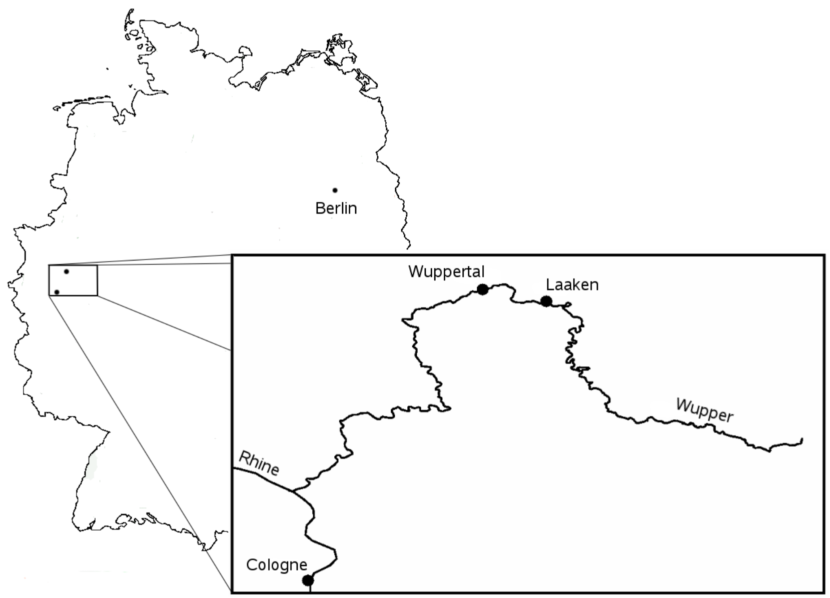
Specification of Landmarks and Forecasting Water Temperature – Water Management in the River Wupper

Göran Kauermann
Center for Statistics
University Bielefeld

Thomas Mestekemper
University Bielefeld

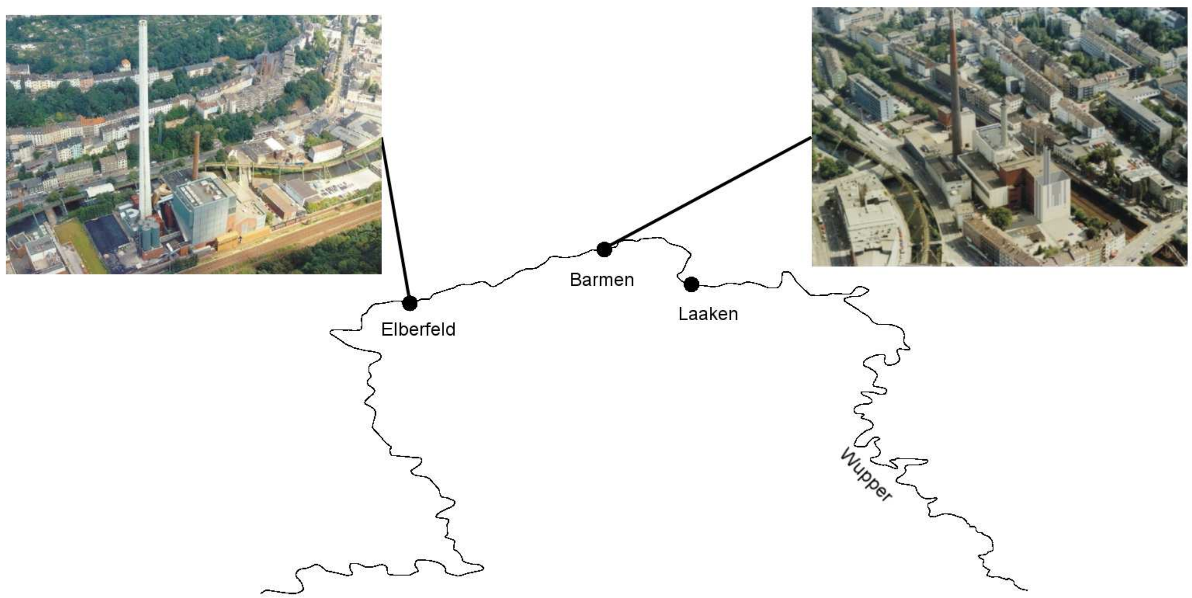
14. August 2008

The River Wupper



The River Wupper

The River Wupper and its Power Plants



The EU Water Framework Directive

Commits European Union member states to achieve good qualitative and quantitative status of water bodies until 2015.

Good surface water status means both, good ecological and chemical status. The first refers to the quality and functioning of the aquatic ecosystem.

For the Wupper this implies:

"Too warm upstream water" $\Longrightarrow$ Reduce electric power production or even shut down power plant

Definition of "Warm Water" depends on the fish life and reproduction cycle and the given threshold may vary over the year.

Outline of Talk

- Forecasting (upstream) Water Temperature
- Specification of Landmarks (Threshold, dependent fish spawning cycle)
- Discussion

Literature on Water Temperature Forecasting

Hydrological Literature:

- Seasonal and daily variations of water temperature are significantly important for aquatic resources. (Caissie et al., 2005, *Hydrological Processes*)
- Two model classes: physical (thermo-dynamic) and stochastic (statistical) models. (Webb et al., 2008, *Hydrological Processes*)

Statistical Literature:

- Functional component models or dynamic factor models. (Cornillon et al., 2008, *CSDA*; Stock & Watson, 2006, *Handbook of Economic Forecasting*)
- Functional Time Series. (Ferraty & Vieu, 2006, *Nonparametric FDA*, Springer-Verlag)

