KERNEL AND WAVELET SMOOTHING: Basic theory and examples of practical data analysis[1)2)]

Vítězslav Veselý
Brno, Czech Republic

1991 Mathematics Subject Classification. Primary: 62G07; **Secondary:** 65D10
Key words and phrases: Kernel smoothing, wavelet smoothing

1. INTRODUCTION

Kernel and wavelet smoothing are two modern tools for data analysis at this time. We shall explain the basic theoretical ideas of both methods and confront the results obtained when processing each of three environmental data series by either method.

Notation.

$\boldsymbol{N}$	...	the set of all natural numbers
$\boldsymbol{Z}$	...	the set of all integers
$\boldsymbol{R}$	...	the set of all real numbers
L_1	...	the Banach space of all complex-valued functions which are absolutely integrable on $\boldsymbol{R}$ in the Lebesgue sense
$L_2(\mathcal{J})$	...	the Hilbert space of all complex-valued measurable functions which are absolutely square-integrable in the Lebesgue sense on the interval $\mathcal{J} \subseteq \boldsymbol{R}$; we denote further $\langle g_1, g_2\rangle = \int_{\mathcal{J}} g_1(x)\overline{g_2(x)}\,dx$ ...the inner product and $\|g\| = \sqrt{\langle g, g\rangle} = \sqrt{\int_{\mathcal{J}} \|g(x)\|^2\,dx}$ ...the norm $L_2 := L_2(\boldsymbol{R})$
$\delta_{i,j}$	...	the Kronecker symbol (= 1 for $i = j$ and zero otherwise)
E	...	the operator of expectation

2. THE BASIC CONCEPT OF KERNEL SMOOTHING

Problem statement

Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be sorted values $a \leq X_1 < X_2 < \ldots < X_n \leq b$ of an independent variable $x \in \boldsymbol{R}$, which are either fixed prescribed or samples of a random variable X, and let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ be samples of a random variable Y observed for $\boldsymbol{X}$ within the additive model

$$Y_i = m(X_i) + E_i, \quad i = 1, 2, \ldots, n,\ n \in \boldsymbol{N}$$

where $m(x) = \mathrm{E}(Y|X = x)$ is an unknown function to be estimated from a set of n measurements of $m(X_i)$ which are loaded with independent and identically distributed errors E_i.
The frequently considered nonparametric estimator $\widehat{m}(x)$ of $m(x)$ is usually the weighted average of response observations $Y_1, Y_2, ..., Y_n$ and a general formula for this estimator is

$$m(x) \approx \widehat{m}(x) = \sum_{i=1}^{n} Y_i\, W_{n,i}(x) \tag{1}$$

where $W_{n,i}(x)$ is a suitable **weight function** depending in general not only on the fitted point x but also on the covariate observations $X_1, X_2, \ldots, X_n$, i.e. $W_{n,i}(x) = W_{n,i}(x, X_1, \ldots, X_n)$. Typically $W_{n,i}(x) \to 0$ as $|x - X_i| \to \infty$. Such an estimator is called a linear smoother because it is linear in the response.
The linear smoothers differ by the choice of the weight functions $W_{n,i}$, $i = 1, 2, ..., n$. (see e.g. [8]), and a very useful method for the choice of weights is the **kernel smoothing**.

Definition.
A real function $K(x)$ is called a **kernel function (kernel)** if $K(x) \in L_1$ and $\int_{-\infty}^{\infty} K(x)\,dx = 1$. We put also $K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right)$, $h > 0$ the **width-modified kernel function** which, clearly, preserves $\int_{-\infty}^{\infty} K_h(x)\,dx = 1$.

There are two popular approaches to constructing kernel estimates.

Definition. (Weight function by Nadaraya-Watson (1964) [12, 14])

$$W_{n,i}^{(1)}(x) := \frac{K_{h_i}(x - X_i)}{\sum_{i=1}^{n} K_{h_i}(x - X_i)} \tag{2}$$

where $K(x)$ is a continuous kernel with fast decay $\lim_{x\to\pm\infty} xK(x) = 0$.

Definition. (Weight function by Gasser-Müller (1979) [7])

Put $s_o = a$, $s_i = \frac{1}{2}(X_i + X_{i+1})$ for $i = 1, 2, \ldots, n-1$ and $s_n = b$.

1) *Research supported by the GA of the Czech Republic under grant number 201/96/0665.*
2) *In: Proceedings and abstracts of ESES'96, Brno (I. Horová, J. Jurečková and J. Vosmanský, ed.), Masaryk University of Brno, Czech Rep., Aug. 1996, pp. 163–174.*

Then

$$W_{n,i}^{(2)}(x) := \int_{s_{i-1}}^{s_i} K_{h_i}(x-t)\,dt \qquad (3)$$

for a kernel $K(x)$.
The former (eq. (2)) is based on the choice of weights by means of direct kernel evaluation — the so-called Nadaraya-Watson estimators (see [8]). The latter (eq. (3)) is based on weights which are a convolution of the kernel with a histogram representing the data (see [11]). It is typical for both of these methods that the weights $W_{n,i}$ depend on some function K called **kernel** and on parameters h_i called **smoothing parameters** or **bandwidths**. The kernel K and the bandwidths h_i determine the local properties of the estimator $\widehat{m}(x)$ at the point x. If the bandwidths $h := h_1 = h_2 = \ldots = h_n$ are **fixed**, depending neither on the location of x nor on the covariate values $X_1, X_2, ..., X_n$, the information provided by the density of data points is not fully incorporated in the estimator. That is why Fan and Gijbels introduced a new type of kernel estimator with **variable bandwidth** (see [6]). Their estimator is based on the Cleveland idea (see [1]) to obtain a linear smoother via a local linear approximation to the regression function. For an expository paper on the variable bandwidth method see [10].
We can also guess the fixed bandwidth by the trial and error method or in some optimal way by minimizing a suitable error measure, the mean-square error estimate being the typical choice.
In addition also the so-called **optimal kernels** may be used. These kernels are compactly supported on $[-1, 1]$ where they coincide with a polynomial of a given degree which smoothly decays to zero at interval bounds -1 and 1 with order of smoothness being prescribed.

