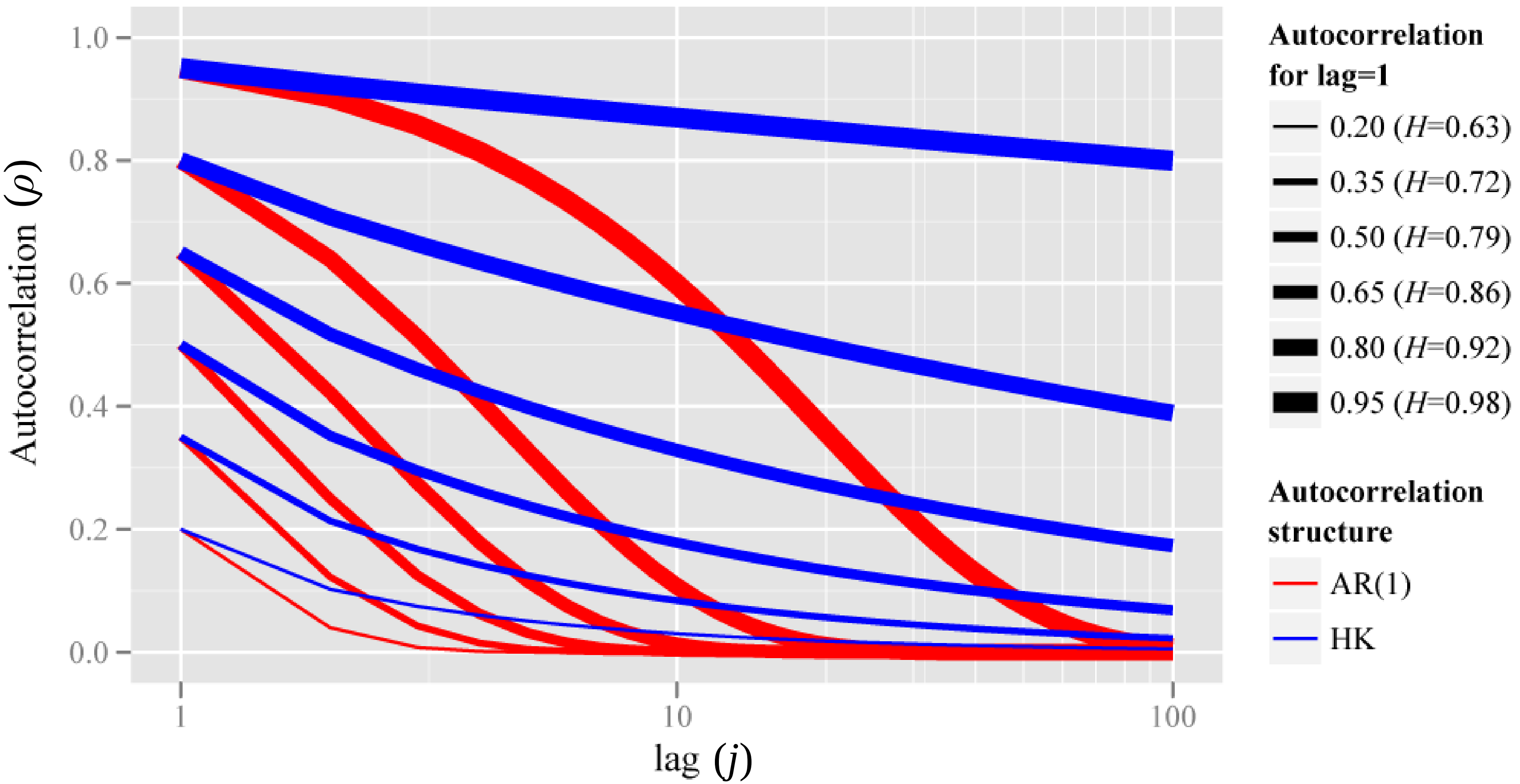
A quick gap-filling of missing hydrometeorological data

Christoforos Pappas[1,2], Simon Michael Papalexiou[2], Demetris Koutsoyiannis[2]

[1]ETH Zurich, Institute of Environmental Engineering, Swiss Federal Institute of Technology
pappas@ifu.baug.ethz.ch