Smooth Cyclic Estimation

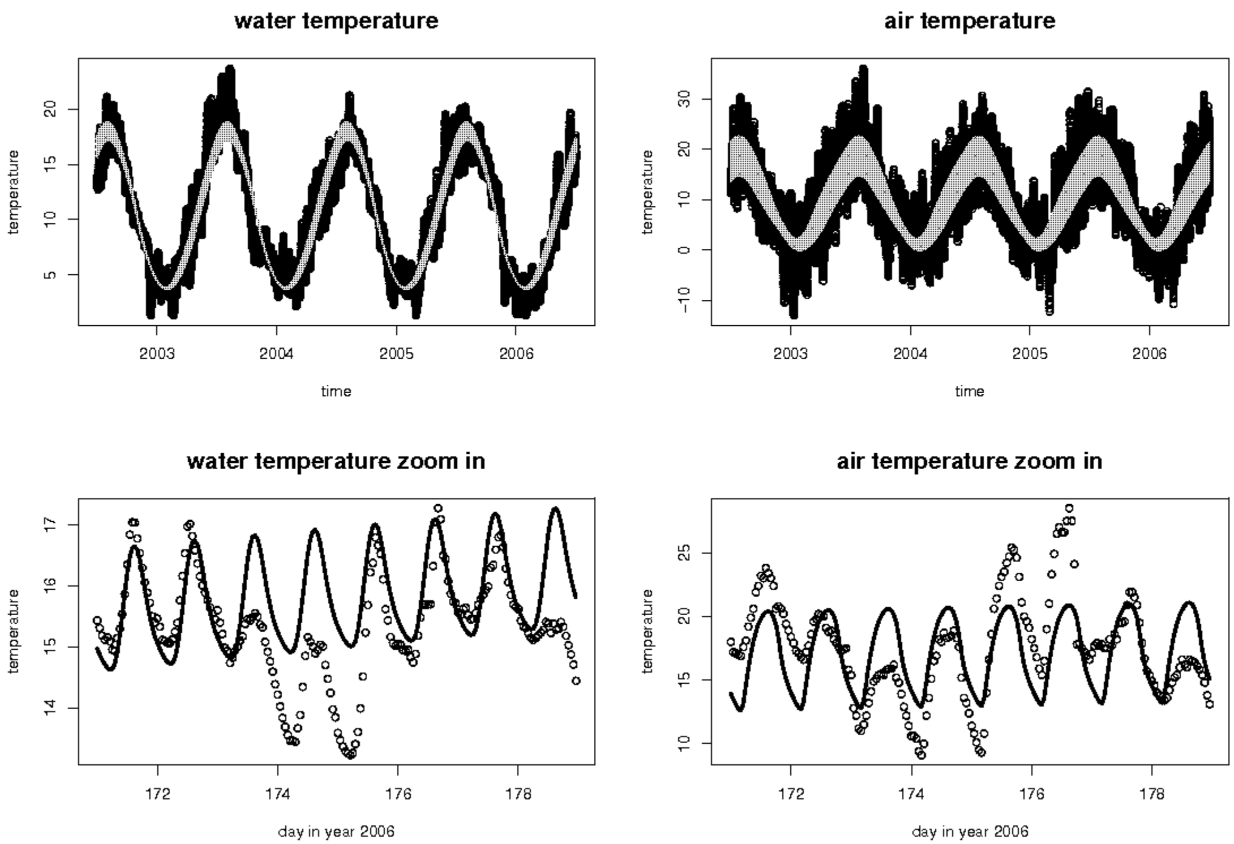
Let index $t = (y, d)$ denote time with year y, day in year d and $\mathbf{w}_t$ and $\mathbf{a}_t$ be a 24-dimensional vectors of the hourly water and air temperature, respectively, which decompose to

$$\mathbf{w}_t = \underset{\substack{\uparrow \\ \text{yearly trend}}}{\boldsymbol{\mu}_w}(d) + \bar{\mathbf{w}}_t, \qquad \mathbf{a}_t = \underset{\substack{\uparrow \\ \text{yearly trend}}}{\boldsymbol{\mu}_a}(d) + \bar{\mathbf{a}}_{yd}.$$

Functions $\boldsymbol{\mu}_w(d)$ and $\boldsymbol{\mu}_a(d)$ are fitted with “wrapped” B-splines, i. e.

$$\lim_{d \to 365+} \widehat{\boldsymbol{\mu}}_w(d) = \lim_{d \to 1-} \widehat{\boldsymbol{\mu}}_w(d).$$

Average Temperatures μ_w and μ_a



Functional Principal Components Decomposition

$\bar{\mathbf{w}}_t$ shall be decomposed to a dynamic factor model, that is, we reduce dimensions by extracting k suitable factors (done by PCA):

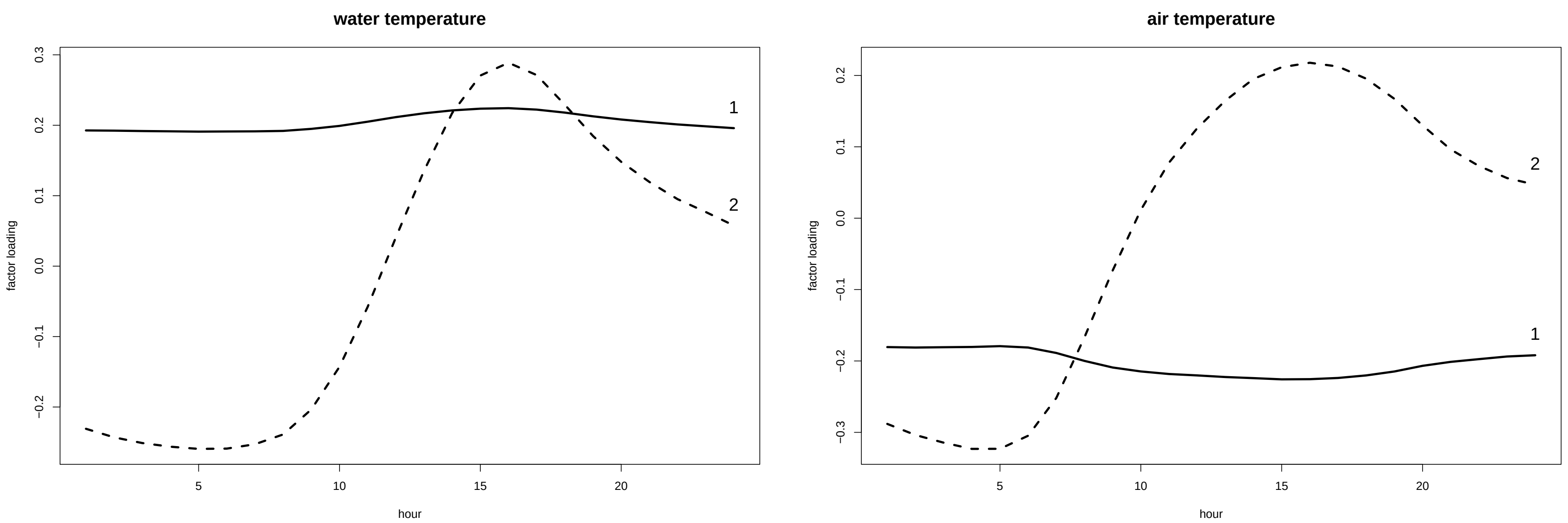
$$\bar{\mathbf{w}}_t = \mathbf{f}_t \Lambda_w^T + \boldsymbol{\epsilon}_{w,t}$$

where Λ_w is a $24 \times k$ dimensional loading matrix, $\mathbf{f}_t$ a k dimensional factor and $\boldsymbol{\epsilon}_{w,t}$ a white noise residual.

