Model-free Source Identification

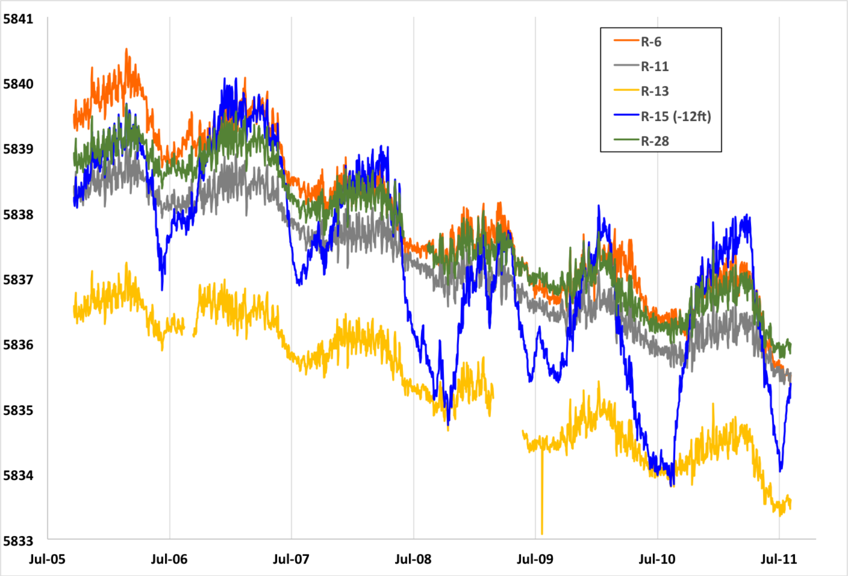
Velimir V. Vesselinov   Boian S. Alexandrov

Los Alamos National Laboratory

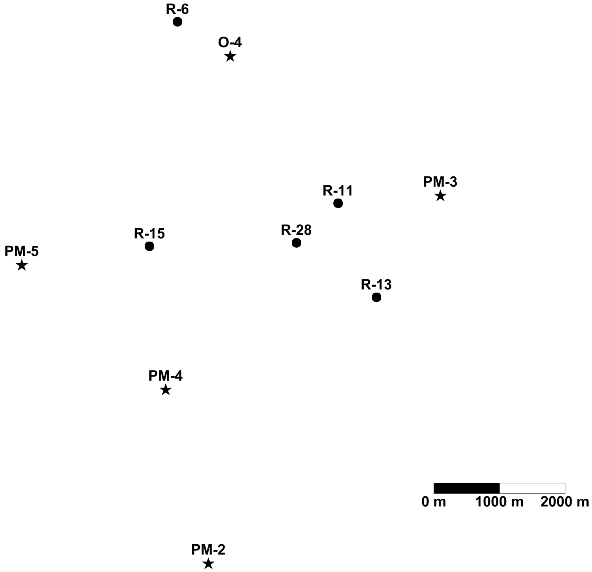
AGU Fall Meeting, San Francisco, December 15, 2014

Unclassified: LA-UR-14-29163

Water-level data



Well locations



Source Identification

- Identification of the physical sources (forcings) causing spatial and temporal variation of state variables
- Variations can be caused by various natural or anthropogenic sources
- The identification of forcings (the source identification) can be crucial for conceptualization and model development
- If the forcings are successfully “unmixed” from the observations, decoupled physics models may then be applied to analyze the propagation of each forcing independently

- Statistical methods ...
- Model inversion ...
- Blind Source Separation (**BSS**) ...
  - unsupervised
  - objective
  - adaptive
  - machine-learning algorithm
  - model-free inversion
  - Jutten and Herault, 1991, Zarzoso and Nandi, 1999

Blind Source Separation (BSS)

- Retrieve the unknown forcing signals (sources) $\mathbf{S}_{p\times r}$ that have produced observation records, $\mathbf{H}_{p\times m}$ with unknown noise (measurement errors) $\mathbf{E}_{p\times m}$:

$$\mathbf{H}_{p\times m} = \mathbf{S}_{p\times r}\mathbf{A}_{r\times m} + \mathbf{E}_{p\times m},$$

  - $\mathbf{A}_{r\times m}$ is unknown "**mixing**" matrix
  - m is the number of the recording sensors (observation wells)
  - r is the number of unknown signals ($m > r$)
  - p is the number of discretized moments in time at which the signals are recorded at the sensors
- The problem is ill-posed and the solutions are non-unique
- **BSS** performs optimization with constraints, such as:
  - maximum variability
  - statistical independence
  - component non-negativity
  - smoothness
  - simplicity, *etc.*