3. THE BASIC CONCEPT OF WAVELET SMOOTHING

Problem statement

Consider the additive model

$$Y(x) = m(x) + e(x), \quad x \in \boldsymbol{R}$$

where $Y(x)$ are observed values, $m(x) \in L_2$ is an unknown real function to be estimated and $e(x)$ is the white noise. In the discrete setting $Y(x)$ are of course observed only at a finite discrete set $x = X_1, X_2, \ldots, X_n$.
On the contrary to kernel smoothing which yields a direct estimate $\widehat{m}(x)$ of $m(x)$, in the case of **wavelet smoothing** (for basic wavelet theory see for example the monographs [2, 9]) we estimate $m(x)$ indirectly by finding estimates $\widehat{c}_{j,k}$ of coefficients $c_{j,k}, j, k \in \boldsymbol{Z}$ in the expansion

$$m(x) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} c_{j,k}\psi_{j,k}(x) \qquad (4)$$

where $\psi_{j,k}(x) = 2^{j/2}\psi\left(\frac{x-k/2^j}{1/2^j}\right)$ is a specific basis in L_2 generated by dilating and shifting a suitable function $\psi(x) \in L_2, ||\psi|| = 1$ which is called **mother wavelet**, clearly $||\psi_{j,k}|| = 1$ is preserved for each j, k. If this basis is orthonormal in L_2 ($\langle\psi_{j,k}, \psi_{l,m}\rangle = \delta_{j,l}\delta_{k,m}$) then $\psi(x)$ is called **orthogonal wavelet** and $c_{j,k}$ are easily computed by

$$c_{j,k} = \langle m, \psi_{j,k}\rangle = \int_{-\infty}^{\infty} m(x)\overline{\psi_{j,k}(x)}\,dx \qquad (5)$$

The wavelet expansions are a non-periodic analogy of Fourier expansions of T-periodic functions in $L_2([a,b]), b-a = T > 0$ where $w(x) = \frac{1}{\sqrt{T}}\exp^{i2\pi x/T}$ is the basic complex wave function playing the role of $\psi(x)$ and generating the orthonormal Fourier basis $\{w_j(x)\}_{j\in\boldsymbol{Z}}, w_j(x) = w(jx)$. Then for T-periodic $m(x) \in L_2([a,b])$ we get its Fourier series expansion

$$m(x) = \sum_{j=-\infty}^{\infty} \langle m, w_j\rangle w_j(x) = \sum_{j=-\infty}^{\infty} c_j \exp^{i2\pi jx/T}$$

where

$$\begin{aligned} c_j &= \frac{1}{\sqrt{T}}\langle m, w_j\rangle = \frac{1}{\sqrt{T}}\int_a^b m(x)\overline{w_j(x)}\,dx \\ &= \frac{1}{T}\int_a^b m(x)\exp^{-i2\pi jx/T}\,dx \end{aligned}$$

are the well-known Fourier coefficients of m.
Hence it is natural to call the wavelet expansion the **wavelet series** and the coefficients $c_{j,k}$ the **wavelet coefficients**. The index j determines the **scaling parameter** $1/2^j$ which is closely related to the j-th harmonic frequency $f_j = j/T$ associated with the j-th Fourier coefficient c_j. Nevertheless, one principal difference between $w(x)$ and $\psi(x)$ is that $\psi \in L_2 \Rightarrow \psi(x) \to 0$ for $x \to \pm\infty$. In addition the theory implies also another natural condition, namely $\psi(x) \in L_1$ and $\int_{-\infty}^{\infty}\psi(x)\,dx = 0$ which says that $\psi(x)$ must change its sign. Altogether $\psi(x)$ is a "damped wave" or **wavelet**. If $\psi(x)$ is completely damped (zero outside of a bounded interval) then we say that $\psi(x)$ is a **compactly supported wavelet**. This property shows to be very useful because each wavelet coefficient $c_{j,k}$ contributes to the wavelet expansion of $m(x)$ only in a neighbourhood of $x = k/2^j$, i.e. its effect is **local** compared with the **global** effect of a Fourier coefficient c_j.

Wavelet smoothing

Due to the linearity of (5) we get

$$c^Y_{j,k} = c_{j,k} + c^e_{j,k}$$

where $c^Y_{j,k}, c_{j,k}$ and $c^e_{j,k}$ are wavelet coefficients in the expansions of $Y(x), m(x)$ and $e(x)$, respectively. The wavelet smoothing techniques are aimed at finding a suitable modification rule $\mu(\cdot)$ such that $\mu(c^Y_{j,k}) = \widehat{c}_{j,k} \approx c_{j,k}$ is a good estimate of $c_{j,k}$. This approach follows again the analogy with Fourier-based filtration where we modify the Fourier coefficients c^Y_j, an important special case being the classical linear filtration where $\mu(c^Y_j) = \mu_j c^Y_j, \mu_j \geq 0$ and $\{\mu_j\}_{j \in \boldsymbol{Z}}$ is the so-called **transfer function of the filter**. Due to the local effect of $c_{j,k}$ for a fixed k the wavelet representation allows one to construct **locally adaptive** filters in this way which is an excellent **new feature** compared with the classical Fourier filters where the effect is global.
There are three wavelet coefficient modification techniques commonly applied.

1. **Positive scaling**

   $$\widehat{c}^{\,\text{pos}}_{j,k} = \mu_{j,k} c^Y_{j,k},\ \mu_{j,k} \geq 0$$

   which is the direct generalization of the transfer function mentioned above.

2. **Hard thresholding**
   All wavelet coefficients which are below a certain threshold level λ are put to zero:

   $$c^{\text{hard}}_{j,k} = \begin{cases} 0 & \text{for} \quad |c_{j,k}| < \lambda \\ c_{j,k} & \text{for} \quad |c_{j,k}| \geq \lambda \end{cases}.$$

3. **Soft thresholding**
   All wavelet coefficients are reduced by a certain threshold level λ:

   $$c^{\text{soft}}_{j,k} = \text{sign}(c_{j,k}) \max(0, |c_{j,k}| - \lambda)$$

Donoho and Johnstone [3, 4, 5] suggested a method for an optimal (in a certain sense) choice of the threshold λ which is either universal or specific for each scaling level j ($\lambda = \lambda_j$). These and other similar methods became known as **wavelet shrinkage** or **wavelet de-noising**.
In practical computation we use the discrete setting where $c_{j,k}$ are evaluated via DWT (**Discrete Wavelet Transform**) which is the natural counterpart of the DFT (Discrete Fourier Transform). The related fast algorithms are known as FWT (**Fast Wavelet Transform**) and FFT (Fast Fourier transform).

