Hydrol. Earth Syst. Sci., 13, 1921–1936, 2009
www.hydrol-earth-syst-sci.net/13/1921/2009/

Hydrology and Earth System Sciences

Hydrological model performance and parameter estimation in the wavelet-domain

B. Schaefli[1] **and E. Zehe**[2]

[1]Faculty of Civil Engineering and Geosciences, Water Resources Section, Delft University of Technology, The Netherlands
[2]Institute of Water and Environment, Dept. for Hydrology and River Basins Management, Technische Uni. München, Germany

Received: 5 March 2009 – Published in Hydrol. Earth Syst. Sci. Discuss.: 20 March 2009
Revised: 28 September 2009 – Accepted: 28 September 2009 – Published: 19 October 2009

Abstract. This paper proposes a method for rainfall-runoff model calibration and performance analysis in the wavelet-domain by fitting the estimated wavelet-power spectrum (a representation of the time-varying frequency content of a time series) of a simulated discharge series to the one of the corresponding observed time series. As discussed in this paper, calibrating hydrological models so as to reproduce the time-varying frequency content of the observed signal can lead to different results than parameter estimation in the time-domain. Therefore, wavelet-domain parameter estimation has the potential to give new insights into model performance and to reveal model structural deficiencies. We apply the proposed method to synthetic case studies and a real-world discharge modeling case study and discuss how model diagnosis can benefit from an analysis in the wavelet-domain. The results show that for the real-world case study of precipitation – runoff modeling for a high alpine catchment, the calibrated discharge simulation captures the dynamics of the observed time series better than the results obtained through calibration in the time-domain. In addition, the wavelet-domain performance assessment of this case study highlights the frequencies that are not well reproduced by the model, which gives specific indications about how to improve the model structure.

1 Introduction

Most hydrological models have parameters that cannot be related to some measurable catchment characteristics and have to be calibrated. Classically, this calibration determines the best parameter values such as the simulations match as closely as possible one or several observed system outputs (for an overview of calibration methods, see, e.g. Gupta et al., 2005). The uncertainties inherent in the simulations of such calibrated models (e.g. Beven and Freer, 2001; Vrugt et al., 2003; Kavetski et al., 2006a) and the question how to reduce them are subject to intense research. Current strategies include a better description and understanding of the uncertainty inherent in the involved natural processes (e.g. Zehe et al., 2005), in the observation of these processes (e.g. Nicótina et al., 2008) or in the mathematical representation of these processes (e.g. Kavetski et al., 2006b). In parallel, the question how to increase the value of observed data through an improved extraction of its information content receives a constantly growing interest (e.g. Herbst and Casper, 2008; Reusser et al., 2008; Yilmaz et al., 2008).

Model parameter estimation is linked to the question how to measure model performance by suitable objective functions. The majority of parameter estimation methods is based on objective functions defined on the residuals, i.e. the difference between the observed and the simulated time series. Most methods minimize the mean squared error, i.e. the sum of the squared residuals (see, e.g. Gupta et al., 2005). By construction, the resulting calibrated model simulation fits well the individual values of the observed reference time series. Such an approach accounts only indirectly for differences in the autocorrelation of the observed and of the simulated times series. Assuming uncorrelated Gaussian residuals and inferring their variance in a full Bayesian approach (e.g. Kavetski et al., 2006a), for example, implicitly minimizes the difference in autocorrelation between the observed and the simulated series: if one of the time series shows a strong autocorrelation, e.g. at lag-1, the residuals cannot be uncorrelated and the assumptions are violated. At least partial explicit assessment occurs in Bayesian methods that assume correlated Gaussian residuals (e.g. Montanari and Toth, 2007; Schaefli et al., 2006) or in methods using calibration objective functions that minimize temporal slope differences between two time series (e.g. Reusser et al., 2008).

Correspondence to: B. Schaefli
(b.schaefli@tudelft.nl)

As a good simulation should mimic the dynamics underlying an observed time series, it is tempting to think that explicitly assessing how well a model reproduces the autocorrelation properties of an observed system response is a promising choice for model calibration. Keeping, furthermore, in mind that time series of hydrological signatures exhibit periodicity at different time scales, model performance measures that are based on spectral information appear rather appealing. A straight forward choice is of course the power-density spectrum of a process which equals the Fourier transform of its autocorrelation function (Priestley, 1981). This idea is indeed not new; Whittle (1953) proposed a method for parameter estimation in the Fourier-domain matching the theoretical power-density spectrum of the model to the estimated power-density spectrum of the process observations. The Whittle estimator has recently been applied to rainfall-runoff models by Montanari and Toth (2007).