[2]National Technical University of Athens, Department of Water Resources, Faculty of Civil Engineering



Exponential ACS : $\rho_j = \rho^{|j|}$

Power-type ACS : $\rho_j = \frac{1}{2}\left[\left(j+1\right)^{2H} + \left(j-1\right)^{2H}\right] - j^{2H}$

- Given that **2×N observations** are available, we want to estimate a **missing value y** :

$$x_{-N}, \ldots, x_{-1}, \boxed{y,} \; x_1, \ldots, x_N$$

- A (linear) **estimate of y** can be expressed as:

$$\underline{y} = w_{-N} x_{-N} + \ldots + w_N x_N + \underline{e}$$

where

x_i : the observed values

w_i : weighting factors

$\underline{e}$: estimation error

- The **Mean Squared Error** of the estimation is then defined as:

$$\mathrm{MSE} := \mathrm{E}\left[e^2\right] = \mathrm{E}\left[\left(y - \underline{y}\right)^2\right]$$

- We examine the following **estimate for *y***:

$$\underline{y} = \frac{\sum_{i=1}^{n} x_{-i} + \sum_{i=1}^{n} x_i}{2n}$$

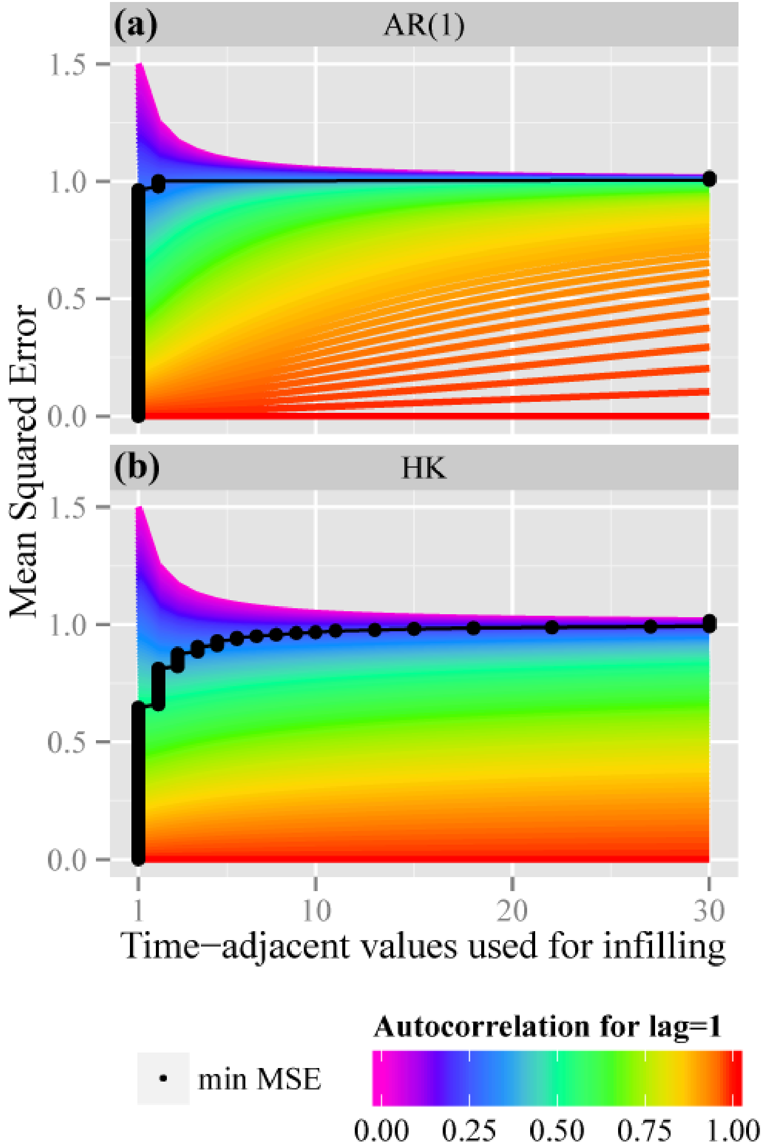
$$x_{-N}, \ldots, x_{-n}, \ldots, x_{-1}, y, x_1, \ldots, x_n, \ldots, x_N$$