4. EXAMPLES OF PRACTICAL DATA ANALYSIS

In this section both kernel and wavelet smoothing will be demonstrated on real time series. All computations were accomplished in MATLAB 4.2c and supported by specialized m-file libraries (toolboxes). While the toolbox for kernel smoothing has been developed by the author himself, the results of wavelet smoothing have been obtained using WavBox 4.3b which is an excellent Wavelet Toolbox (218 m-files, 850 kB) of Carl Taswell from the Stanford University, CA, USA [13]. In particular for wavelet shrinkage we have applied the WavBox function `wdenoise` with the threshold estimator DJE (Donoho-Johnstone-Estimator) which yields a separate threshold for each scaling level. As the mother wavelet the compactly supported orthogonal least asymmetric wavelet of order 8 from the Daubechies family has been chosen. For more details the list of main WavBox object properties follows:

DataDimension = 1
MappingStructureType = DWT
MappingStructureClass = DWT
ObjectStructureType = DWT
ObjectStructureClass = DWT
FilterClass = ORTH
FilterFamily = DOLA
NumberVoices = 1
AnalysisFilterParameter = 8
SynthesisFilterParameter = 8
ConvolutionVersion = CPF
FilterName = DOLA16
FilterLength = 16

Description of data sets

Data set 1 (size 90):
Mean autumn atmospheric temperatures measured in Hurbanov, Czech republic in 1903–1992.
Data set 2 (size 168):
Seasonal deviations of cloudiness from the mean observed in the Northern Croatia in 1951–1992.
Data set 3 (size 912):
Monthly mean flow on the river Morava observed in Kroměříž, Czech Republic in 1916–1991.

Description of figures showing the smoothing results

Figure 1:
Plots of raw data — data set 1 (top), data set 2 (middle), data set 3 (bottom).
Figures 2–4:
Plots of kernel smoothed data sets 1, 2 and 3 , respectively using the weights (2) with the quartic kernel

$$K(x) = \begin{cases} \frac{15}{16}(1-x^2)^2 & \text{for} \quad x \in [-1,1] \\ 0 & \text{for} \quad x \notin [-1,1] \end{cases}$$

and three bandwidths: optimal (top), small (middle) resulting in undersmoothing and large (bottom) resulting in oversmoothing.
Figures 5–7:
Plots of wavelet processed data sets 1, 2 and 3, respectively showing the DJE smoothed data (top) along with the wavelet coefficients for the raw data

(bottom left) and smoothed data (bottom right). In the bottom plots $c_{j,k}$ are shown at positions related horizontally to k and vertically to j in reverse order of level numbers (the finest scaling level 1 corresponds to the largest j).

Figure 8:
Multiresolution analysis of the data set 3. The six plots when ordered row-by-row by the top-down and left-right method show data smoothed via clearing all wavelet coefficients at levels 1,1–2, ..., 1–6, respectively. The plots visualize the step-by-step decrease in the resolution when still more coefficients at successive levels are put to zero.

5. CONCLUSION

There are two basic factors which control the smoothing operation. First, both methods allow for a wide choice among various kernel or wavelet shapes. Second, both methods are yielding a great flexibility in the choice of the smoothing strategy (fixed or variable bandwidth, thresholding method or some other specific manipulation with the wavelet coefficients). For example the figures 2–4 show that the right choice of the bandwidth magnitude is crucial to the final effect. Both the variable bandwidth and a clever modification of wavelet coefficients are surely a powerful tool how to adapt to the local data behaviour. However the grand problem is to find the 'best' procudure just for the data we have.

Both methods offer certain universal procedures (optimal bandwidth, DJE thresholding) which should give an optimal result. But these optimality criteria are hard to compare, they are more related to the method itself than to the data being processed for which only certain general assumptions should be satisfied which cannot be usually exactly verified (except in simulations where everything works well). Observe that the optimal results of wavelet smoothing from figures 5–7 are not in a good agreement with those of optimal kernel smoothing in figures 2–4, respectively. In case of the data set 1 we see that kernel smoothing with bandwidth three times smaller than the optimal value gives a result which is clearly closer to the optimal wavelet result. The plots for the data set 2 exhibit opposite behaviour giving better agreement with a bandwidth two times larger than the optimal one. The data sets 2 and 3 seem to be extreme cases towards the DJE thresholding. The wavelet smoothing of the data set 2 (Fig. 6) is nearly total (close to zero mean) saying that only a negligible portion of useful information was detected. The data set 3 is the other extreme (Fig. 7) exhibiting a negligible smoothing effect. Although kernel smoothing with optimal bandwidth follows this trend, too (Fig. 3 and 4) we have not obtained such extreme results. So the question about credibility of the results is evident and one is advised to be very careful with universal techniques without exploiting any additional information about the data being processed.

References

[1] W. S. Cleveland, *Robust locally weighted regression and smoothing scatterplots*, J. Amer. Statist. Assoc. **74** (1979), 829–836.

[2] Ingrid Daubechies, *Ten lectures on wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 61, SIAM, Philadelphia, Pennsylvania, 1992.

[3] David L. Donoho and Iain M. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, Technical report, Stanford University, Department of Statistics, Stanford CA 94305, July 1994.

[4] ———, *Ideal spatial adaptation via wavelet shrinkage*, Biometrika **81** (1994), 425–455.

[5] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian and D. Picard, *Wavelet shrinkage: Asymptopia ?*, J. Royal Statist. Soc. B **57** (1995), no. 2, 301–337.

[6] J. Fan and I. Gijbels, *Variable bandwidth and local linear regression smoother*, The Annals of Statistics **20** (1992), no. 4, 2008–2036.

[7] T. Gasser and H. G. Müller, *Kernel estimation of regression functions*, In: Smoothing Techniques for Curve Estimation (Berlin) (T. Gasser nad M. Rosenblatt, ed.), Lecture Notes in Mathematics, vol. 757, Springer-Verlag, 1979, pp. 23–68.

[8] W. Härdle, *Applied nonparametric regression*, Cambridge University Press, Cambridge, 1990.

[9] Charles K. Chui, *An introduction to wavelets*, Wavelet Analysis and Its Applications, vol. 1, Academic Press, Inc., San Diego, CA, 1992.

[10] Jaroslav Michálek, *Kernel smoothing with variable bandwidth*, In: Proceedings of the summer school MATLAB'94, Velké Karlovice (I. Horová, ed.), Masaryk University, Brno, Czech Rep., August 1994, in print.

[11] Hans-Georg Müller, *Nonparametric regression analysis of longitudinal data*, Lecture Notes in Statistics, vol. 46, Springer-Verlag, Berlin, 1988.