Whittle's Fourier-domain estimator is a consistent approximation of the classical time-domain likelihood. For infinitely long time series, it will, thus, yield the same result as time-domain estimation (see, e.g. Hannan, 1973; Yao and Brockwell, 2006). However, as shown by Contreras-Cristán et al. (2006), it can produce unreliable estimates for non-Gaussian processes or show an important loss of efficiency if the autocorrelation of the process is high.

1.1 Objectives of this paper

The overall idea is to present a new performance measure to assess how closely the time-varying frequency content of a simulated time series matches the time-varying frequency content of the observed series. This objective function is based on a continuous wavelet transform that yields a representation of the time-varying frequency content of an observed time series – as opposed to a Fourier transform, where the moment of occurrence of the different frequencies is not preserved. The wavelet transform is particularly useful for application to natural processes, such as discharge, that integrate various time-varying small scale processes at a larger spatial scale and that, thus, have time-varying autocorrelation properties. This time-variation of the hydrological response of an ecosystem is, of course, partly induced by the time-variation of the relevant input processes, e.g. precipitation and temperature. Furthermore, the rainfall-runoff response is essentially nonlinear, including threshold behavior (Zehe et al., 2007; Blöschl and Zehe, 2005). The input frequencies are thus nonlinearly filtered by the catchment and its biotic (e.g. vegetation) and abiotic characteristics. In glaciated catchments, the real-world case study of this paper, the overall time-variability is particularly pronounced since discharge is induced by a combination of rainfall, ice and snowmelt (see Fig. 1a).

We will give evidence that an objective function that measures explicitly how well the time-varying frequency content of an observed series is reproduced can provide important

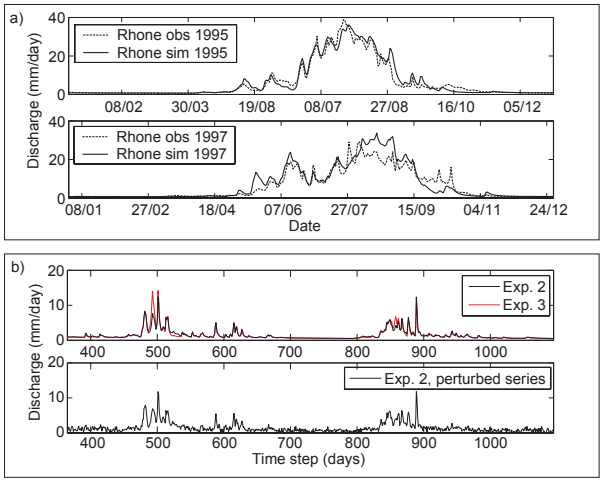

Fig. 1. **(a)** Observed discharge time series of the Rhone River at Gletsch and corresponding (time-domain) calibrated time series (Nash value L_N=1-R_N=0.94), for the years 1995 (top) and 1997 (bottom); **(b)** synthetic discharge time series generated with GSM-SOCONT, unperturbed series for experiments 2 and 3 (top), perturbed series for experiment 2 (bottom).

and new pieces of information to the puzzle of understanding performance and structural deficits of hydrological models. Such an objective function allows, furthermore, estimation of model parameters. We intend to show that such an approach represents a very valuable opportunity compared to parameter estimation in the time domain. As the suggested approach depends crucially on how similarity between estimated wavelet-power spectra is defined, this will be discussed in detail in Sect. 3 after an introduction to continuous wavelet transform in Sect. 2. We then illustrate the advantages and drawbacks of parameter estimation in the wavelet-domain through simple examples and synthetic case studies, i.e. using synthetic data generated either with a simple statistical model or with a conceptual, reservoir-based rainfall-runoff transformation model (Sects. 4 and 5). Finally, we apply the wavelet-domain objective function to parameter estimation of the GSM-SOCONT (Schaefli et al., 2005) model for a highly glacierized catchment in the Swiss Alps. Based on this case study, we discuss the potential of wavelet-domain calibration and performance analysis and show how it can contribute to improve the structure of hydrological models (Sect. 5). The main conclusions and open questions are summarized in Sect. 6.