Accordingly for the air temperature we extract h suitable factors:

$$\bar{\mathbf{a}}_t = \mathbf{g}_t \Lambda_a^T + \boldsymbol{\epsilon}_{a,t}$$

Fitted Principal Components



The Dynamic Factor Model

Using the backshift operator $\Delta_{a,b}\mathbf{f}_t = (\mathbf{f}_{t-a}, \ldots, \mathbf{f}_{t-b})$. We assume an autoregressive model for the factor $\mathbf{f}_t$:

$$\mathbf{f}_t = (\Delta_{1,p}\mathbf{f}_t)\boldsymbol{\beta}_f + (\Delta_{0,q}\mathbf{g}_t)\boldsymbol{\beta}_g + \boldsymbol{\epsilon}_{f,t}.$$

This implies that $\mathbf{f}_t$ depends on:

- water temperature factors of the p previous days
- air temperature factors of the q previous days
- the current day air temperature factors.

Note: In a forecasting setting the last point is only available as meteorological forecast.

Estimation of the factors $\mathbf{f}_t$ and $\mathbf{g}_t$

We want to compare three different approaches to estimate the factors.

1) We start with a quite simple Least Squares estimation method where the facor loadins are taken as

$$\widehat{\mathbf{f}}_t = \bar{\mathbf{w}}_t \Lambda_w \quad \text{and} \quad \widehat{\mathbf{g}}_t = \bar{\mathbf{a}}_t \Lambda_a$$

Pro: The remaining parameters $\boldsymbol{\beta}_f$ and $\boldsymbol{\beta}_g$ can easily be found using least squares regression.

Con: The resulting estimates are not Maximum Likelihood-based.

We need to incorporate our stochastic models in the estimation method.

Estimation of the factors f_t and g_t (continued)

2) In a Maximum Likelihood approach we assume that the residuals in the former mentioned models follow normal distributions:

$$\boldsymbol{\epsilon}_{w,t} \sim \mathsf{N}\big(\mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}_w^2)\big) \quad \text{and} \quad \boldsymbol{\epsilon}_{f,t} \sim \mathsf{N}\big(\mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}_f^2)\big).$$

3) We incorporate a stochastic autoregressive model for the air temperature, as well, in a Full Maximum Likelihood estimation method:

$$\mathbf{g}_t = (\Delta_{1,\tilde{q}}\mathbf{g}_t)\tilde{\boldsymbol{\beta}}_g + \boldsymbol{\epsilon}_{g,t}$$

asuming $\quad \boldsymbol{\epsilon}_{a,t} \sim \mathsf{N}\big(\mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}_a^2)\big) \quad$ and $\quad \boldsymbol{\epsilon}_{g,t} \sim \mathsf{N}\big(\mathbf{0}, \mathrm{diag}(\boldsymbol{\sigma}_g^2)\big).$

The unknown parameters $\theta = (\boldsymbol{\beta}_f, \boldsymbol{\beta}_g, \boldsymbol{\sigma}_f^2, \boldsymbol{\sigma}_w^2)$ and $\tilde{\theta} = (\theta, \tilde{\boldsymbol{\beta}}_g, \tilde{\boldsymbol{\sigma}}_g^2, \tilde{\boldsymbol{\sigma}}_a^2)$ are now estimated using an EM-algorithm.

Model Selection (in progress)

In order to select the best performig model we divide our dataset in a training and a forecasting sample. To measure the model quality one could, for example, make use of the Mean Squared Prediction Error defined by:

$$\text{MSPE} = \frac{1}{n}\sum_{t=i}^{n}(\mathbf{w}_t - \widehat{\mathbf{w}}_t)(\mathbf{w}_t - \widehat{\mathbf{w}}_t)^T.$$

We have to select:

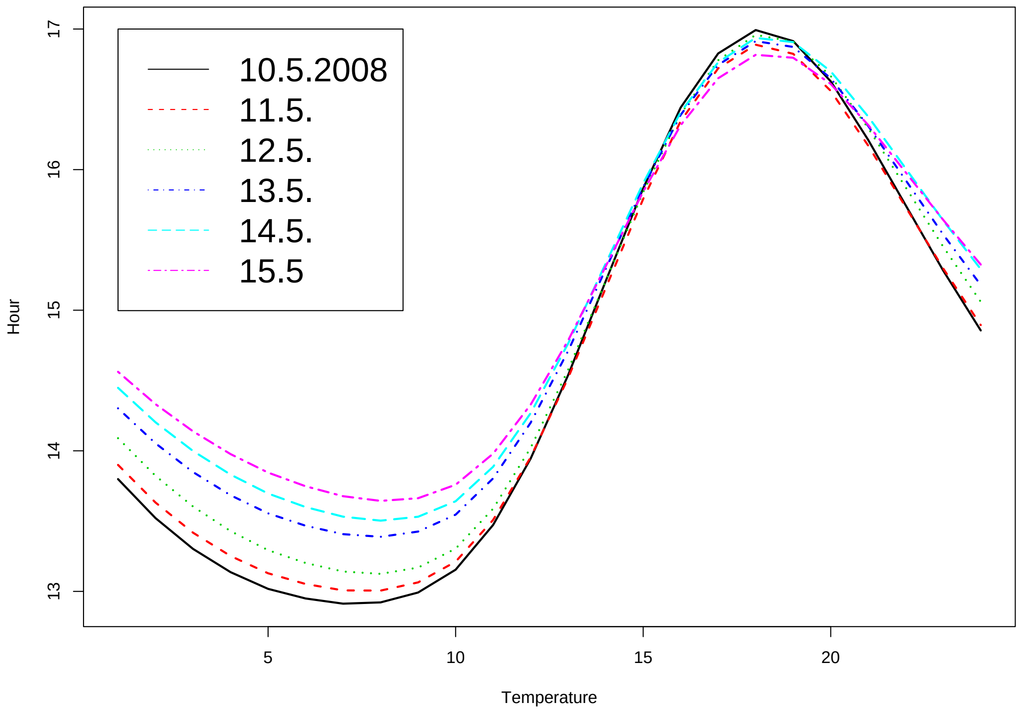
- k and h; the optimal number of factors for water and air temperature, respectively,

