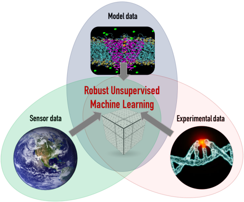
Machine Learning Analyses of Climate Data and Models

Velimir V. Vesselinov (monty) (vvv@lanl.gov), **Daniel O'Malley**, **Boian Alexandrov**
Los Alamos National Laboratory, NM, USA

Stefan Kollet, **Carina Furuso**, **Klaus Gorgen**
Forschungszentrum Juelich, Germany



EWRA 2019

LA-UR-XX-XXXXX

Nonnegative Matrix/Tensor Factorization

- We have developed a series of novel **unsupervised** Machine Learning (ML) methods
- Our unique ML methods are based in **matrix/tensor factorization** coupled with custom **k-means clustering** and **nonnegativity/sparsity** constraints:
  - **NMFk**: Nonnegative **Matrix** Factorization
  - **NTFk**: Nonnegative **Tensor** Factorization
- **NMFk** / **NTFk** are capable to efficiently process large datasets (GB/TB's) utilizing GPU's & TPU's
- **NMFk** / **NTFk** have been applied to analyze a series of real-world analyses

- **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

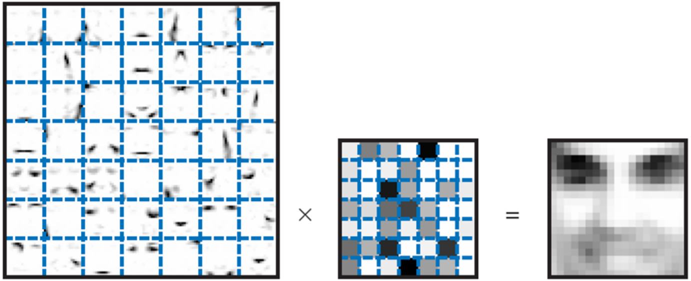
  Cannot discover something that we do not already know

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

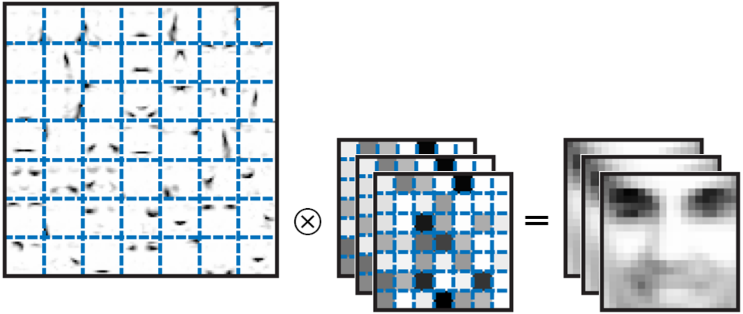
  Example: Identify features that distinguish images of animals (e.g., cats, dogs, horses, etc.); without prior information or training

  Can discover hidden (latent) features in the data (unknown/unutilized)

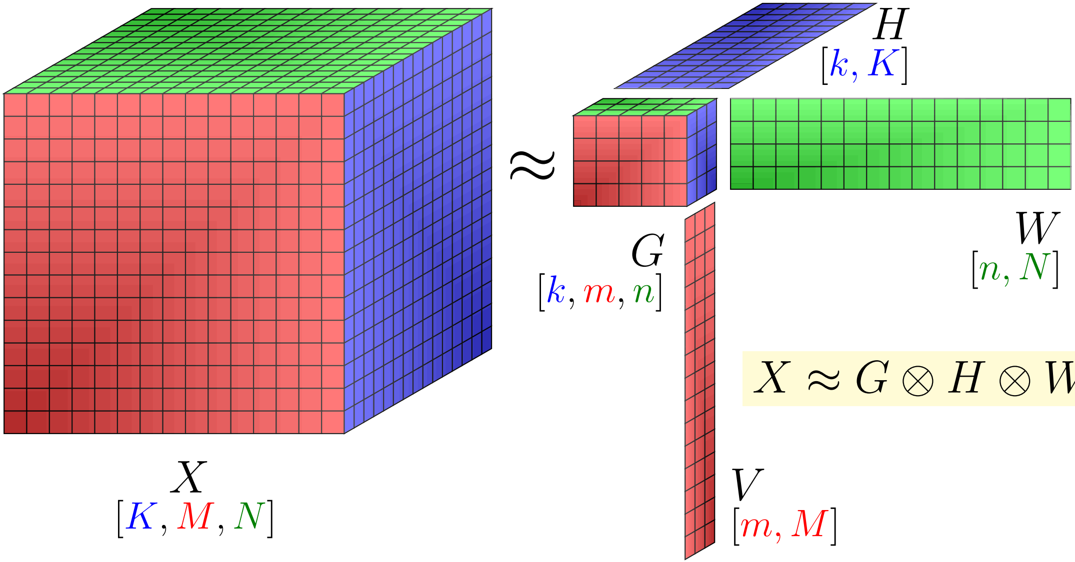
NMF: Nonnegative Matrix Factorization



NTF: Nonnegative Tensor Factorization

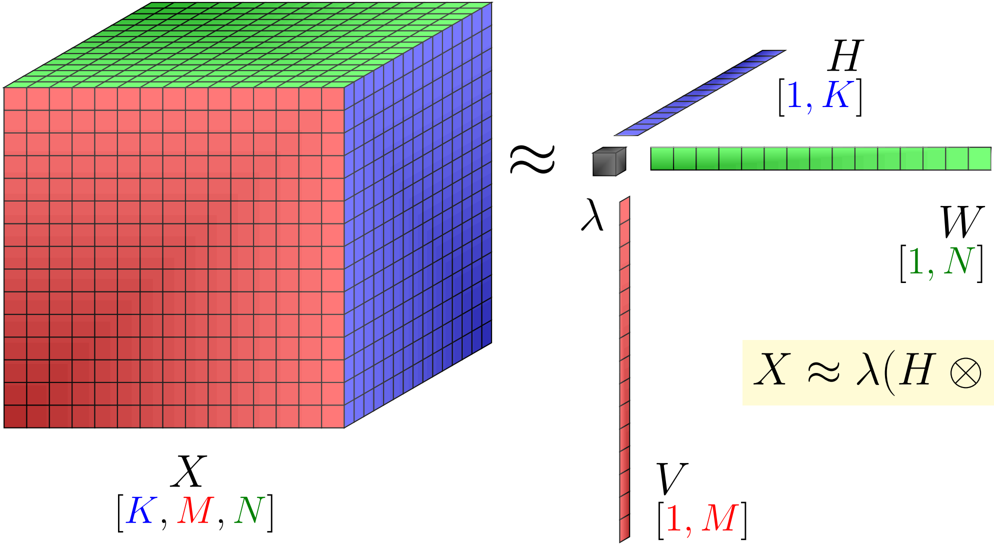


Tucker Tensor Decomposition (3D): Rank-60 Multirank-(3,4,5)



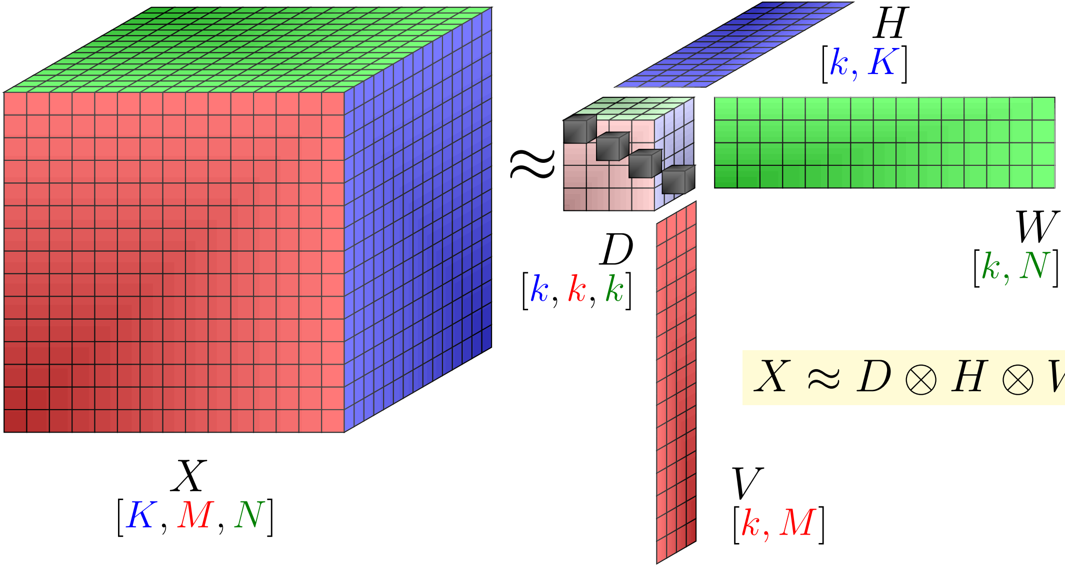
$$X \approx G \otimes H \otimes W \otimes V$$

Canonical Polyadic Tensor Decomposition (3D): Rank-1



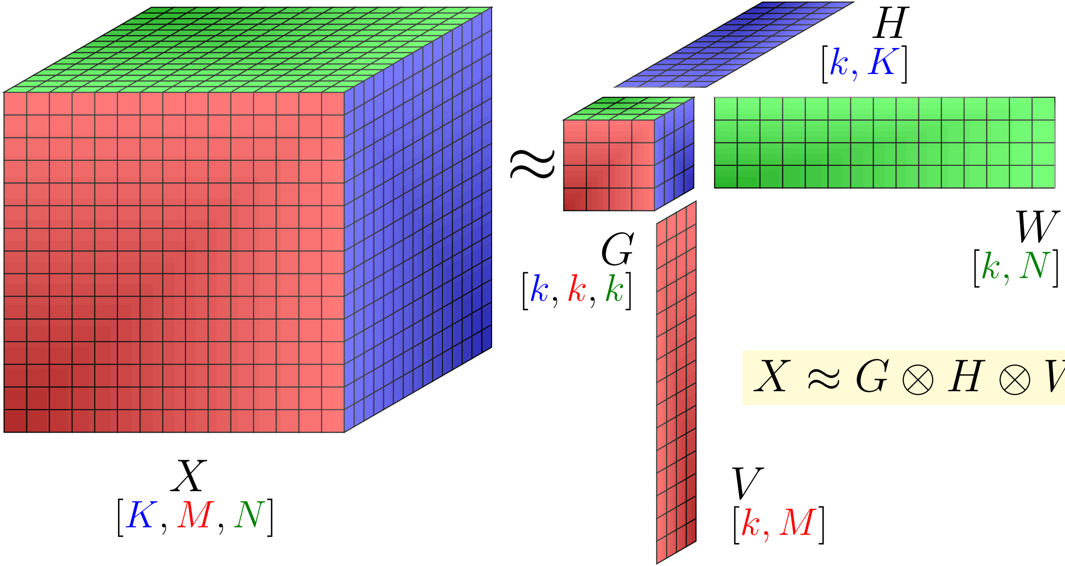
$$X \approx \lambda(H \otimes W \otimes V)$$

Canonical Polyadic Tensor Decomposition (3D): Rank-4



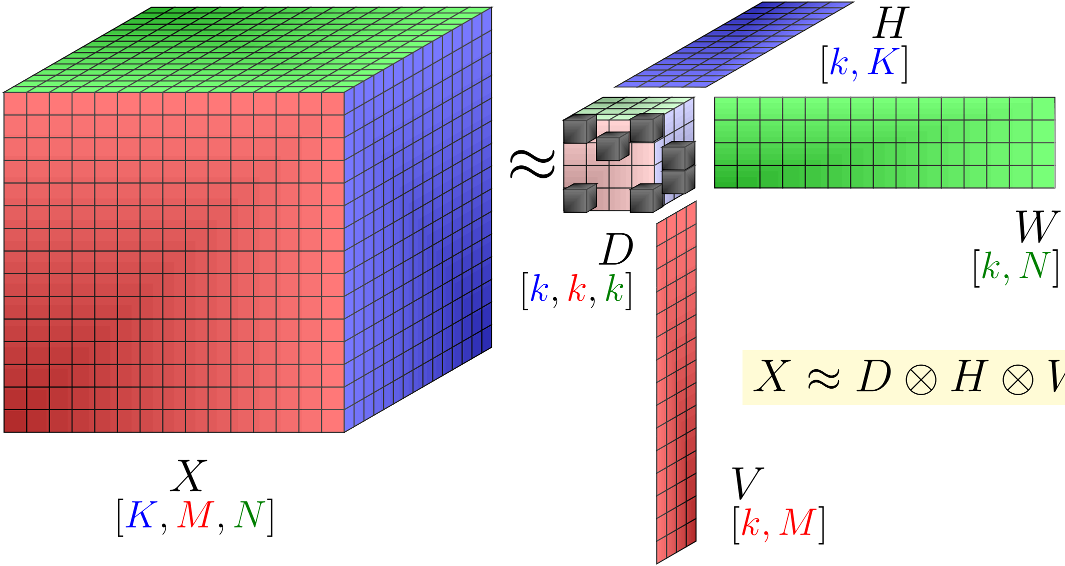
$$X \approx D \otimes H \otimes W \otimes V$$

Tucker Tensor Decomposition (3D): Rank-64 Multirank-(4,4,4)



$$X \approx G \otimes H \otimes W \otimes V$$

Tucker Tensor Decomposition (3D): Rank-7 Multirank-(3,3,4)



$$X \approx D \otimes H \otimes W \otimes V$$

- **Rank-1 Tensor**: a single tensor product of vectors
- **Tensor Rank**: smallest number r of rank-1 tensors under Canonical Polyadic Decomposition
- **Tensor Multi-Rank**: smallest dimensions of core tensor G under Tucker Decomposition (it always exists)
- **Tensor Rank** is always equal to the rank of Tucker tensor core: $rank(X) = rank(G)$
- **Nonnegative Tensor Rank**: smallest number r of rank-1 tensors under Nonnegative Canonical Polyadic Decomposition
- **Nonnegative Tensor Multi-Rank**: smallest dimension of core tensor G under Nonnegative Tucker Decomposition (existence cannot be guaranteed)
- **Nonnegative Tensor Rank** may not be equal to the rank of Nonnegative Tucker tensor core: $rank_+(X) \leq rank_+(G)$
- **Nonnegative Tensor Rank**: smallest number r of nonzero entries in tensor core under Nonnegative Tucker Decomposition

Tensor Rank and Tensor Decomposition

- **Rank-1 Tensor**: a single tensor product of vectors
- **Tensor Rank**: smallest number r of rank-1 tensors under Canonical Polyadic Decomposition
- **Tensor Multi-Rank**: smallest dimensions of core tensor G under Tucker Decomposition (it always exists)
- **Tensor Rank** is always equal to the rank of Tucker tensor core: $rank(X) = rank(G)$
- **Nonnegative Tensor Rank**: smallest number r of rank-1 tensors under Nonnegative Canonical Polyadic Decomposition
- **Nonnegative Tensor Multi-Rank**: smallest dimension of core tensor G under Nonnegative Tucker Decomposition (existence cannot be guaranteed)
- **Nonnegative Tensor Rank** may not be equal to the rank of Nonnegative Tucker tensor core: $rank_+(X) \leq rank_+(G)$
- **Nonnegative Tensor Rank**: smallest number r of nonzero entries in tensor core under Nonnegative Tucker Decomposition

Tensor Rank and Tensor Decomposition

- **Rank-1 Tensor**: a single tensor product of vectors
- **Tensor Rank**: smallest number r of rank-1 tensors under Canonical Polyadic Decomposition
- **Tensor Multi-Rank**: smallest dimensions of core tensor G under Tucker Decomposition (it always exists)
- **Tensor Rank** is always equal to the rank of Tucker tensor core: $rank(X) = rank(G)$
- **Nonnegative Tensor Rank**: smallest number r of rank-1 tensors under Nonnegative Canonical Polyadic Decomposition
- **Nonnegative Tensor Multi-Rank**: smallest dimension of core tensor G under Nonnegative Tucker Decomposition (existence cannot be guaranteed)
- **Nonnegative Tensor Rank** may not be equal to the rank of Nonnegative Tucker tensor core: $rank_+(X) \leq rank_+(G)$
- **Nonnegative Tensor Rank**: smallest number r of nonzero entries in tensor core under Nonnegative Tucker Decomposition

Tensor Rank and Tensor Decomposition

- **Rank-1 Tensor**: a single tensor product of vectors
- **Tensor Rank**: smallest number r of rank-1 tensors under Canonical Polyadic Decomposition
- **Tensor Multi-Rank**: smallest dimensions of core tensor G under Tucker Decomposition (it always exists)
- **Tensor Rank** is always equal to the rank of Tucker tensor core: $rank(X) = rank(G)$
- **Nonnegative Tensor Rank**: smallest number r of rank-1 tensors under Nonnegative Canonical Polyadic Decomposition
- **Nonnegative Tensor Multi-Rank**: smallest dimension of core tensor G under Nonnegative Tucker Decomposition (existence cannot be guaranteed)
- **Nonnegative Tensor Rank** may not be equal to the rank of Nonnegative Tucker tensor core: $rank_+(X) \leq rank_+(G)$
- **Nonnegative Tensor Rank**: smallest number r of nonzero entries in tensor core under Nonnegative Tucker Decomposition