2 Continuous wavelet spectral analysis

Wavelet analysis, initially formalized by Grossmann and Morlet (1984), is the most recent solution to overcome the main shortcoming of the Fourier transform that identifies the frequencies present in a signal but not their moment

of occurrence. Wavelet analysis, in turn, results in a time-frequency (or time-scale) representation of the signal. Instead of decomposing a signal into constituent harmonic functions as in Fourier analysis, wavelet analysis transforms a signal into scaled and translated versions of an original (mother) wavelet. Compared to a simple windowed Fourier transform, as suggested by Gabor (1946), wavelet transform has the main advantage of adjusting intrinsically the resolution to the analyzed scale (e.g. Daubechies, 1992).

In hydrology, continuous wavelet transform became popular in different types of applications; it is for example used to characterize river regimes and to detect how discharge is related to climate variability indices (e.g. Labat, 2005) or to qualitatively analyze how certain features of the meteorological input time series are transferred to the hydrological system output (e.g. Gaucherel, 2002; Lafrenière and Sharp, 2003; Schaefli et al., 2008). Lane (2006) was the first to use it to investigate rainfall-runoff models, namely to investigate the impact of perturbing single model parameters on the resulting hydrographs.

Even though wavelet spectral analysis has found a wide spread application, few papers present all the mathematical details which we judge to be necessary to understand this paper and to interpret the results. Therefore, the following section might seem rather detailed to the reader with a background in wavelet spectral analysis.

2.1 Continuous wavelet transform

Given a stochastic process $X(t)$, its wavelet transform $W_g[\tau, s|X(t)]$ at time τ and scale s with respect to the chosen wavelet $g(t)$ is

$$W_g[\tau, s|X(t)] = \int \frac{1}{c(s)} g^*\left(\frac{t-\tau}{s}\right) X(t)dt \tag{1}$$

where g^* denotes the complex conjugate of g and $c(s)$ is a normalization constant (see Sect. 2.3). For a detailed discussion of continuous wavelet transform (CWT) and for example the requirements on the wavelet $g(t)$, we refer the reader to the comprehensive literature (e.g. Daubechies, 1992; Holschneider, 1998).

The choice of the wavelet $g(t)$ depends on the type of application. In geosciences applications, the Morlet wavelet is frequently used (for a short discussion of how to choose a wavelet for hydrological applications, see Schaefli et al., 2008):

$$g_m(\theta) = \exp(i\omega_0\theta)\exp\left(\frac{-\theta^2}{2}\right) \tag{2}$$

where $i=\sqrt{-1}$ and $\theta=(t-\tau)/s$. The parameter ω_0 adjusts the time/scale resolution. In the present application, we use ω_0=6, a choice that has empirically been shown to work well for geosciences applications (Labat, 2005; Si and Zeleke, 2005; Torrence and Compo, 1998).

For a Morlet wavelet, the relationship between the scale s and the frequency f reads as (e.g. Holschneider, 1998):

$$\frac{1}{f} = \frac{4\pi s}{\omega_0 + \sqrt{2+\omega_0^2}} \tag{3}$$

Therefore, for ω_0=6, $f \approx 1/s$.

It is important to note that the CWT transforms a time series from one to two dimensions (time and scale). This transformation re-uses the same original information several times and results, therefore, in a considerable amount of redundancies. The inherent correlations of a CWT, given by the reproducing kernel (e.g. Holschneider, 1998), make statistical analysis of wavelet-power spectra a non-trivial task (Maraun et al., 2007; Schaefli et al., 2008). They represent a fundamental difference to estimated Fourier power-density spectra where neighboring frequencies are asymptotically independant.

2.2 Wavelet-power spectrum

Analogue to Fourier analysis, the wavelet-power spectrum (WPS) is defined as the wavelet transform of the autocovariance function, which for a nonstationary process $X(t)$ can be written as (e.g. Shumway and Stoffer, 2006):

$$\begin{aligned}&\mathrm{acv}[\ell, \eta|X(t)] \equiv \\ &\mathrm{E}\left[(X(\eta) - \mathrm{E}[X(\eta)])\,\mathrm{conj}\,(X(\eta+\ell) - \mathrm{E}[X(\eta+\ell)])\right]\end{aligned} \tag{4}$$

where η is the time argument of the autocovariance function and ℓ is the lag from time η. $\mathrm{E}[\cdot]$ is the expected value and "conj" denotes here the complex conjugate (elsewhere denoted by *). For simplicity of notation, let's assume in the following that $X(t)$ is a zero-mean process, i.e. $\mathrm{E}[X(t)]$=0 for all t. The WPS does becomes (e.g. Holschneider, 1998):

$$\begin{aligned}S_g[\tau, s|X(t)] &\equiv W_g\left\{\tau, s\left|\mathrm{E}\left[X(\eta)\mathrm{conj}\,(X(\eta+\ell))\right]\right.\right\} \\ &= \mathrm{E}\left[W_g[\tau, s|X(\eta)]\,W_g^*[\tau, s|X(\eta+\ell)]\right]\end{aligned} \tag{5}$$

