identified sources (groundwater types) Jan-Dec 2008

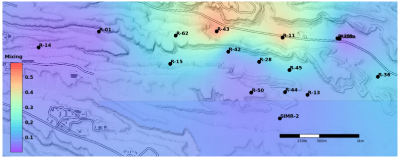
Source 7: (background)



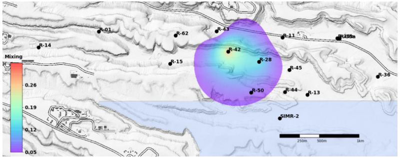
Source 3: Cl^-, Ca, Mg (background)



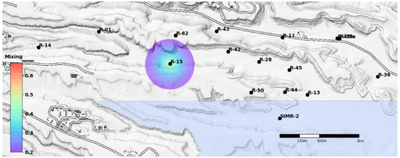
Source 4: NO_3



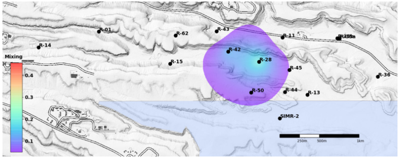
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



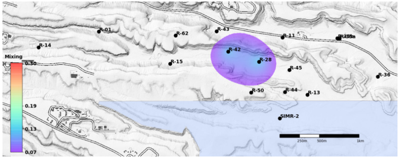
Source 2: ClO_4



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

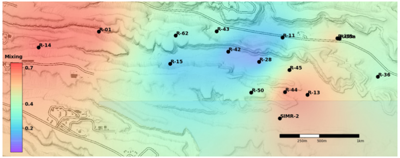


Source 6: Cl^-, Ca, Mg, and SO_4

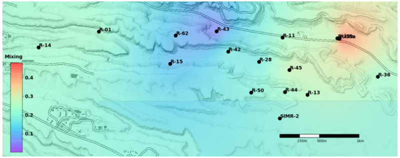


NTFk identified sources (groundwater types) Jan-Dec 2009

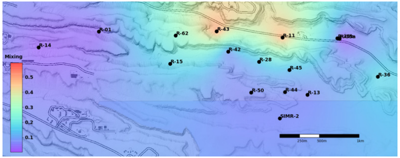
Source 7: (background)



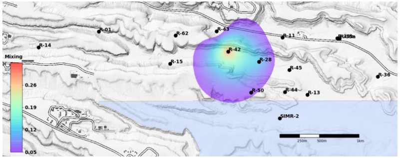
Source 3: Cl^-, Ca, Mg (background)



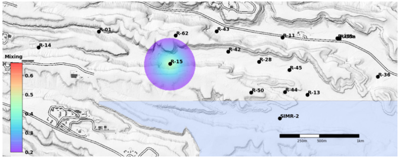
Source 4: NO_3



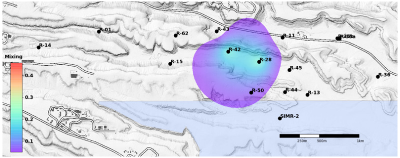
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



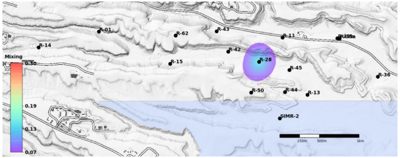
Source 2: ClO_4



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

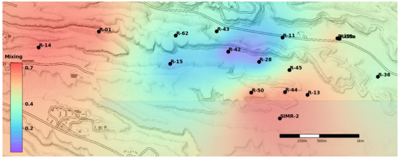


Source 6: Cl^-, Ca, Mg, and SO_4

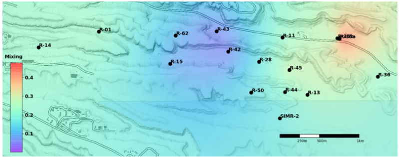


NTFk identified sources (groundwater types) Jan-Dec 2010

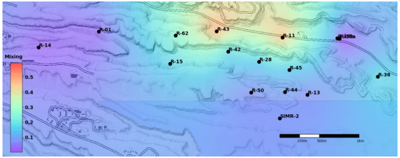
Source 7: (background)



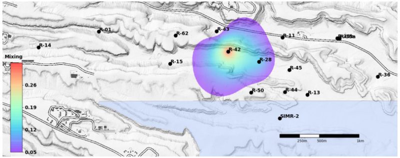
Source 3: Cl^-, Ca, Mg (background)



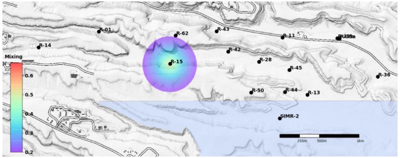
Source 4: NO_3



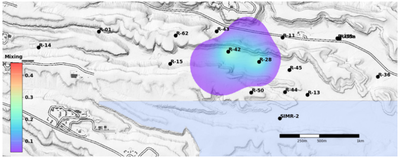
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



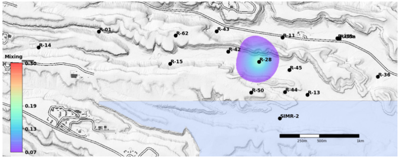
Source 2: ClO_4



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

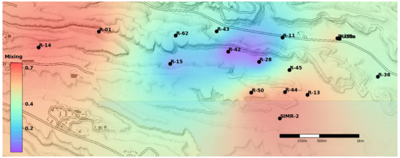


Source 6: Cl^-, Ca, Mg, and SO_4

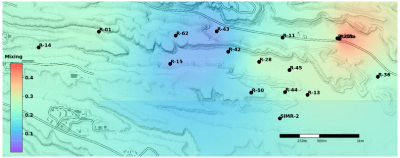


NTFk identified sources (groundwater types) Jan-Dec 2011

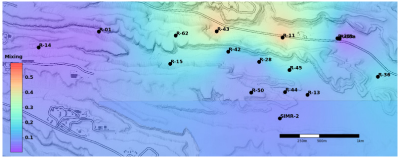
Source 7: (background)



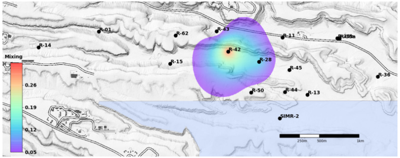
Source 3: Cl^-, Ca, Mg (background)



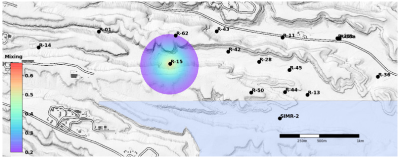
Source 4: NO_3



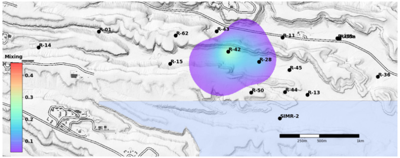
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



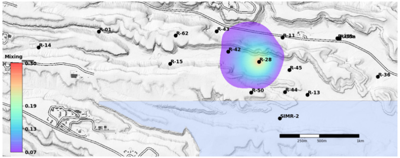
Source 2: ClO_4



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

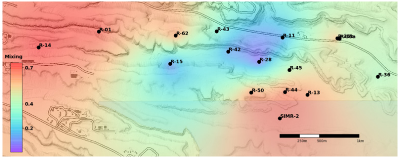


Source 6: Cl^-, Ca, Mg, and SO_4

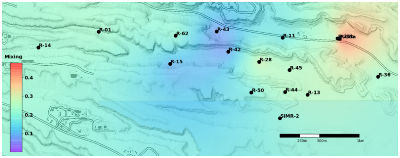


NTFk identified sources (groundwater types) Jan-Dec 2012

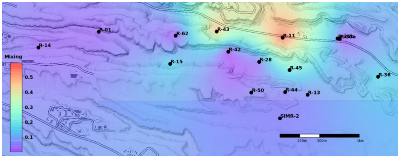
Source 7: (background)



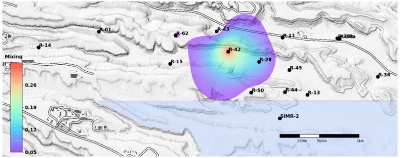
Source 3: Cl^-, Ca, Mg (background)



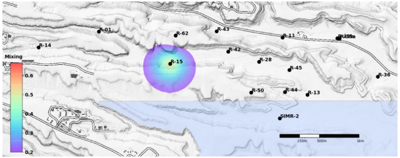
Source 4: NO_3



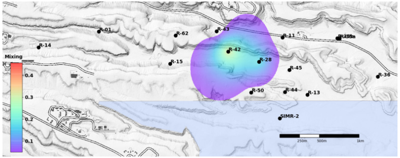
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



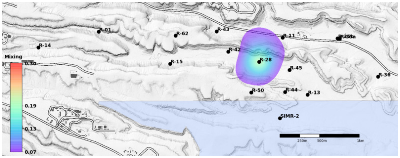
Source 2: ClO_4



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

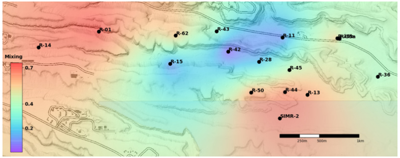


Source 6: Cl^-, Ca, Mg, and SO_4

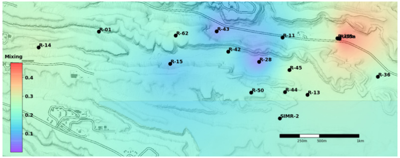


NTFk identified sources (groundwater types) Jan-Dec 2013

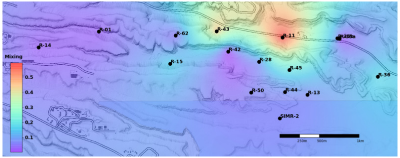
Source 7: (background)



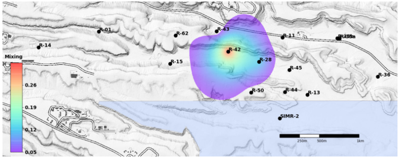
Source 3: Cl^-, Ca, Mg (background)



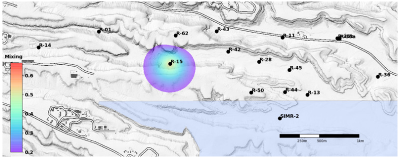
Source 4: NO_3



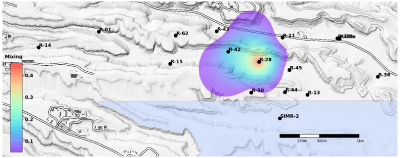
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