Tensor Rank and Tensor Decomposition

- **Rank-1 Tensor**: a single tensor product of vectors
- **Tensor Rank**: smallest number r of rank-1 tensors under Canonical Polyadic Decomposition
- **Tensor Multi-Rank**: smallest dimensions of core tensor G under Tucker Decomposition (it always exists)
- **Tensor Rank** is always equal to the rank of Tucker tensor core: $rank(X) = rank(G)$
- **Nonnegative Tensor Rank**: smallest number r of rank-1 tensors under Nonnegative Canonical Polyadic Decomposition
- **Nonnegative Tensor Multi-Rank**: smallest dimension of core tensor G under Nonnegative Tucker Decomposition (existence cannot be guaranteed)
- **Nonnegative Tensor Rank** may not be equal to the rank of Nonnegative Tucker tensor core: $rank_+(X) \leq rank_+(G)$
- **Nonnegative Tensor Rank**: smallest number r of nonzero entries in tensor core under Nonnegative Tucker Decomposition

Tensor Rank and Tensor Decomposition

- **Rank-1 Tensor**: a single tensor product of vectors
- **Tensor Rank**: smallest number r of rank-1 tensors under Canonical Polyadic Decomposition
- **Tensor Multi-Rank**: smallest dimensions of core tensor G under Tucker Decomposition (it always exists)
- **Tensor Rank** is always equal to the rank of Tucker tensor core: $rank(X) = rank(G)$
- **Nonnegative Tensor Rank**: smallest number r of rank-1 tensors under Nonnegative Canonical Polyadic Decomposition
- **Nonnegative Tensor Multi-Rank**: smallest dimension of core tensor G under Nonnegative Tucker Decomposition (existence cannot be guaranteed)
- **Nonnegative Tensor Rank** may not be equal to the rank of Nonnegative Tucker tensor core: $rank_+(X) \leq rank_+(G)$
- **Nonnegative Tensor Rank**: smallest number r of nonzero entries in tensor core under Nonnegative Tucker Decomposition

Tensor Rank and Tensor Decomposition

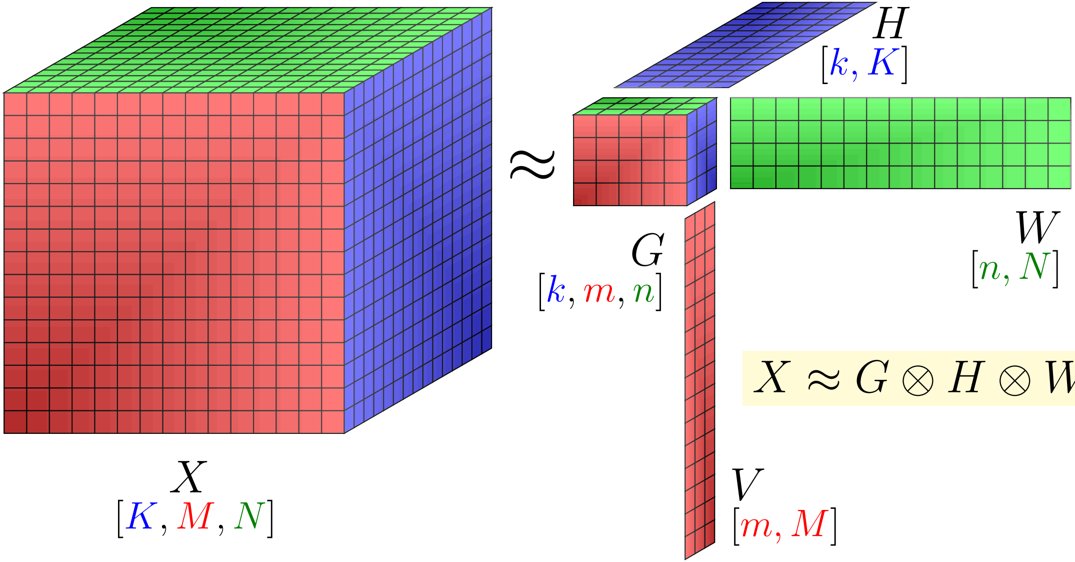
- **Rank-1 Tensor**: a single tensor product of vectors
- **Tensor Rank**: smallest number r of rank-1 tensors under Canonical Polyadic Decomposition
- **Tensor Multi-Rank**: smallest dimensions of core tensor G under Tucker Decomposition (it always exists)
- **Tensor Rank** is always equal to the rank of Tucker tensor core: $rank(X) = rank(G)$
- **Nonnegative Tensor Rank**: smallest number r of rank-1 tensors under Nonnegative Canonical Polyadic Decomposition
- **Nonnegative Tensor Multi-Rank**: smallest dimension of core tensor G under Nonnegative Tucker Decomposition (existence cannot be guaranteed)
- **Nonnegative Tensor Rank** may not be equal to the rank of Nonnegative Tucker tensor core: $rank_+(X) \leq rank_+(G)$
- **Nonnegative Tensor Rank**: smallest number r of nonzero entries in tensor core under Nonnegative Tucker Decomposition

Tensor Rank and Tensor Decomposition

- **Rank-1 Tensor**: a single tensor product of vectors
- **Tensor Rank**: smallest number r of rank-1 tensors under Canonical Polyadic Decomposition
- **Tensor Multi-Rank**: smallest dimensions of core tensor G under Tucker Decomposition (it always exists)
- **Tensor Rank** is always equal to the rank of Tucker tensor core: $rank(X) = rank(G)$
- **Nonnegative Tensor Rank**: smallest number r of rank-1 tensors under Nonnegative Canonical Polyadic Decomposition
- **Nonnegative Tensor Multi-Rank**: smallest dimension of core tensor G under Nonnegative Tucker Decomposition (existence cannot be guaranteed)
- **Nonnegative Tensor Rank** may not be equal to the rank of Nonnegative Tucker tensor core: $rank_+(X) \leq rank_+(G)$
- **Nonnegative Tensor Rank**: smallest number r of nonzero entries in tensor core under Nonnegative Tucker Decomposition

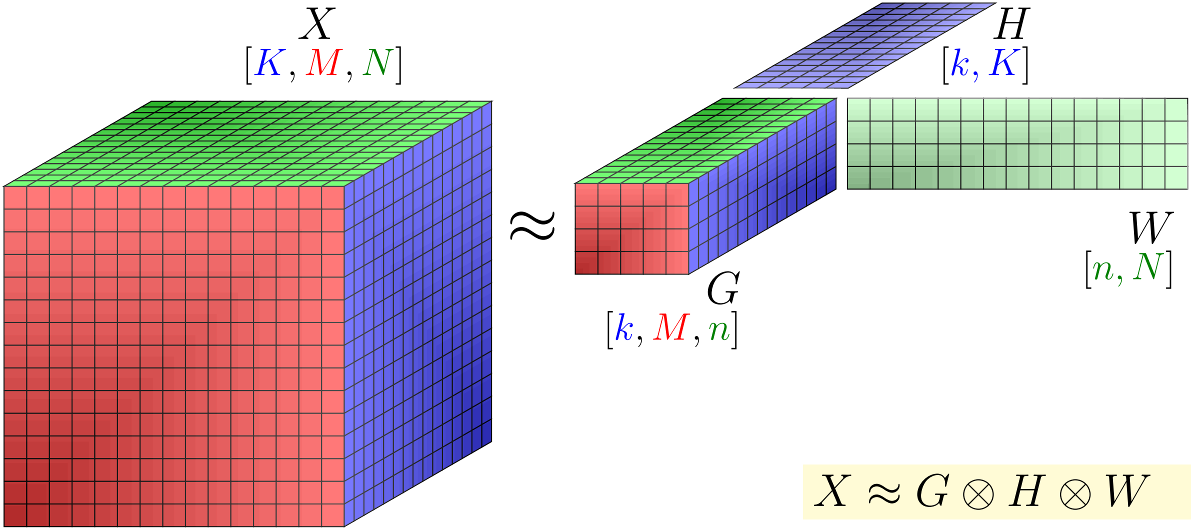
Tensor Rank and Tensor Decomposition

- **Rank-1 Tensor**: a single tensor product of vectors
- **Tensor Rank**: smallest number r of rank-1 tensors under Canonical Polyadic Decomposition
- **Tensor Multi-Rank**: smallest dimensions of core tensor G under Tucker Decomposition (it always exists)
- **Tensor Rank** is always equal to the rank of Tucker tensor core: $rank(X) = rank(G)$
- **Nonnegative Tensor Rank**: smallest number r of rank-1 tensors under Nonnegative Canonical Polyadic Decomposition
- **Nonnegative Tensor Multi-Rank**: smallest dimension of core tensor G under Nonnegative Tucker Decomposition (existence cannot be guaranteed)
- **Nonnegative Tensor Rank** may not be equal to the rank of Nonnegative Tucker tensor core: $rank_+(X) \leq rank_+(G)$
- **Nonnegative Tensor Rank**: smallest number r of nonzero entries in tensor core under Nonnegative Tucker Decomposition



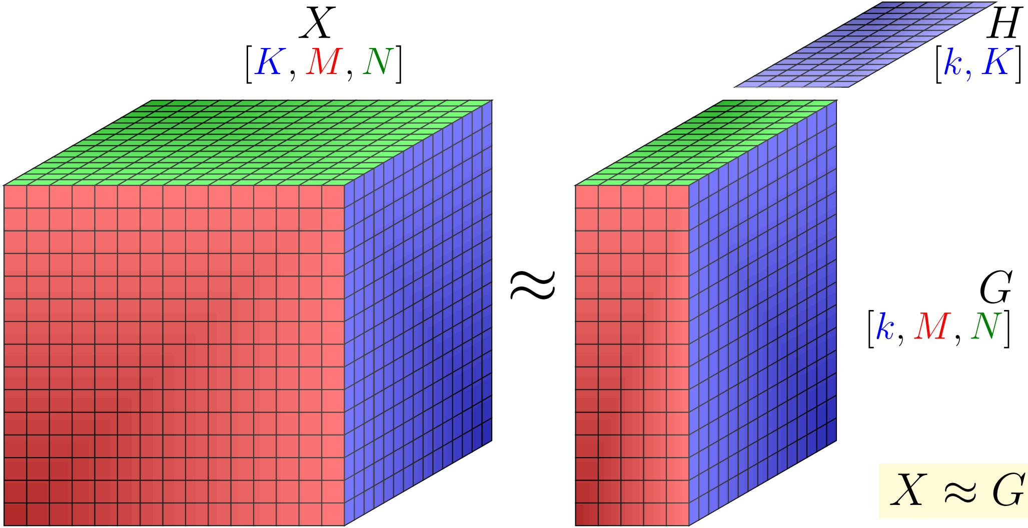
$$X \approx G \otimes H \otimes W \otimes V$$

Tucker Tensor Decomposition (3D): Tucker-2 (three possible alternatives)



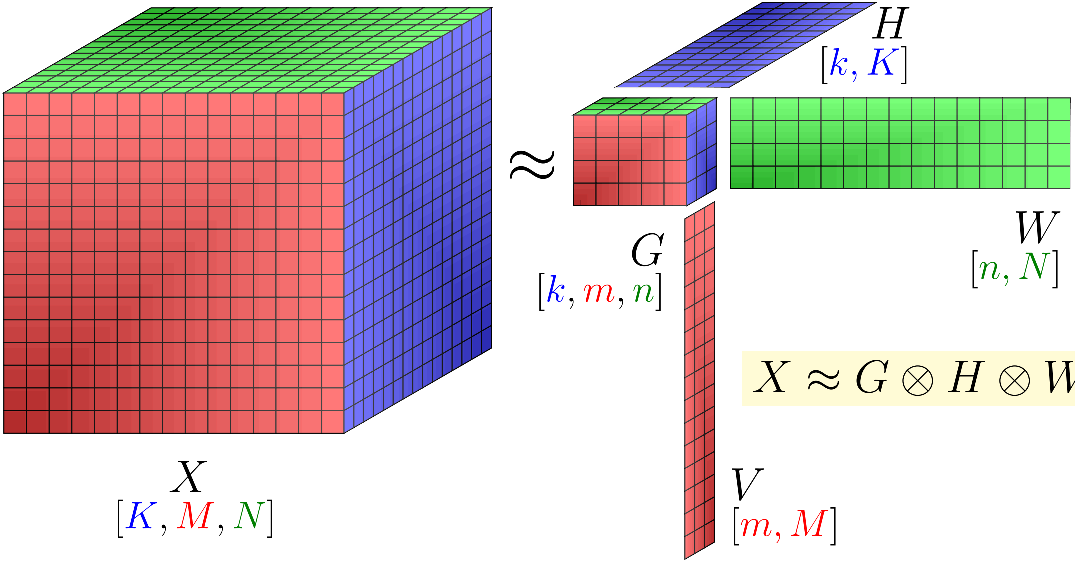
$$X \approx G \otimes H \otimes W$$

Tucker Tensor Decomposition (3D): Tucker-1 (three possible alternatives)



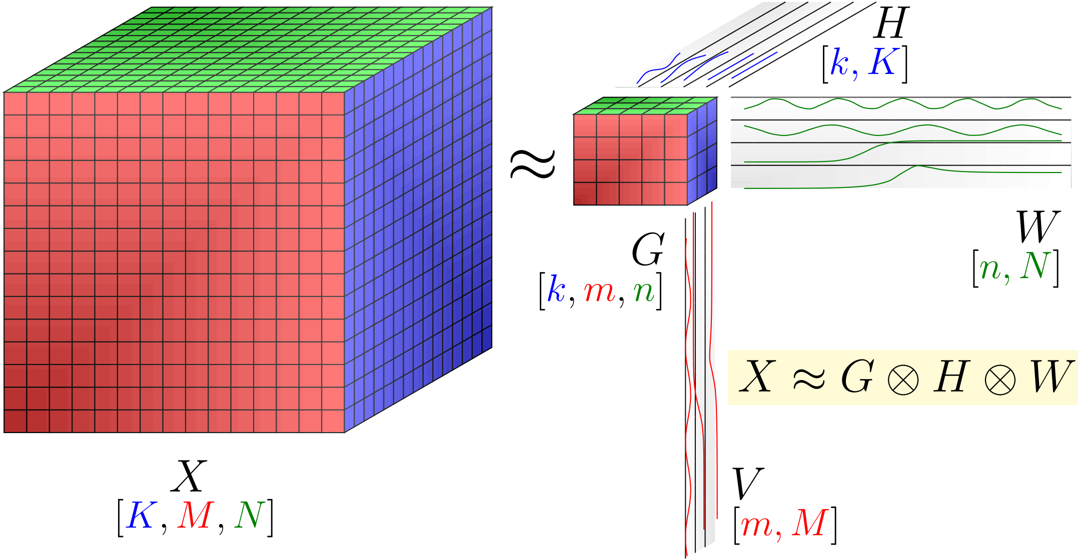
$$X \approx G \otimes H$$

Tucker Tensor Decomposition (3D): Rank-60 Multirank-(3,4,5)



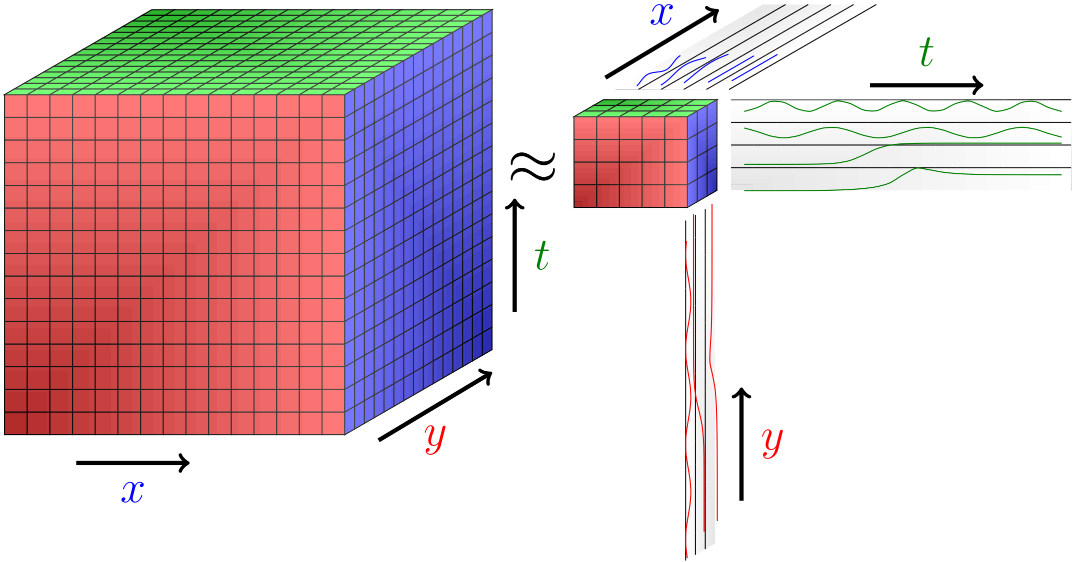
$$X \approx G \otimes H \otimes W \otimes V$$

Tucker Tensor Decomposition (3D): Feature extraction



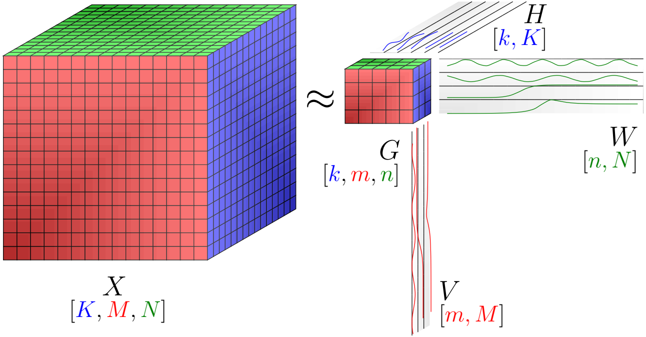
$$X \approx G \otimes H \otimes W \otimes V$$