Blind source separation methods

- **ICA**: Independent Component Analysis
  - Maximizing the statistical independence of the retrieved forcings signals in $\mathbf{S}$ (i.e. the matrix columns are expected to be independent) by maximizing some high-order statistics for each source signal, such as the kurtosis or negentropy (negative entropy).
  - The main idea behind **ICA** is that, while the probability distribution of a linear mixture of sources in $\mathbf{H}$ is expected to be close to a Gaussian (the Central Limit Theorem), the probability distribution of the original independent sources is expected to be non-Gaussian.
- **NMF**: Non-negative Matrix Factorization
  - Non-negativity constraint on the original sources in $\mathbf{S}$ and their mixing components in $\mathbf{A}$
  - As a result, the observed data are representing only additive signals that cannot cancel mutually.
  - Additivity and non-negativity requirements lead to a sparseness in both the signal $\mathbf{S}$ and mixing $\mathbf{A}$ matrices

- Generate random initial guesses for $\widetilde{\mathbf{S}}$ and $\widetilde{\mathbf{A}}$
- Update $\widetilde{\mathbf{S}}$ and $\widetilde{\mathbf{A}}$ (η is a small arbitrary positive constant):

$$a^*_{j,i} \leftarrow a_{j,i} \frac{\left[\tilde{\mathbf{S}}^{\mathbf{T}}\mathbf{H}\right]_{j,i}}{\left[\tilde{\mathbf{S}}^{\mathbf{T}}\tilde{\mathbf{S}}\tilde{\mathbf{A}}\right]_{j,i} + \eta}, \qquad s^*_{q,j} \leftarrow s_{q,j} \frac{\left[\mathbf{H}\tilde{\mathbf{A}}^{*\mathbf{T}}\right]_{q,j}}{\left[\tilde{\mathbf{S}}\tilde{\mathbf{A}}^{*}\tilde{\mathbf{A}}^{*\mathbf{T}}\right]_{q,j} + \eta}$$

- Loop till some convergence criteria are satisfied (*e.g.*, based on objective function and/or number of iterations).

Improved NMF (NMF + k-means = NMFk)

- We propose an improved **NMF** coupled with **k-means** analysis (we call it **NMFk**)
- Perform a series of **NMF** analyses for a series of different predetermined initial guesses for the number of sources ($r = 1, 2, ..., m$).
- For each r value, perform **NMF** runs with a series (n) of different random initial guesses $\widetilde{\mathbf{S}}$ and $\widetilde{\mathbf{A}}$ (the total number of solutions is $\frac{m*(m+1)}{2} \times n$).
- All the obtained solutions for given r ($r \times n$) are **k-means** clustered based on the cosine similarity between the estimated sources using k-means analysis where $k = r$ (k-means clustering is performed m times)
- The optimal number of sources is identified based on the Objective function $\mathcal{O}$ and Silhouette width $\mathcal{C}$

- Objective function $\mathcal{O}$ based on Frobenius norm:

$$\mathcal{O} = \frac{1}{2}\left(\left\|\mathbf{H} - \widetilde{\mathbf{S}} * \widetilde{\mathbf{A}}\right\|_F\right)^2 = \sum_{i=1}^{m}\sum_{q=1}^{p}\left(h_{q,i} - \sum_{j=1}^{r}\widetilde{s}_{q,j}\widetilde{a}_{j,i}\right)^2 .$$

NMFk criteria: Silhouette width

- Cosine distance (cosine similarity) ρ representing the similarity between any two forcing signals j_1 and j_2 ($\widetilde{s}_{q,j_1}$ and $\widetilde{s}_{q,j_2}$):

$$\rho(j_1, j_2) = 1 - \frac{\sum_{q=1}^{p} \widetilde{s}_{q,j_1} \widetilde{s}_{q,j_2}}{\sqrt{\sum_{q=1}^{p} (\widetilde{s}_{q,j_1})^2} \sqrt{\sum_{q=1}^{p} (\widetilde{s}_{q,j_2})^2}}.$$

- Silhouette value (c_d) for each solution based on the cosine similarity

$$c_d = \frac{R_{in,d} - R_{out,d}}{max\left[R_{in,d}, R_{out,d}\right]}, \forall\, d = 1, ..., n \times r$$