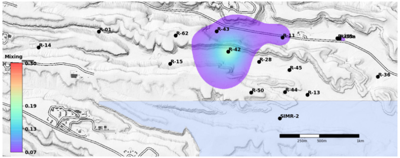
Source 2: ClO_4



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

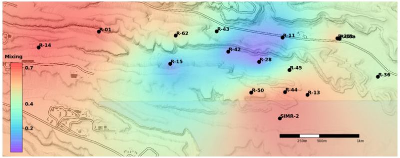


Source 6: Cl^-, Ca, Mg, and SO_4

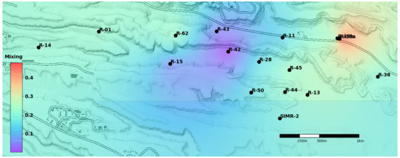


NTFk identified sources (groundwater types) Jan-Dec 2014

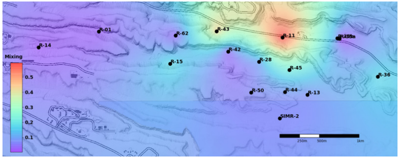
Source 7: (background)



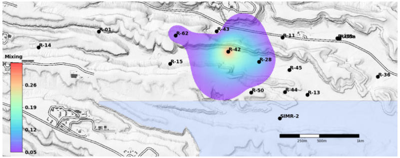
Source 3: Cl^-, Ca, Mg (background)



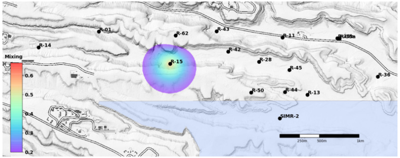
Source 4: NO_3



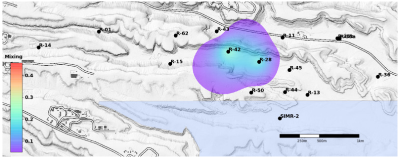
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



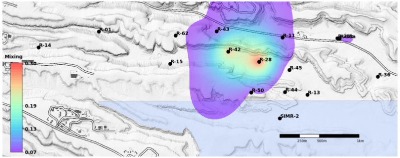
Source 2: ClO_4



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

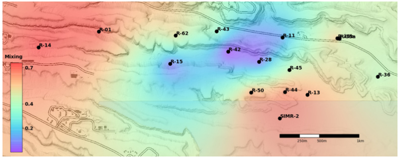


Source 6: Cl^-, Ca, Mg, and SO_4

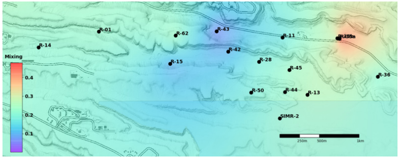


NTFk identified sources (groundwater types) Jan-Dec 2015

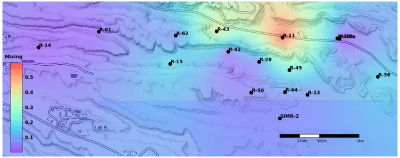
Source 7: (background)



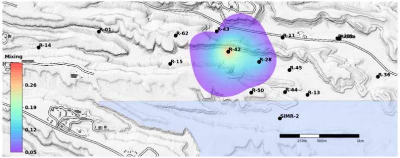
Source 3: Cl^-, Ca, Mg (background)



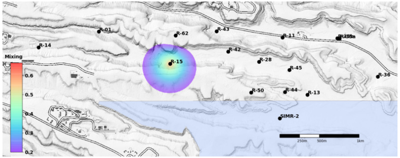
Source 4: NO_3



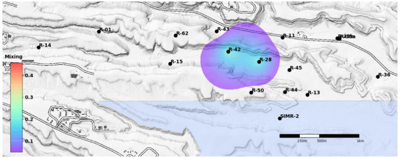
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



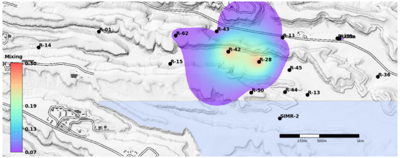
Source 2: ClO_4



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

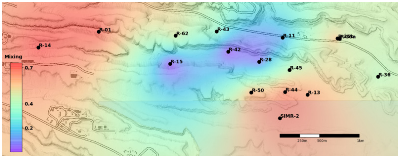


Source 6: Cl^-, Ca, Mg, and SO_4

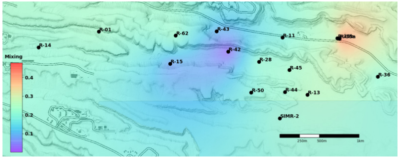


NTFk identified sources (groundwater types) Jan-Dec 2016

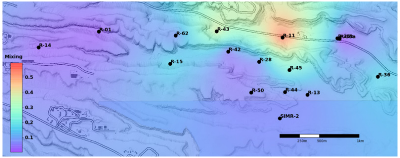
Source 7: (background)



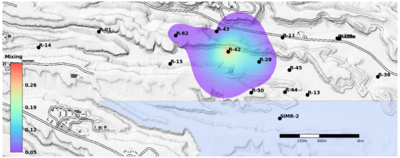
Source 3: Cl^-, Ca, Mg (background)



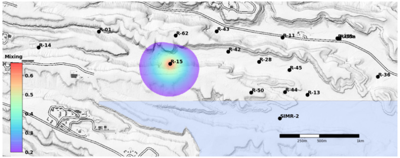
Source 4: NO_3



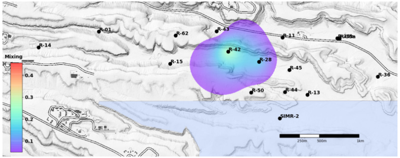
Source 1: Cr, Cl^-, NO_3, Ca, Mg, and SO_4



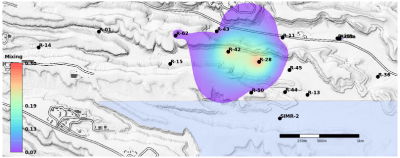
Source 2: ClO_4



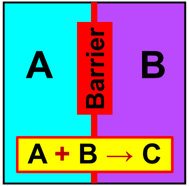
Source 5: ^3H, Cr, Cl^-, Ca, Mg, and SO_4



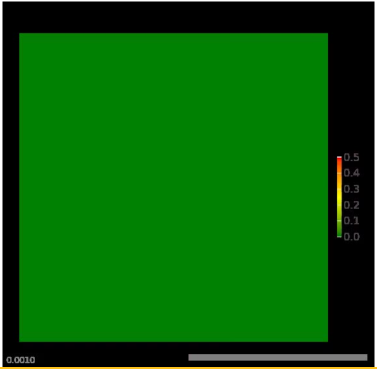
Source 6: Cl^-, Ca, Mg, and SO_4

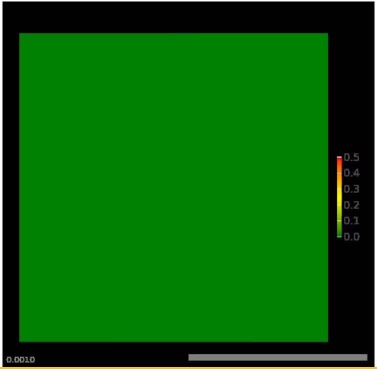


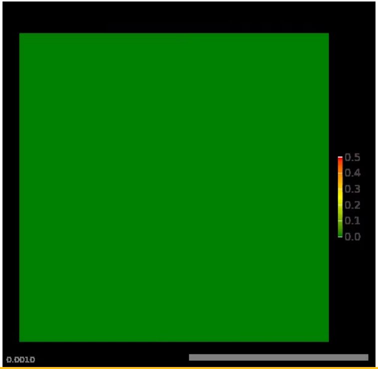
Fluid mixing

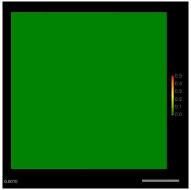


- We want to find how time/space behavior of C concentrations is controlled by the simulated physics processes
- > 2000 simulations of C concentrations in time/space for a series of model parameters impacting fluid mixing; 3 example predictions:

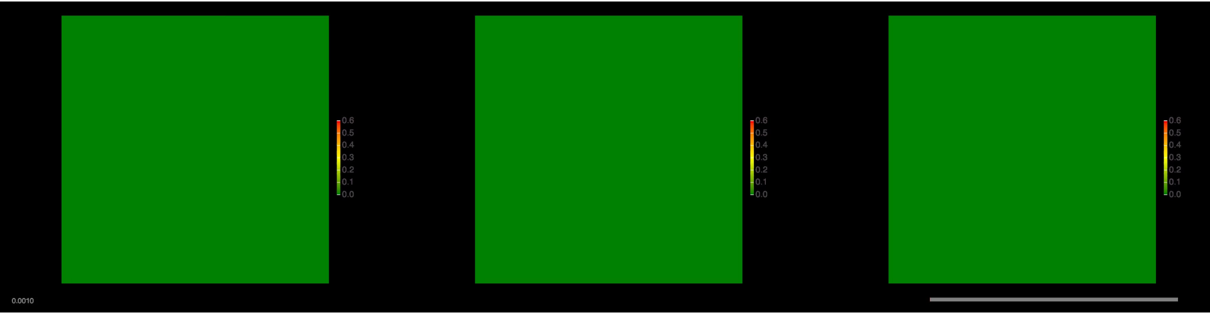








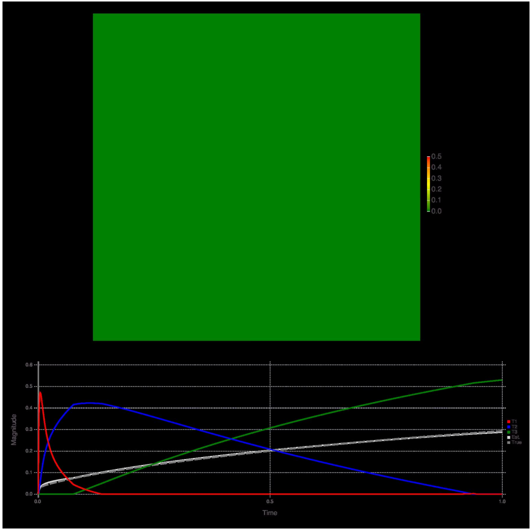
- > 200 GB simulation data compressed to ≈ 70 MB (compression $\approx 4 \times 10^{-4}$)
  Here, $(1000 \times 81 \times 81) \rightarrow (3 \times 12 \times 13)$
- **NTFk** processed all the data and extracted the dominant time/space features (**processes** / **vortices**)