Tucker Tensor Decomposition (3D): Feature extraction



Tensor Decomposition

- Tucker/CPD decomposition is achieved through minimization
- Nonnegativity and sparsity constraints applied
- Optimal number of features $[k, m, n]$ is estimated through k-means clustering of a series minimization solutions with random initial guesses



NMFk / NTFk Analyses

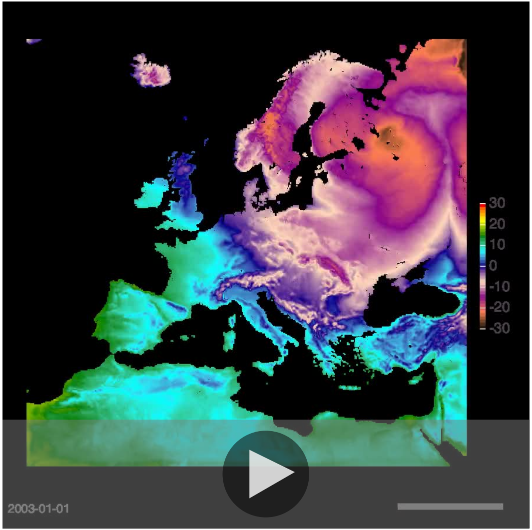
- **Field Data**:
  - Groundwater contaminant migration
  - US Climate
  - Geothermal
  - Seismic
- **Lab Data**:
  - X-ray Spectroscopy
  - UV Fluorescence Spectroscopy
  - Microbial population analyses
- **Operational Data**:
  - LANSCE: Los Alamos Neutron Accelerator
  - Hydrocarbon (oil/gas) production
- **Model Data**:
  - Reactive mixing $A + B \to C$
  - Phase separation of co-polymers
  - Molecular Dynamics of proteins
  - Lattice-Boltzmann simulations of fluid displacement
  - Europe Climate modeling

NMFk / NTFk Publications

- Vesselinov, Munuduru, Karra, O'Maley, Alexandrov, Unsupervised Machine Learning Based on Non-Negative Tensor Factorization for Analyzing Reactive-Mixing, **Journal of Computational Physics**, 2019.
- Vesselinov, O'Malley, Alexandrov, Nonnegative Tensor Factorization for Contaminant Source Identification, **Journal of Contaminant Hydrology**, 2018.
- 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.
- O'Malley, Vesselinov, Alexandrov, Alexandrov, Nonnegative/binary matrix factorization with a D-Wave quantum annealer, **PLOS ONE**, 2018.
- Vesselinov, O'Malley, Alexandrov, Contaminant source identification using semi-supervised machine learning, **Journal of Contaminant Hydrology**, 2017.
- Alexandrov, Vesselinov, Blind source separation for groundwater level analysis based on nonnegative matrix factorization, **WRR**, 2014.

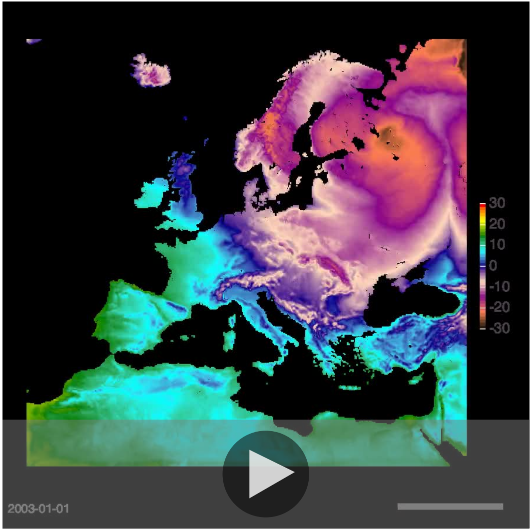
Climate model of Europe

- **TerrSysMP**: coupled climate simulator (CLM/ParFlow) of atmospheric, surface and subsurface processes

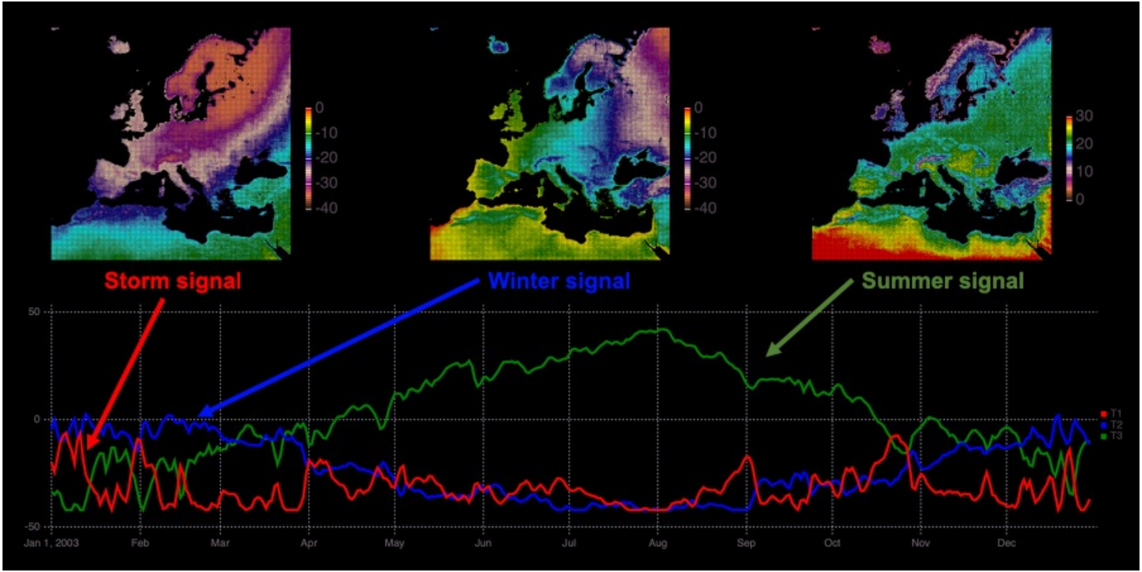


Climate model of Europe: 2003 air temperature

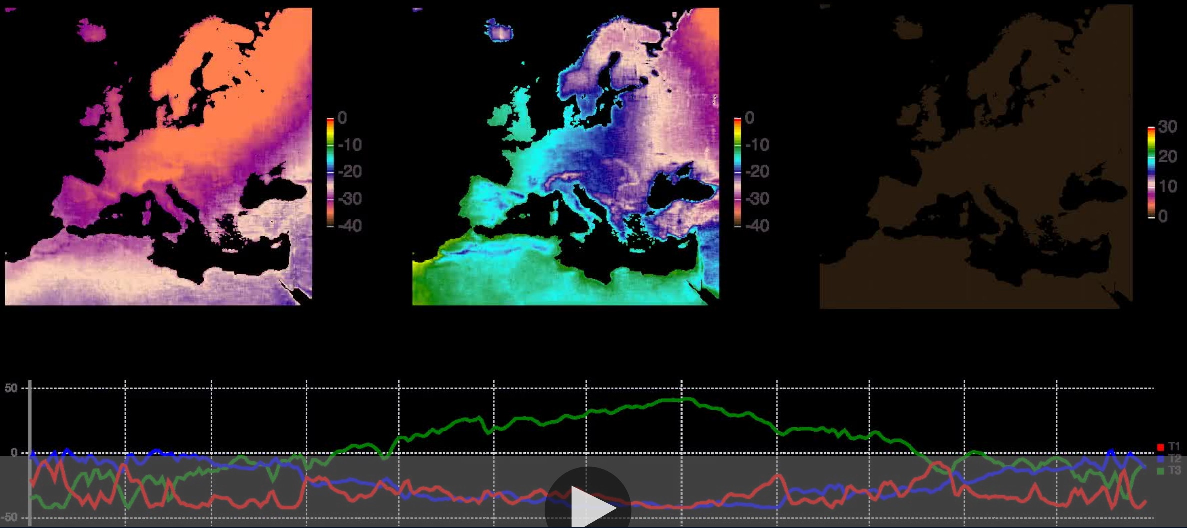
- Daily fluctuations in the air temperature $[^{\circ}C]$
- Tensor: $(424 \times 412 \times 365)$ $(columns \times rows \times days)$
- **NTF**k applied to extract dominant hidden (latent) features



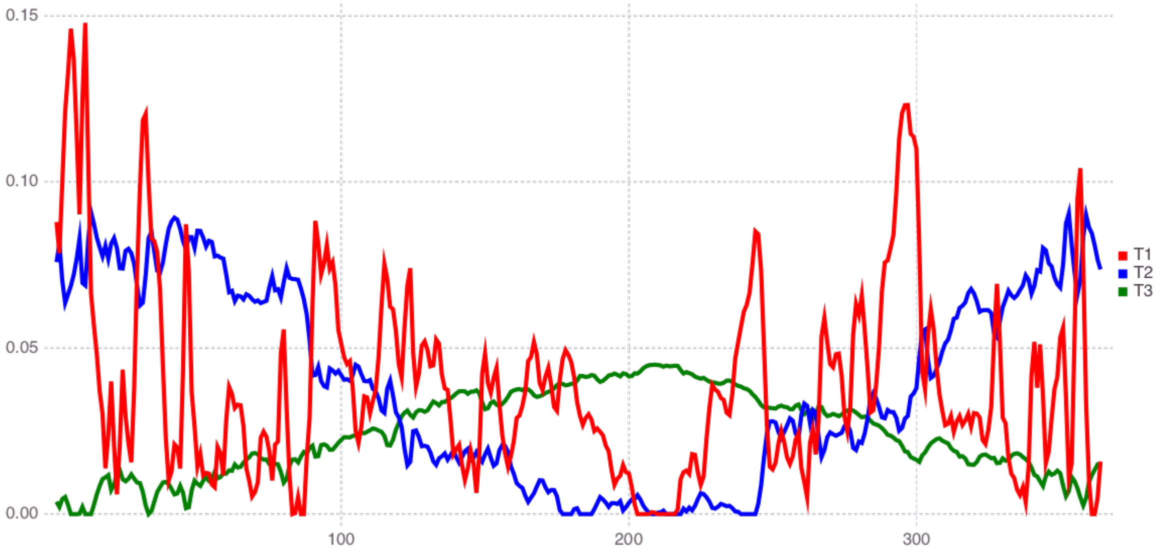
Climate model of Europe: 2003 temperature fluctuations represented by 3 features



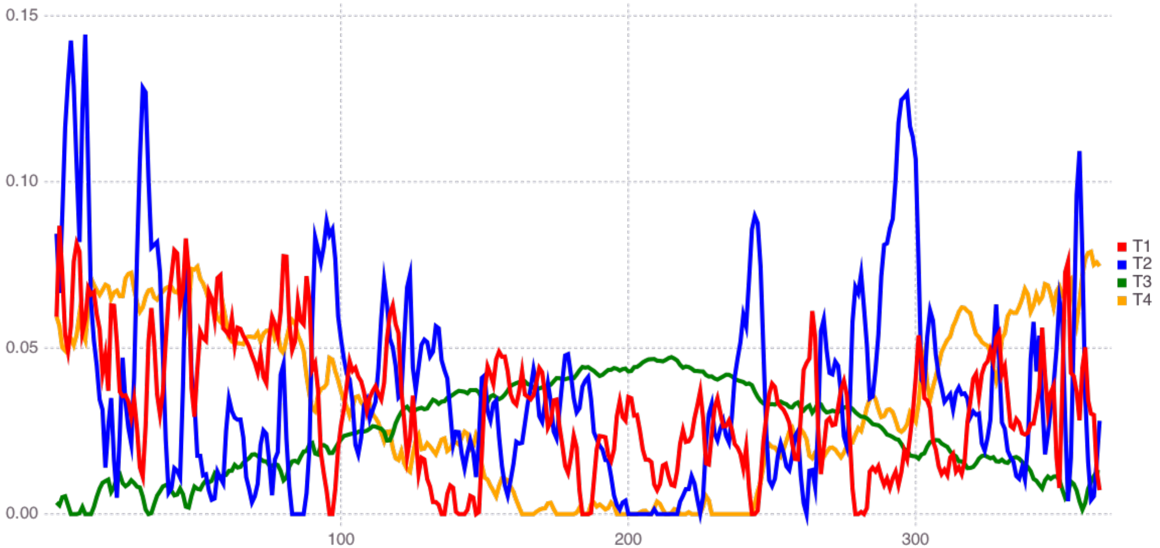
Climate model of Europe: 2003 temperature fluctuations represented by 3 features



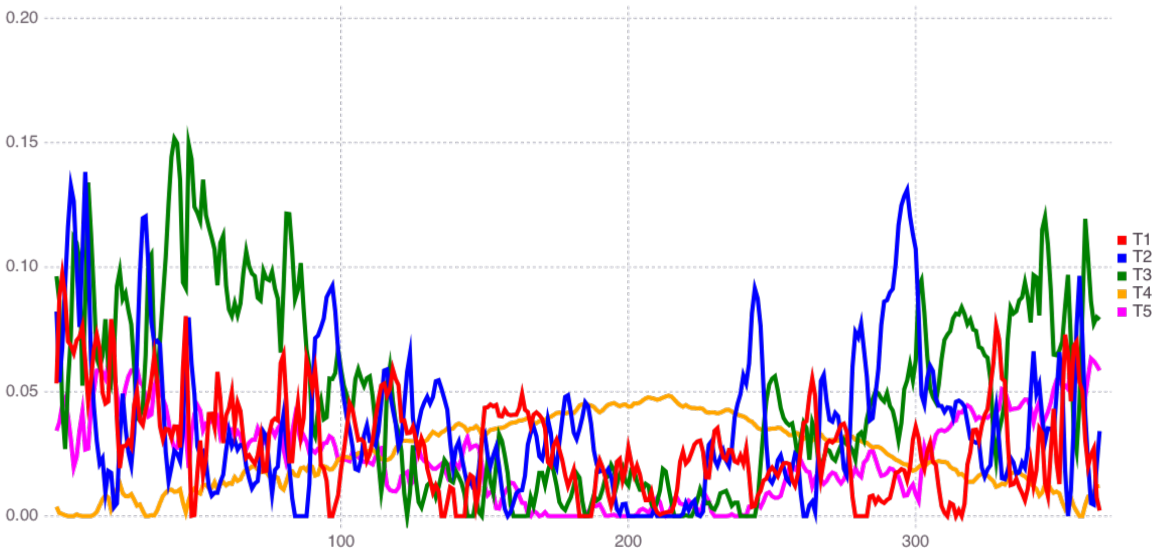
Climate model of Europe: 2003 temperature fluctuations represented by 3 features



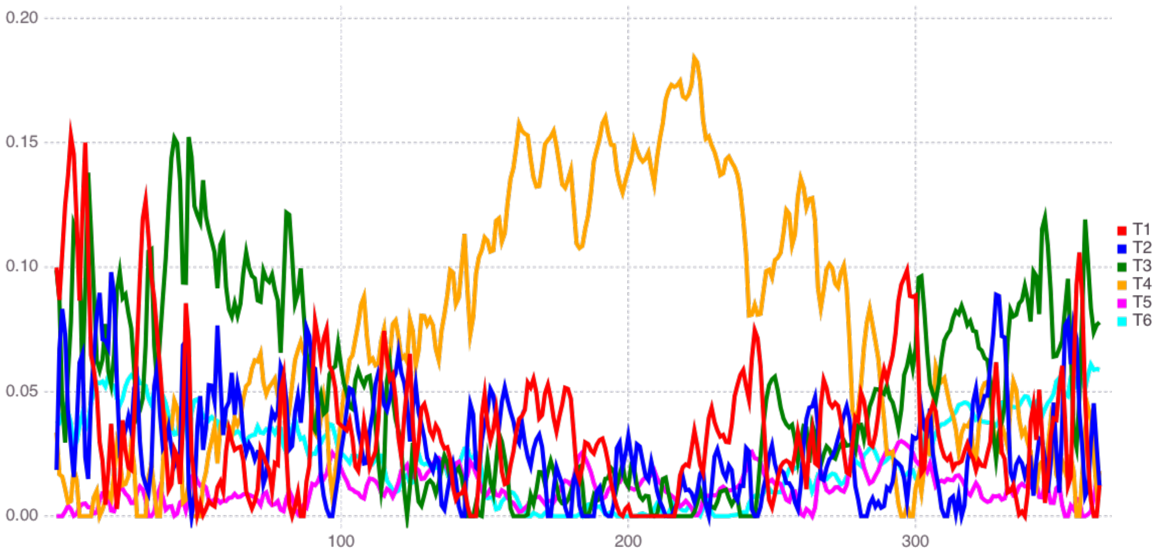
Climate model of Europe: 2003 temperature fluctuations represented by 4 features