- p and q; the optimal number of time lags for water and air temperature, respectively, (we treat $\tilde{q} = 2$ as fixed)

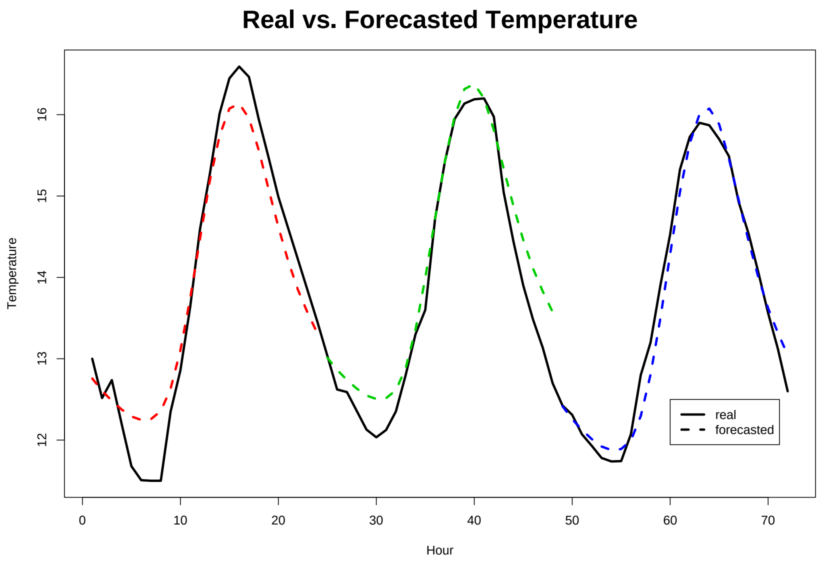
- the optimal estimation method.

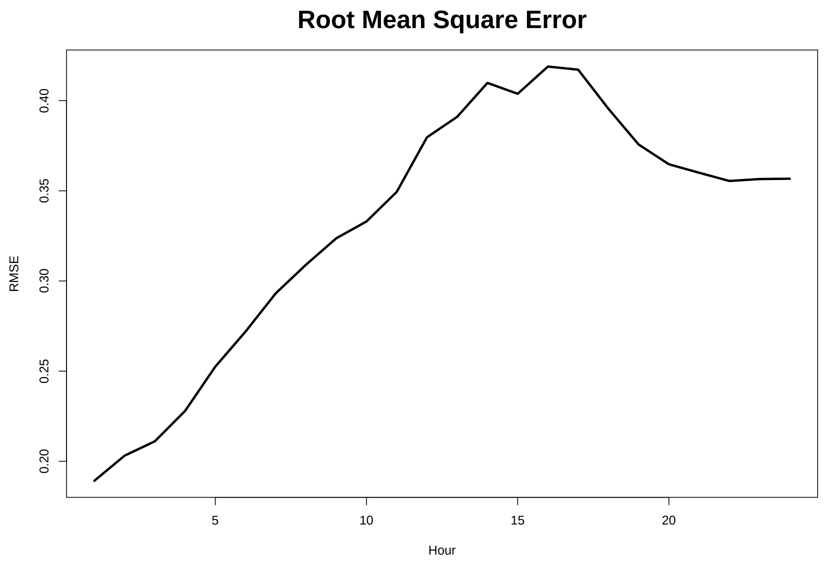
Demonstration

Warm spring days over Whitsun 2008



Demonstration





Multiple Day Forecast

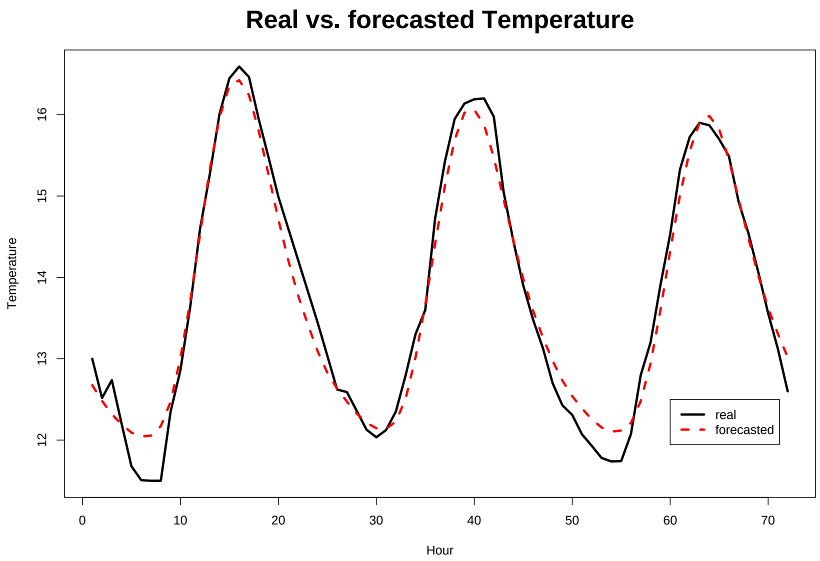
Multiple day forecasts show discontinuities.