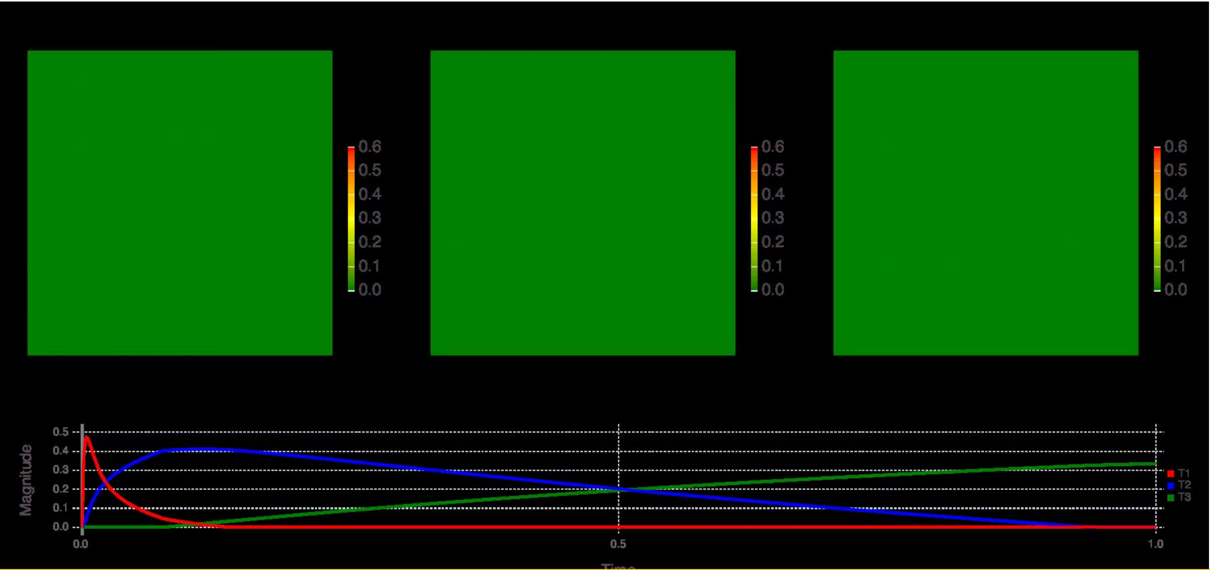
Advection **Dispersion** **Diffusion**

NTFk results

- T1: Advection
- T2: Dispersion
- T3: Diffusion

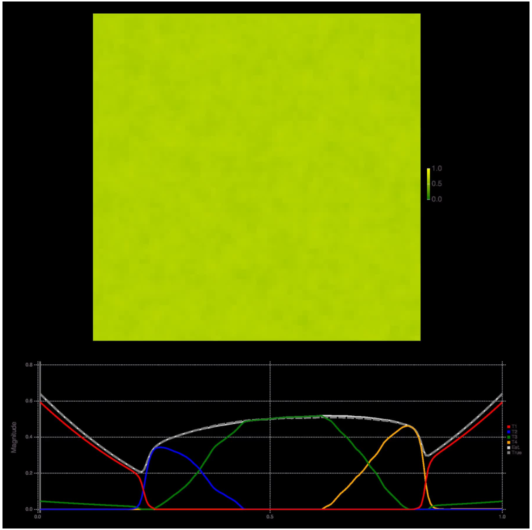


NTFk results



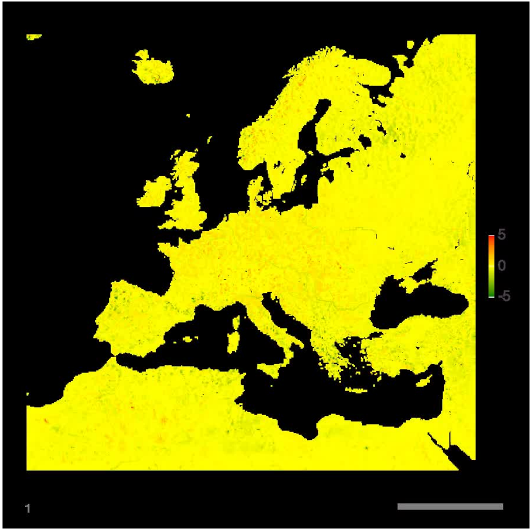
Polymer-chain folding

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

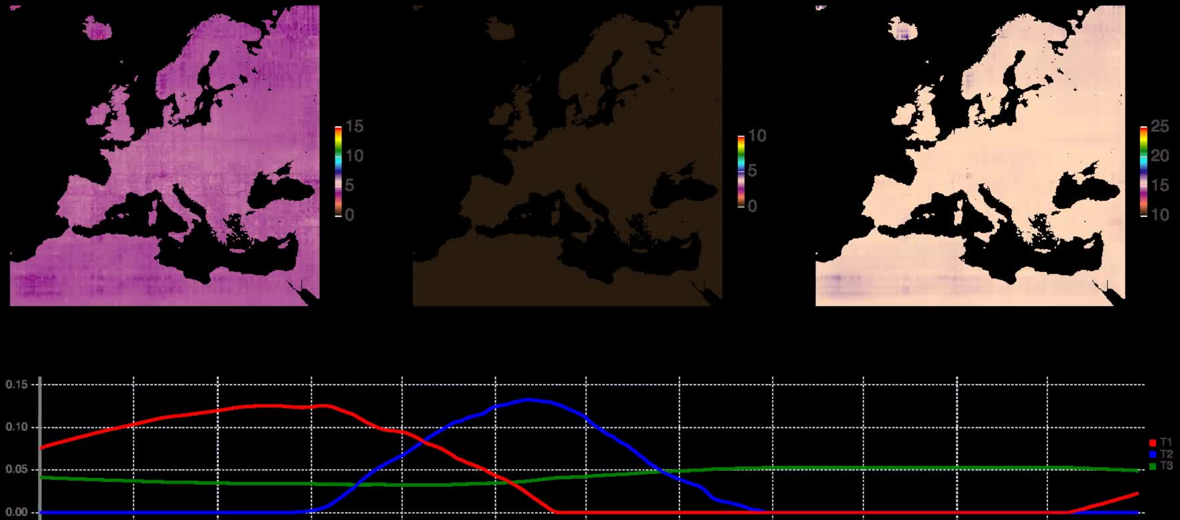


« Play/Pause

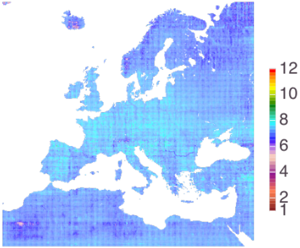
- fluctuations in the water-table levels
- **NTF*k*** extracts seasonal changes and dominant infiltration signals
- $(424 \times 412 \times 365) \rightarrow (? \times ? \times ?)$

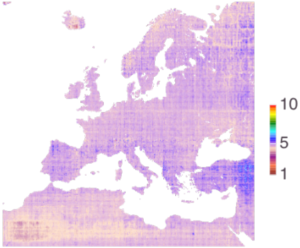


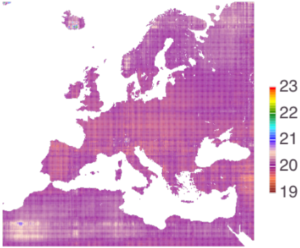
Climate model of Europe: water-table fluctuations represented by 3 signals

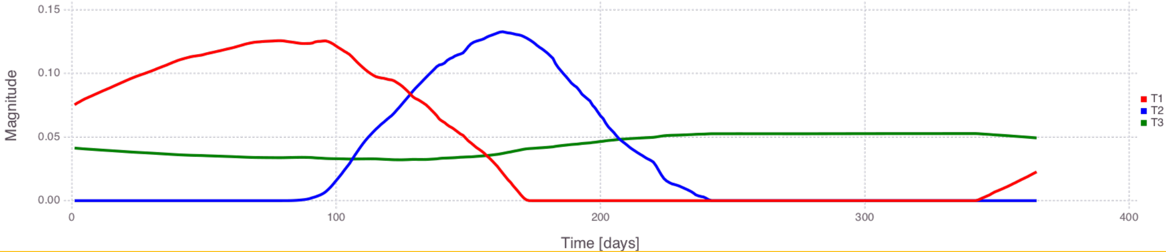


Climate model of Europe: maximum water-table fluctuations for each signal (3)