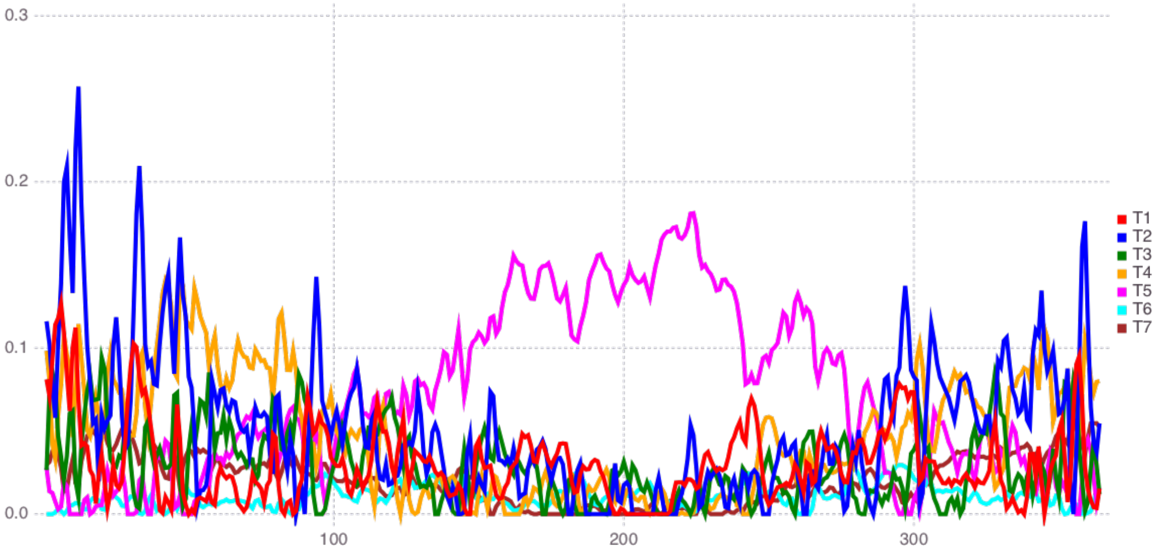
Climate model of Europe: 2003 temperature fluctuations represented by 5 features



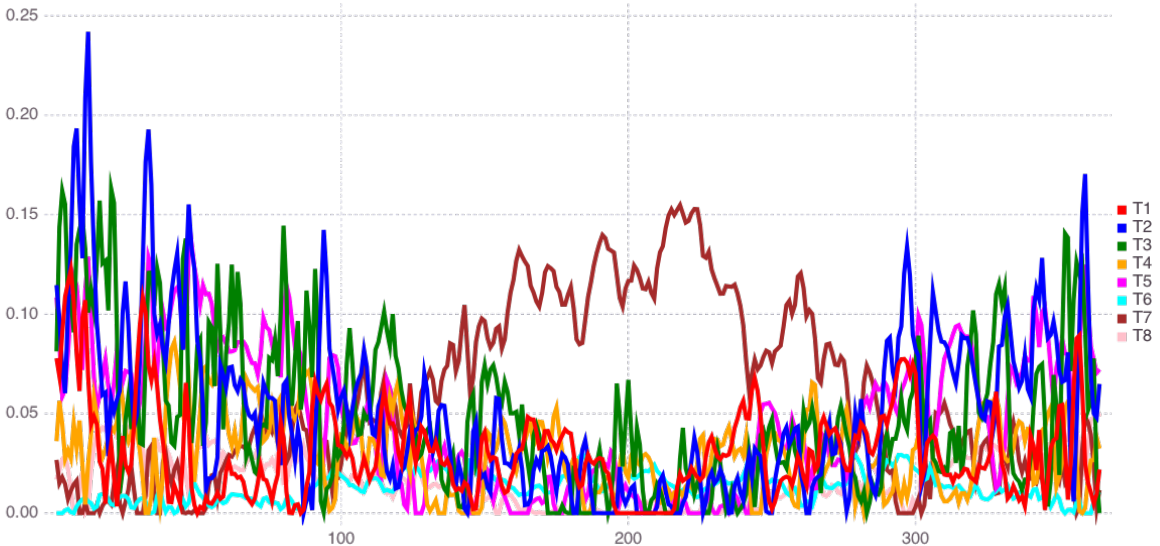
Climate model of Europe: 2003 temperature fluctuations represented by 6 features



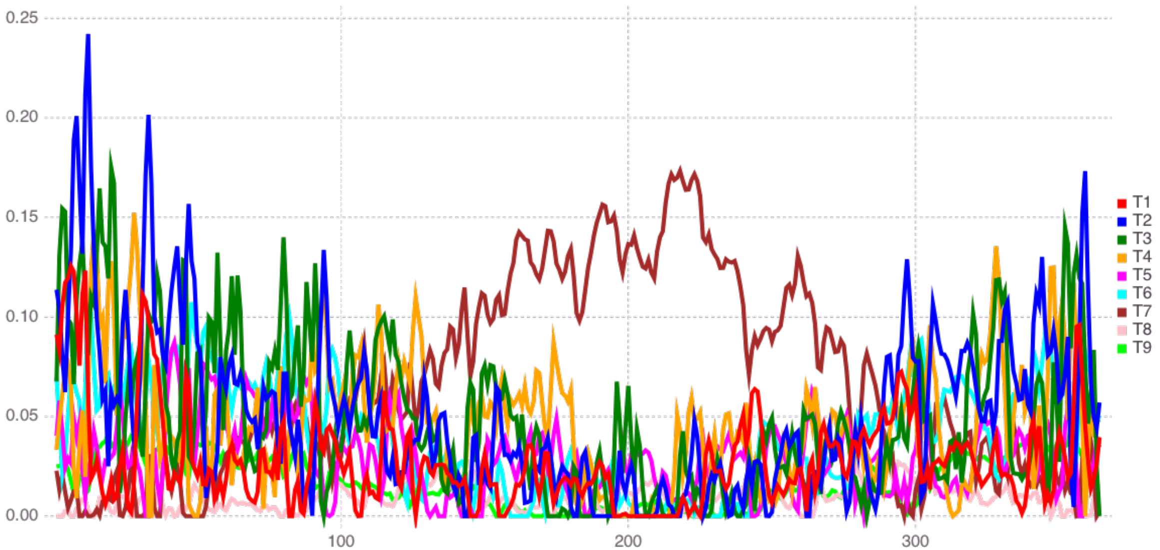
Climate model of Europe: 2003 temperature fluctuations represented by 7 features



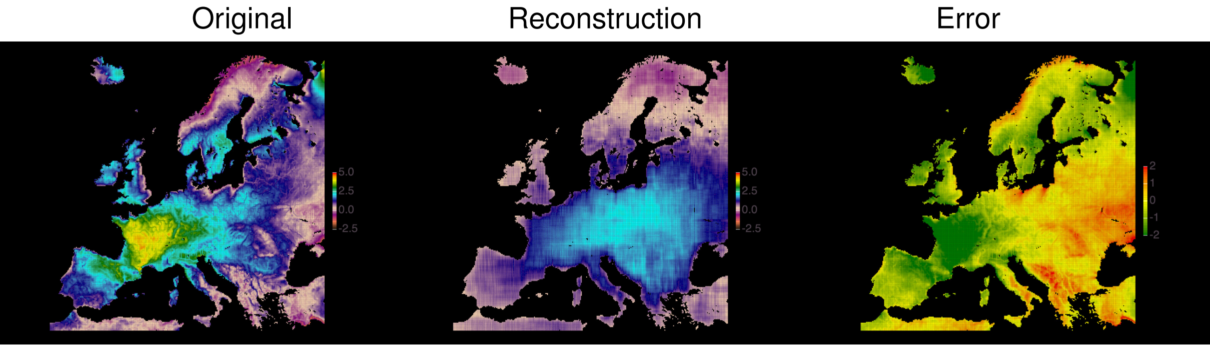
Climate model of Europe: 2003 temperature fluctuations represented by 8 features



Climate model of Europe: 2003 temperature fluctuations represented by 9 features

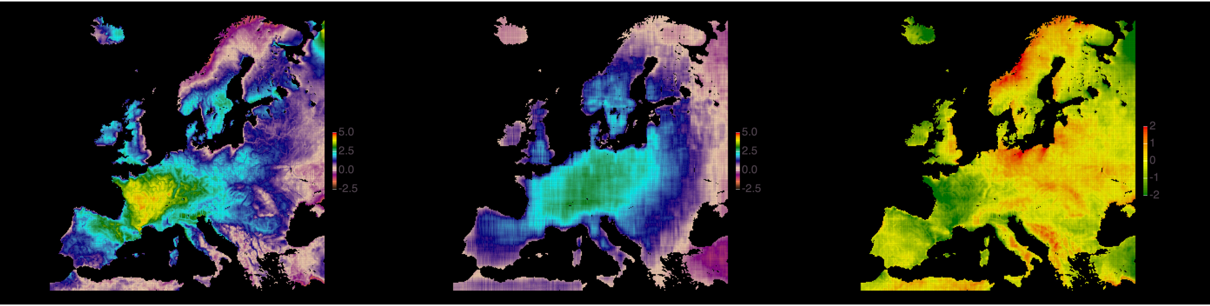


Climate model of Europe: 2003 air temperature reconstruction by 3 features

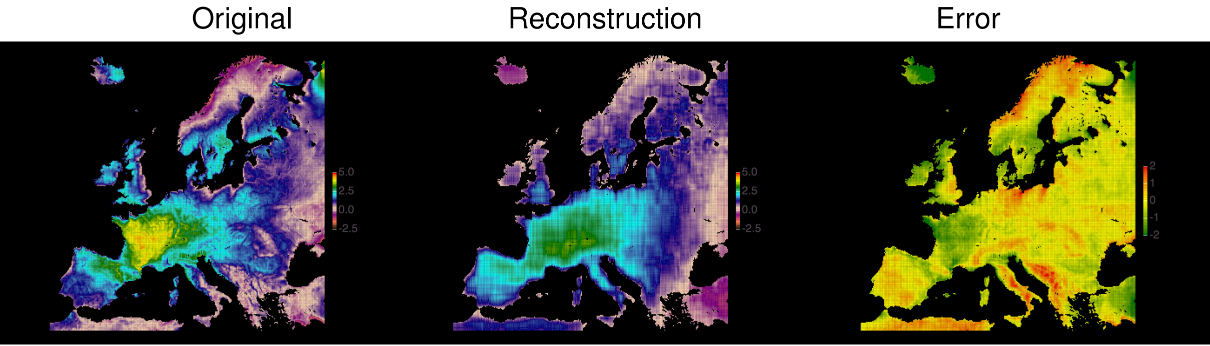


Climate model of Europe: 2003 air temperature reconstruction by 4 features

Original | Reconstruction | Error

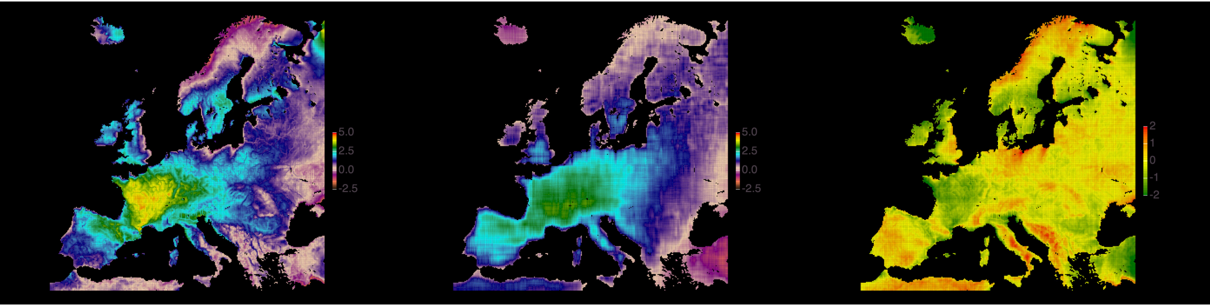


Climate model of Europe: 2003 air temperature reconstruction by 5 features

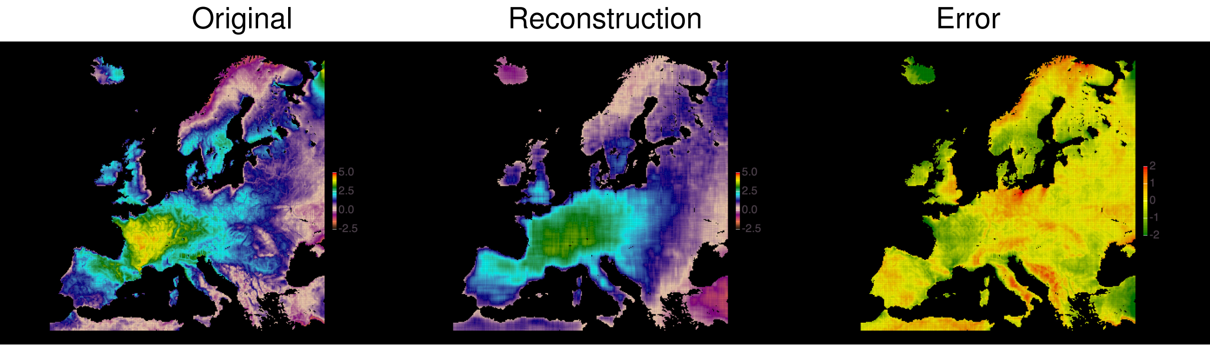


Climate model of Europe: 2003 air temperature reconstruction by 6 features

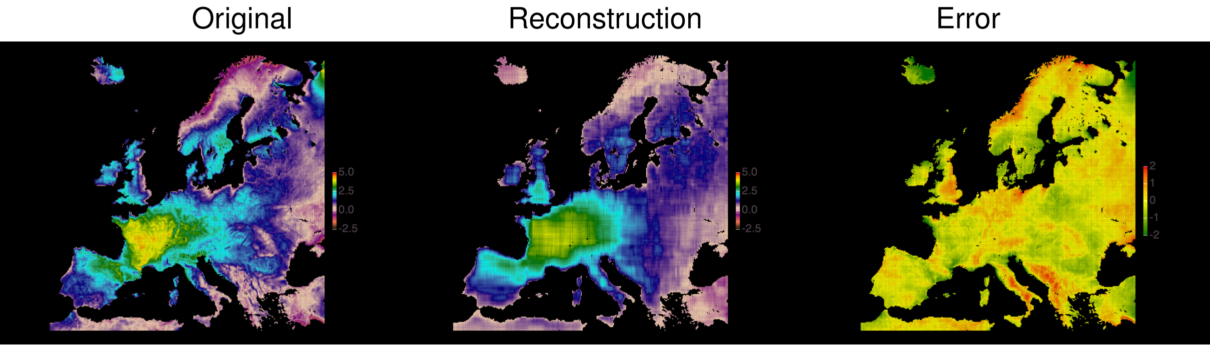
Original | Reconstruction | Error



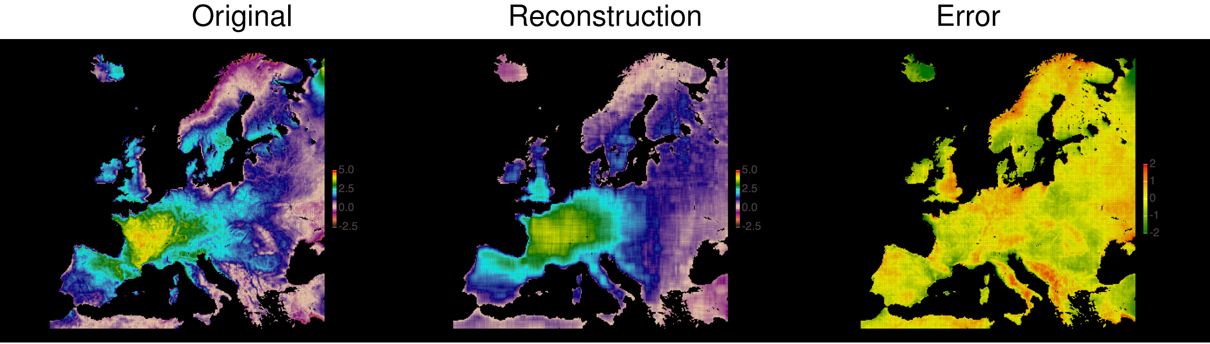
Climate model of Europe: 2003 air temperature reconstruction by 7 features



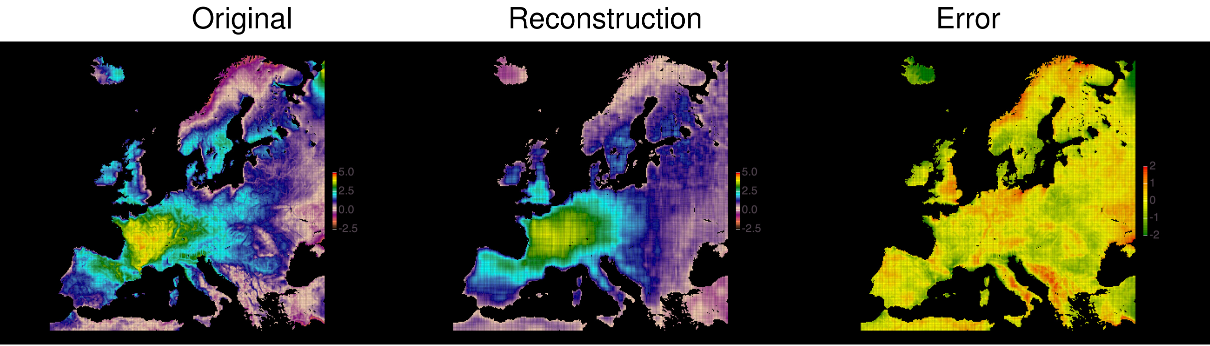
Climate model of Europe: 2003 air temperature reconstruction by 8 features