This last equation is often written in the following short form:

$$S_g[\tau, s|X(t)] \equiv \mathrm{E}\left[\left|W_g[\tau, s|X(t)]\right|^2\right] \tag{6}$$

The exact WPS of observed or simulated processes is generally unknown; we can estimate it based on the CWT of observed process realizations (observed time series):

$$\hat{S}_g\left[\tau, s|x^{(m)}(t), \hat{\mu}(t)\right] = \left\langle\left|W_g[\tau, s|x^{(m)}(t) - \hat{\mu}(t)]\right|^2\right\rangle \tag{7}$$

where $\hat{S}_g\left[\tau, s|x^{(m)}(t), \hat{\mu}(t)\right]$ is an estimator of $S_g[\tau, s|X(t)]$. $\langle\cdot\rangle$ denotes the averaging operator, $\mathrm{x}^{(m)}(t)$ is a matrix containing m realizations (time series) of the process $X(t)$ and $\hat{\mu}(t)$ is an estimator of the expected value of $X(t)$. In practice, an estimator of the true WPS is often

obtained based on a single realization $x(t)=x^{(1)}(t)$ of length N:

$$\hat{S}_g\left[\tau, s|x(t)\right] = \left|W_g\left[\tau, s|x(t)-\hat{\mu}\right]\right|^2 \tag{8}$$

where the estimator $\hat{\mu}$ is obtained as the average of the realization, i.e. $\hat{\mu}=\frac{1}{N}\sum_{k=1}^{N} x(k)$.

In analogy to Fourier analysis, this estimator $\hat{S}_g\left[\tau, s|x(t)\right]$ is called the wavelet periodogram. It is computed at a finite number of scales between the lowest and the highest resolvable scales that depend on the sampling time step and the number of sampled time steps. The selected scales are usually:

$$s_i = s_0 2^{\frac{i-1}{N_{\text{voice}}}} \tag{9}$$

with i=1,.., $N_{\text{voice}}\times N_{\text{octave}}+1$. The lowest calculated scale is s_0 corresponding to a frequency lower than or equal to the Nyquist frequency (i.e. half the sampling frequency, see, e.g., Priestley, 1981) and the highest scale is $s_0\times 2^{N_{\text{octave}}}$ where N_{octave} denotes the number of octaves (i.e. powers of two), and N_{voice} the number of voices (i.e. calculated scales) per octave.

It is important to note that wavelet analysis has the subtlety that, since neighboring points in time and scale are correlated, the wavelet periodogram looks smooth even if the fluctuations around the true wavelet-power spectrum are not smaller than for a (Fourier) periodogram (see Maraun and Kurths, 2004). Accordingly, as for the Fourier periodogram, the wavelet periodogram has to be smoothed to obtain a consistent estimator of the true wavelet-power spectrum (see, e.g. Maraun and Kurths, 2004). For the present application that uses the difference between the wavelet periodograms of two time series for model calibration, the consistency of the estimator is, however, of no relevance.

2.3 Normalization of the wavelet transform

The choice of the normalization constant in Eq. (1) is not unambiguous. It can in principle be chosen arbitrarily (Kaiser, 1994, p. 62) and just as in Fourier analysis, different conventions are in use.

To compute the wavelet-power-based performance criteria, we use the L^2-norm preserving normalization $c(s)=\sqrt{s}$, which ensures that (Kaiser, 1994, p. 63)

$$\left\|\frac{1}{c(s)}g\left(\frac{t-\tau}{s}\right)\right\|^2 = \int_{-\infty}^{\infty}\left|\frac{1}{\sqrt{s}}g\left(\frac{u-\tau}{s}\right)\right|^2 du$$
$$= \int_{-\infty}^{\infty}|g(v)|^2\,dv = \|g(t)\|^2 = cst \tag{10}$$

3 Wavelet periodogram – based performance assessment

3.1 Visual inspection

Wavelet periodograms (i.e. estimated wavelet-power spectra) have the potential to efficiently distinguish between time series that seem to be similar in the time-domain but that have a (locally) different frequency content and, thus, locally different autocorrelation properties. We illustrate this potential based on the wavelet periodograms of the simulated and of the observed daily discharge series from the Rhone River at Gletsch. Further details on these time series are given in Sect. 4. The simulated series appears to be very similar to the observed series (Fig. 1a), with a linear correlation of 0.97 or, using the classical hydrological performance criterion proposed by Nash and Sutcliffe (1970), a Nash value of 0.94 (see Sect. 4).