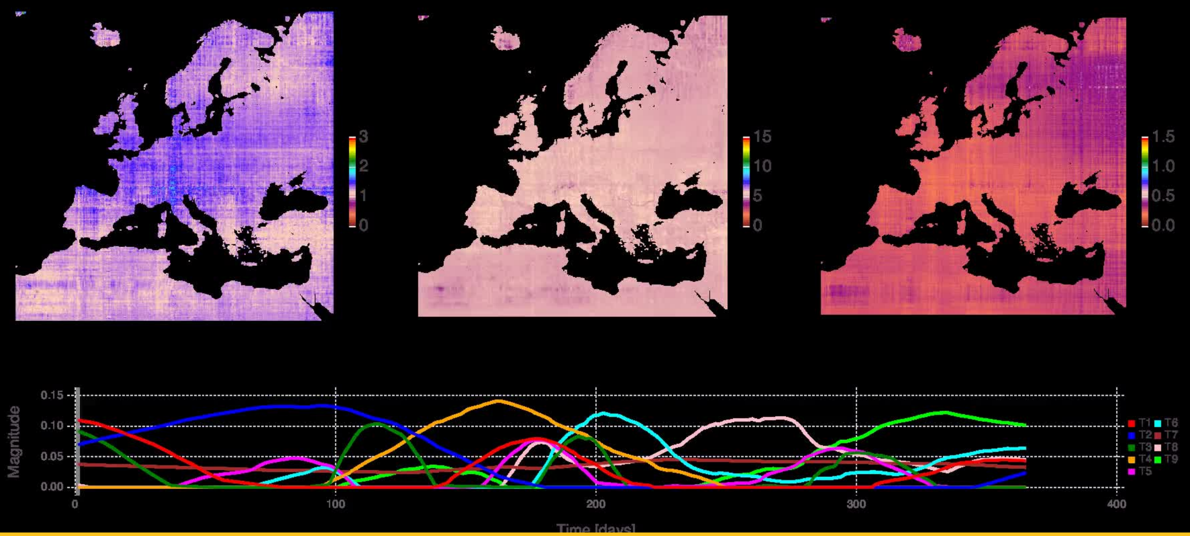




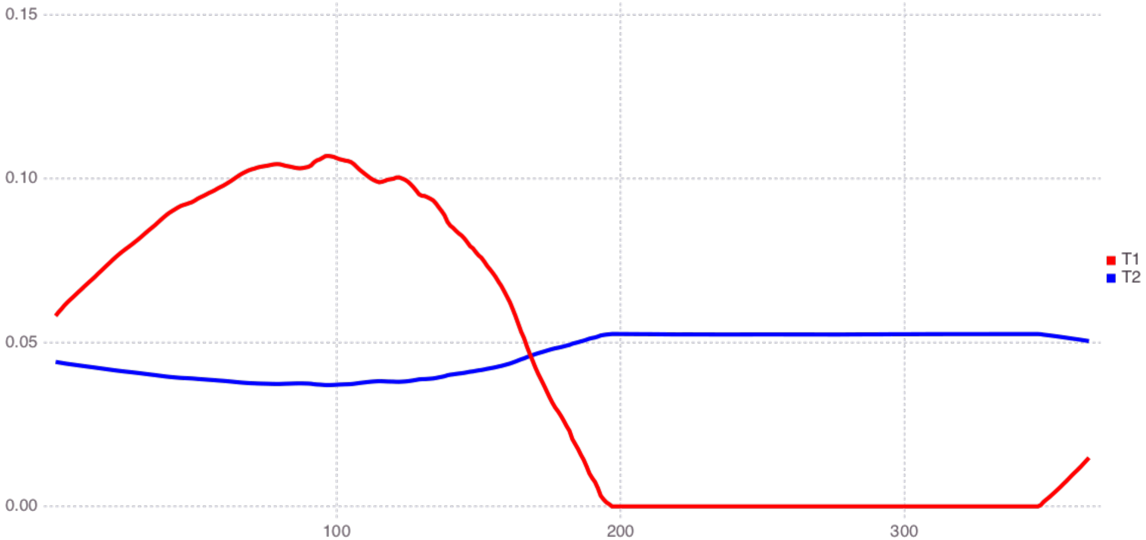




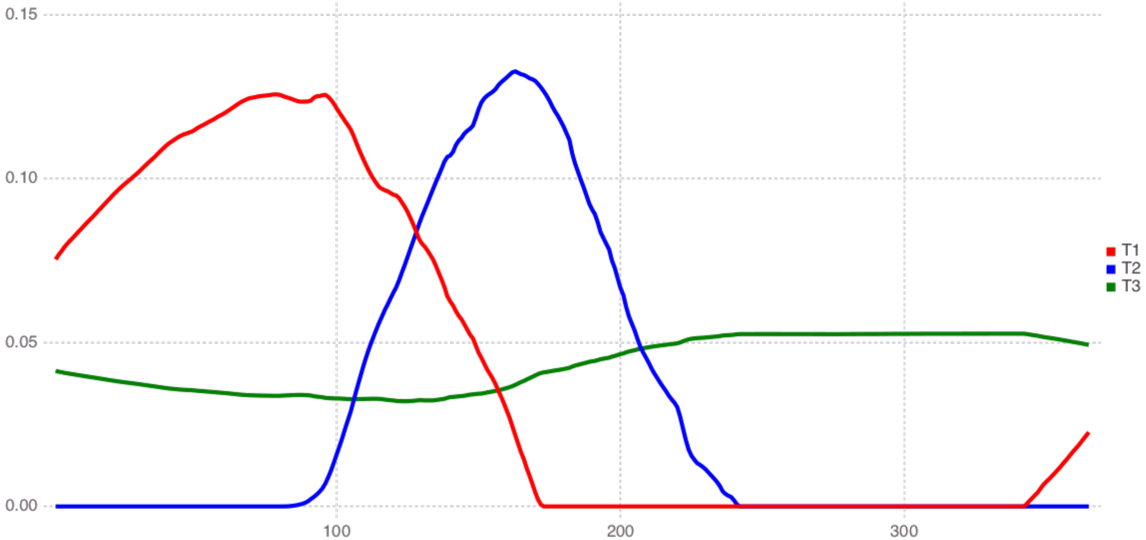
Climate model of Europe: water-table fluctuations represented by 9 signals



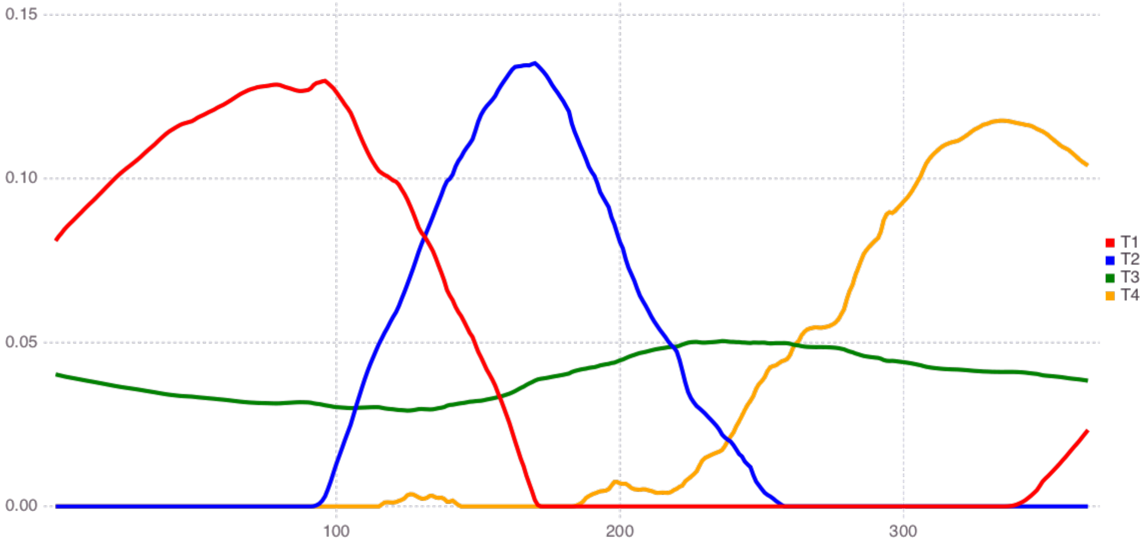
Climate model of Europe: Water-table fluctuations represented by 2 signals



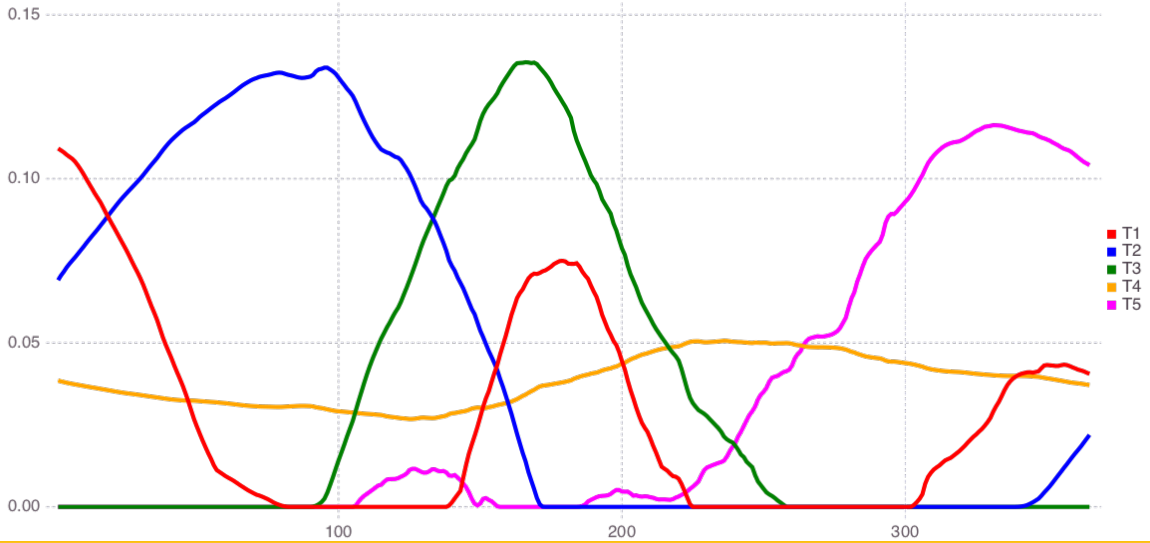
Climate model of Europe: Water-table fluctuations represented by 3 signals



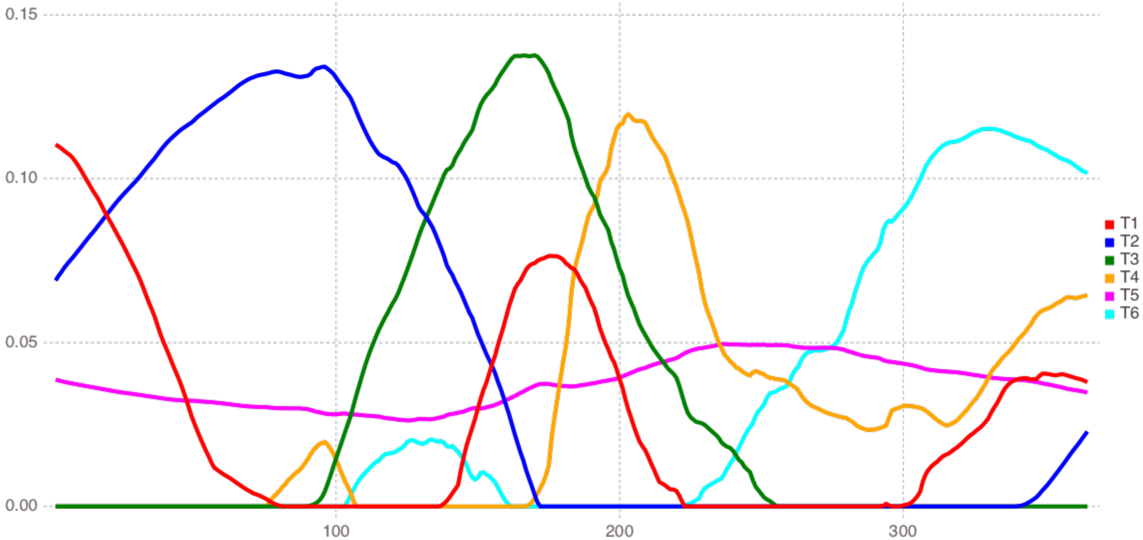
Climate model of Europe: Water-table fluctuations represented by 4 signals