Climate model of Europe: 2003 air temperature reconstruction by 9 features

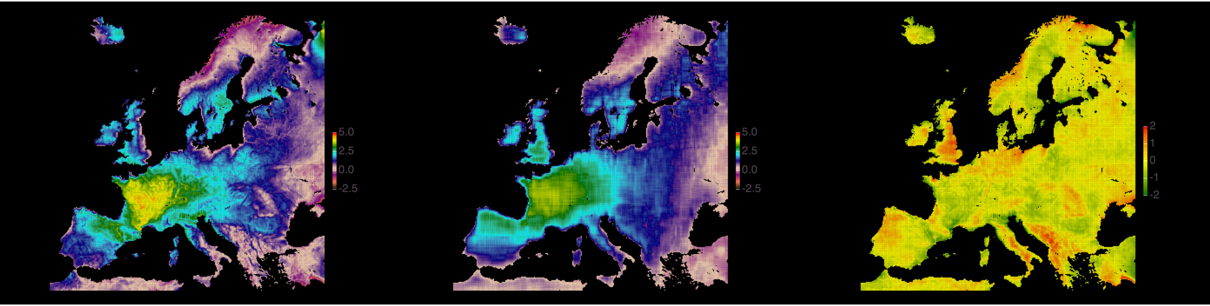


Climate model of Europe: 2003 air temperature reconstruction by 10 features



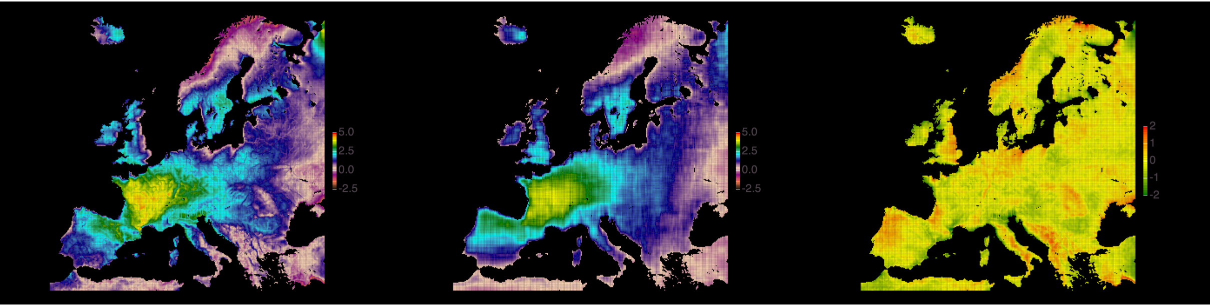
Climate model of Europe: 2003 air temperature reconstruction by 15 features

Original | Reconstruction | Error

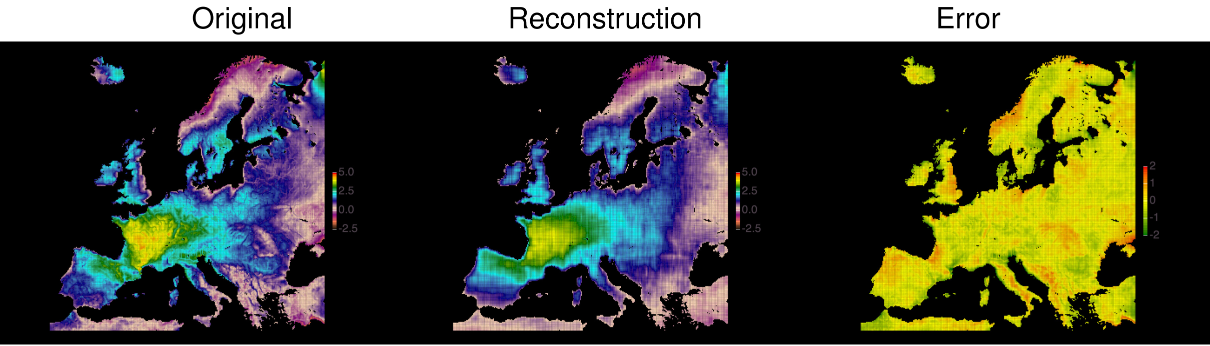


Climate model of Europe: 2003 air temperature reconstruction by 20 features

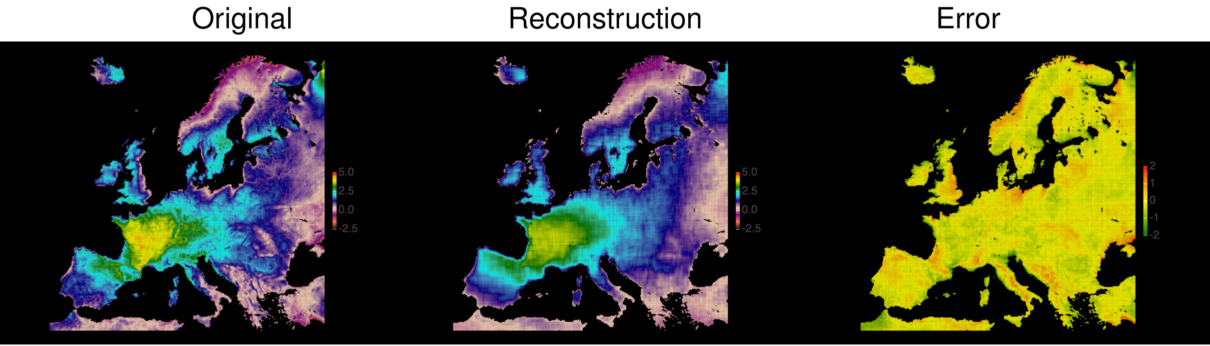
Original | Reconstruction | Error



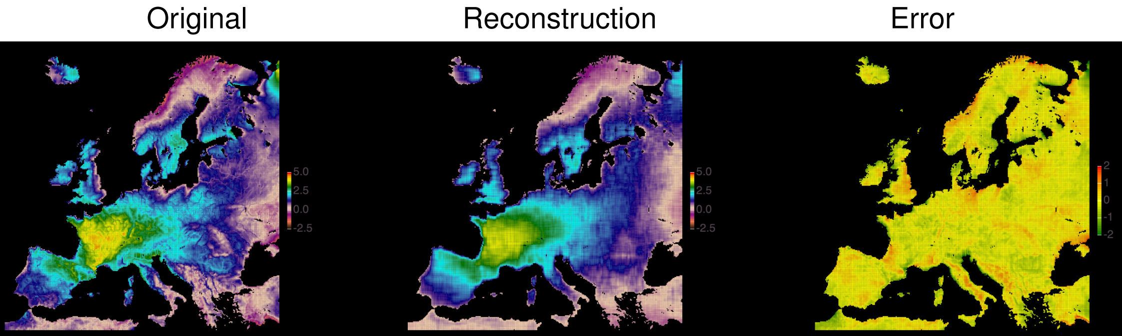
Climate model of Europe: 2003 air temperature reconstruction by 25 features



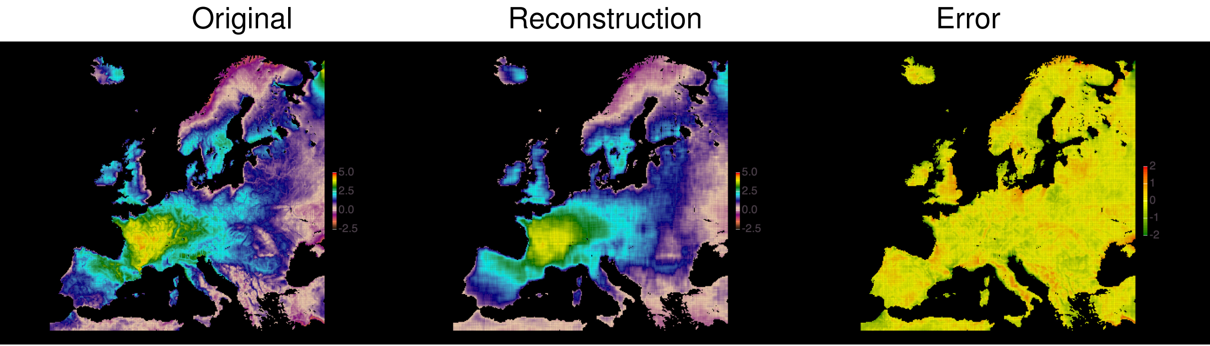
Climate model of Europe: 2003 air temperature reconstruction by 30 features



Climate model of Europe: 2003 air temperature reconstruction by 35 features

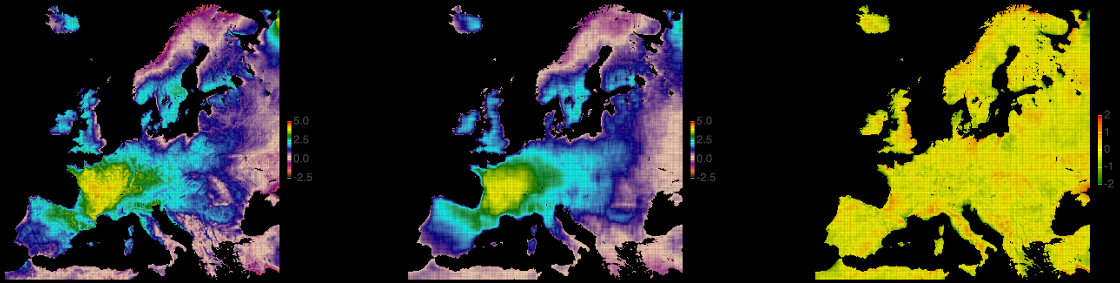


Climate model of Europe: 2003 air temperature reconstruction by 40 features

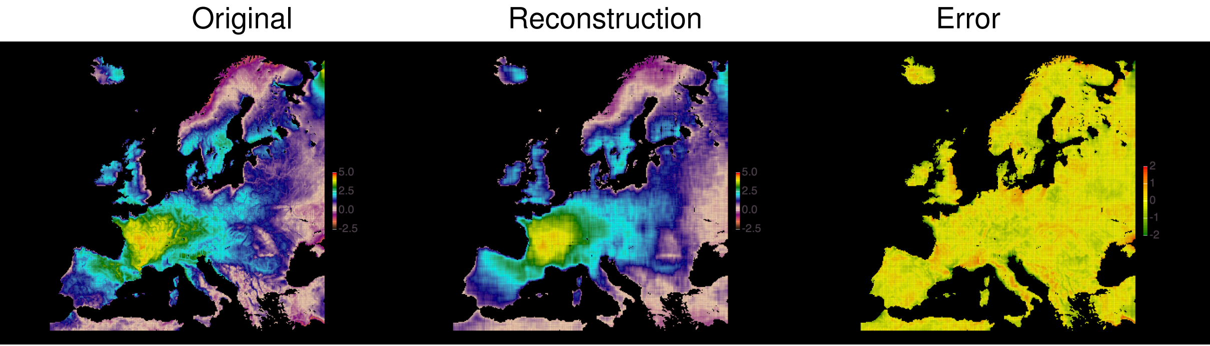


Climate model of Europe: 2003 air temperature reconstruction by 45 features

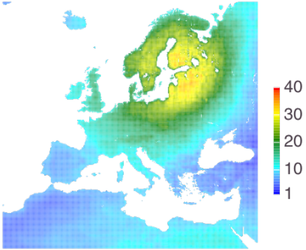
Original | Reconstruction | Error

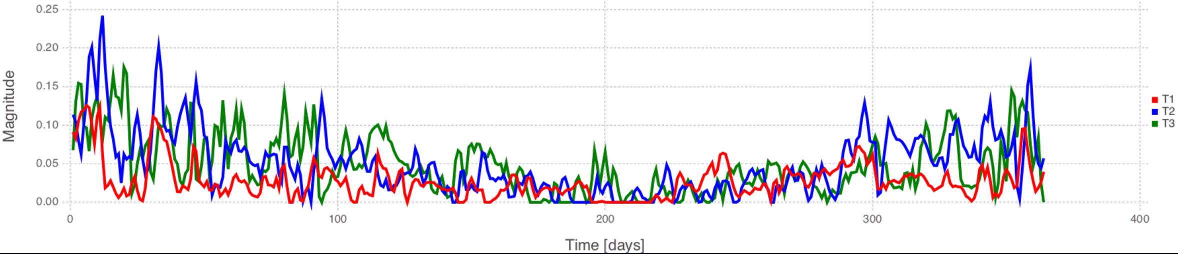


Climate model of Europe: 2003 air temperature reconstruction by 50 features

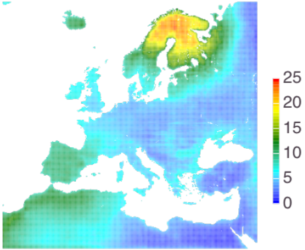


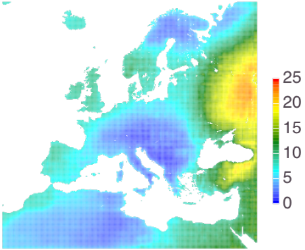
Climate model of Europe: 2003 temperature fluctuations represented by 9 features