- Assuming that the underlying process is (weakly) **stationary**, the **MSE** of the estimation is given by:

$$\text{MSE} := \text{E}\left[e^2\right] = \text{E}\left[\left(y - \underline{y}\right)^2\right] = \text{E}\left[\left(y - \frac{\sum_{i=1}^{n} x_{-i} + \sum_{i=1}^{n} x_i}{2n}\right)^2\right]$$

$$= \frac{1}{2}\left(\frac{\sigma}{n}\right)^2 \left[(2n+1)\left(n - 2\sum_{i=1}^{n} \rho_i\right) + \sum_{i=1}^{2n} (2n+1-i)\,\rho_i\right]$$

- Which is the **optimal** (i.e., minMSE) number of **neighbouring values (*n*)** that should be used**?**



$$\underline{y} = \frac{\sum_{i=1}^{n} x_{-i} + \sum_{i=1}^{n} x_i}{2n}$$

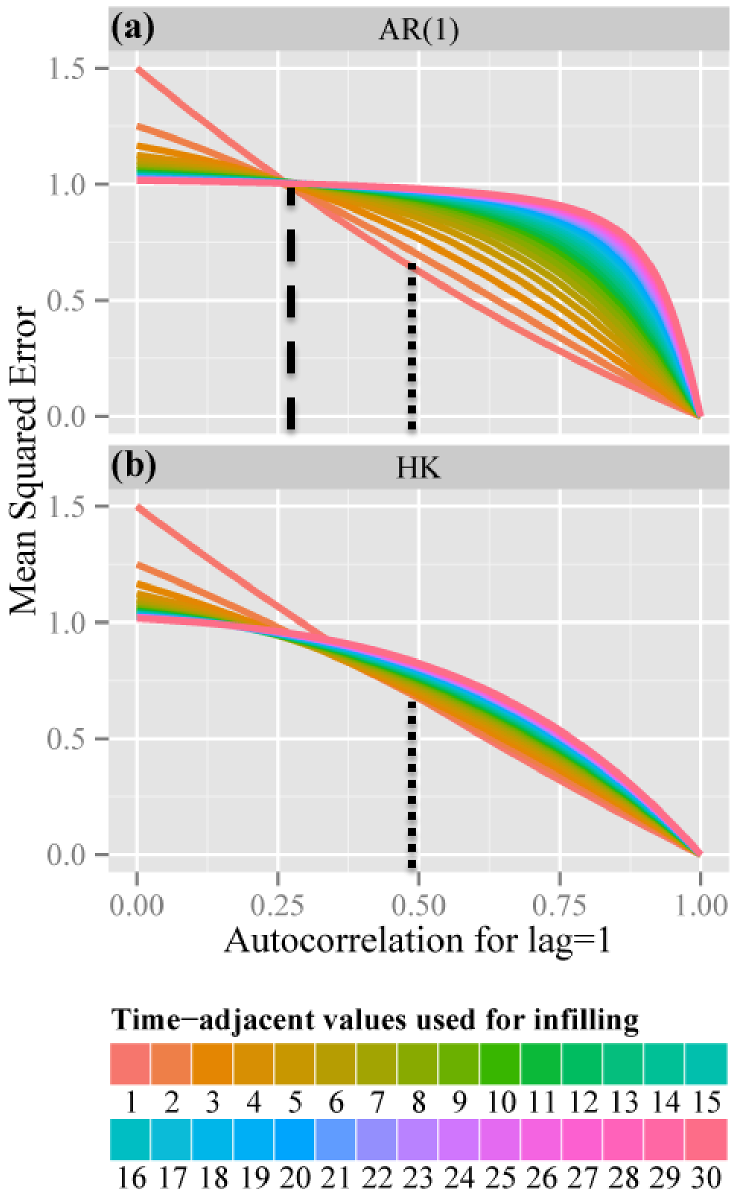
- **AR(1)**

For a wide range of **lag-1 autocorrelations**, the strictly local average (i.e., ***n*=1**) provides the **minMSE**.