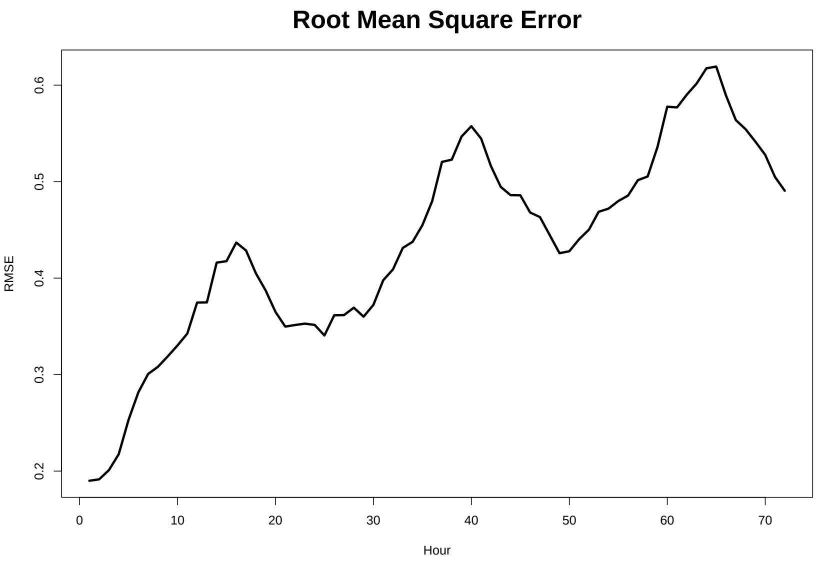
Solution: To achieve a continuous m day forecast we divide our time axis into time intervals of length m, i. e.

$$\mathbf{w}_t^m = \mathbf{w}_{\tilde{t}} = \left(\mathbf{w}_{yd1}, \ldots, \mathbf{w}_{yd24}, \mathbf{w}_{y(d+1)1}, \ldots \mathbf{w}_{y(d+m)24}\right)$$

The above models are re-fitted in analogy to the 24h case.

Demonstration





Comparison to other modelling approaches

We compared our Least Squares model to three approaches to model the daily maximum temperature presented in Cassie et al. (1998, *Can. J. Civ. Eng.*)

1. $\bar{w}_t^{\text{max}} = (\Delta_{0,2}\bar{a}_t^{\text{max}})\boldsymbol{\beta}^1 + \epsilon_t^1$ resulted in an RMSE of 1.295°C.

2. $\bar{w}_t^{\text{max}} = (\Delta_{1,2}\bar{w}_t^{\text{max}})\boldsymbol{\beta}^2 + K\,\bar{a}_t^{\text{max}}$ resulted in an RMSE of 2.439°C.

3. $\bar{w}_t^{\text{max}} = \frac{\zeta_0}{1-\delta_1 B}\bar{a}_t^{\text{max}} + \frac{1}{1-\phi_1 B}n_t$ resulted in an RMSE of 1.018°C.

For $p = 2, q = 1, k = h = 3$ and $\tilde{q} = 2$ our Least Squares Model yielded an RMSE of 0.42°C.

Finding Seasonal Pattern

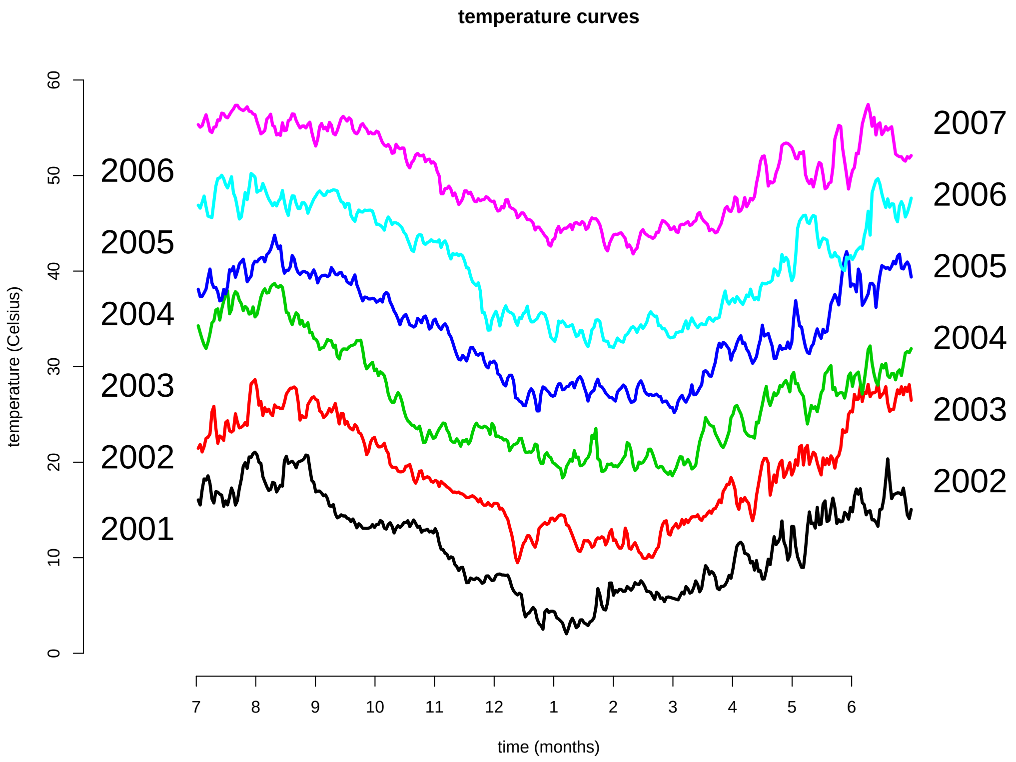
Besides forecasting is the specification of seasonal pattern an important issue, since:

- Water temperature has to stay below ecologically justified thresholds to preserve the fish populations.
- Threshold values depend on season, or more precisely on reproduction cycle of fish.
- Seasons can vary like an early spring or late summer.
- What is the “reference year”?

Literature in 'Warping' and 'Landmark Specification'

- Landmark specification in growth curves. (Kneip & Gasser, 1992, *Annals of Statistics*; Gasser & Kneip, 1995, *JASA*)
- Automatic Warping (or self-modelling). (Ramsay & Li, 1998, *JRSS B*; Gervini & Gasser, 2004, *JRSS B*)
- We need an “online” warping, as data arrives over time.