[12] E. A. Nadaraya, *On estimating regression*, Theory Probab. Appl. **9** (1964), 141–142.

[13] Carl Taswell, *WavBox 4.3b (Wavelet Toolbox for MATLAB 4.2c)*, January 1996, (Internet URL: http://www.wavbox.com).

[14] G. S. Watson, *Smooth regression analysis*, Sankhya **A26** (1964), 359–372.

Author's address: V. Veselý, Masaryk University, Janáčkovo nám. 2a, 662 95 Brno, Czech Republic (email: vesely@math.muni.cz)

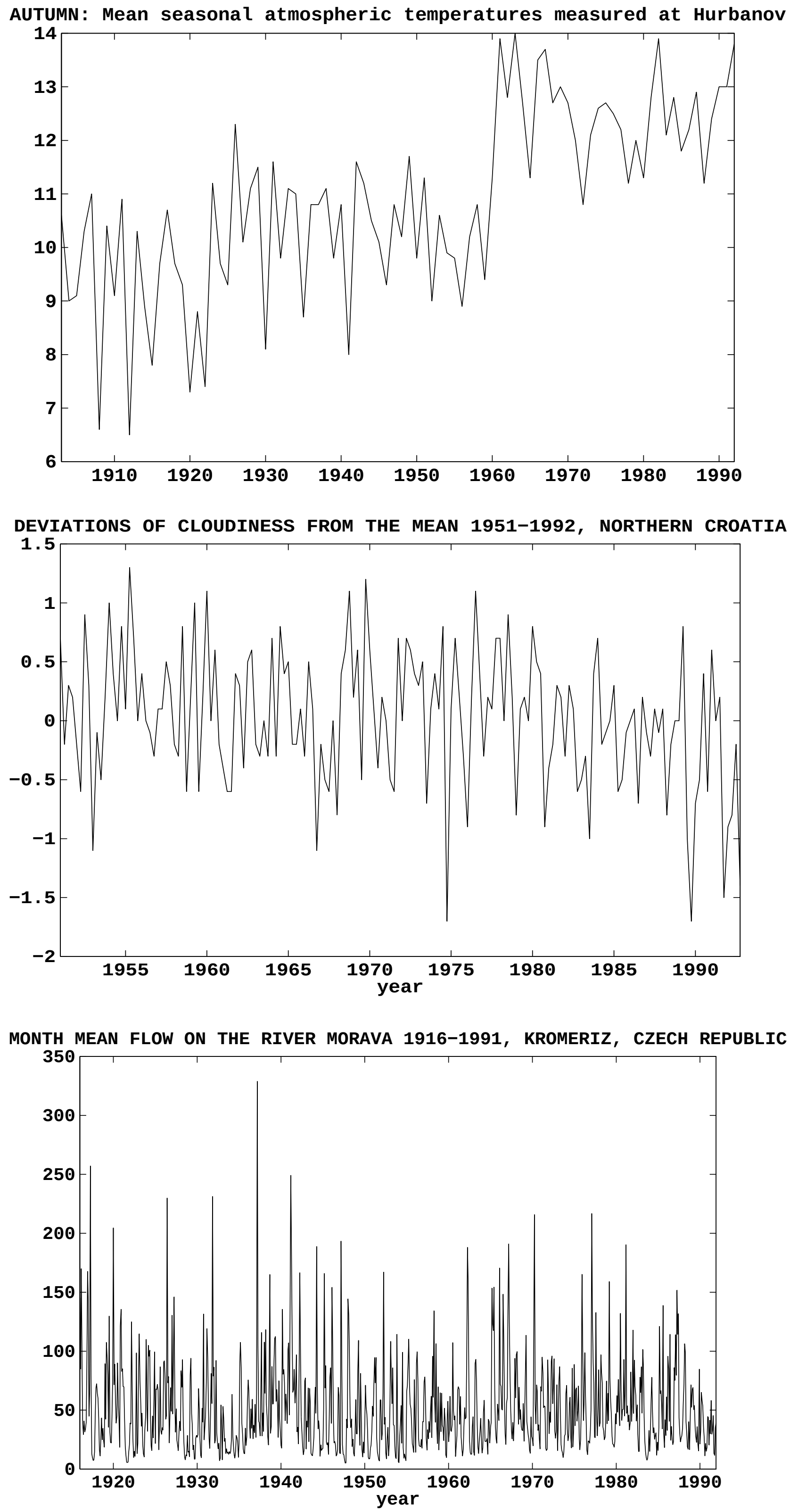

Figure 1: Plots of raw data sets 1,2 and 3

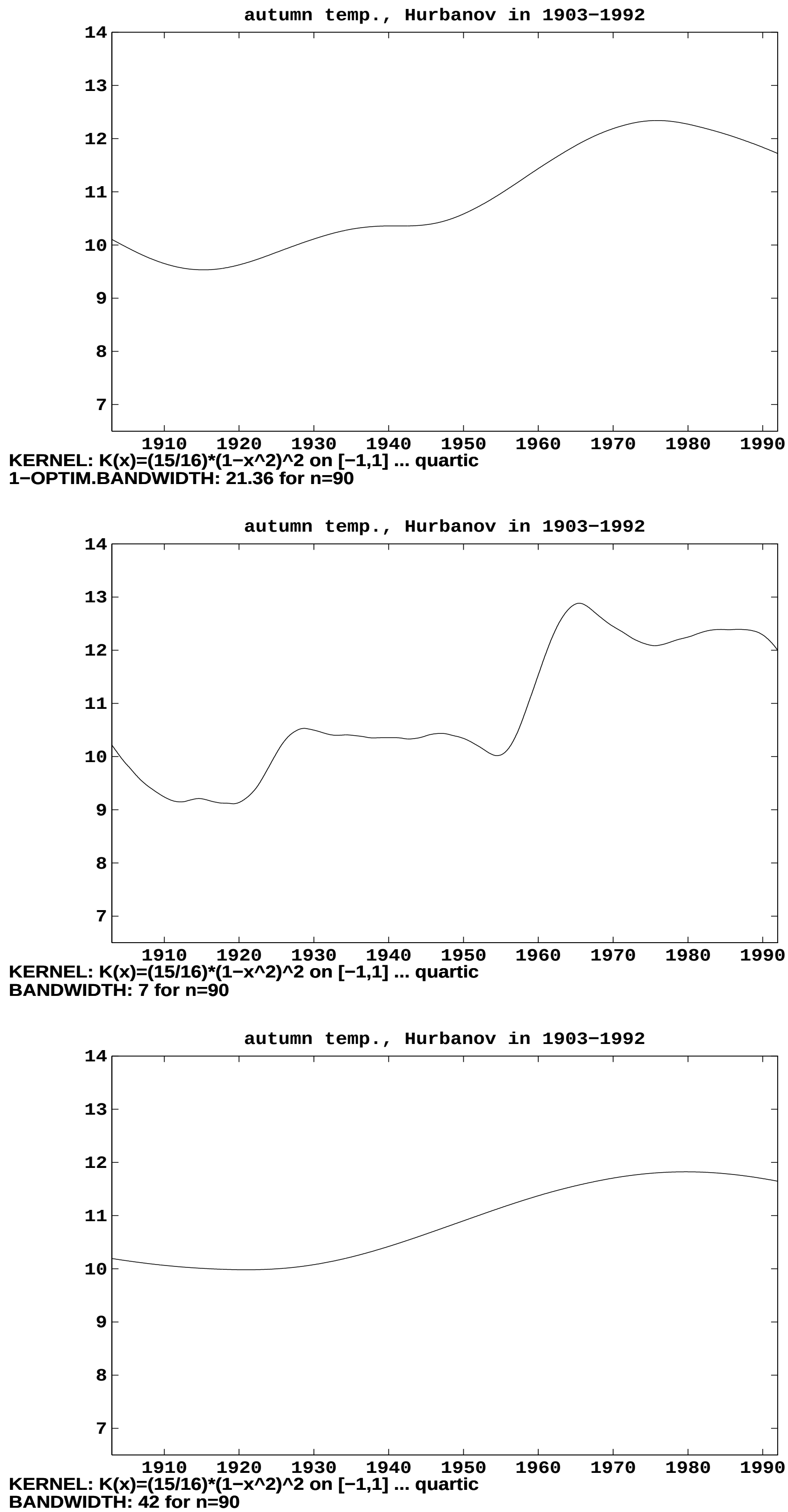

Figure 2: Data set 1 smoothed with optimal, small and large bandwidth (top-down)

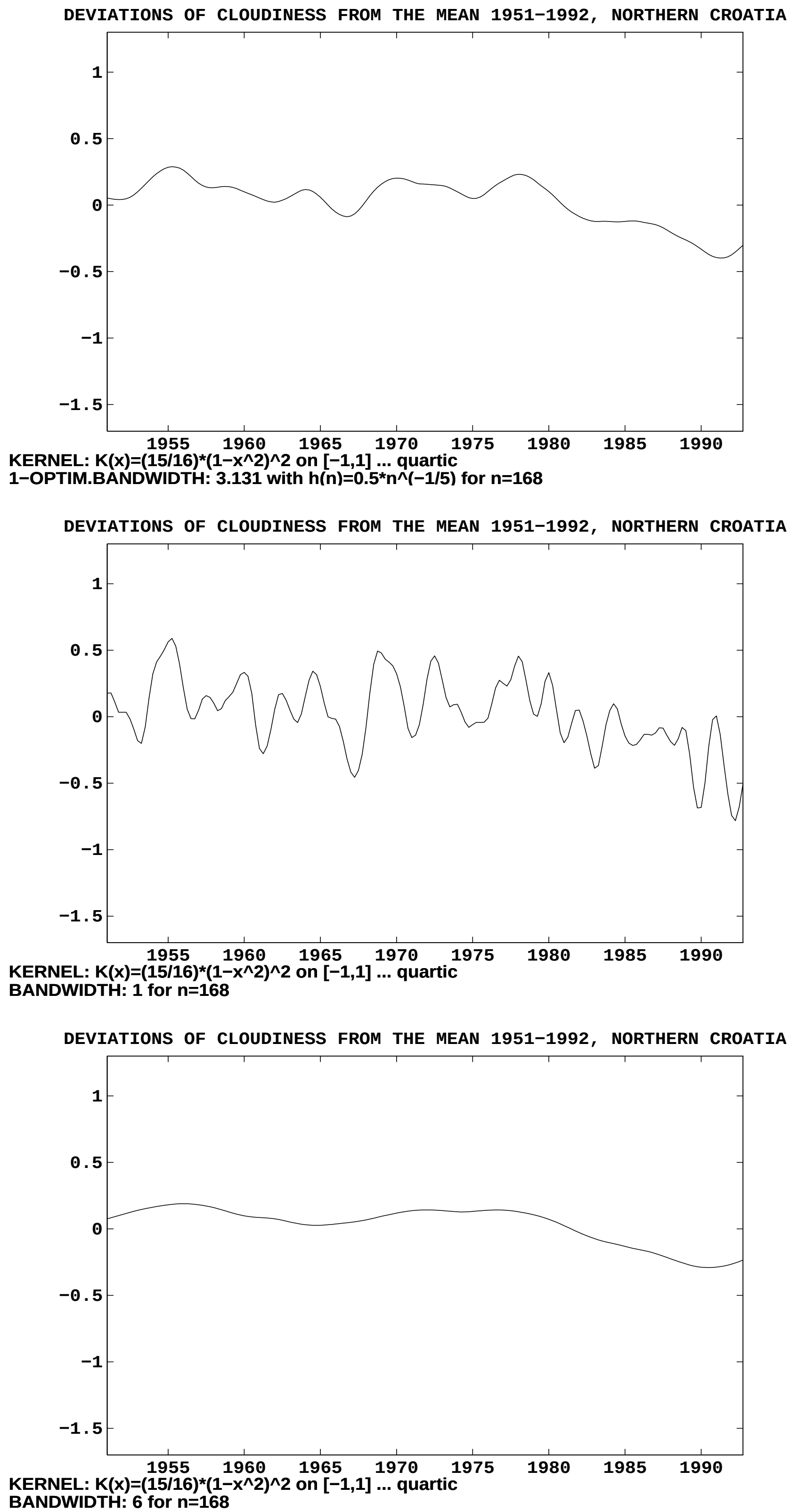

Figure 3: Data set 2 smoothed with optimal, small and large bandwidth (top-down)