- **HK**

As the **lag-1 autocorrelation increases**, the time-adjacent values (***n***) required for a **minMSE** gradually **decrease**.

See also
Dialynas et al. (2010)
http://itia.ntua.gr/en/docinfo/981/
Pappas (2010)
http://itia.ntua.gr/en/docinfo/1065/



$$\underline{y} = \frac{\sum_{i=1}^{n} x_{-i} + \sum_{i=1}^{n} x_i}{2n}$$

Markovian property:
"The future does not depend on the past when the present is known" [*Papoulis*, 1965, p.535].

Optimal Local Average			
Short-term persistence -AR(1)-		Long-term persistence -HK-	
$\rho \leq 0.25$	n=$n_{\max}$	$\rho \leq 0.3$	n=$n_{\max}$
$0.25 < \rho \leq 0.28$	n=2	$0.30 < \rho \leq 0.32$	n=4
		$0.32 < \rho \leq 0.38$	n=3
$\rho > 0.28$	n=1	$0.38 < \rho \leq 0.51$	n=2
		$\rho > 0.51$	n=1

ρ: lag-one autocorrelation coefficient
n: time-adjacent values used for the infilling
n_{max}: all the available observed values, i.e., total/sample average

For **both ACS** (exponential or power-type) when $\boldsymbol{\rho}$**>0.51** the strictly local average (**n=1**) provides the **minMSE**.

- **Generalization** of the OLA methodology, so that **information** from both **local** and **global average** will be used according to the **lag-1 autocorrelation**.

- We examine the following **estimate for *y***:

$$\underline{y} = \lambda \frac{\sum_{i=-N}^{N} x_i}{2N} + (1-\lambda)\frac{x_{-1}+x_1}{2}$$

Total (sample) average — **Local** (strictly) average

where λ **is the weighting factor** for the **total (sample) average** and the **local (strictly) average**.

- Parameter λ reflects the strength of the temporal autocorrelation:

low values → **high** correlation
high values → **low** correlation

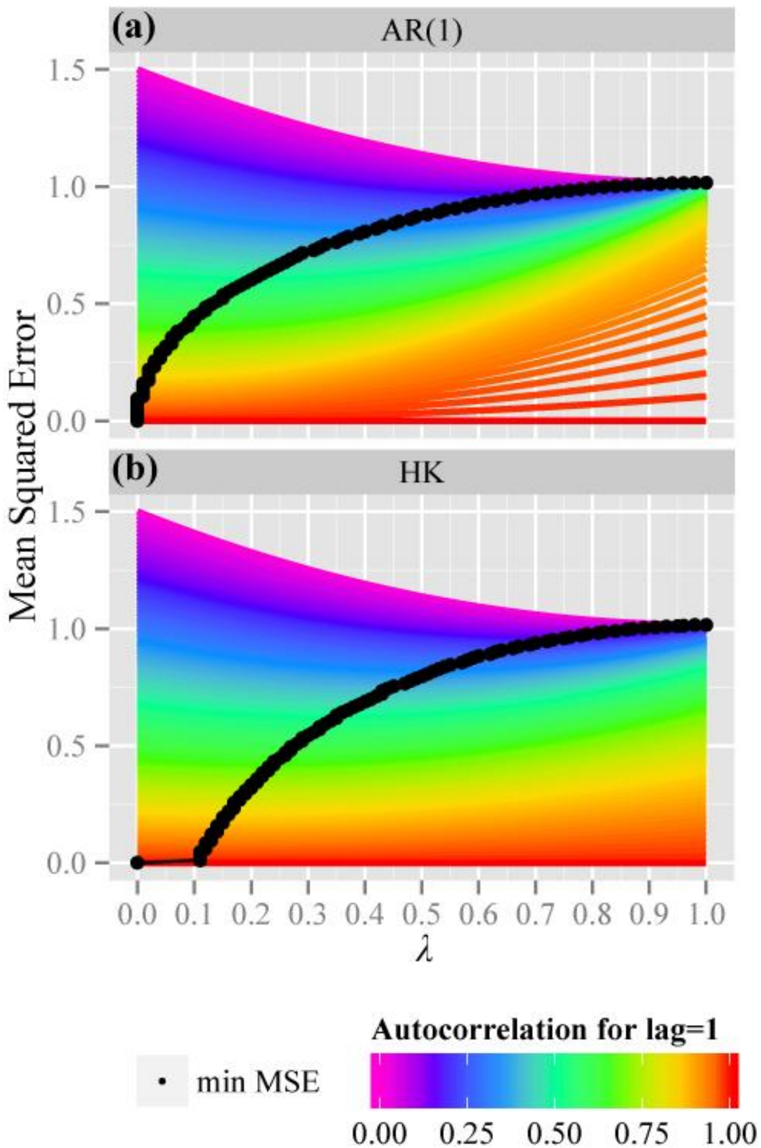
See also:
Pappas (2010)
http://itia.ntua.gr/en/docinfo/1065/