Structure of Water temperature



Different modelling for landmark registration

Let $t = (y, d, h)$ where h is the hour in day d.

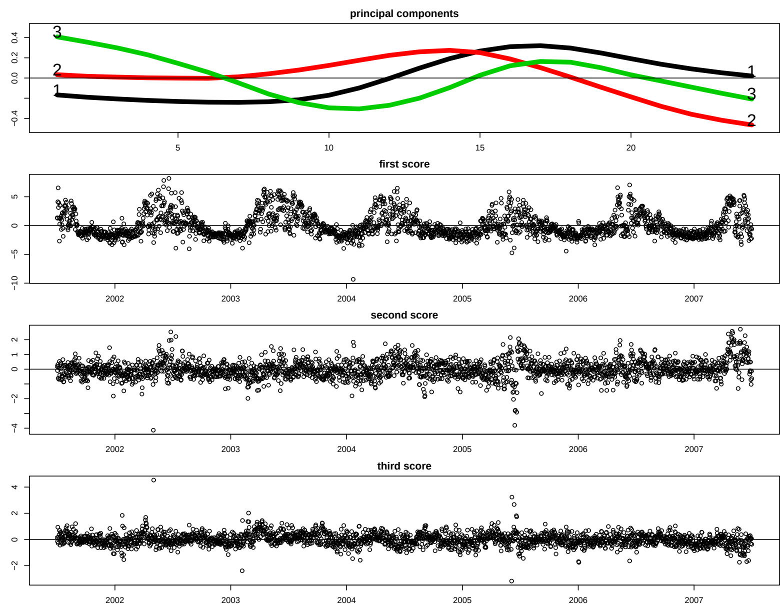
$$
\begin{array}{lllcll}
\text{water:} & w_t = w_{ydh} & = & \bar{w}_{yd} & + & x_{ydh} \\
 & & & \uparrow & & \uparrow \\
 & & & \text{daily avg. temp.} & & \text{residual} \\
 & & & \downarrow & & \downarrow \\
\text{air:} & a_t = a_{ydh} & = & \bar{a}_{yd} & + & z_{ydh}
\end{array}
$$

A principal component analysis is run on the residuals x_{ydh} and z_{ydh} after substracting the mean daily temperature course:

$$x_{ydh} = \mu_x(d) + \bar{x}_{ydh} \quad \text{and} \quad z_{ydh} = \mu_z(d) + \bar{z}_{ydh}.$$

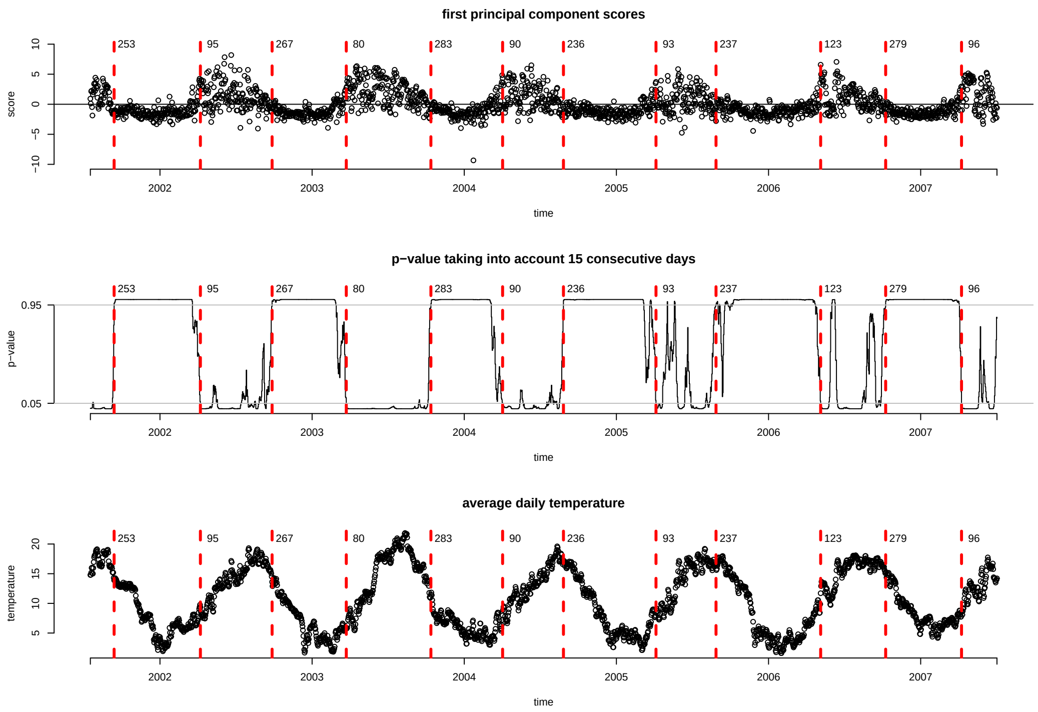
Seasonal Pattern in PCA coefficients

$$x_{ydh} = \mu_x(h) + \sum_{k=1}^{K_x} f_{yd,k} \lambda_{x,k}(h)$$



Landmark based on First PCA Score

We check, whether $\mathrm{H}_0 : \mathrm{E}(f_{yd,1}) \leq 0$ is rejected.