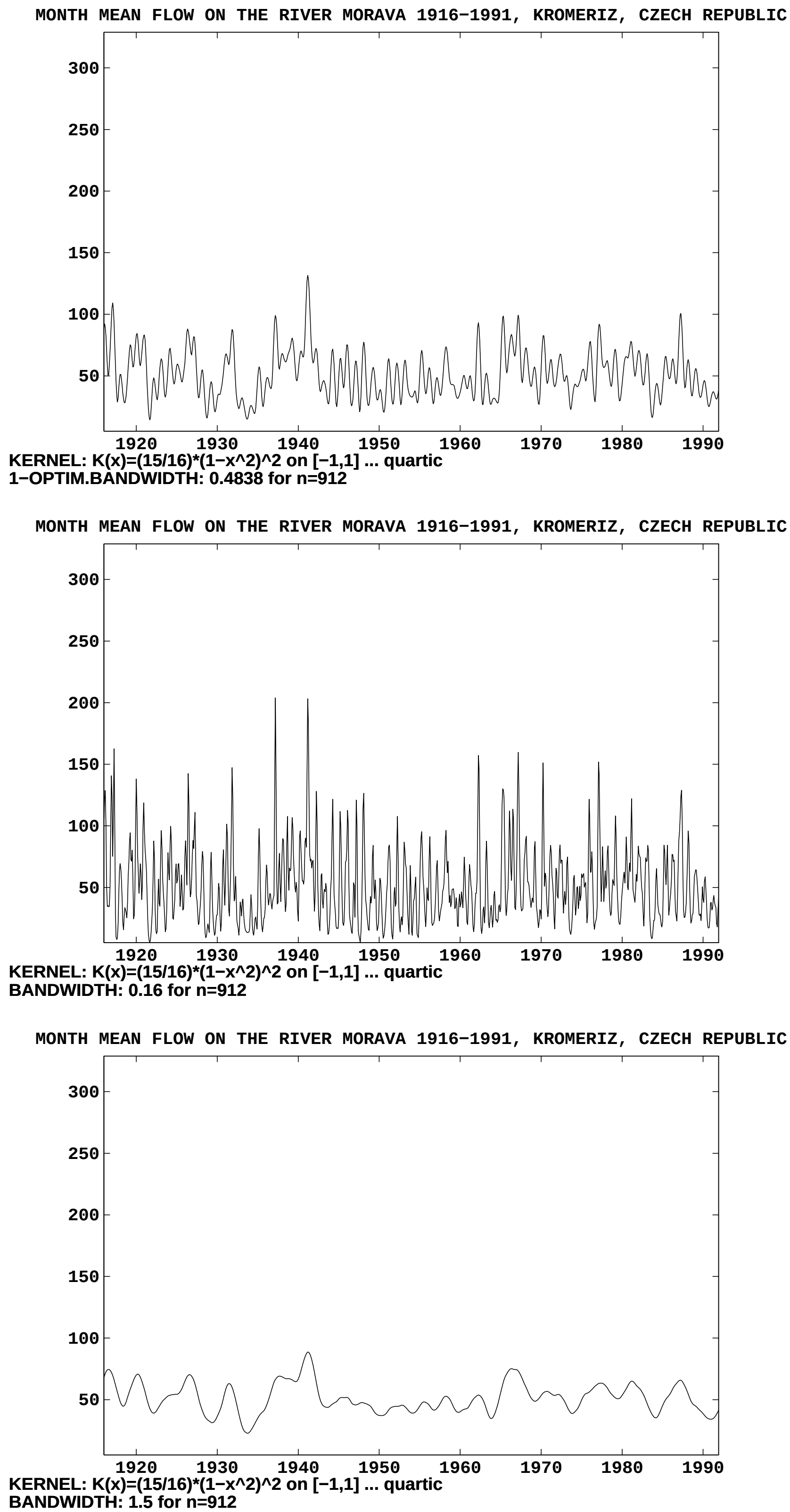

Figure 4: Data set 3 smoothed with optimal, small and large bandwidth (top-down)