The visual interpretation of the 2-D wavelet periodograms (Fig. 2a and b) is difficult and error prone (Maraun and Kurths, 2004; Maraun et al., 2007). Therefore, Fig. 2c and d show the so-called wavelet bands that correspond to the scale-average wavelet-power over given ranges of scales, where the scale-average power is defined as the scale-weighted sum of the wavelet-power (Torrence and Compo, 1998). The bands are normalized by the variance of the time series. As Fig. 2c and d in conjunction with Fig. 1a illustrate, such a plot reveals differences that are not readily seen in the original data. We see namely that for the band of scales 64 days to 128 days, the calibrated model does not correctly reproduce the dynamics. These differences would also be visible in a detailed inspection of the time series, by comparing the weekly, monthly and seasonal statistics or by subtracting the series from each other and analyzing the obtained "residuals". However, inspecting wavelet bands provides several views of the signal at the same time and has the main advantage of yielding a rapid overview over the differences.

3.2 Wavelet performance measure

We propose a method for the estimation of model parameters in the wavelet-domain. It is based on the following hypotheses: (i) the dynamics of two processes are similar if their time-varying autocorrelation properties are similar; (ii) these autocorrelation properties can be estimated based on the wavelet periodograms of process realizations. For model calibration, this translates into the assumption that the more similar the wavelet periodograms of a simulated and an observed time series are, the better the model mimics the behavior of the natural system.

Quantifying the similarity between the wavelet periodograms requires the definition of a distance metric that measures how different the periodograms are at a give time step. The overall distance of the two periodograms can then be expressed as the mean distance over all time steps. The