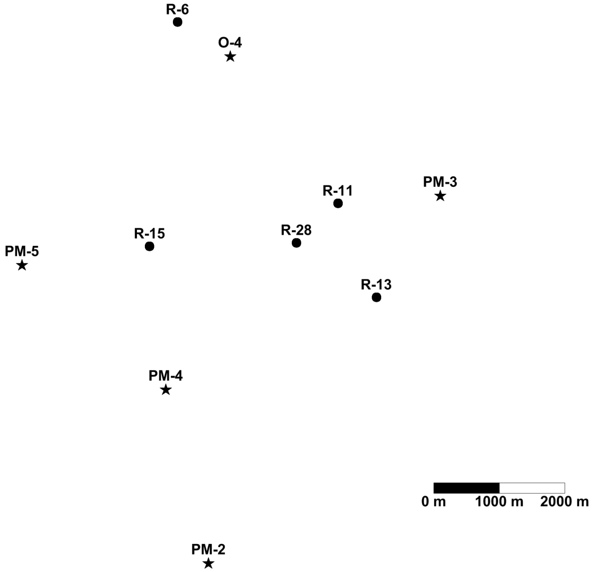
- $R_{in,d} = \mathrm{E}\langle \rho(j_{in}, j_d) \rangle$ similarity with solutions within the cluster
- $R_{out,d} = \mathrm{E}\langle \rho(j_{out}, j_d) \rangle$ similarity with solutions outside the cluster
- If $c_d \to 1$, the element is appropriately clustered; if $c_d \approx 0$, the element is between two clusters (if $c_d \to -1$, the clustering failed).

- Silhouette width $\mathcal{C}$ of k-means results for a given r is:

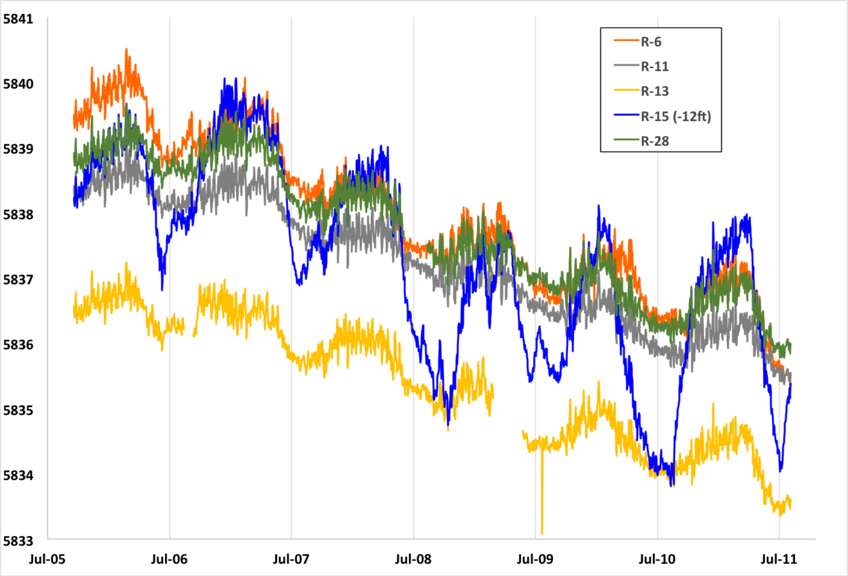
$$\mathcal{C} = \mathrm{E}\langle c_d \rangle,\ d = 1, ..., n \times r$$

- The optimal number of sources is identified based on the Objective function $\mathcal{O}$ and Silhouette width $\mathcal{C}$ for a given r.