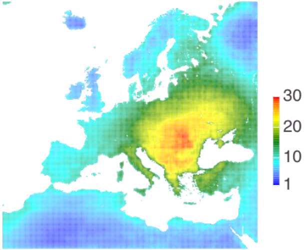


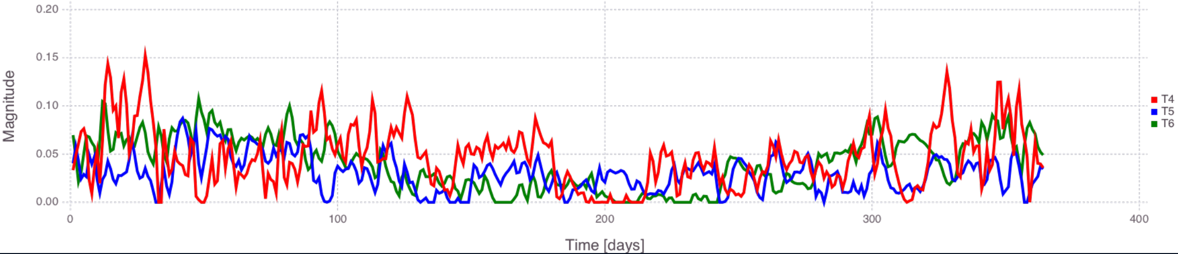


Climate model of Europe: 2003 temperature fluctuations represented by 9 features

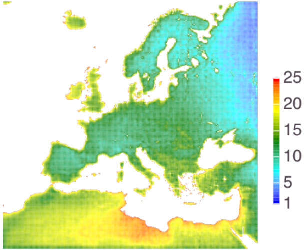


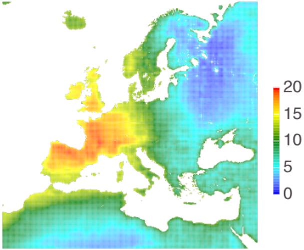


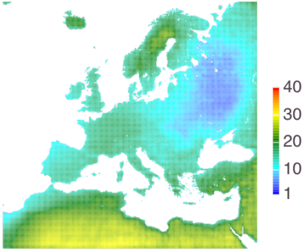


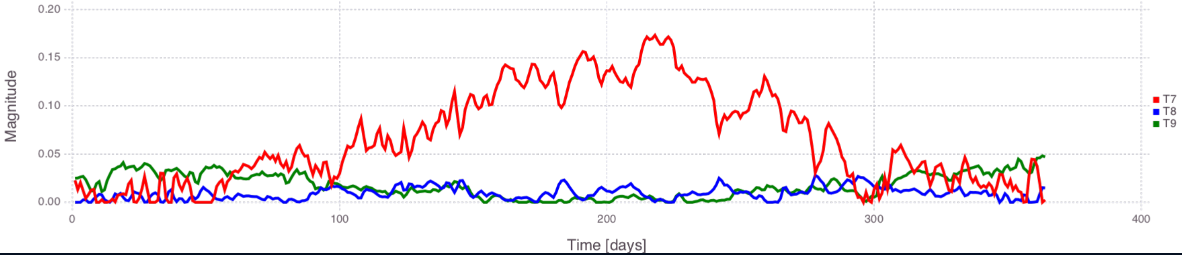


Climate model of Europe: 2003 temperature fluctuations represented by 9 features



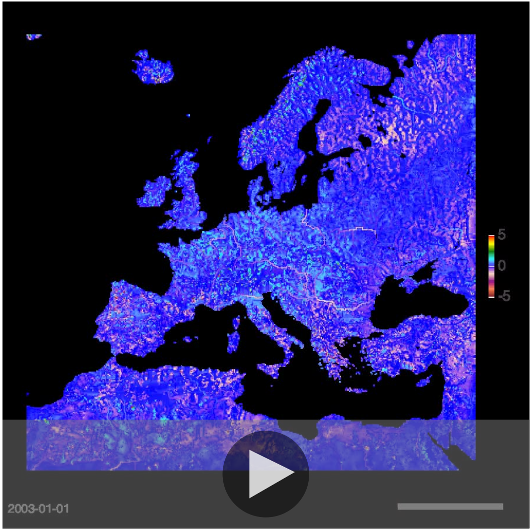




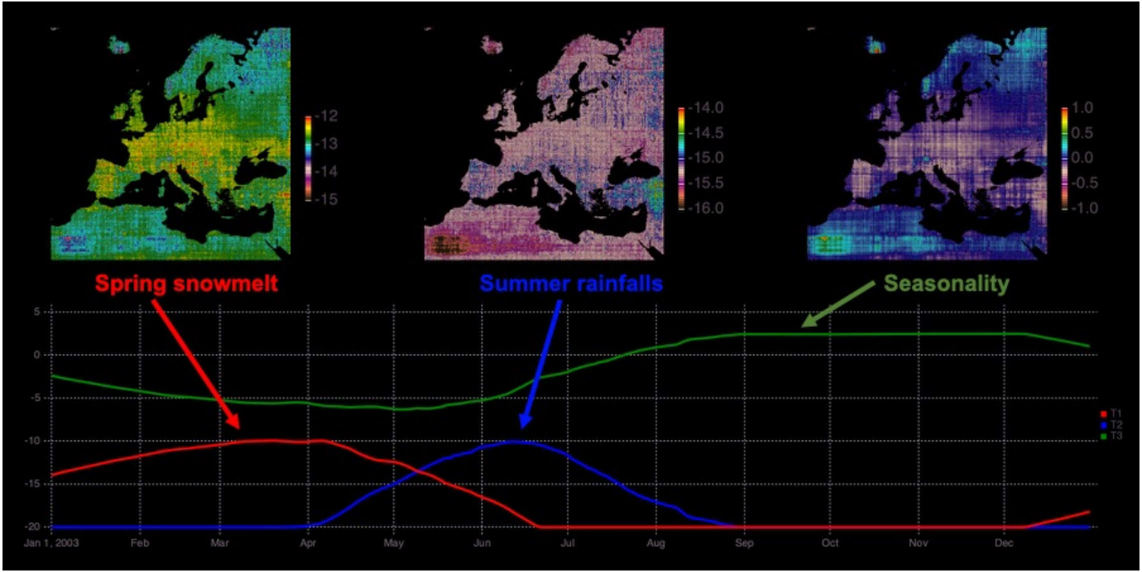


Climate model of Europe: 2003 water-table depth

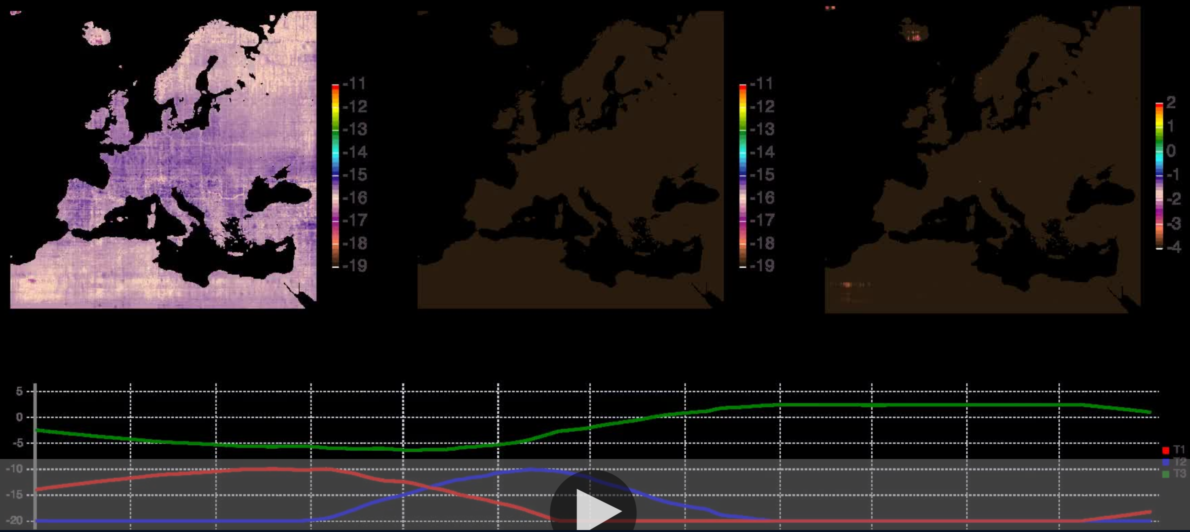
- Daily fluctuations in the water-table depth $[m]$
- Tensor: $(424 \times 412 \times 365)$ $(columns \times rows \times days)$
- **NTF**k applied to extract dominant hidden (latent) features



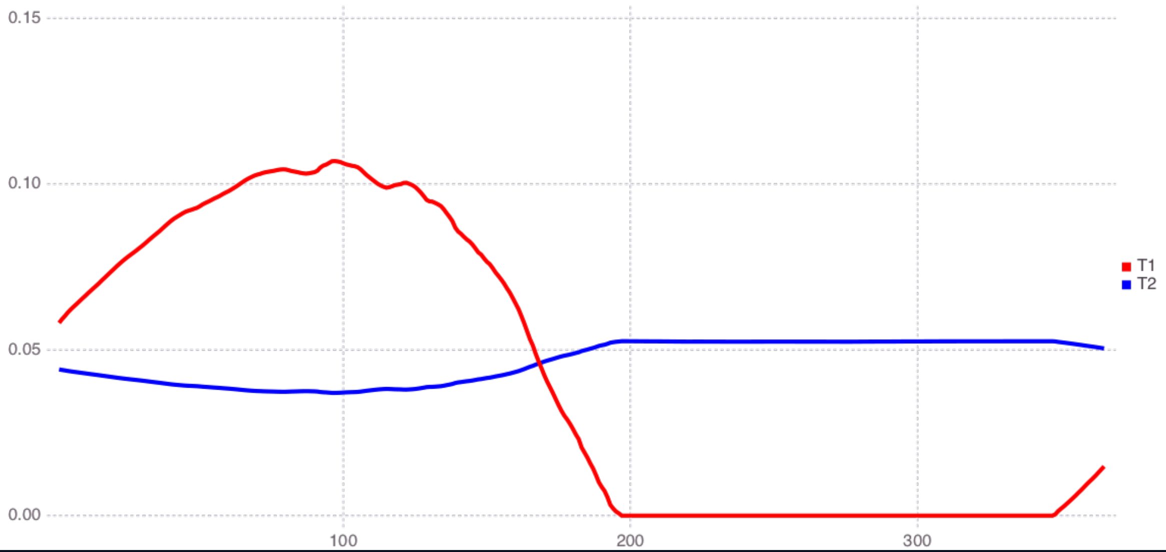
Climate model of Europe: 2003 water-table fluctuations represented by 3 features



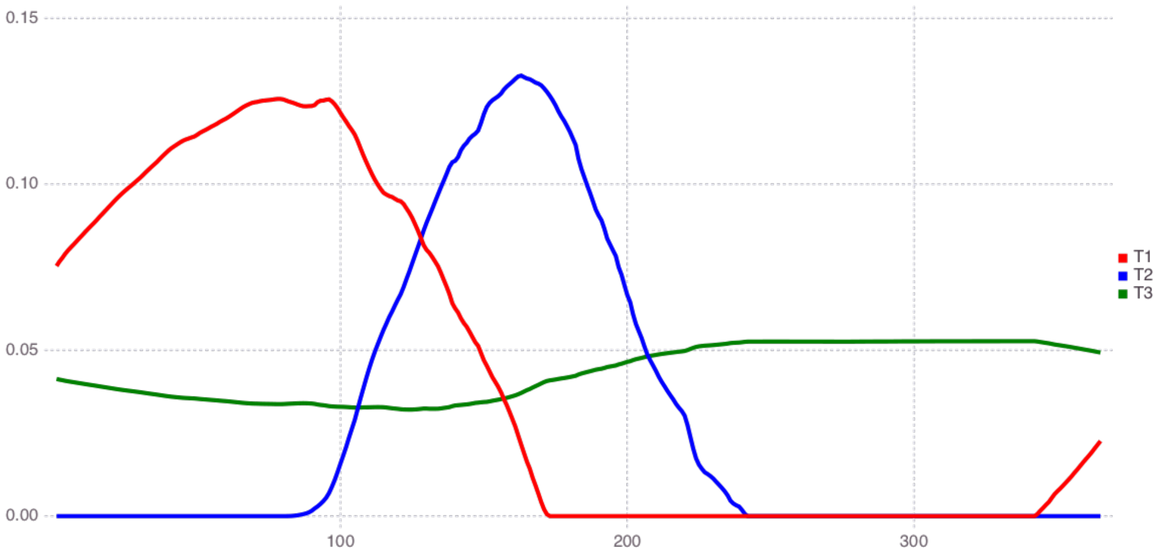
Climate model of Europe: 2003 water-table fluctuations represented by 3 features



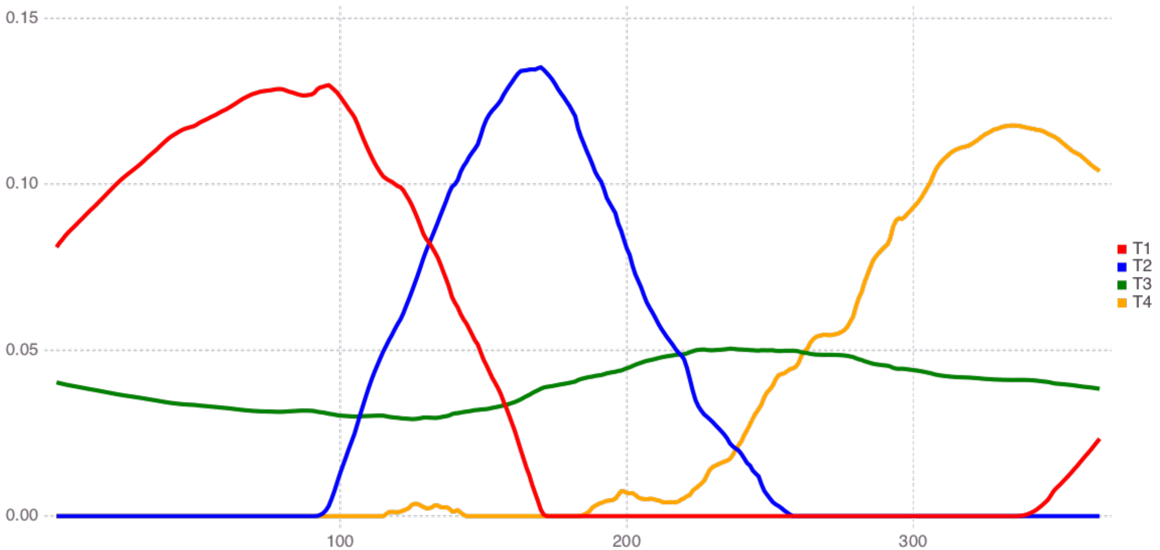
Climate model of Europe: 2003 water-table fluctuations represented by 2 features



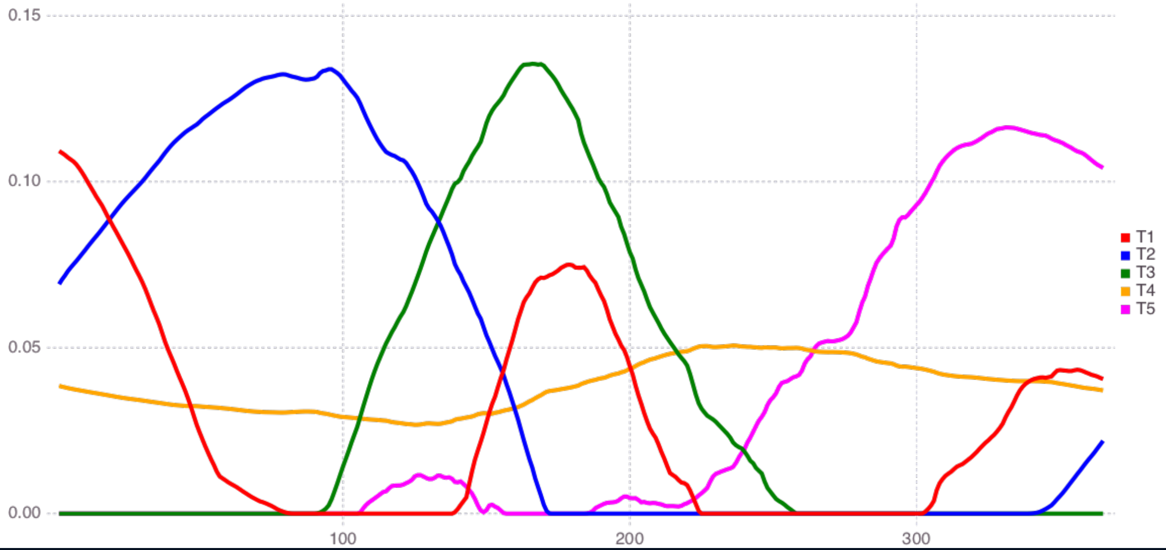
Climate model of Europe: 2003 water-table fluctuations represented by 3 features



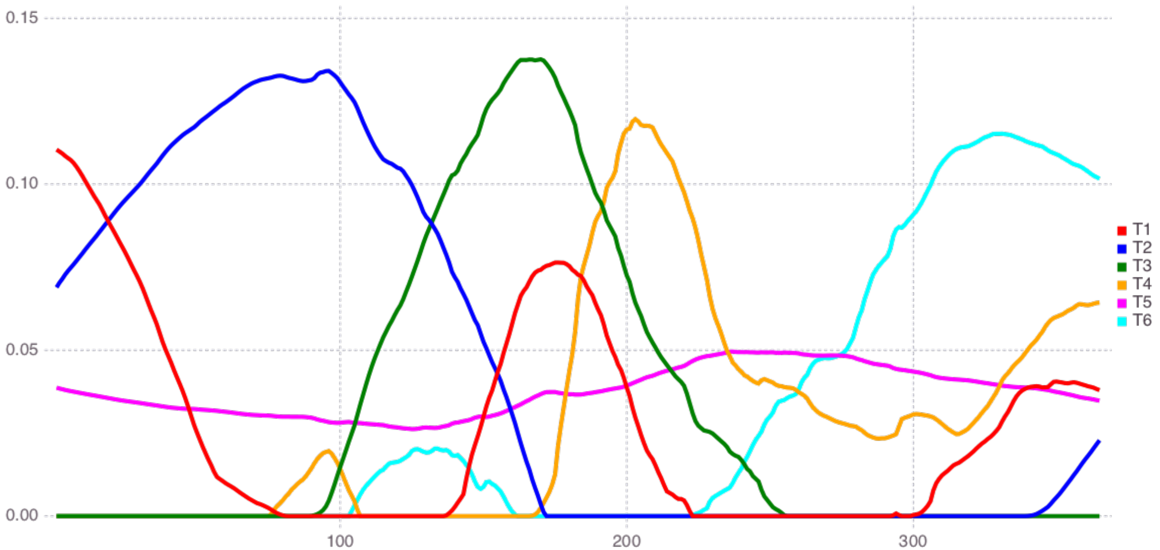
Climate model of Europe: 2003 water-table fluctuations represented by 4 features



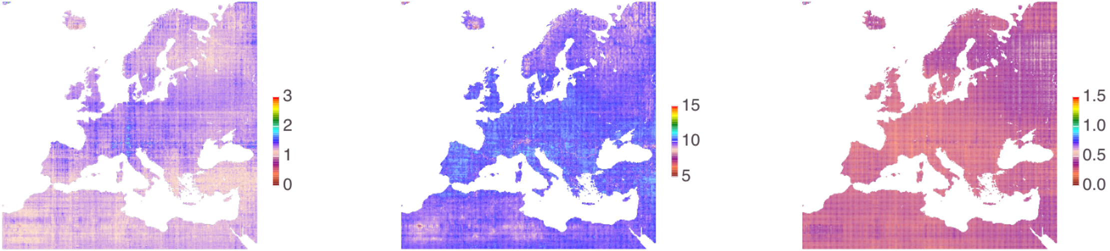
Climate model of Europe: 2003 water-table fluctuations represented by 5 features

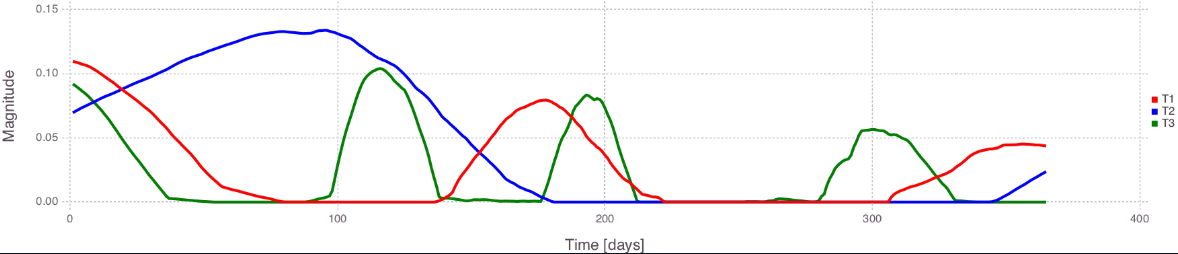


Climate model of Europe: 2003 water-table fluctuations represented by 6 features

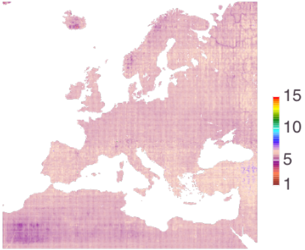


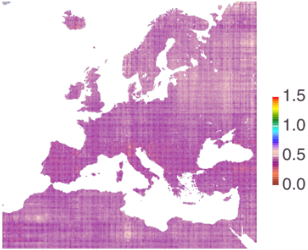
Climate model of Europe: 2003 water-table fluctuations represented by 9 features

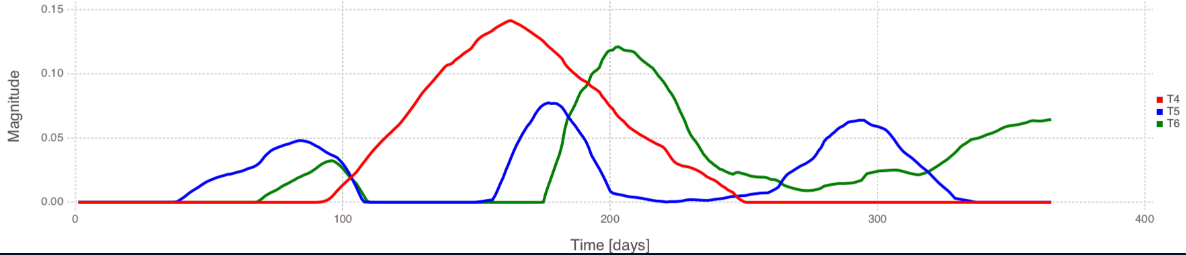




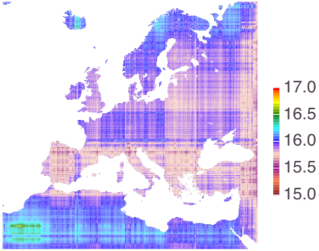
Climate model of Europe: 2003 water-table fluctuations represented by 9 features