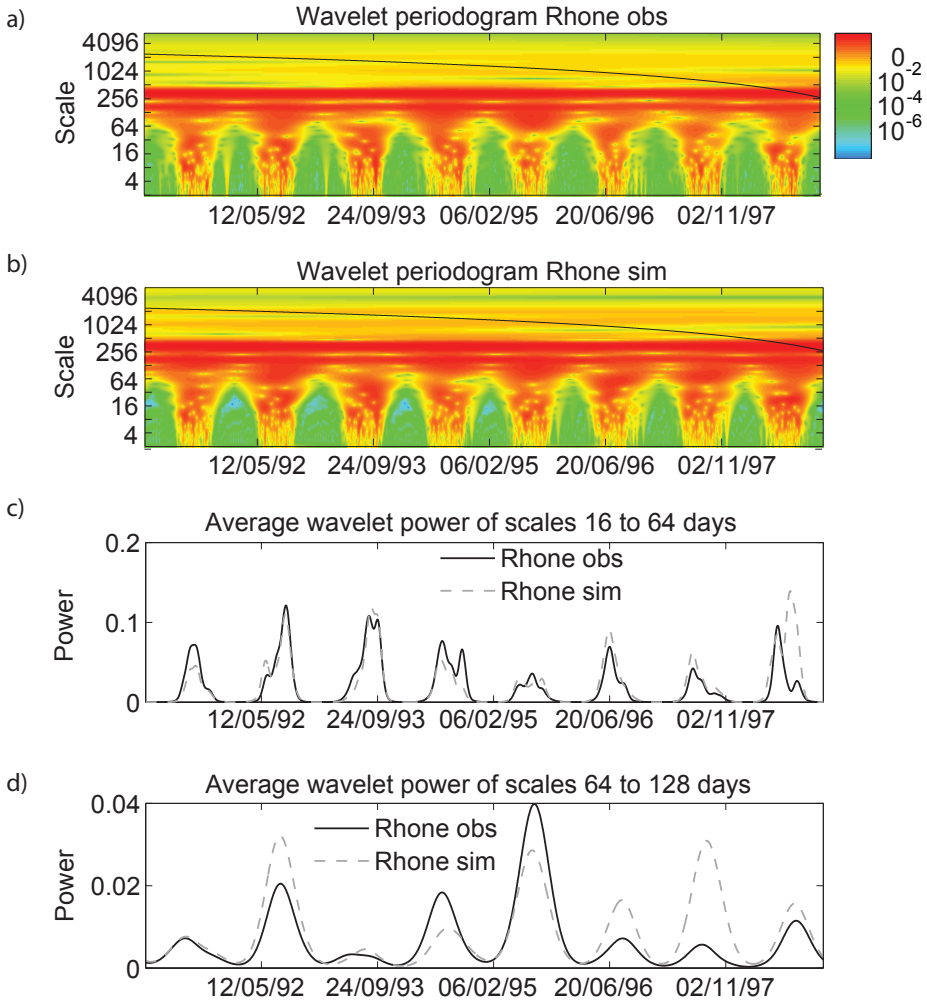

Fig. 2. Estimated wavelet-power for the Rhone case study: **(a)** wavelet periodogram of the observed discharge (zoom on the 2nd half of the available time series, black line: cone of influence), **(b)** wavelet periodogram of the simulated discharge, **(c)** average estimated wavelet-power of both series between the scales of 16 days and 64 days, **(d)** same as (c) but between the scales of 64 days and 128 days.

choice of this distance metric has to take into account the fact that in the wavelet periodogram neighboring scales and neighboring time steps are correlated.

We use a metric similar to the Kolmogorov-Smirnov distance, which is classically used to measure the distance of the probability distributions of two samples and which equals the maximum distance between the cumulative distribution functions. This metric is particularly useful to measure whether at a given time step t, the power of the observed and of the simulated wavelet periodogram is similarly distributed over all scales: it is sensitive to the shape of the power distribution over the scales but compared to a squared error-based metric, it is much less sensitive to slight shifts in peaks and to the chosen normalization constant in Eq. (1).

The cumulative wavelet-periodogram $\hat{C}_g[\tau, s|x(t)]$ is computed as:

$$\hat{C}_g[\tau, s|x(t)] = \sum_{k=s_0}^{s} \hat{S}_g[\tau, k|x(t)] \tag{11}$$

where $s=s_0, \ldots, s_{\max}(\tau)$. $s_{\max}$ is the maximum scale analyzed at each time step. This maximum scale varies from time step to time step because of the edge effects. In CWT, the area of the wavelet periodogram that is influenced by edge effects is called "cone of influence". In the present paper, we exclude edge effects by fixing a cone of influence that equals the e-folding time of the wavelet, which is defined as the time at which the wavelet-power drops to $1/e^2$ and which is a measure of the wavelet width at a given scale. For a Morlet wavelet, this equals $\sqrt{2}s$ (for a discussion, see Torrence and Compo, 1998).

The Kolmogorov-Smirnov distance $D_g\,[\tau|x(t),\,y(t)]$ at time step τ between two process realizations $x(t)$ and $y(t)$ becomes:

$$D_g\,[\tau|x(t),\,y(t)] = \max \left| \frac{\hat{C}_g\,[\tau,\,s|x(t)]}{\hat{C}_g\,[\tau,\,s=s_{\max}|x(t)]} - \frac{\hat{C}_g\,[\tau,\,s|y(t)]}{\hat{C}_g\,[\tau,\,s=s_{\max}|y(t)]} \right| \tag{12}$$

where $\frac{\hat{C}_g[\tau,s|x(t)]}{\hat{C}_g[\tau,s=s_{\max}|x(t)]}=\frac{\sum_{k=s_0}^{s}\hat{S}_g[\tau,k|x(t)]}{\sum_{k=s_0}^{s_{\max}}\hat{S}_g[\tau,k|x(t)]}$ is the normalized cumulative wavelet periodogram of the process realization $x(t)$ at time step τ.

A good simulation should have a wavelet periodogram that fits the periodogram of the observed series at all time steps. Accordingly, the overall wavelet periodogram efficiency criterion, R_W, averages $D_g\,[\tau|x(t),\,y(t)]$ over all time steps. For an observed time series $y(t)$ and the corresponding simulated series $x(t|\varphi)$ this becomes

$$R_W\,[\boldsymbol{\varphi}|x(t|\boldsymbol{\varphi}),\,y(t)] = \frac{1}{N}\sum_{\tau=1}^{N} D_g\,[\tau|x(t|\boldsymbol{\varphi}),\,y(t)] \tag{13}$$

where N is the total number of time steps and $\boldsymbol{\varphi}$ a vector containing all model parameters. R_W takes values between 0 and 1 and has to be minimized during calibration.