- Assuming that the underlying process is (weakly) **stationary**, the **MSE** of the estimation is then defined as:

$$\text{MSE} := \text{E}\left[e^2\right] = \text{E}\left[\left(y - \underline{y}\right)^2\right] = \text{E}\left[\left(y - \left(\lambda \frac{\sum_{i=-N}^{N} x_i}{2N} + (1-\lambda)\frac{x_{-1} + x_1}{2}\right)\right)^2\right]$$

- After some **algebraic** manipulations:

$$\text{MSE} = \frac{1}{2}\sigma^2(3 - 4\rho_1 + \rho_2) - 2\lambda\sigma^2\left[\frac{1}{N}\sum_{i=1}^{N}\rho_i - \frac{1}{2N}\left(\sum_{i=1}^{N-1}\rho_i - \sum_{i=2}^{N+1}\rho_i + 1\right) - \rho_1 + \frac{\rho_2}{2} + 0.5\right]$$
$$+\lambda^2\sigma^2\left[\frac{1}{2N^2}\left(2\sum_{i=1}^{N-1}(N-i)\rho_i + \sum_{i=2}^{N+1}(i-1)\rho_i + \sum_{i=N+2}^{2N}(2N+1-i)\rho_i + N\right) + \frac{\rho_2}{2} + \frac{1}{2} - \frac{1}{N}\left(\sum_{i=1}^{N-1}\rho_i + \sum_{i=2}^{N+1}\rho_i + 1\right)\right]$$



$$\underline{y} = \lambda \frac{\sum_{i=-N}^{N} x_i}{2N} + (1-\lambda)\frac{x_{-1} + x_1}{2}$$

- **AR(1) & HK**

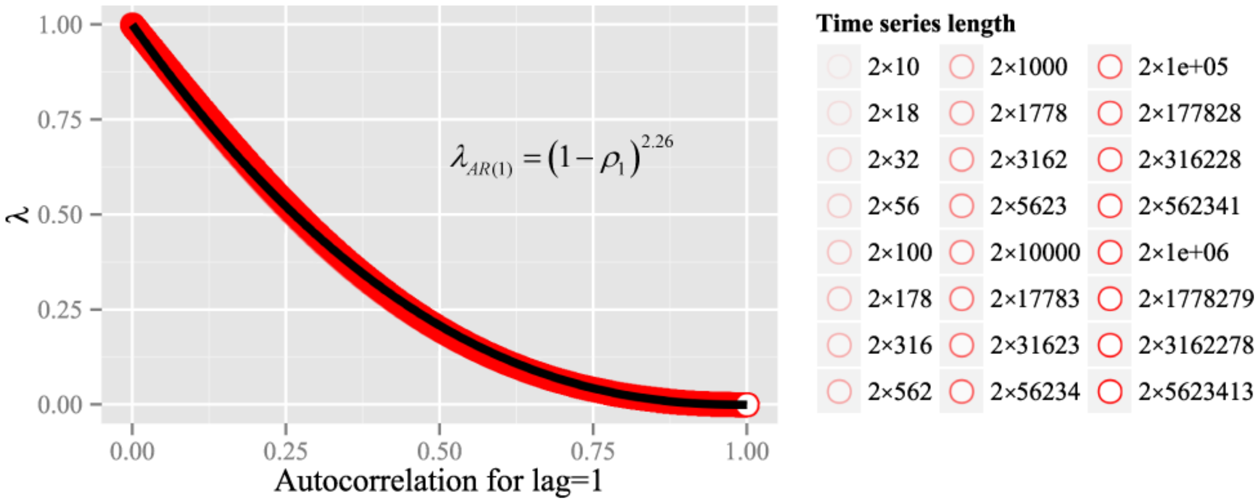
As **lag-1 autocorrelation** increases, the contribution of the local average increases (i.e., lower values of $\boldsymbol{\lambda}$).