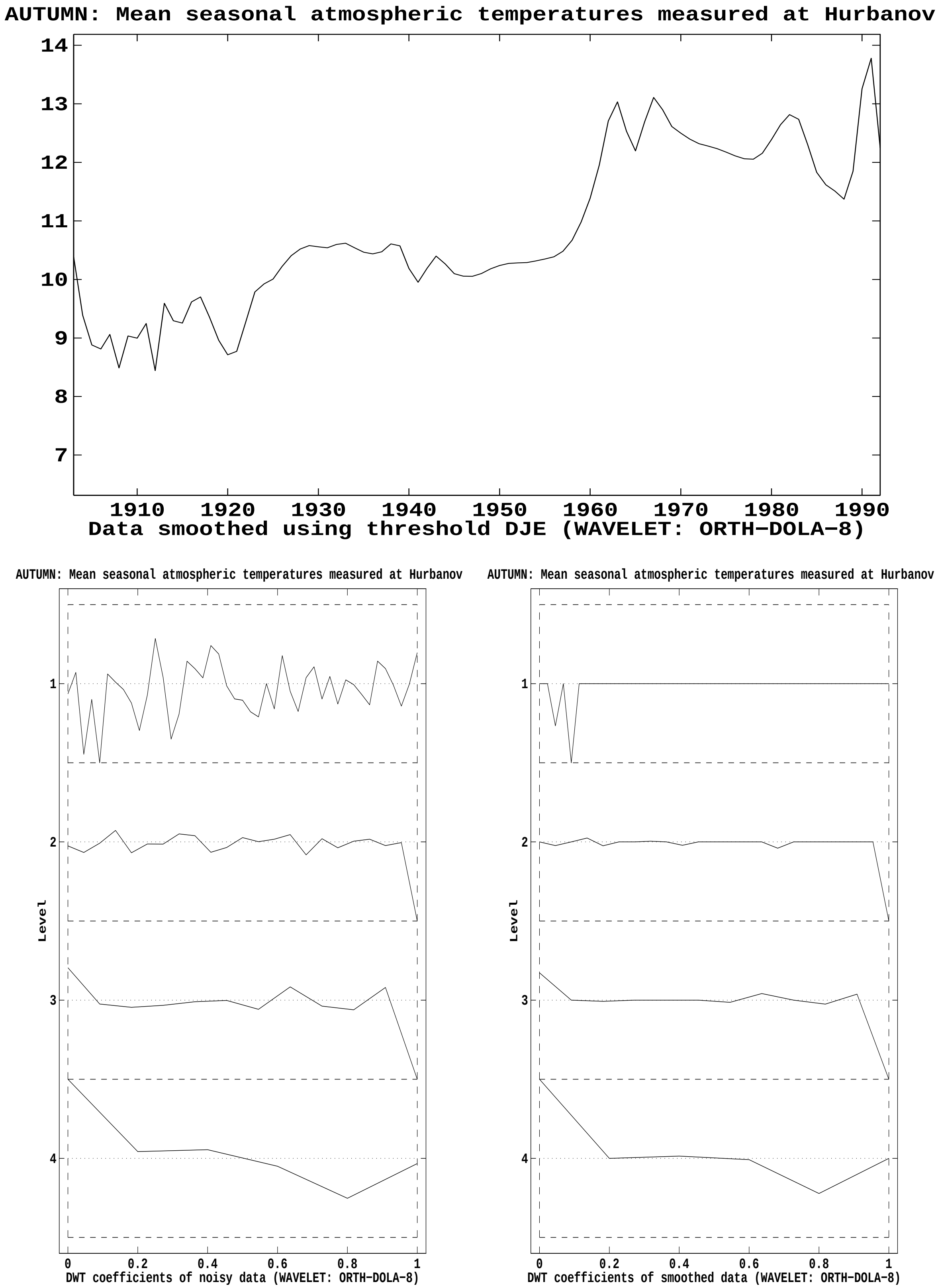

Figure 5: Data set 1 smoothed via wavelet shrinkage of type DJE

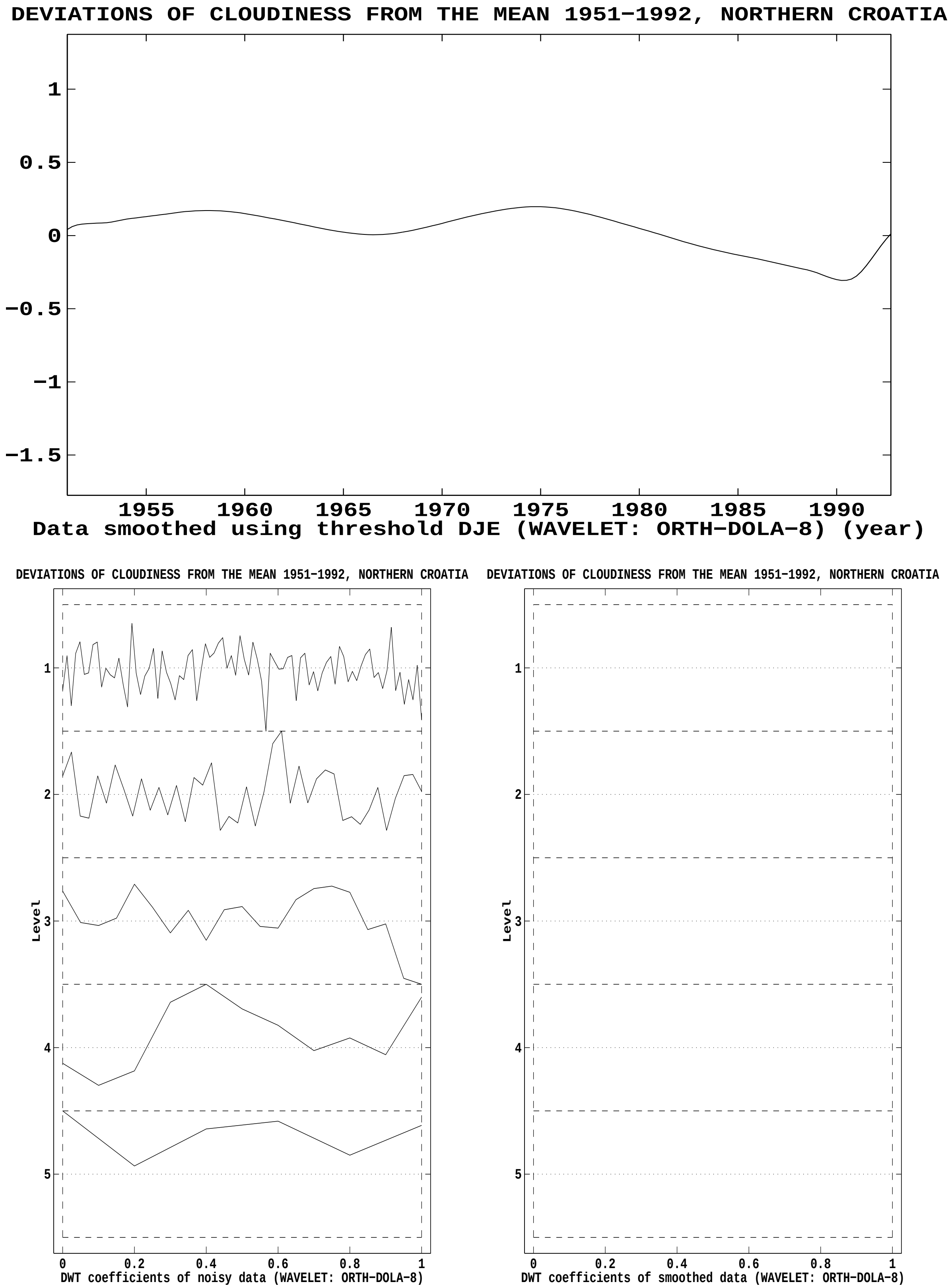

Figure 6: Data set 2 smoothed via wavelet shrinkage of type DJE