In order to be applicable to parameter estimation, a distance metric has to fulfill formal requirements (e.g. Weisstein, 2008). For some general distance metric $D(A,B)$ between A and B it has to hold that (i) $D(A,\,B)\geq 0,\ \forall A,\,B$, ii) $D(A,\,B)=D(B,\,A)$ (iii) $D(A,\,B)=0$ if and only if $A=B$ and (iv) $D(A,\,C)\leq D(A,\,B)+D(B,\,C)$. For the wavelet-based distance measure $D_g\,[\tau\,|x(t),\,y(t)]$, it follows directly from its definition that conditions (i) and (ii) hold. Condition (iv), the so-called triangle inequality, holds since the maximum distance between two monotonically increasing functions between 0 and 1 can never be bigger than the sum of the maximum distances between these two functions and a third function. Since $R_W\,[\cdot]$ results from a simple average of $D_g\,[\tau\,|x(t),\,y(t)]$ over all τ, conditions (i), (ii) and (iv) also hold for $R_W\,[\cdot]$. Condition (iii) does not necessarily hold for $D_g\,[\tau\,|x(t),\,y(t)]$: two process realizations $x(t)$ and $y(t)$ could, in theory, have locally exactly the same distribution of wavelet-power without having $x(t)=y(t),\ \forall t$. However, it holds that $R_W\,[\cdot]=0$ if and only if $D_g\,[\tau\,|x(t),\,y(t)]=0\ \forall\tau$ which implies $x(t)=y(t)$. We conclude that $R_W\,[\cdot]$ satisfies the formal conditions of a distance metric.

R_W measures whether the wavelet-power content at every time-step is distributed similarly in a simulated series and a reference series, i.e. it measures differences in the autocorrelation properties at a given time step. Accordingly, it does not explicitly measure differences in the mean or in the variance of two time series.

As in every parameter estimation procedure, preserving the mean, or in physical terms the mass balance, is a very important criterion to accept or reject simulations and the underlying model. Traditional time-domain calibration ensures preservation of the mean either through the assumptions on the residual distribution (e.g. zero-mean Gaussian residuals, Kavetski et al., 2006b) or through explicit exclusion of parameter sets leading to a too high bias between the observation and the simulation (see, e.g., Montanari and Toth, 2007). We retain this last solution by deteriorating the wavelet performance criterion R_W of a given simulation if its bias exceeds a certain tolerance factor. The exact value of this tolerance factor has to be fixed empirically. For perfect model situations where the true (and hence unbiased) simulation exists, the tolerance factor does not affect the best identified parameter set but restrains the search space. For real-world applications, this restriction of the search space might influence the best identified parameter set since it excludes not mass conservative, i.e. physically meaningless parameter sets.

In the present study, we use a tolerance factor of 10% for all case studies. For the real-world application, this choice is in line with the semi-automatic calibration method suggested by Schaefli et al. (2005). In general, the value of the tolerance factor should reflect the available information about the observational uncertainties of the different terms of the water balance. The exact penalization procedure based on this tolerance factor is discussed in Sect. 4.3.3.

For general stochastic processes with stationary mean, variance and autocorrelation properties, these are a priori unrelated properties and a good process model should preserve them. Discharge processes, on which we focus in the present paper, have a time-variable mean, variance and autocorrelation (see an example in Fig. 1). Since discharge results from a time-variable combination of different hydrological processes (infiltration, snowmelt etc.), these statistical properties are strongly related. For such processes, as our empirical results show, preserving the mean and the time-varying autocorrelation properties ensures the preservation of the process variance.

We would like to add here that in the statistics literature, wavelet-based estimators have been proposed in the 1990s to estimate long-memory parameters (see Velasco, 1999, p. 107) but their statistical properties are analyzed only recently (e.g. Moulines et al., 2008). As the corresponding estimation problems are very different from the scope of the present paper, we do not discuss them here.

4 Case studies

4.1 Synthetic case studies

We designed a number of synthetic case studies to highlight and discuss the potential of wavelet-domain calibration and performance analysis. These case studies correspond to

model calibration experiments where the reference discharge series are generated with known parameter values. Three different sets of synthetic discharge series are used. For experiment 1, we use the realization of an ARMAX process. For experiments 2 and 3, we use a realization of the hydrological model GSM-SOCONT, which is also used in the real-world case study.