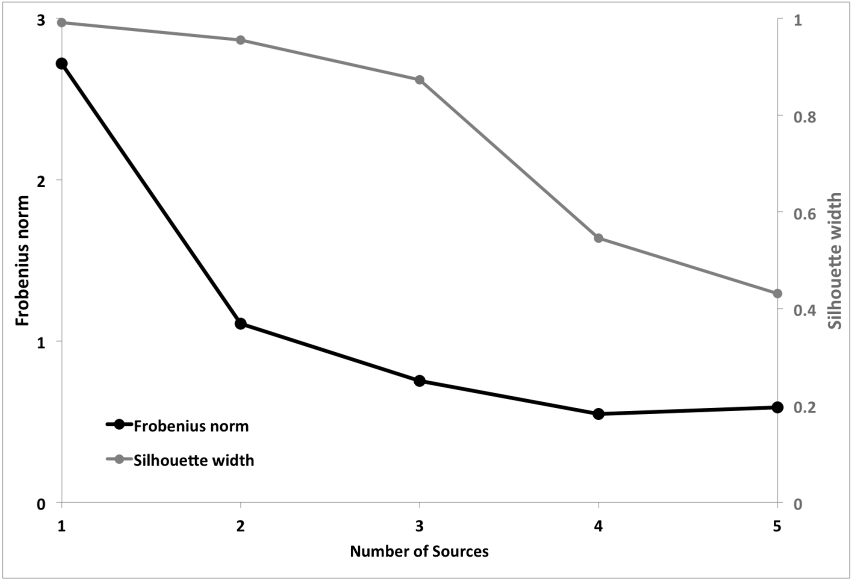
Well locations



Water-level data

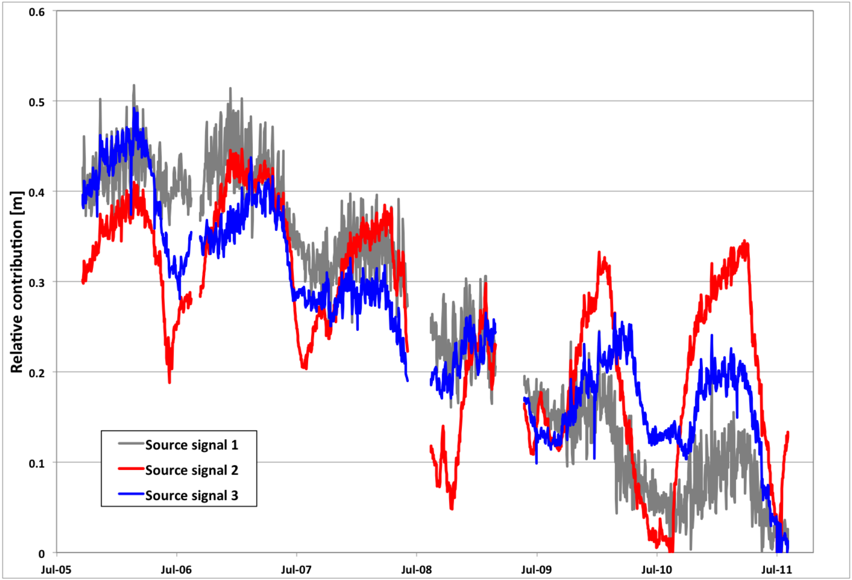


Stability



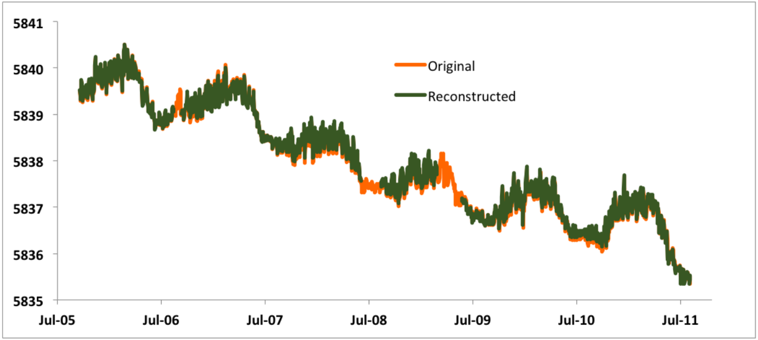
Frobenius norm
Silhouette width
Number of Sources
Frobenius norm
Silhouette width
0
1
2
3
0
0.2
0.4
0.6
0.8
1
1
2
3
4
5

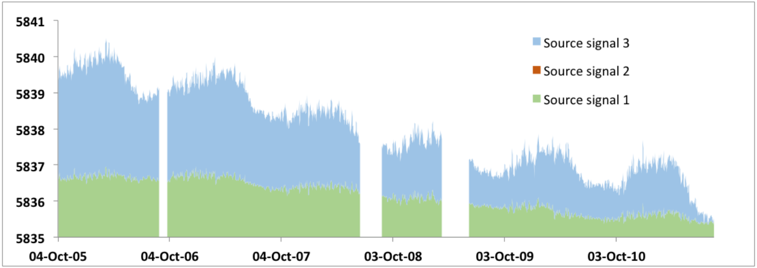
Signals



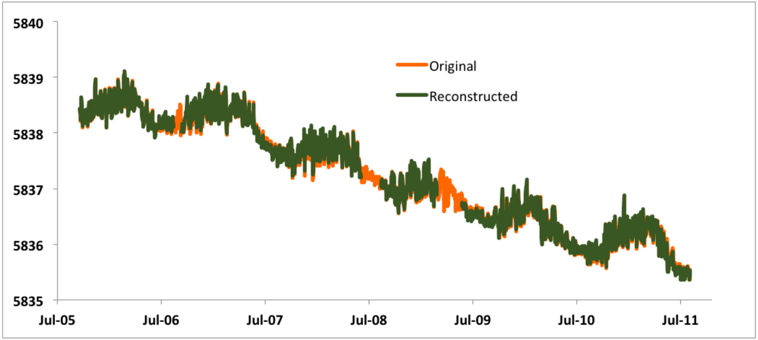
Relative contribution [m]
0.6
0.5
0.4
0.3
0.2
0.1
0
Jul-05
Jul-06
Jul-07
Jul-08
Jul-09
Jul-10
Jul-11
Source signal 1
Source signal 2
Source signal 3