- **HK**

It *takes time* for the **HK process** to **reveal** its **properties**.

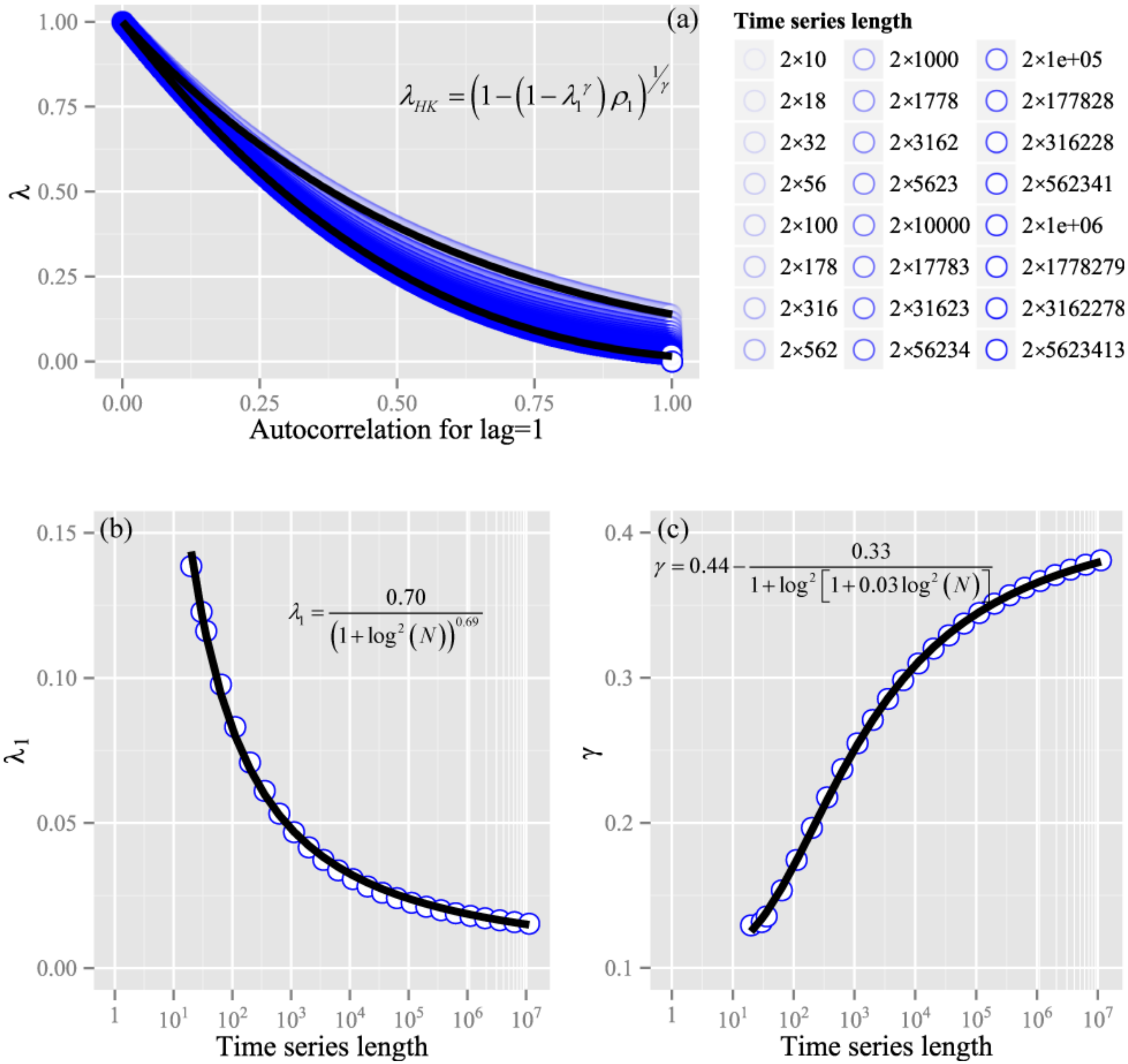
- The influence of **sample size (N)** :