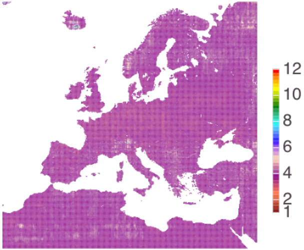


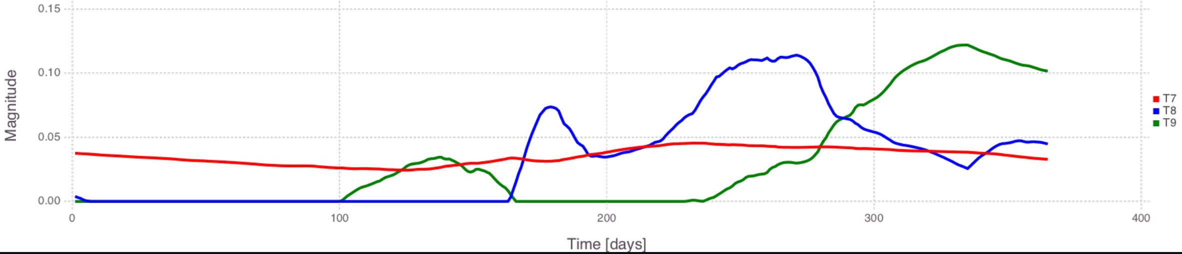




Climate model of Europe: 2003 water-table fluctuations represented by 9 features

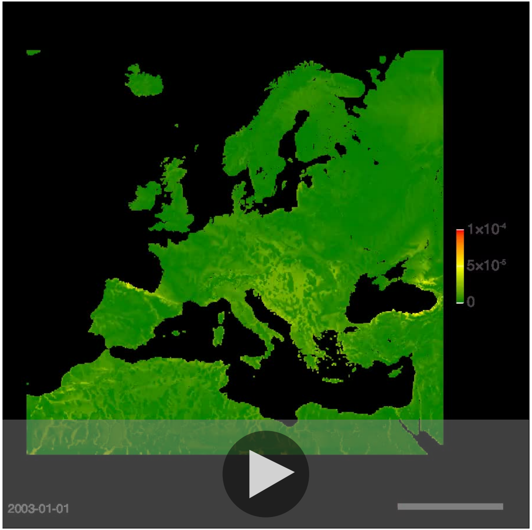




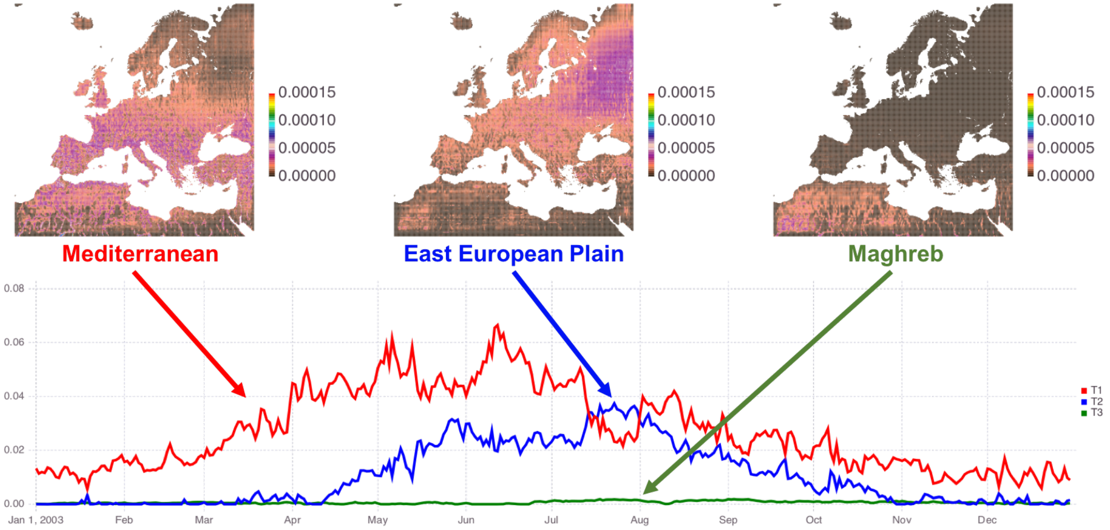


Climate model of Europe: 2003 evaporation

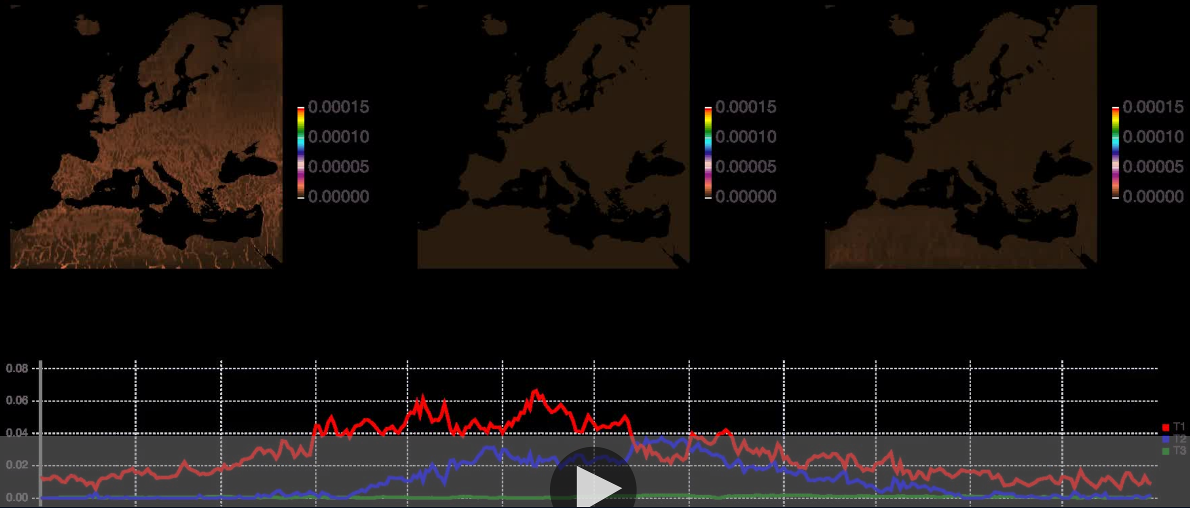
- fluctuations in evaporation $[kg/(m^2 s)]$
- $(424 \times 412 \times 365)$
  $(columns \times rows \times days)$
- **NTFk** extracts spatial and temporal footprints of dominant features



Climate model of Europe: 2003 evaporation fluctuations represented by 3 features

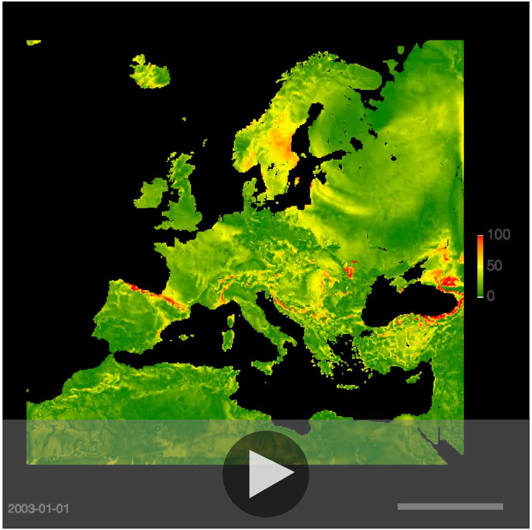


Climate model of Europe: 2003 evaporation fluctuations represented by 3 features

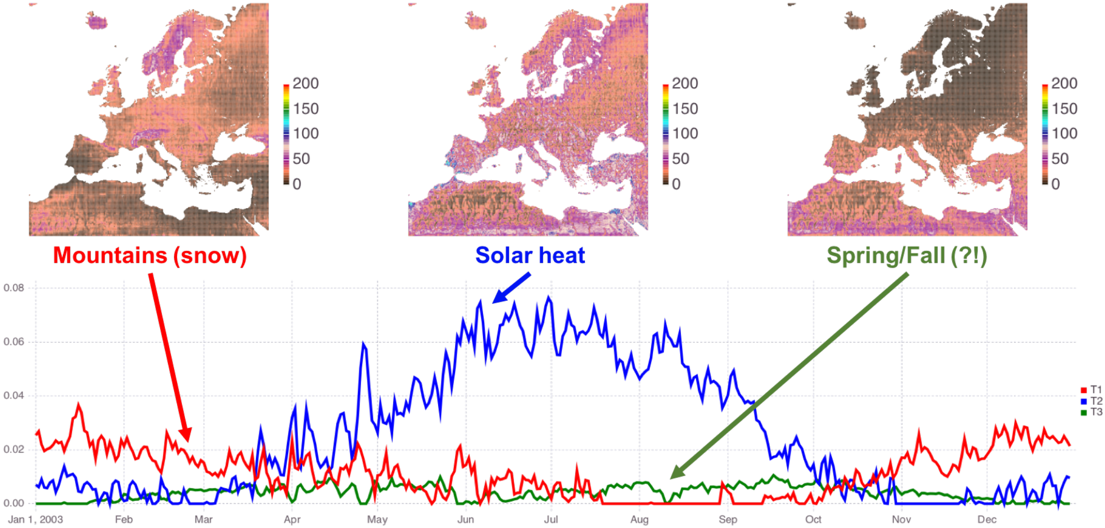


Climate model of Europe: 2003 surface upward sensible heat flux

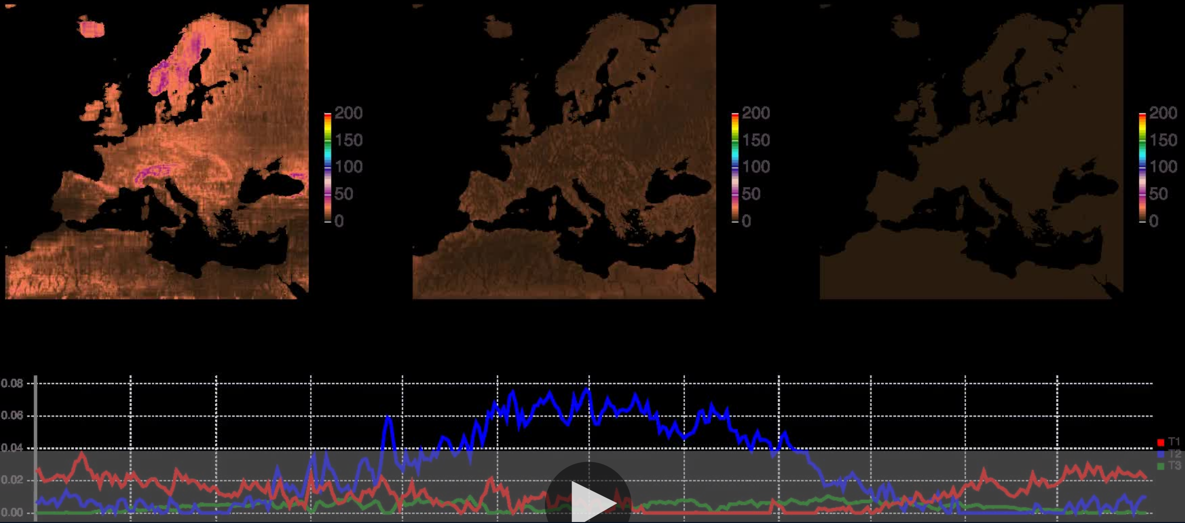
- fluctuations in surface upward sensible heat flux $[W/m^2]$
- $(424 \times 412 \times 365)$
  $(columns \times rows \times days)$
- **NTFk** extracts spatial and temporal footprints of dominant features



Climate model of Europe: 2003 heat flux fluctuations represented by 3 features

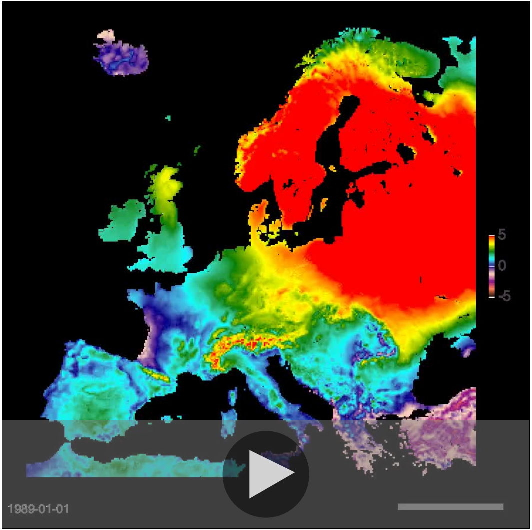


Climate model of Europe: 2003 heat flux fluctuations represented by 3 features

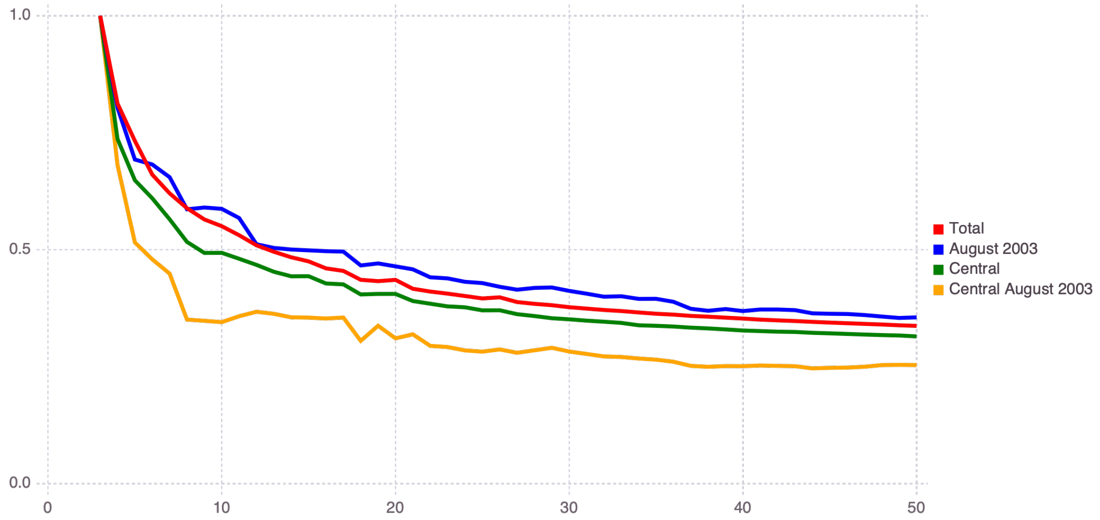


Climate model of Europe: air temperature 1989-2017

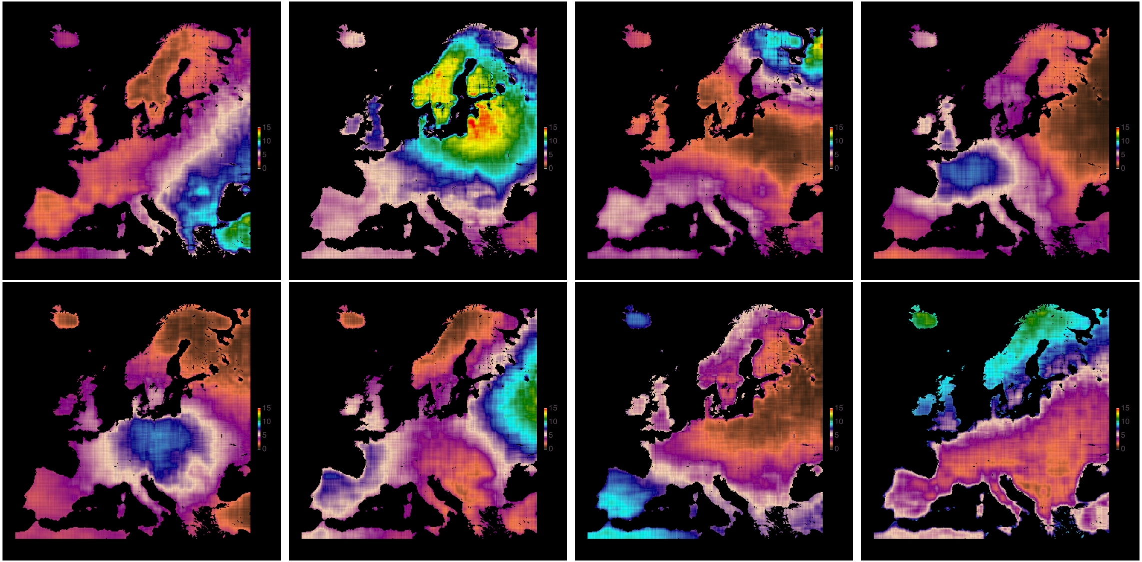
- Monthly fluctuations in the air temperature from 1989 to 2017 [$^\circ C$]
- Tensor: $(316 \times 316 \times 348)$ $(columns \times rows \times months)$
- **NTFk** applied to extract dominant hidden (latent) features