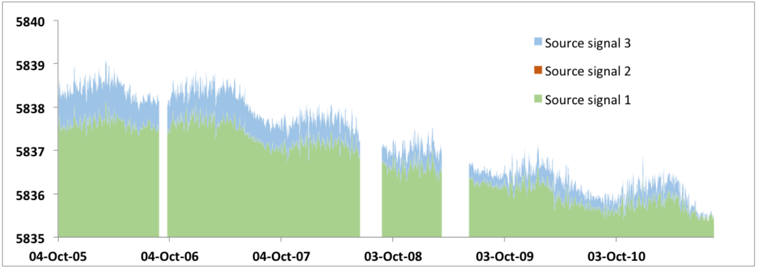
Reconstruction of R-6



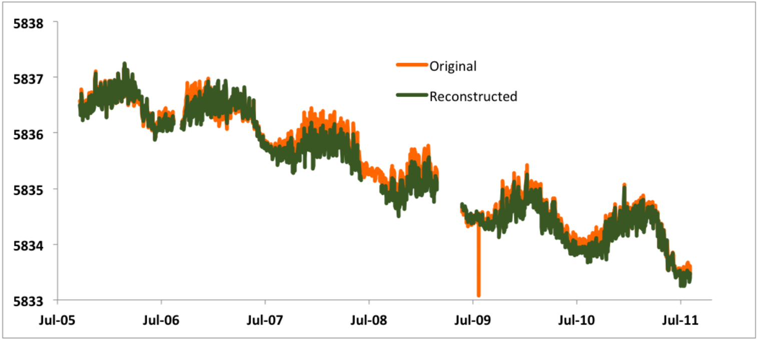


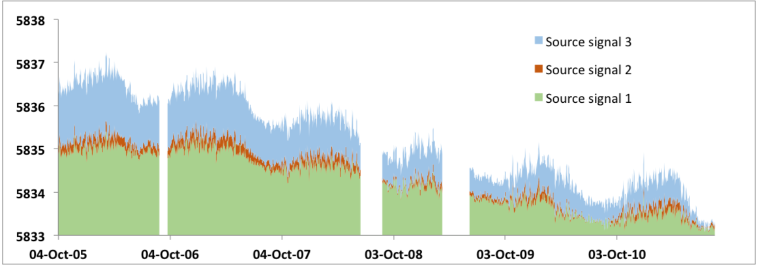
Reconstruction of R-11



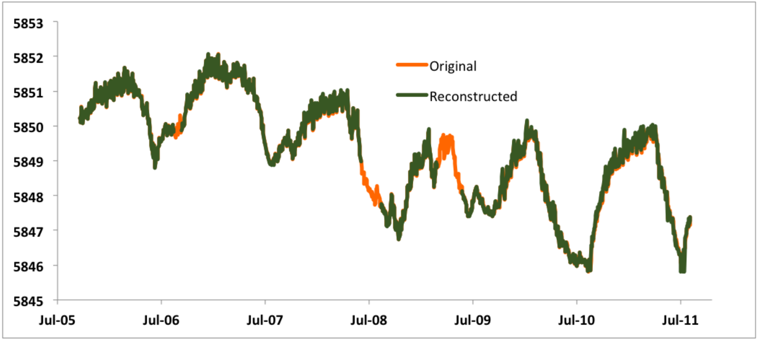


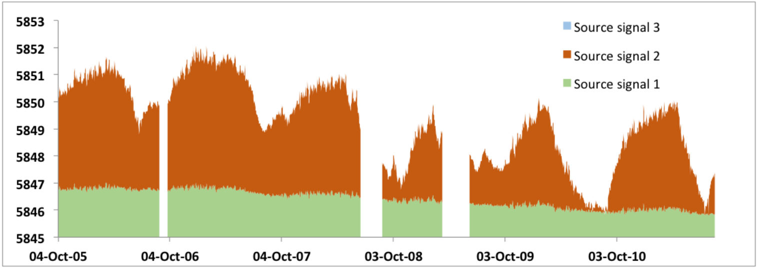
Reconstruction of R-13