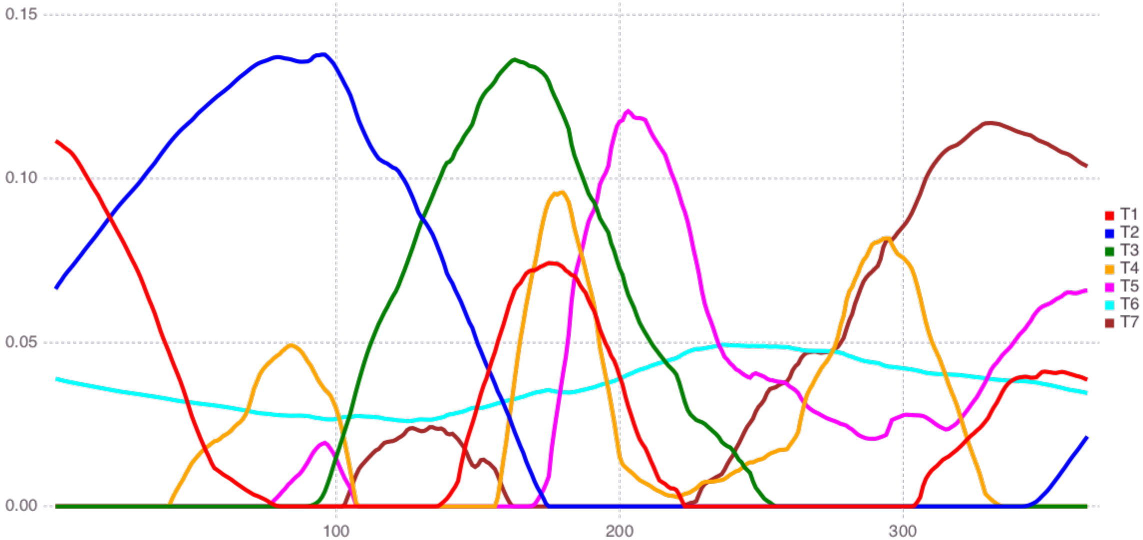
Climate model of Europe: Water-table fluctuations represented by 5 signals



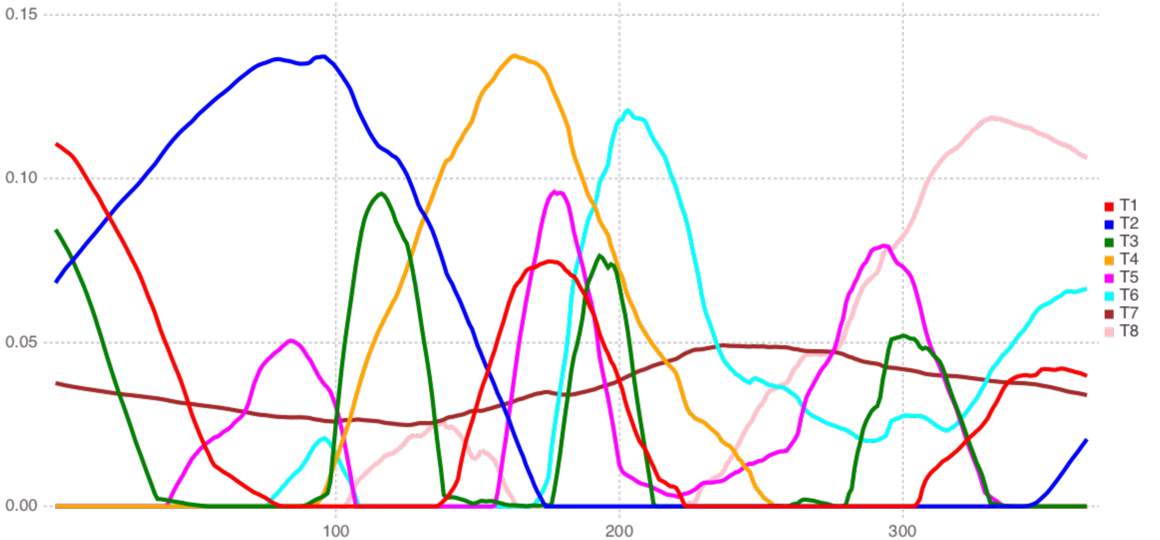
Climate model of Europe: Water-table fluctuations represented by 6 signals



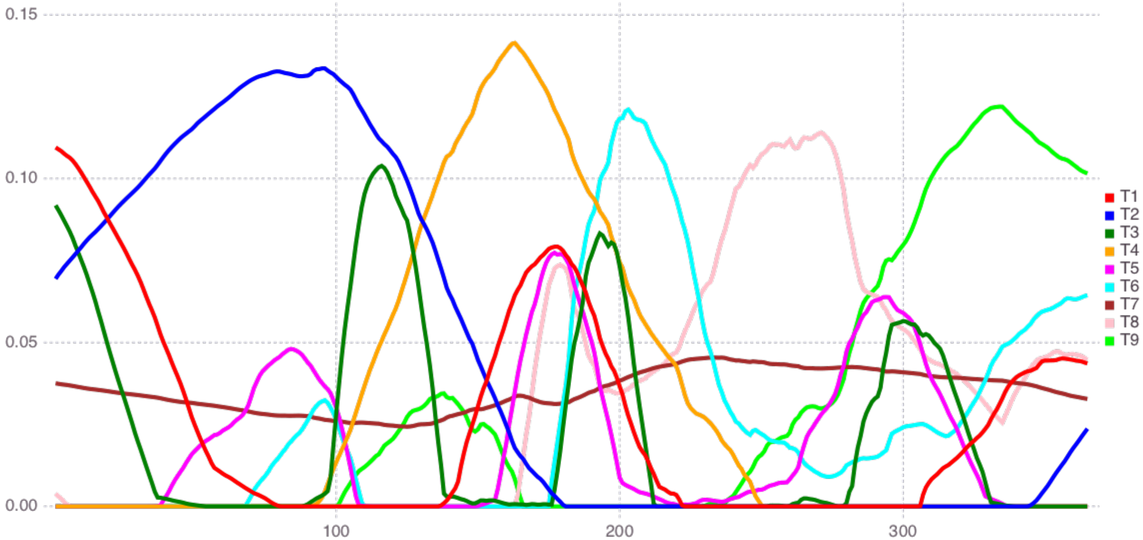
Climate model of Europe: Water-table fluctuations represented by 7 signals



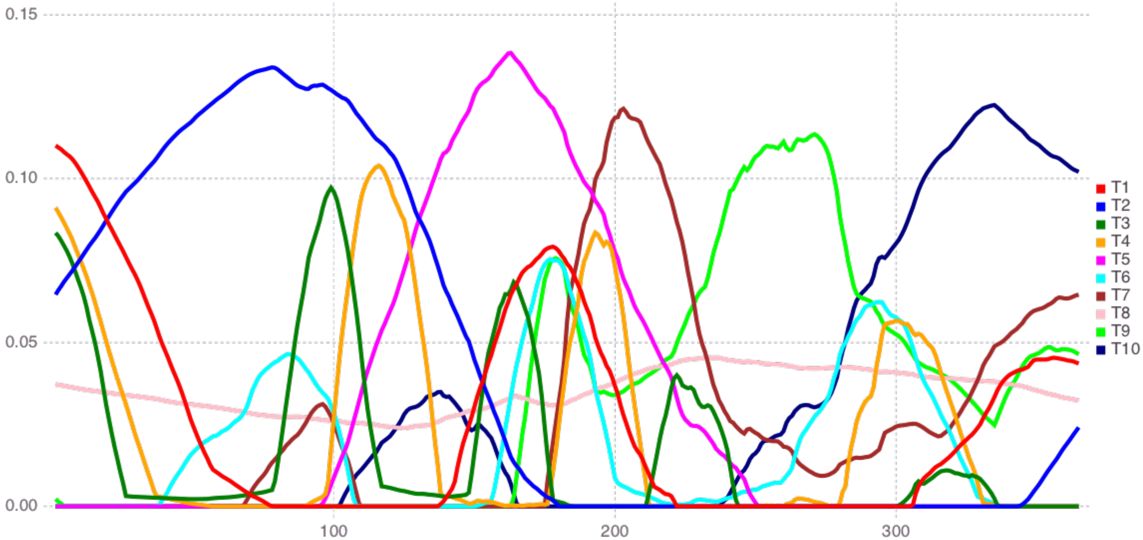
Climate model of Europe: Water-table fluctuations represented by 8 signals



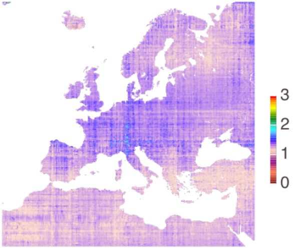
Climate model of Europe: Water-table fluctuations represented by 9 signals

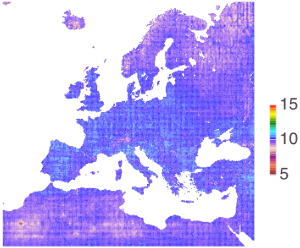


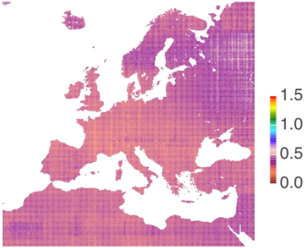
Climate model of Europe: Water-table fluctuations represented by 10 signals

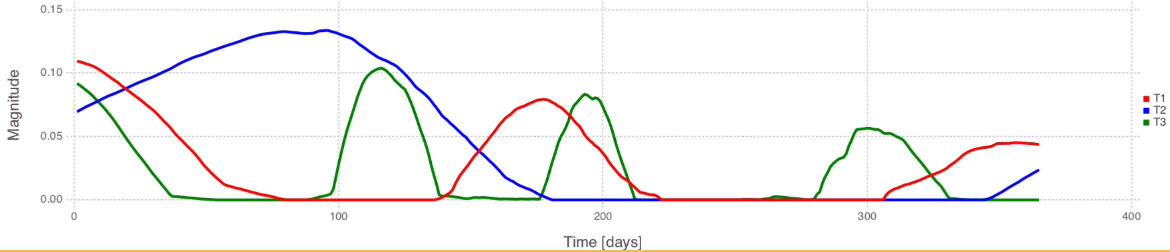


Climate model of Europe: maximum water-table fluctuations for each signal (9)