$$\mathrm{MSE} = \frac{1}{2}\sigma^2(3-4\rho_1+\rho_2) - 2\lambda\sigma^2\left[\frac{1}{N}\sum_{i=1}^{N}\rho_i - \frac{1}{2N}\left(\sum_{i=1}^{N-1}\rho_i - \sum_{i=2}^{N+1}\rho_i + 1\right) - \rho_1 + \frac{\rho_2}{2} + 0.5\right]$$
$$+\lambda^2\sigma^2\left[\begin{array}{l}\frac{1}{2N^2}\left(2\sum_{i=1}^{N-1}(N-i)\rho_i + \sum_{i=2}^{N+1}(i-1)\rho_i + \sum_{i=N+2}^{2N}(2N+1-i)\rho_i + N\right) \\ +\frac{\rho_2}{2} + \frac{1}{2} - \frac{1}{N}\left(\sum_{i=1}^{N-1}\rho_i + \sum_{i=2}^{N+1}\rho_i + 1\right)\end{array}\right]$$



$$\underline{y} = \lambda_{AR(1)} \frac{\sum_{i=-N}^{N} x_i}{2N} + (1-\lambda_{AR(1)}) \frac{x_{-1} + x_1}{2}$$

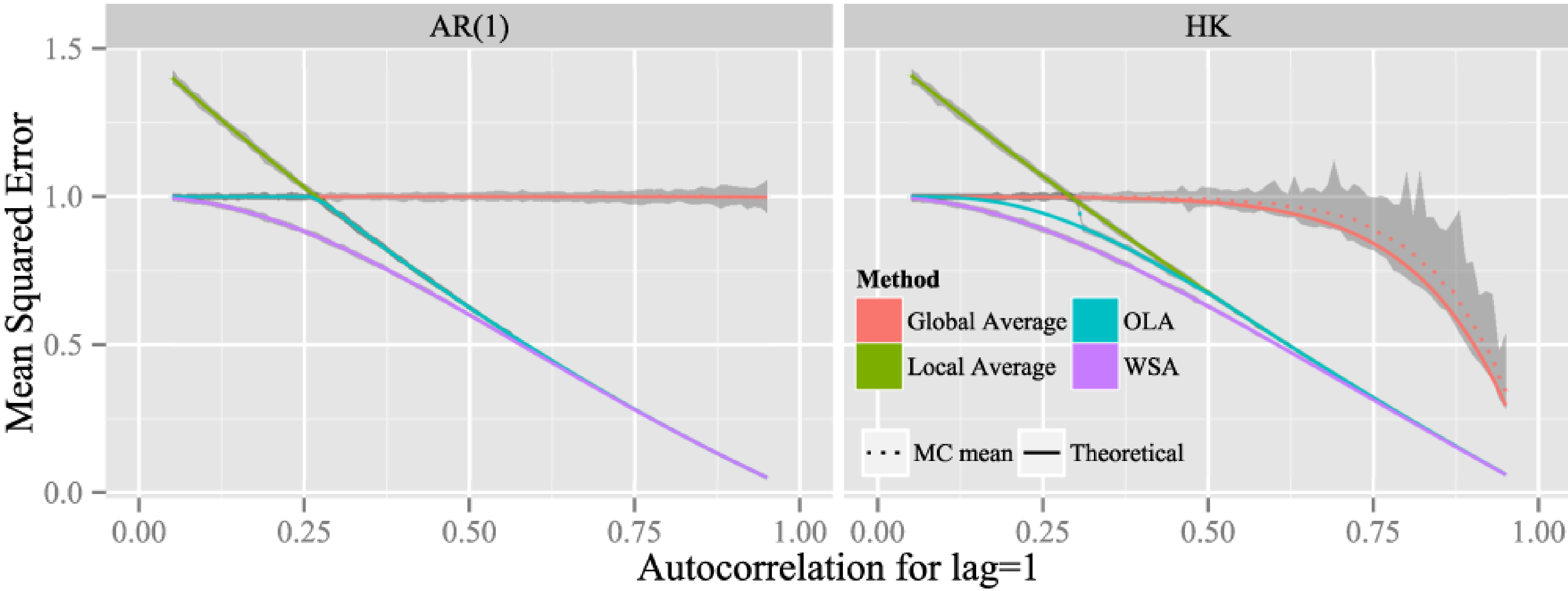
- For the case of **exponential ACS**, the $\boldsymbol{\lambda}$-$\boldsymbol{\rho_1}$ relationship does **not vary** significantly with time series length $\boldsymbol{N}$.