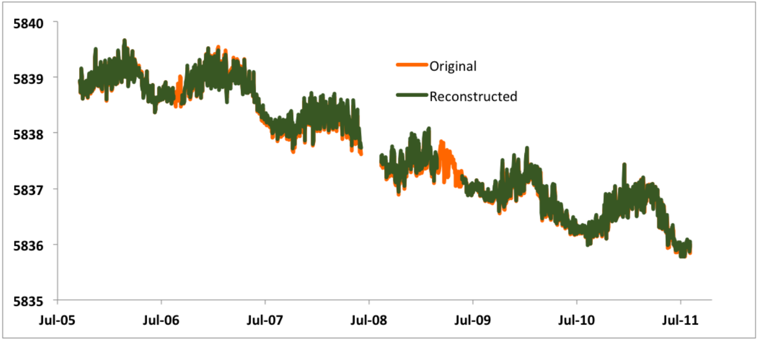


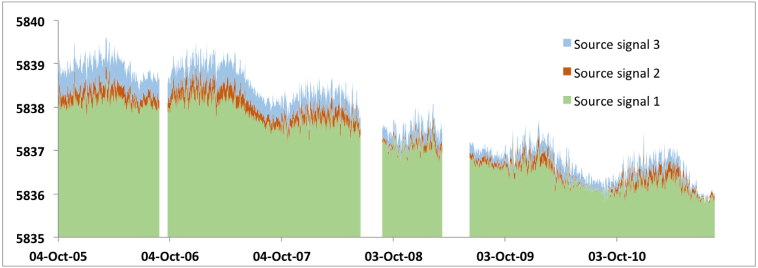
Reconstruction of R-15



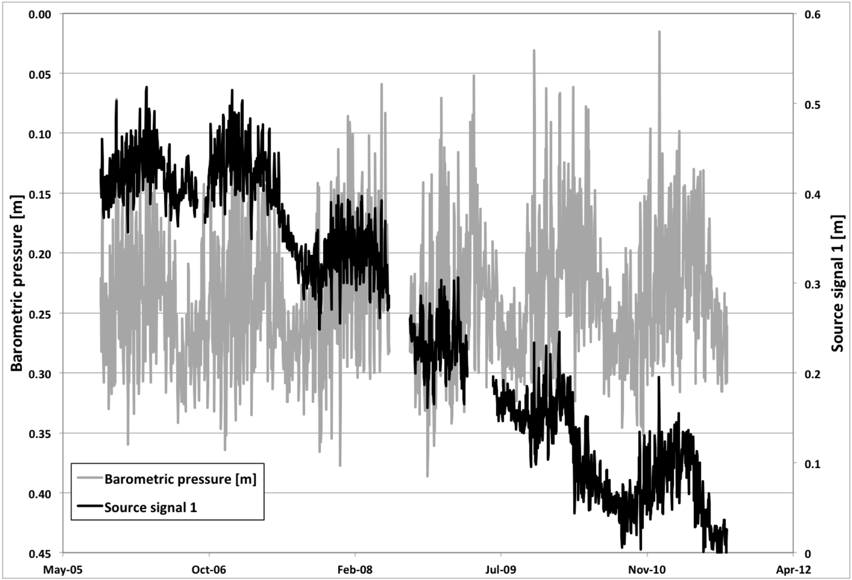


Reconstruction of R-28

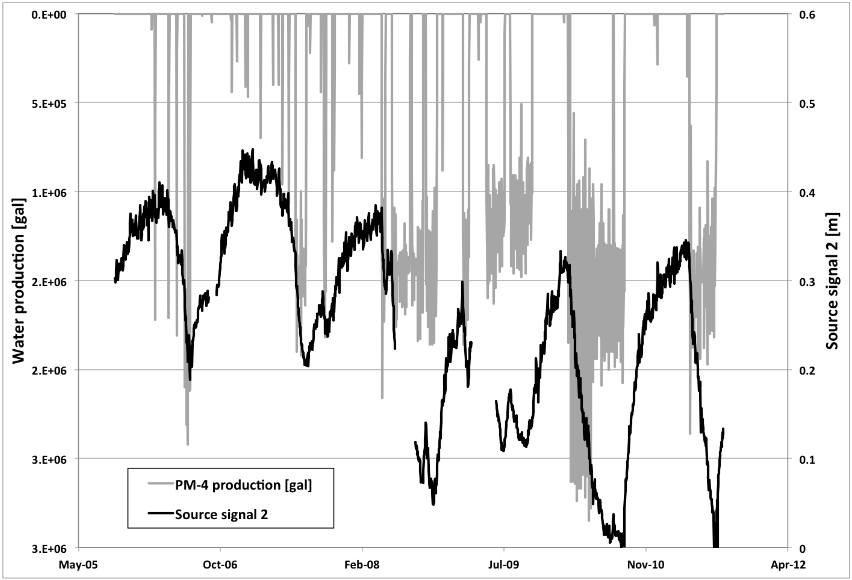




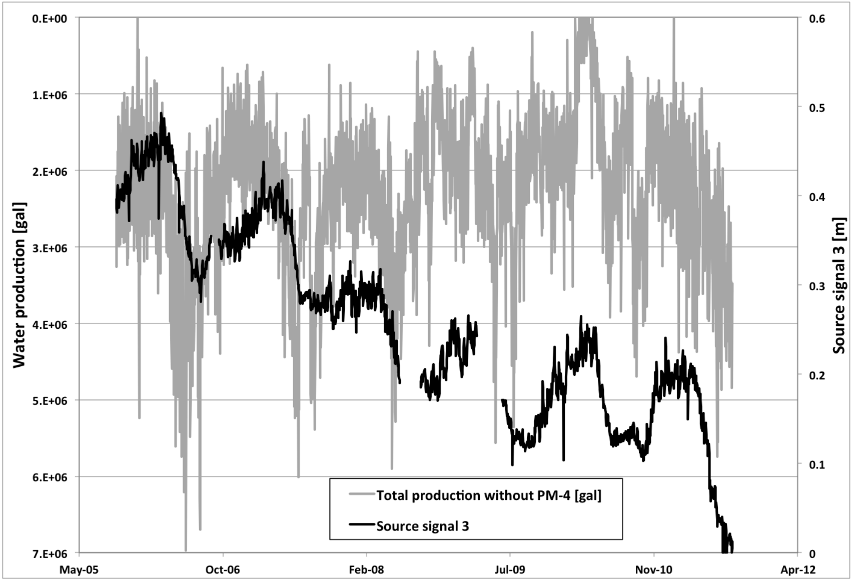
Signal #1 - Barometric pressure