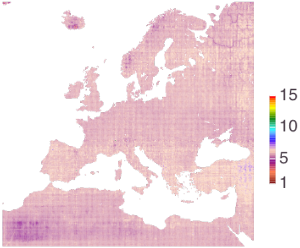


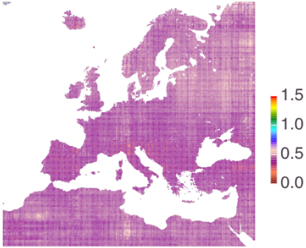


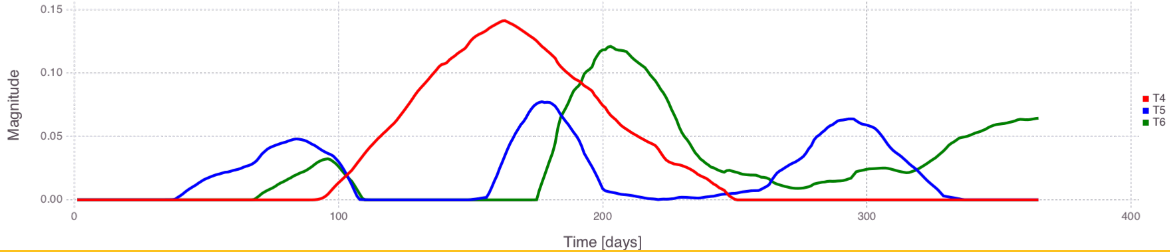




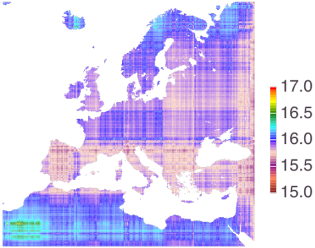
Climate model of Europe: maximum water-table fluctuations for each signal (9)

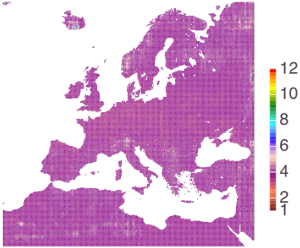


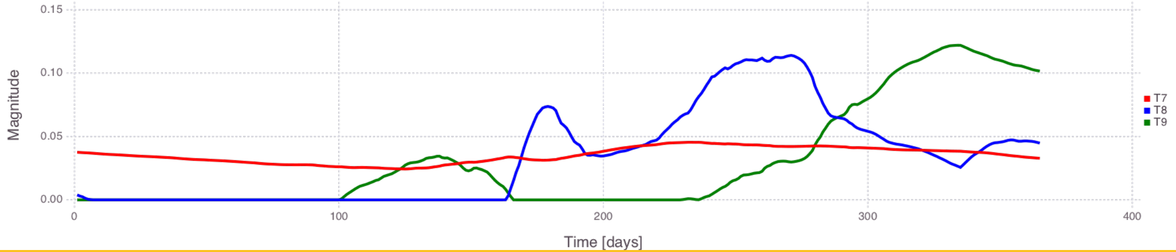




Climate model of Europe: maximum water-table fluctuations for each signal (9)

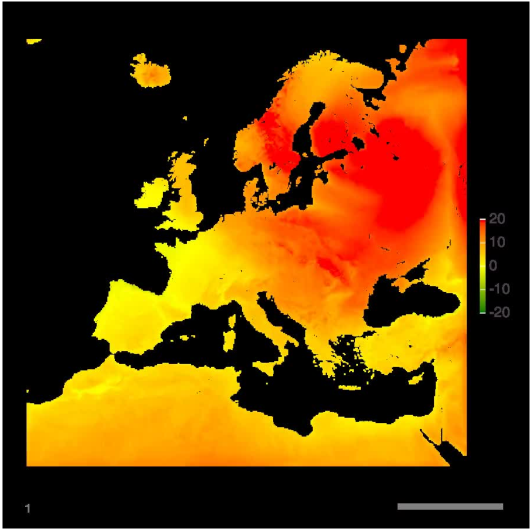




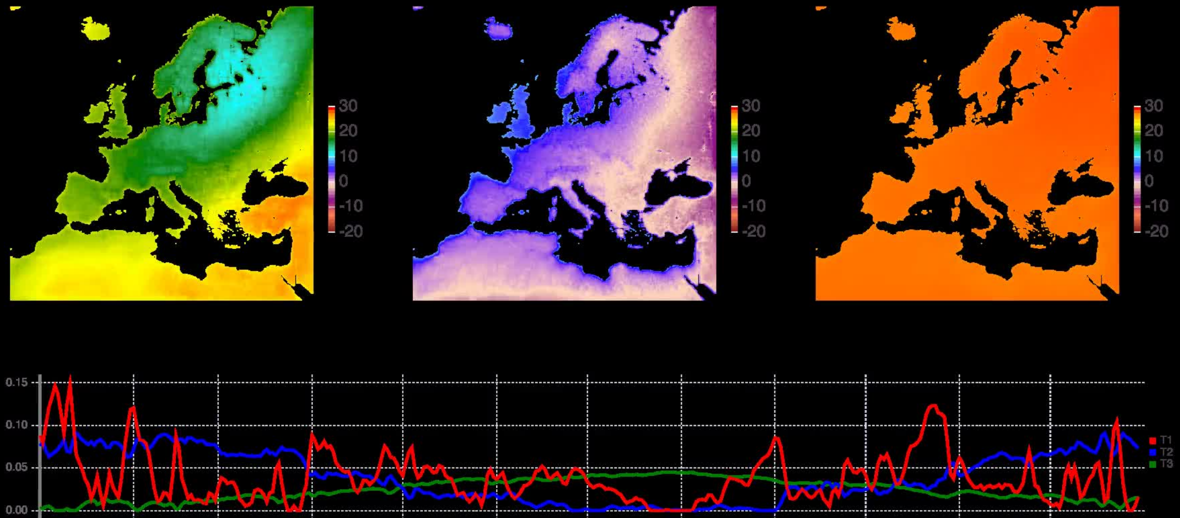


Climate model of Europe: air-temperature fluctuations

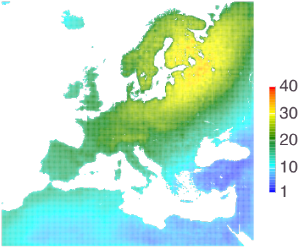
- fluctuations in the air temperature
- $(424 \times 412 \times 365) \rightarrow (? \times ? \times ?)$

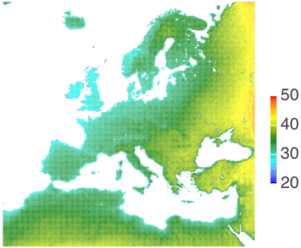


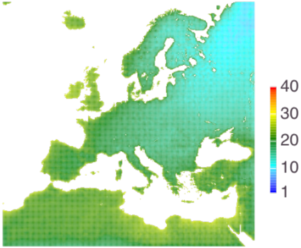
Climate model of Europe: air temperature fluctuations represented by 3 signals

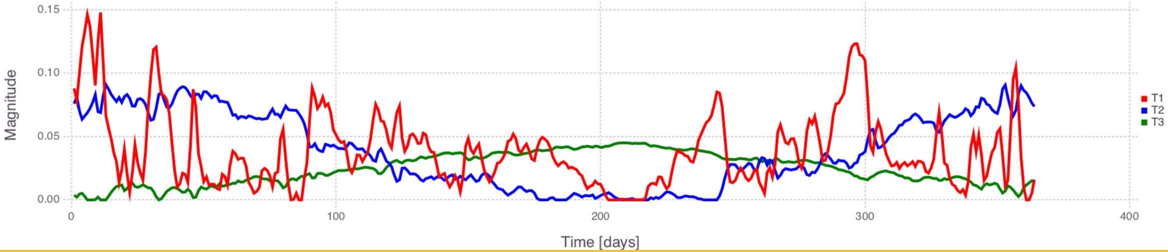


Climate model of Europe: maximum air temperature fluctuations for each signal (3)