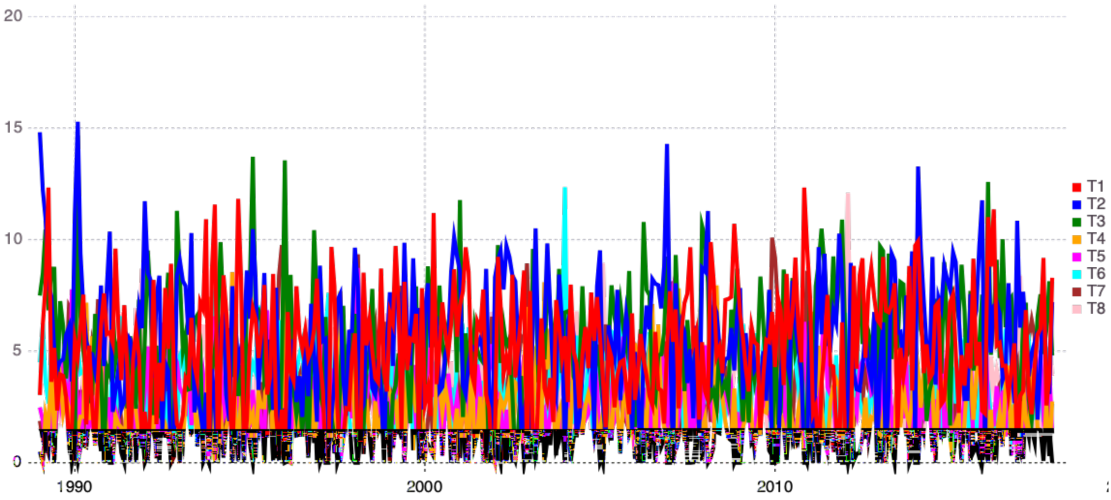
Climate model of Europe: air temperature reconstruction errors



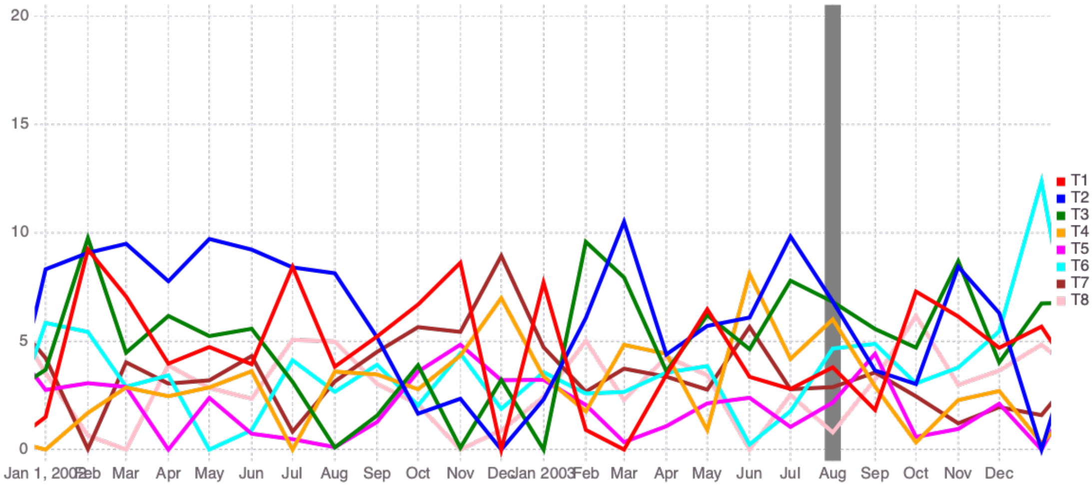
Climate model of Europe: air temperature features (8)



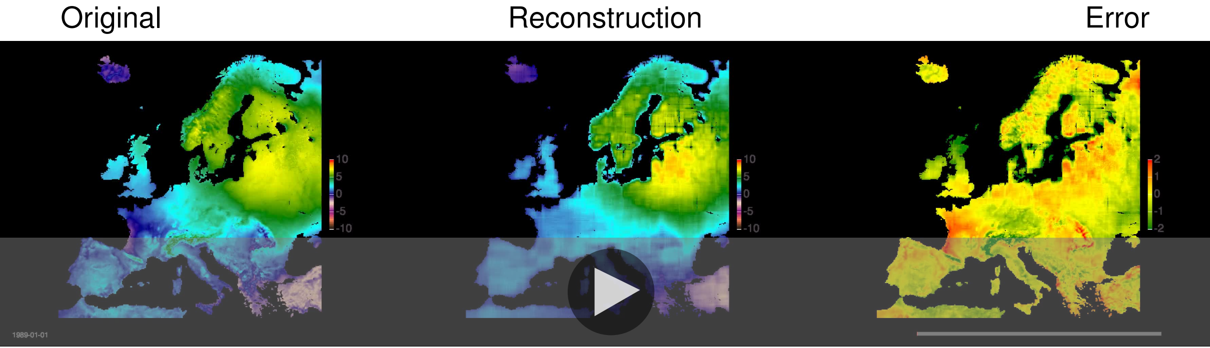
Climate model of Europe: air temperature features (8) 1989-2017



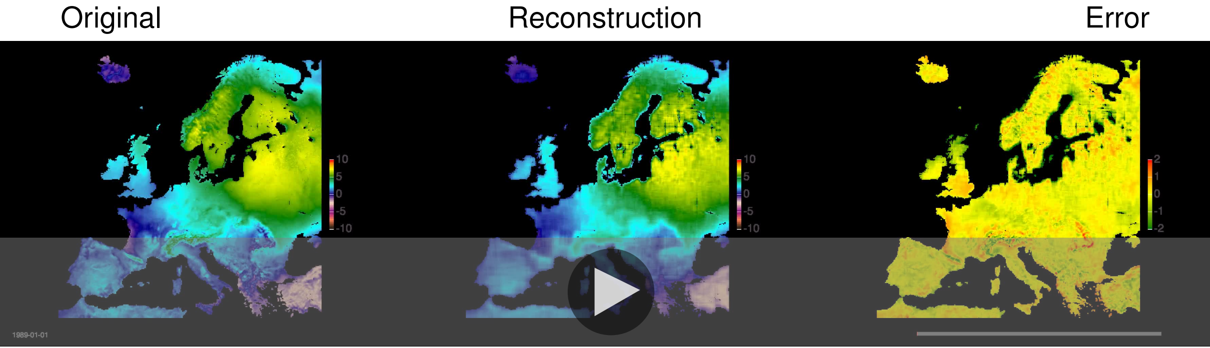
Climate model of Europe: air temperature features (8) 2002-2003



Climate model of Europe: air temperature reconstruction by 8 features

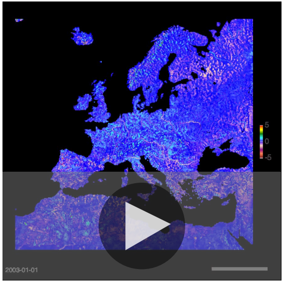


Climate model of Europe: air temperature reconstruction by 50 features

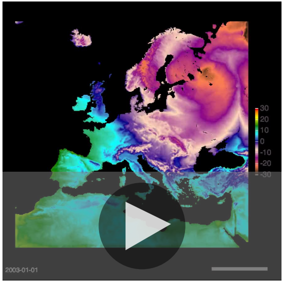


Climate model of Europe: analyze all model outputs (>40) simultaneously

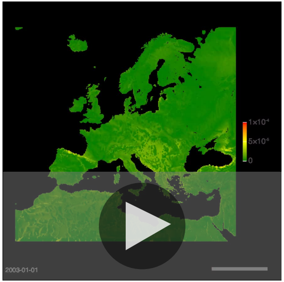
Water-table depth



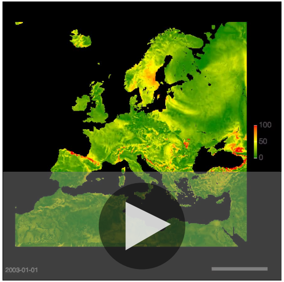
Temperature



Evaporation



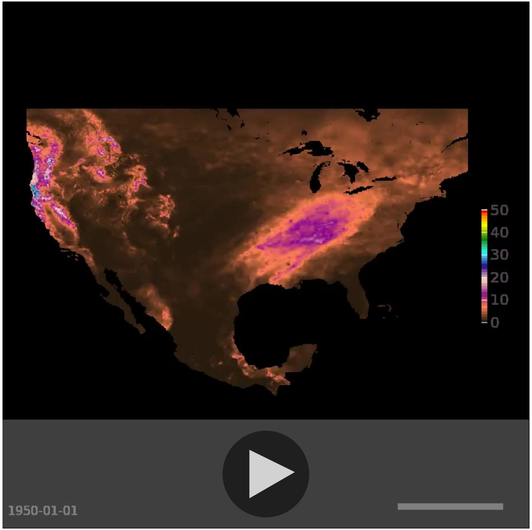
Sensible heat flux



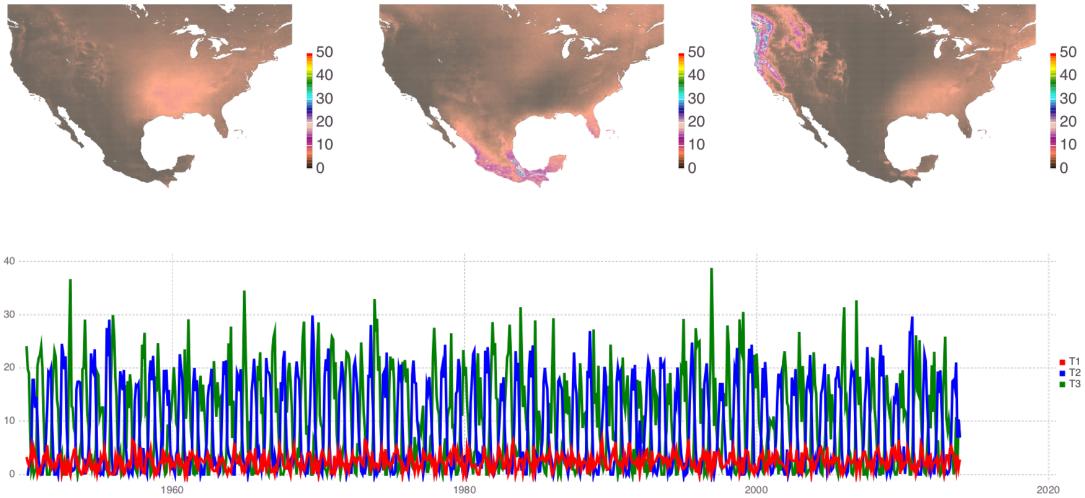
- Find interconnections and common dominant features in model outputs
- Evaluate impacts of different model setups (initial conditions, resolution, etc.)
- Find dominant processes impacting model predictions
  (e.g., occurrence of heat waves, climate impacts on groundwater resources, climate impacts of subsurface processes on atmospheric conditions)

US Climate data: precipitation

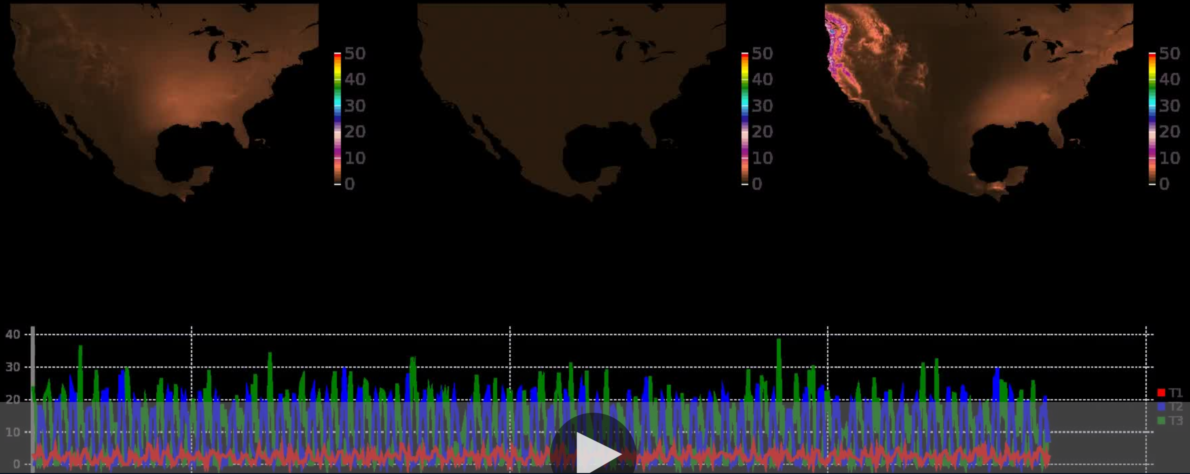
- NOAA data set of observed monthly averaged precipitation with spatial resolution of 6km (1/16Â°)
- fluctuations in monthly averaged precipitation [inches]
- $(928 \times 614 \times 768)$
  $(columns \times rows \times months)$



US precipitation: NTFk results

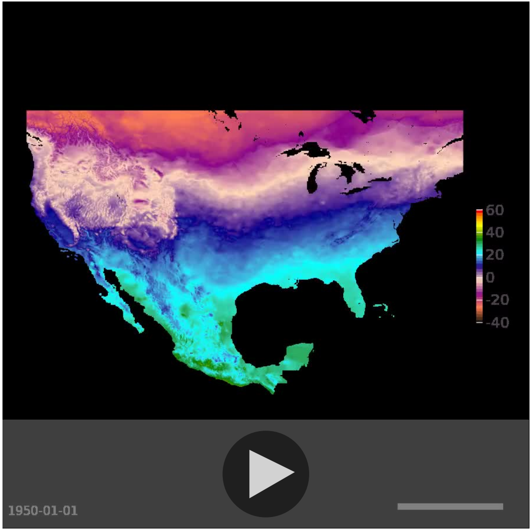


US precipitation: NTFk results

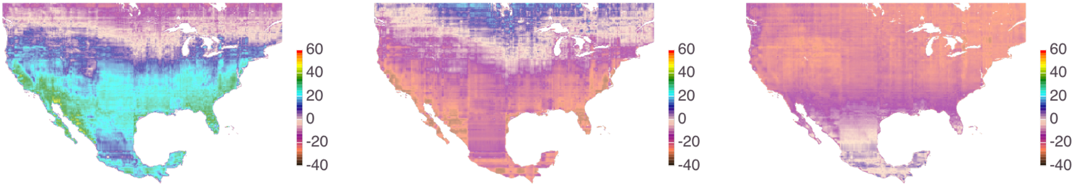


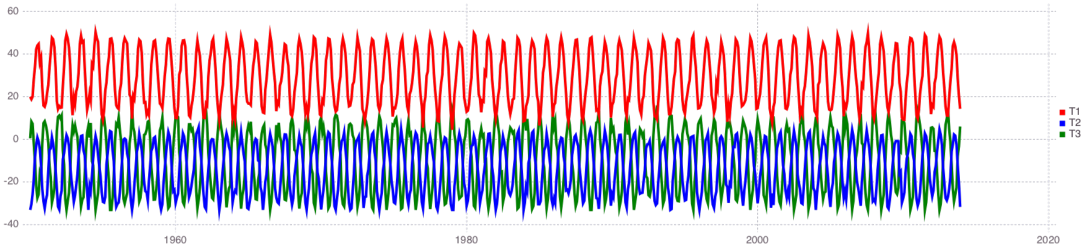
US Climate data: temperature

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

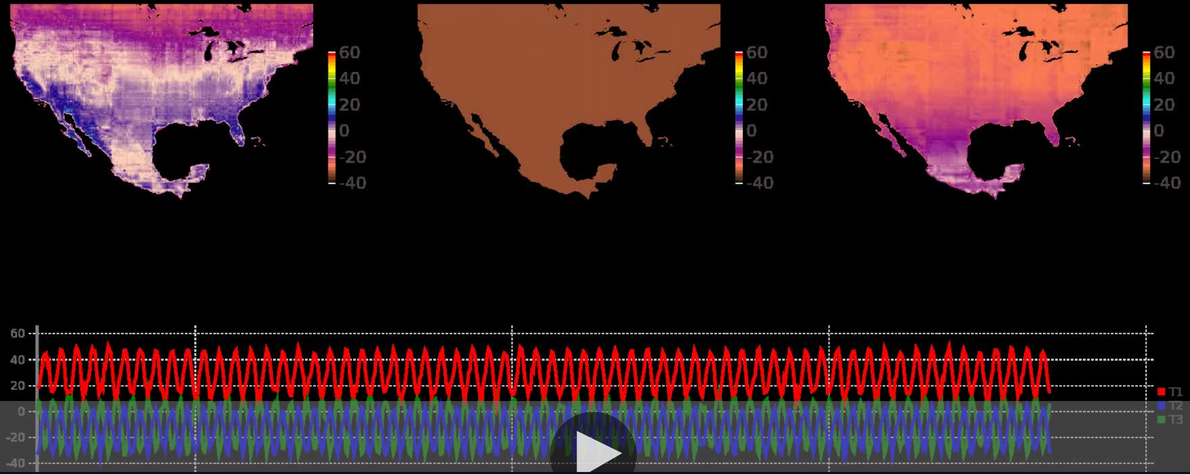


US maximum monthly temperatures: NTFk results

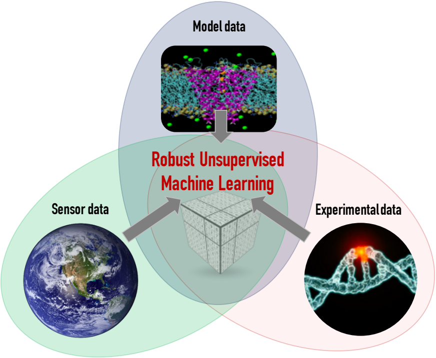




US maximum monthly temperatures: NTFk results



- 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
- Our goal is to further tests our algorithms on diverse datasets



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

- **NMFk**
- **NTFk**
- **MADS**: Model-Analyses & Decision Support
  `http://mads.gitlab.io` `http://madsjulia.github.io/Mads.jl`
- Examples:
  `http://madsjulia.github.io/Mads.jl/Examples/blind_source_separation`
- Slide decks / publications: `http://monty.gitlab.io`