Correlation between Water and Air Temperature

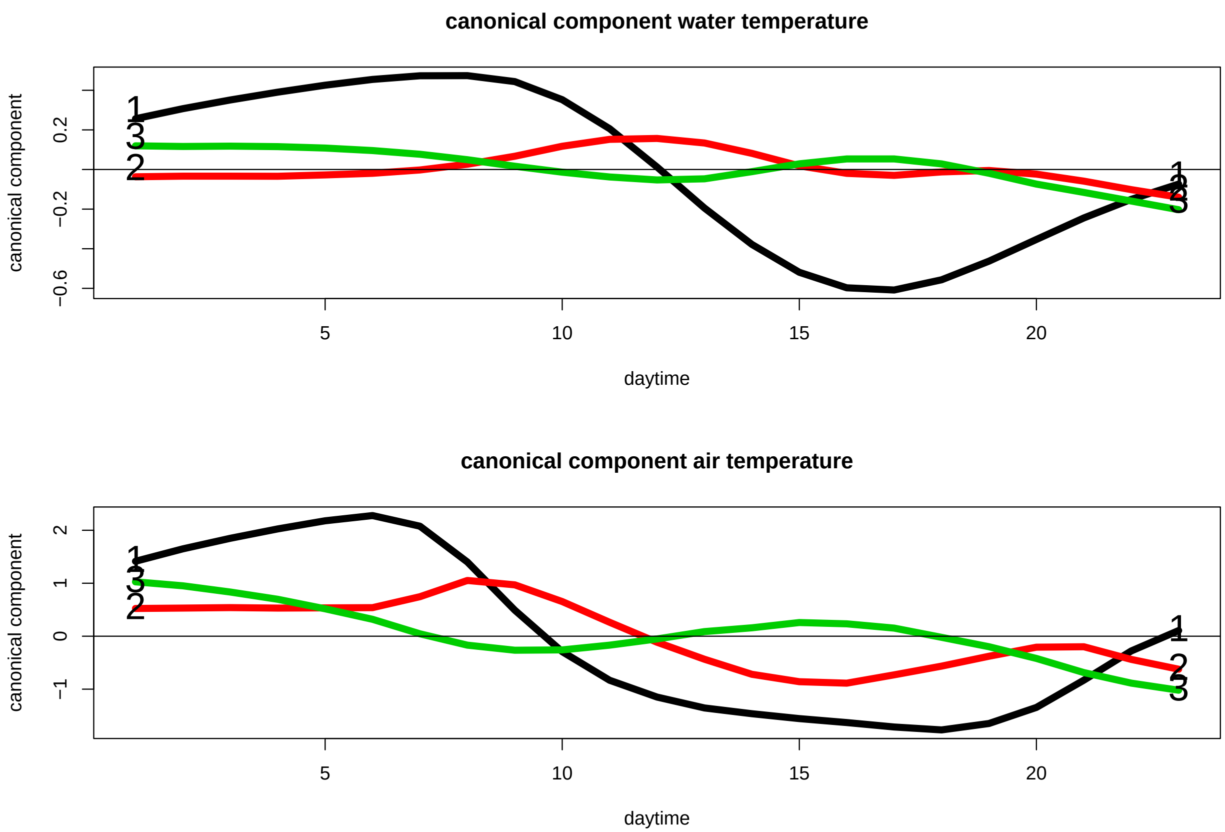
Water: $$x_{ydh} = \mu_x(h) + \sum_{k=1}^{K} f_{yd,k} \lambda_{x,k}(h)$$

Air: $$z_{ydh} = \mu_z(h) + \sum_{k=1}^{K} g_{yd,k} \lambda_{z,k}(h)$$

Canonical correlation:
For coefficient vectors δ_k^T and γ_k^T we obtain the maximal correlation between water and air temperature, i. e.