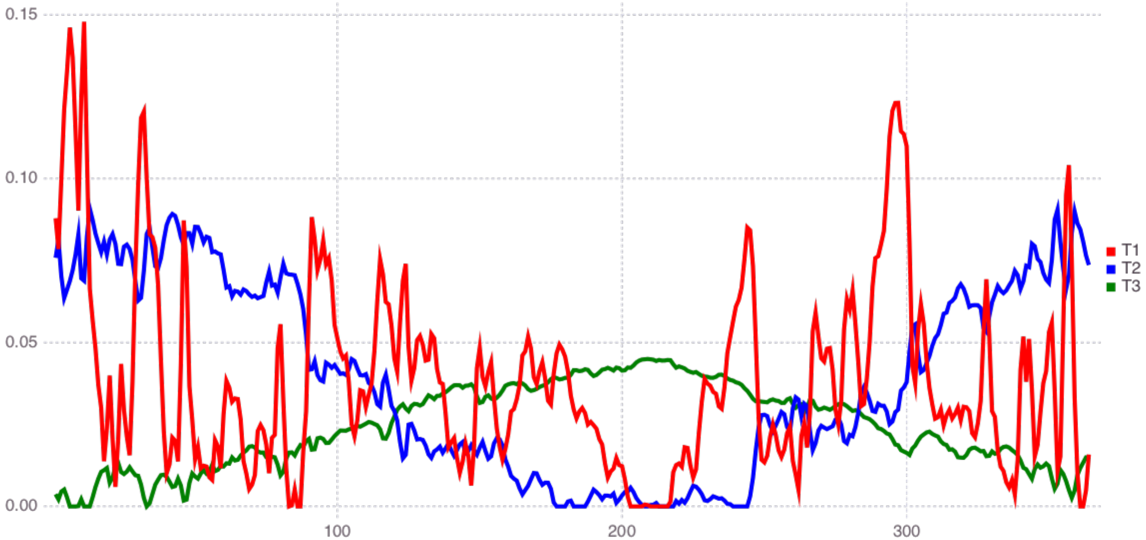




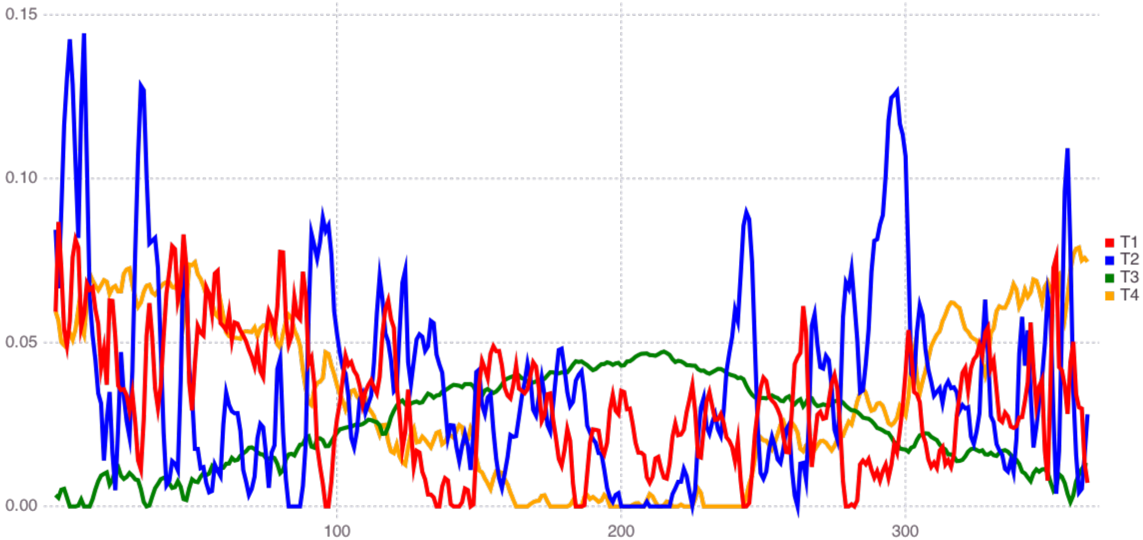




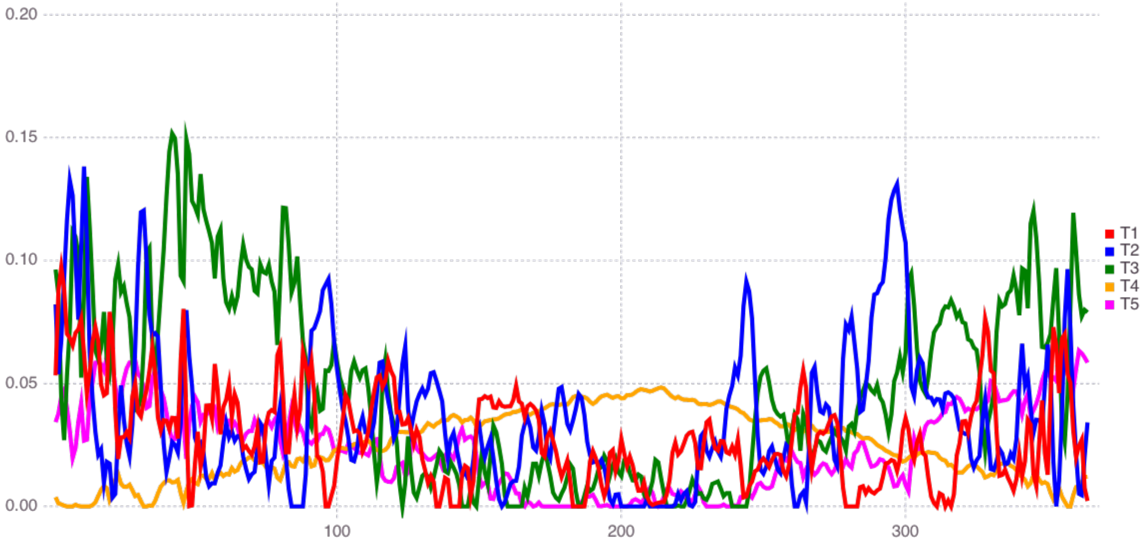
Climate model of Europe: Air temperature fluctuations represented by 3 signals



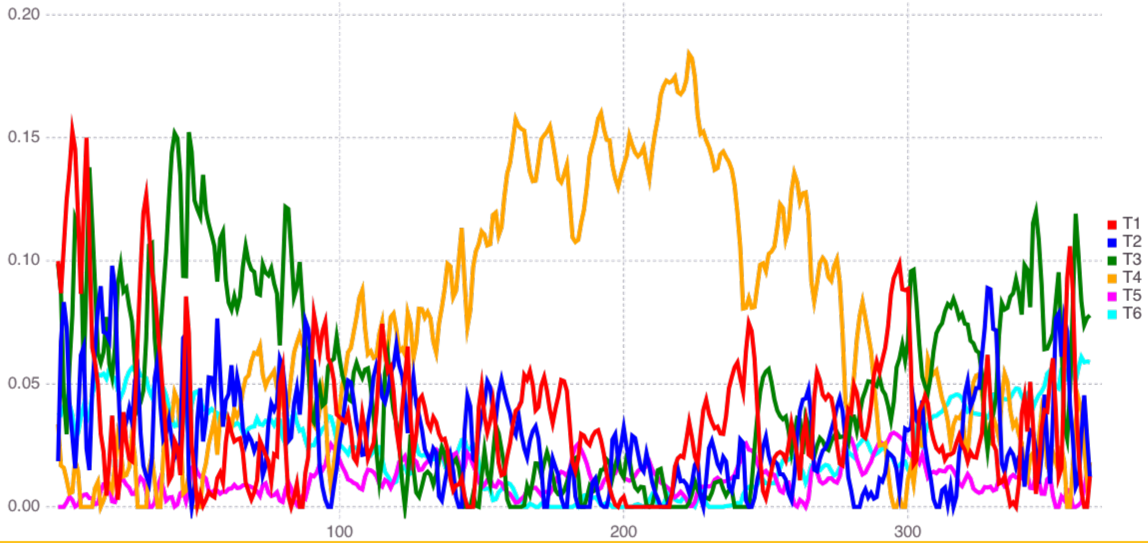
Climate model of Europe: Air temperature fluctuations represented by 4 signals



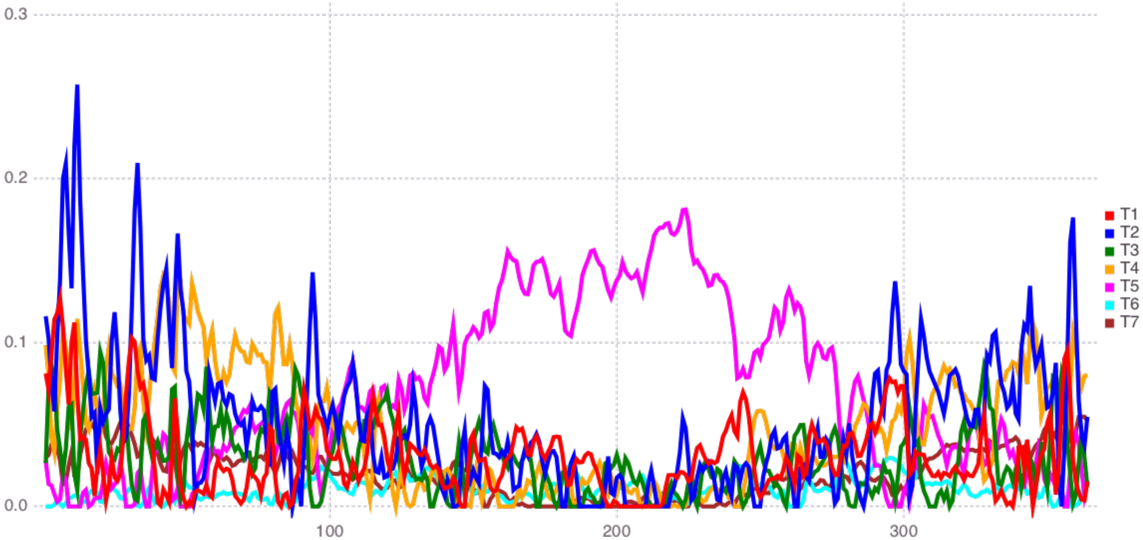
Climate model of Europe: Air temperature fluctuations represented by 5 signals



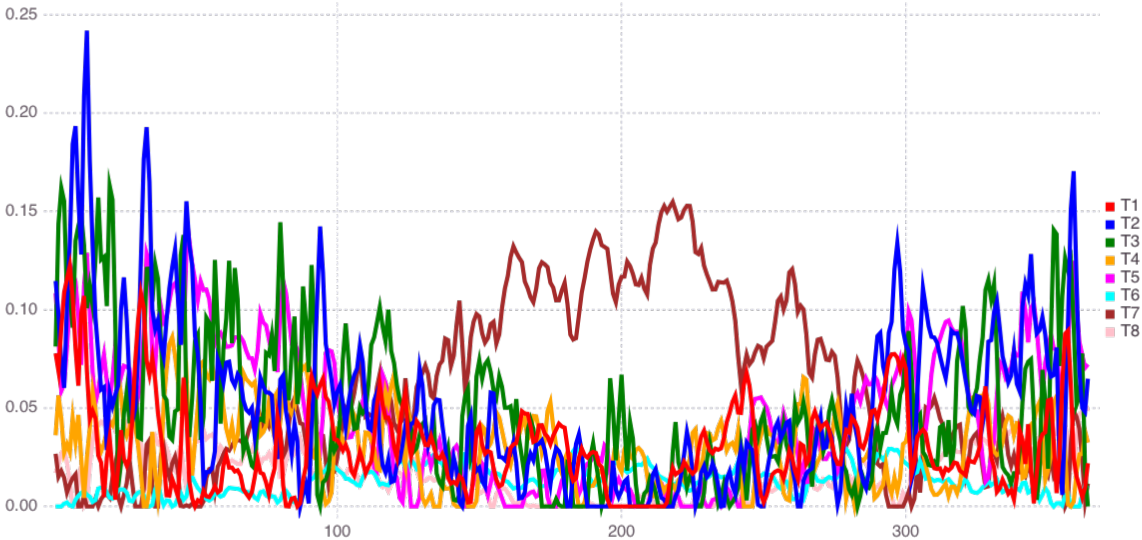
Climate model of Europe: Air temperature fluctuations represented by 6 signals