Signal #2 - PM-4 water-supply pumping



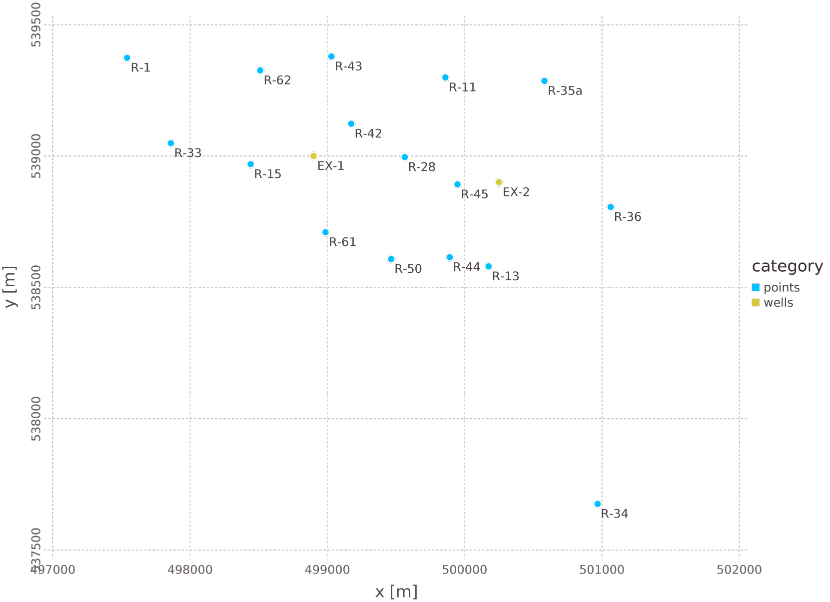
Signal #3 - All the other pumping wells (without PM-4)



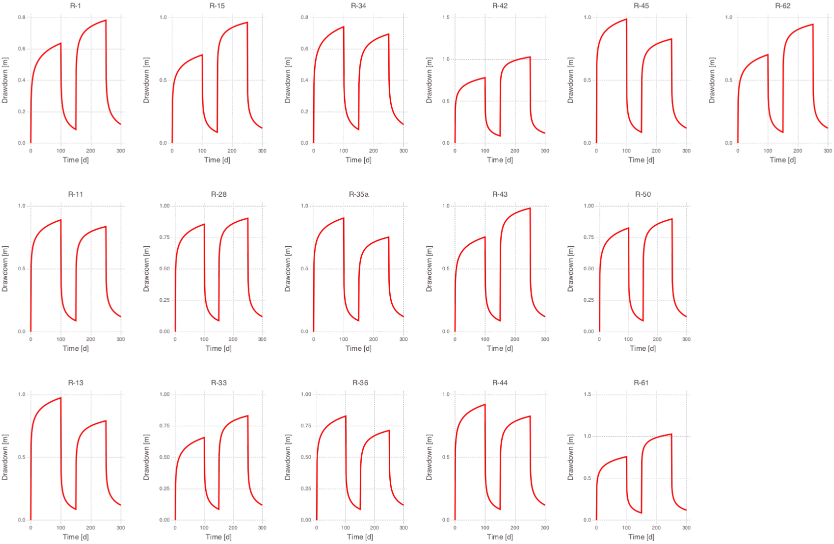
Source locations

NMFk can be extended to identify the location of the sources (pumping wells) as well ...