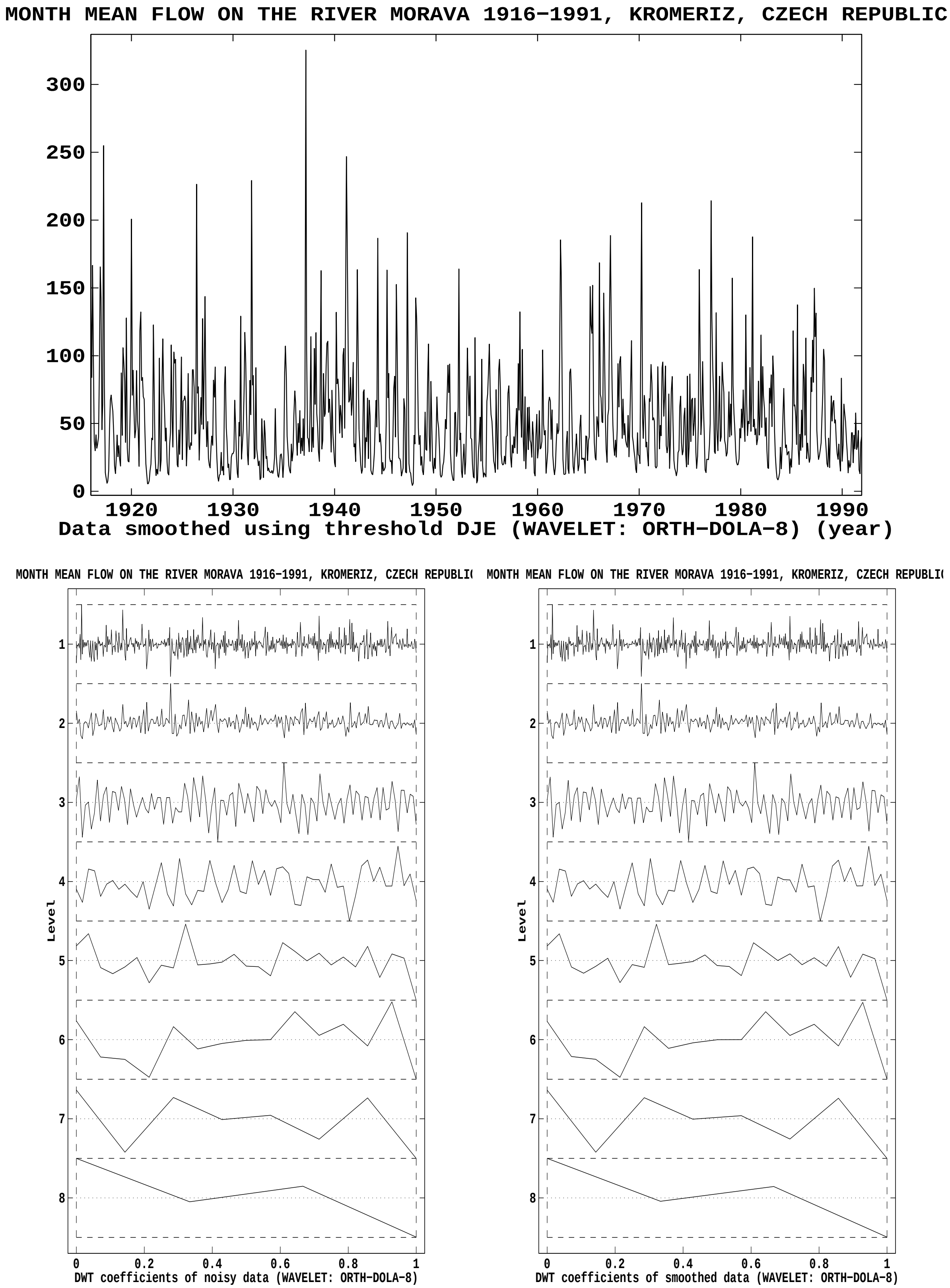

Figure 7: Data set 3 smoothed via wavelet shrinkage of type DJE

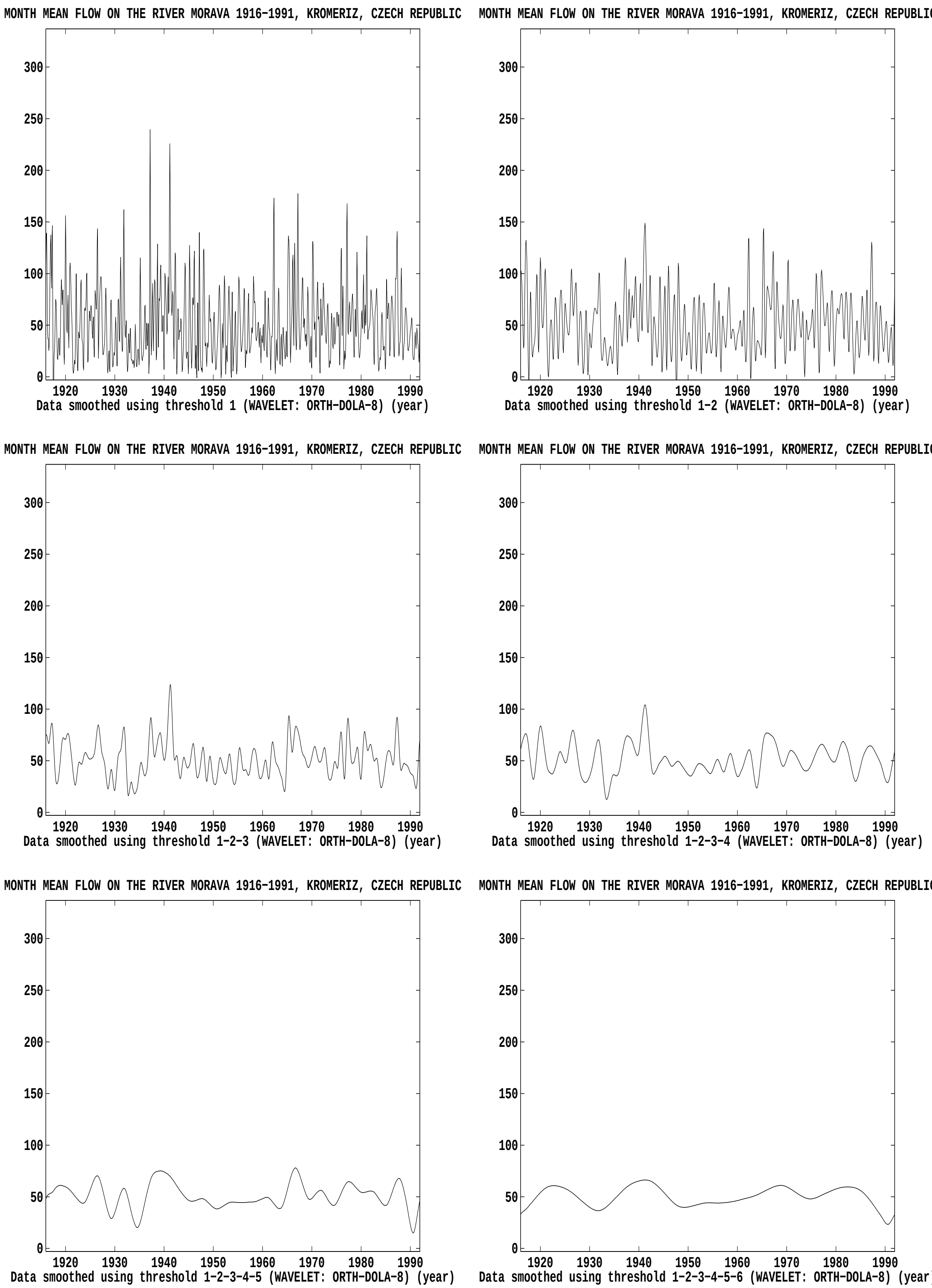

Figure 8: Multiresolution analysis of the data set 3