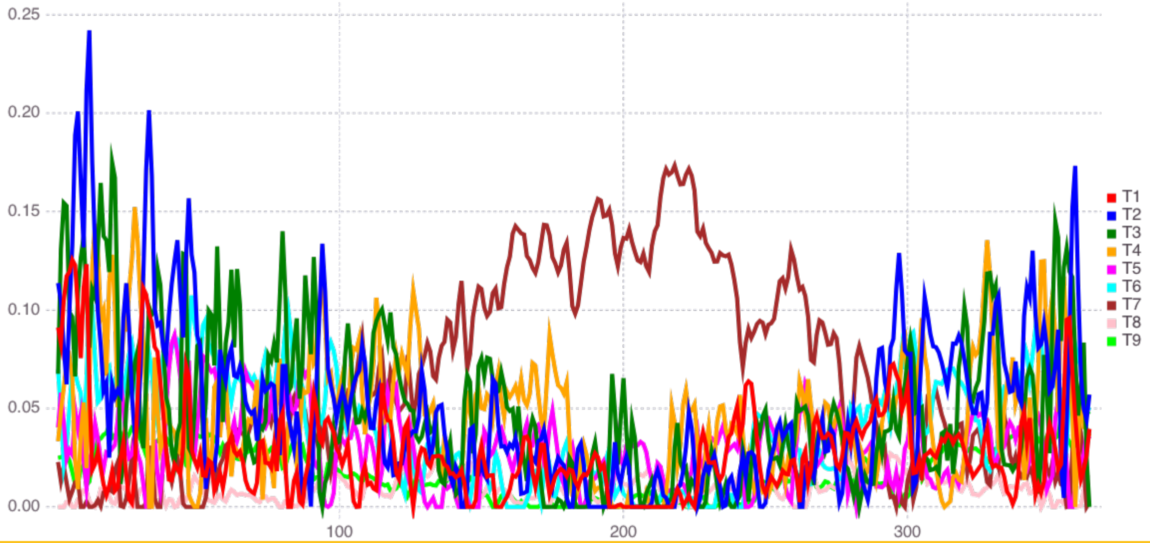
Climate model of Europe: Air temperature fluctuations represented by 7 signals



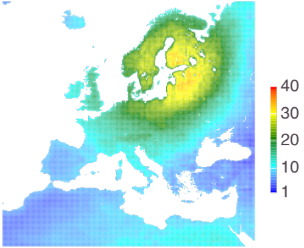
Climate model of Europe: Air temperature fluctuations represented by 8 signals

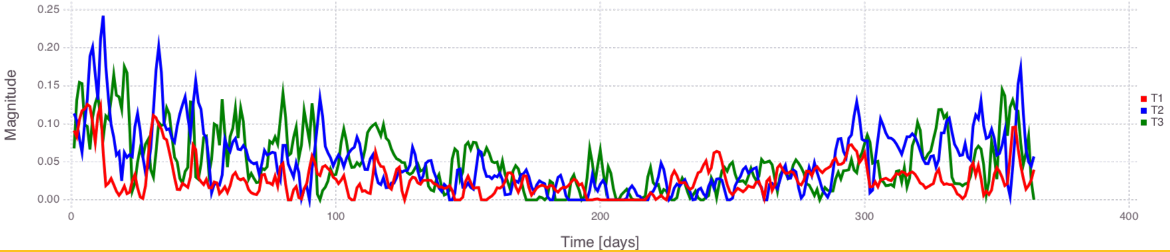


Climate model of Europe: Air temperature fluctuations represented by 9 signals

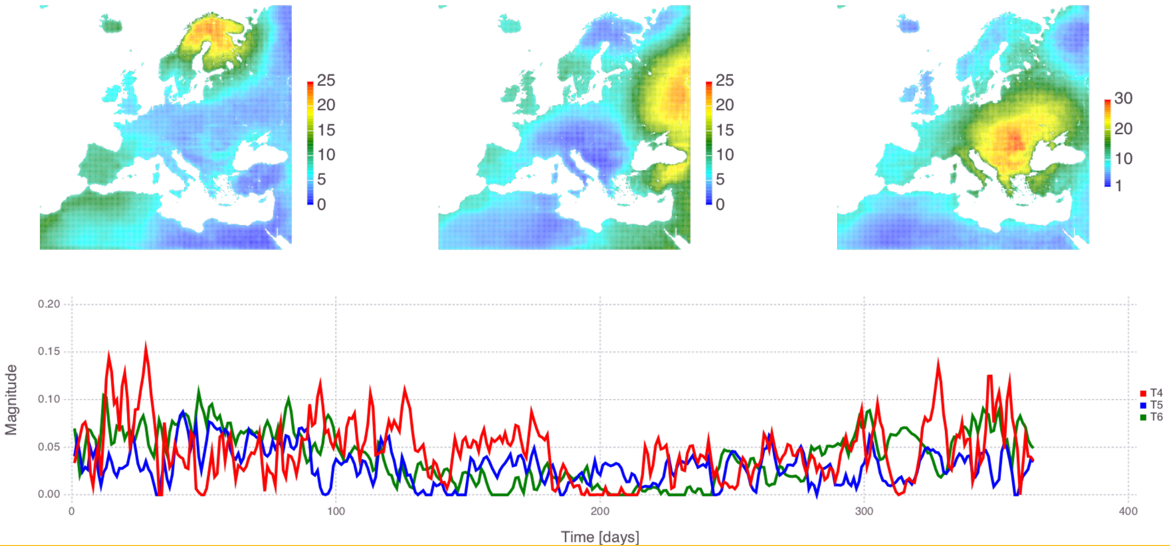


Climate model of Europe: maximum air temperature fluctuations for each signal (9)

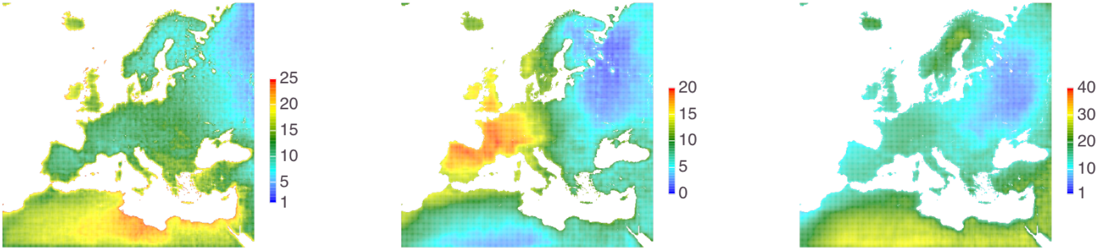


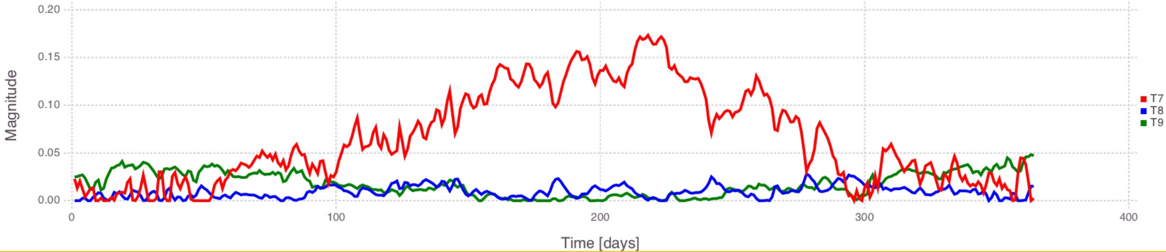


Climate model of Europe: maximum air temperature fluctuations for each signal (9)

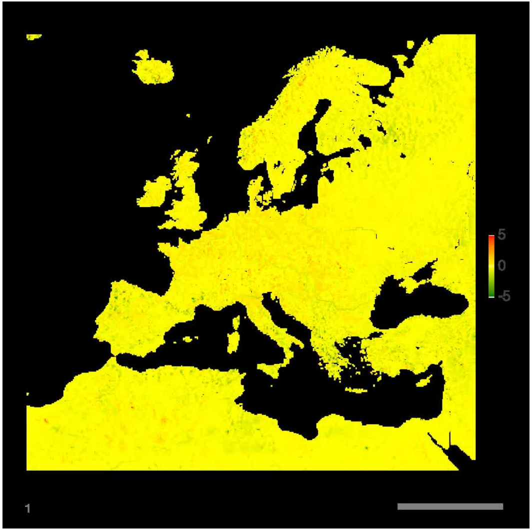


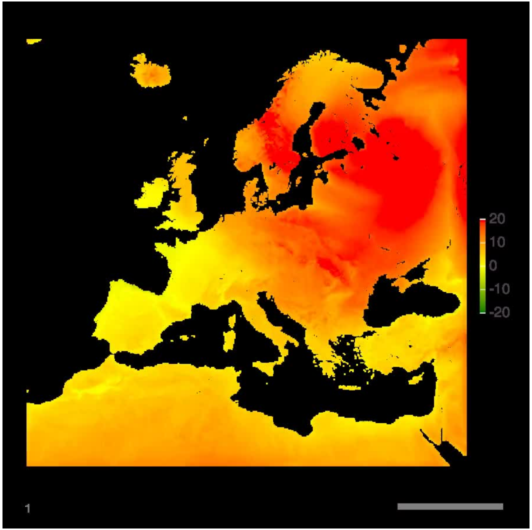
Climate model of Europe: maximum air temperature fluctuations for each signal (9)





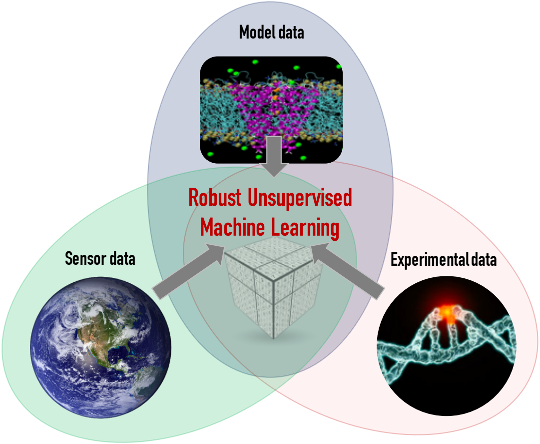
Climate model of Europe: water-table vs air-temperature fluctuations





Summary

- We have developed a series of novel unsupervised ML methods based on Nonnegative Factorization (Matrices/Tensors)
- These ML methods have been used to solve various real-world problems
- Some of our ML analyses brought breakthrough discoveries (especially related to human cancer research)
- We have developed a series of ML computational tools for solving big-data problems using high-performance computing (HPC)



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

- NMFk + ShiftNMFk + GreenNMFk **(patent)**
- NTFk
- 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`