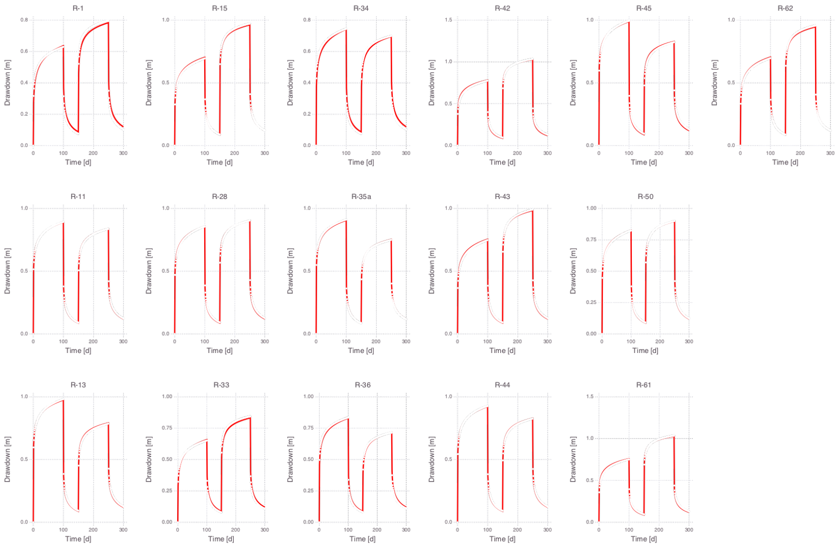
Synthetic test problem (based on the LANL site)



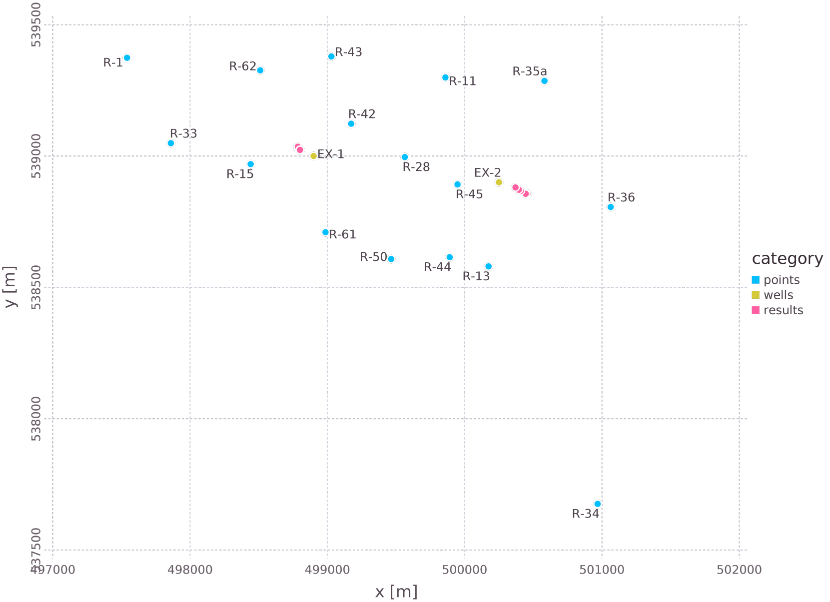
Synthetic test problem: Water-level input



Synthetic test problem: Water-level matches



Synthetic test problem: Estimated source locations



Conclusions

- **NMF**k combines the strengths of standard NMF (Non-negative Matrix Factorization) and k-means clustering for characterization of unknown forcing signals in observed state variables
- **NMF**k can be applied to unmix transients in groundwater levels
- **NMF**k can be applied to find the source locations
- Additional work is needed to add known forcing signals (e.g. linear decline) in the analyses
- Future work can address uncertainty associated with estimated forcing signals
- Future work will also provide coupling between **NMF**k with physics-based inverse models

Thank you!

- Alexandrov & Vesselinov, Blind source separation for groundwater level analysis based on non-negative matrix factorization, Water Resources Research, doi: 10.1002/2013WR015037, 2014.
- **NMFk** is coded in Julia