$$\max \mathsf{Cor}(\delta_k^T x_t, \gamma_k^T z_t), k = 1, 2, \ldots$$

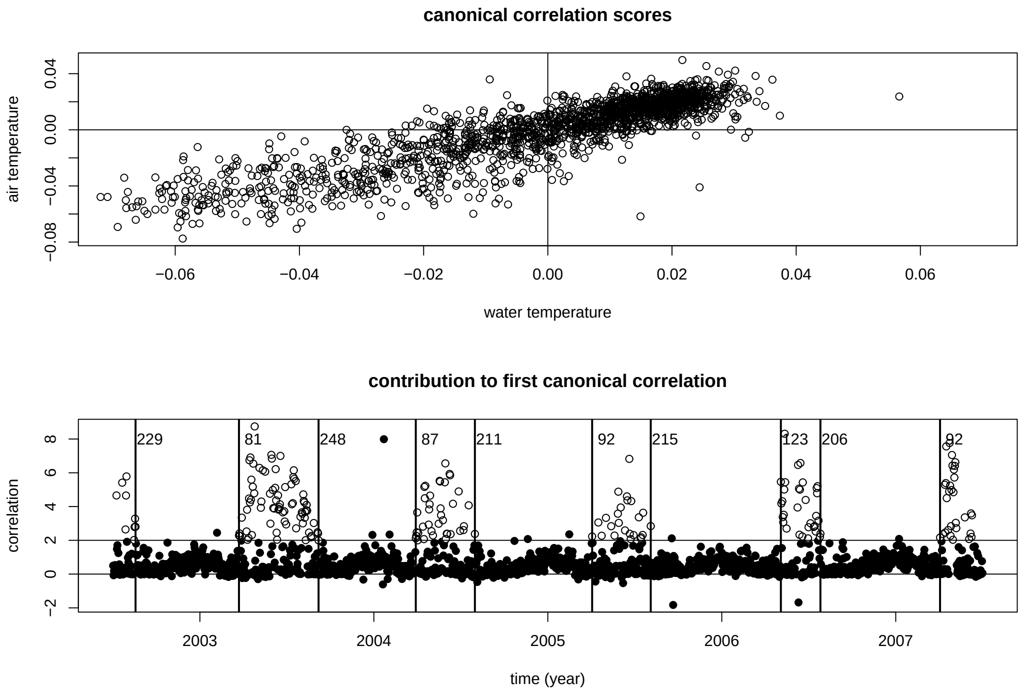
Canonical Correlation Landmark



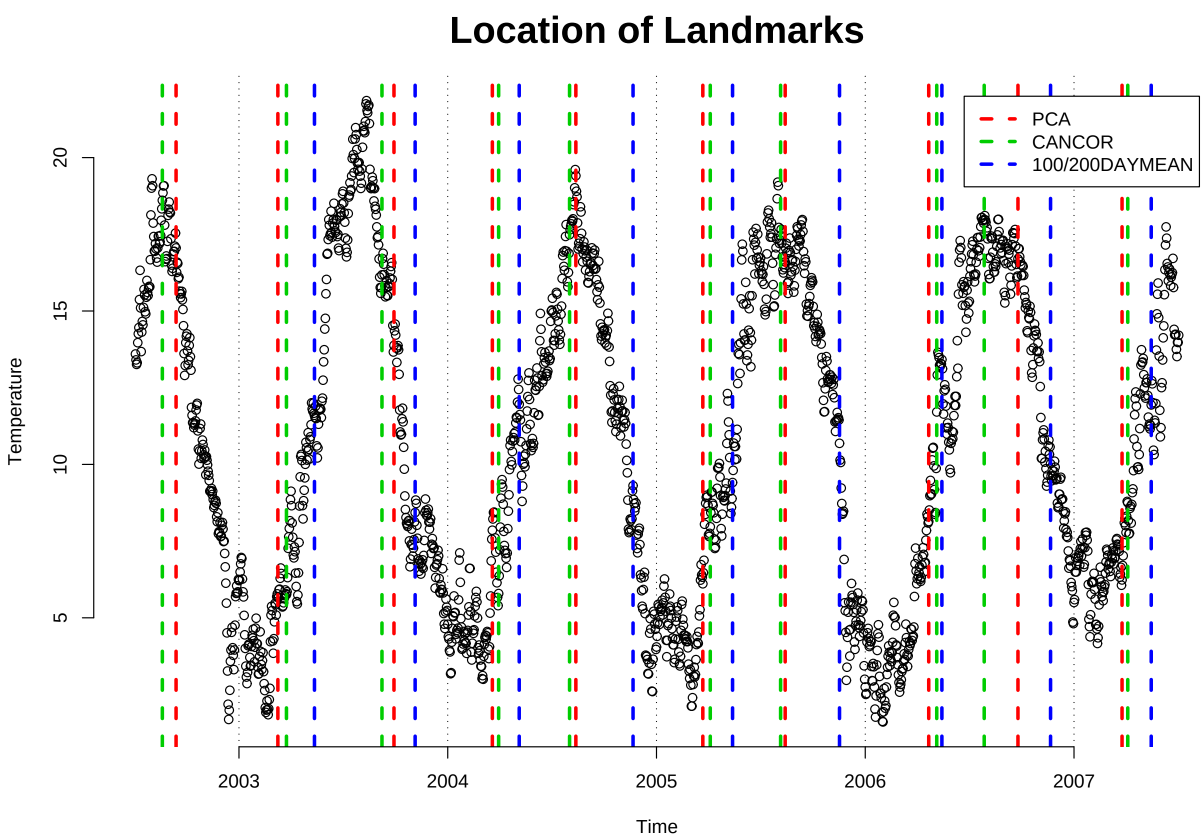
Canonical Correlation Contributions

We look at the canonical correlation:

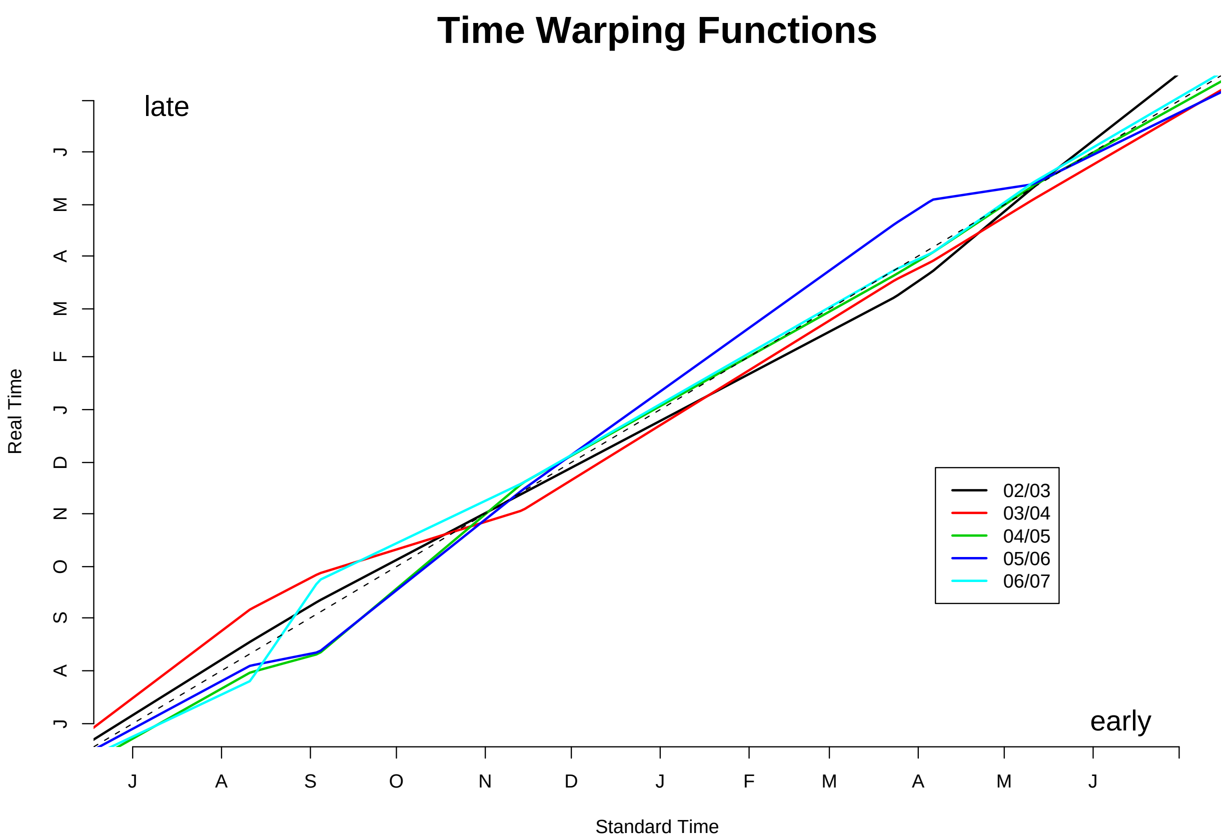
water: $\omega_t = \delta_1^T x_t$ air: $\nu_t = \gamma_1^T z_t$ both: $\omega_t \cdot \nu_t$



Plotting the Landmarks



Warping the Years



Discussion

- Analysis on Forecasting of Water Temperature is an important issue (and is getting even more important based on new EU laws).
- The issue is not fully covered by classical and newer approaches in time series analysis.
- Finding landmarks for seasonal variation is relevant from an ecological point of view.
- More to do: Compare our time warp results to observed fish spawning cycles.