$$\underline{y} = \lambda_{HK}\frac{\sum_{i=-N}^{N} x_i}{2N} + (1-\lambda_{HK})\frac{x_{-1}+x_1}{2}$$

- For the case of **power-type ACS**, the $\boldsymbol{\lambda}$-$\boldsymbol{\rho_1}$ relationship **varies** significantly with time series length $\boldsymbol{N}$.

- To circumvent this issue, the $\boldsymbol{\lambda}$-$\boldsymbol{\rho_1}$ is approximated using two additional parameters ($\boldsymbol{\lambda_1}$, $\boldsymbol{\gamma}$).


AR(1)
HK
1.5
1.0
0.5
0.0
Mean Squared Error
0.00
0.25
0.50
0.75
1.00
Autocorrelation for lag=1
Method
Global Average
Local Average
OLA
WSA
MC mean
Theoretical

- We provide a **definitive argument against** the effortless use of **global (sample) mean** for infilling hydrometeorological (i.e., correlated) data.

- Local average ($\boldsymbol{n=1}$) is preferable for:
  - **Markovian** processes with $\boldsymbol{\rho > 0.28}$
  - **HK** processes with $\boldsymbol{\rho > 0.51}$

Tobler's first law in geography:
"Everything is related to everything else, but near things are more related than distant things" [*Tobler*, 1970]

- A generalized framework, described by the **Weighted Sum of local and total Average** (WSA), is developed and its advantages are demonstrated.

- The **WSA** methodology is therefore tailored for a **quick** infilling of **sporadic gaps** in hydrometeorological time series.

- Dialynas, Y., P. Kossieris, K. Kyriakidis, A. Lykou, Y. Markonis, C. Pappas, S.M. Papalexiou, and D. Koutsoyiannis, Optimal infilling of missing values in hydrometeorological time series, *European Geosciences Union General Assembly 2010, Geophysical Research Abstracts, Vol. 12*, Vienna, EGU2010-9702, European Geosciences Union, 2010.

- Papoulis, A. (1965), *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill.

- Pappas, C., *Optimal infilling of missing hydrometeorological data using time-adjacent observations*, Diploma thesis, 226 pages, Department of Water Resources and Environmental Engineering – National Technical University of Athens, October 2010.

- Tobler, W. (1970), A computer movie simulating urban growth in the Detroit region, *Economic Geography*, *46*, 234–240.

Many thanks to…

European Geosciences Union General Assembly 2010

Vienna, Austria, 2-7 May 2010

Session HS5.5: Stochastics in hydrometeorological processes: from point to global spatial scales and from minute to climatic time scales

Optimal infilling of missing values in hydrometeorological time series

Y. Dialynas, P. Kossieris, K. Kyriakidis, A. Lykou, Y. Markonis, C. Pappas, S.M. Papalexiou and D. Koutsoyiannis

Department of Water Resources and Environmental Engineering
National Technical University of Athens
(www.itia.ntua.gr)

Thank you!

Christoforos Pappas[1,2], Simon Michael Papalexiou[2], Demetris Koutsoyiannis[2]

[1]ETH Zurich, Institute of Environmental Engineering, Swiss Federal Institute of Technology (pappas@ifu.baug.ethz.ch)

[2]National Technical University of Athens, Department of Water Resources, Faculty of Civil Engineering