Experiment 2 uses the classical model structure, whereas experiment 3 is based on a slightly modified model version including a time-varying parameter. Details about all synthetic experiments are given hereafter. Additional illustrative toy examples as well as a synthetic case study with the well-known HYMOD model (e.g. Schaefli and Gupta, 2007) can be found in (Schaefli and Zehe, 2009).

4.1.1 Input time series

The synthetic experiments have been designed to illustrate the differences between parameter estimation in the time-domain and in the wavelet spectral domain. We therefore use as external forcing a nonstationary precipitation series which is obtained by joining two precipitation series that have different statistical properties. To have a realistic situation, these two individual series are surrogate series generated based on the precipitation series observed at the station Bourg St. Pierre between 1903 and 1999, located in the Southern Swiss Alps (1620 m a.s.l., 7.21° E, 45.95° N), which is also used also for the real-world case study. The precipitation in this area is known to have undergone a substantial modification over the last century (Frei and Schär, 2001; Schmidli and Frei, 2005). The generation of the nonstationary rainfall series involves the following steps: (i) Generate a 250 days surrogate series based on the first 20 years of observed precipitation. The surrogate series is generated using the so-called Iterative Amplitude Adjusted Fourier Transform (IAAFT) algorithm (Schreiber and Schmitz, 2000). This is a classical method to obtain surrogates by first taking the Fourier transform of a time series, replacing the phases by randomly drawn phases and then completing the inverse Fourier transform. (ii) Generate a 250 days surrogate series based on the last 20 years of observed precipitation. (iii) Contract both series.

For experiments 2 and 3, we generate a longer series, 10 years, and use the first half for calibration and the second half for validation.

4.1.2 Output time series

For experiment 1, the used ARMAX process is:

$$y(t) = a \cdot y(t-1) + b_1 z(t-n_k) + b_2 z(t-1-n_k) + b_3 z(t-2-n_k) \quad (14)$$

where t is the time step, $z(t)$ is the input variable (in our case precipitation), n_k is the delay parameter that is set to 4 and φ=[a, b_1, b_2, b_3] are the parameters to be inferred. The reference exact series is generated using the following parameters: φ=[a, b_1, b_2, b_3]=[−0.85,0.080, 0.018, 0.029]. The resulting series is perturbed with uncorrelated Gaussian noise having zero mean and standard deviation 0.4, corresponding to 25% of the standard deviation of the generated $y(t)$.

Experiments 2 and 3 are based on a reference discharge series simulated with the hydrological model GSM-SOCONT (Schaefli et al., 2005) (see also Sect. 4.) using the same precipitation series as in experiment 1. We assume that there is no glacier cover and use a temperature time series corresponding to a low land station (the station called Grono, 380 m a.s.l., 9.15° E, 46.25° N). This makes the discharge time series explicitly distinct from the real-world case study (see Sect. 4.2); in particular there is a less strong annual cycle of the discharge (see Fig. 1b).

For experiment 3, we generate a reference series with GSM-SOCONT having a time-variable snowmelt parameter (see Table 4) and then calibrate the model with a constant snowmelt parameter on this reference series. This experiment illustrates a typical example of model misspecification.

For both experiments 2 and 3, the synthetic realizations are perturbed by adding white noise before the parameter calibration process (see results section for details).

4.2 Real-world case study

For the real-world case study, we use the GSM-SOCONT (Schaefli et al., 2005) model, which is a conceptual precipitation-runoff transformation model for high mountainous catchments having an ice-melt component. The discharge is simulated separately for the glacier part and the non-glacier part and within each part separately for 5 elevation bands. We apply it to a gauging station of the Rhone River located in Gletsch, in the Southern Swiss Alps (8.36° E, 46.56° N). This catchment is highly glacierized (around 50% of the surface covered by glaciers) and has a mean altitude of around 2700 m. For a more detailed description and the used meteorological input time series, see (Schaefli et al., 2005). We use the period 1981 to 1990 for calibration and 1991 to 1999 for validation. The meteorological conditions during these two periods where quite different. During the first period, there was in particular quite extensive snowfall (during this period the number of increasing glaciers in the Swiss Alps was much higher than during the 1990s, e.g. Herren et al., 2002). As a result, the hydrological regime of these two periods is quite distinct. The first period has its maximum mean monthly discharge in July, the second period in August.

4.3 Parameter estimation

4.3.1 Reference performance criteria

For comparison purposes, we use the classical squared error-based Nash-Sutcliffe efficiency measure (Nash and Sutcliffe, 1970), called hereafter Nash value:

$$L_N\left[\boldsymbol{\varphi}|x(t|\boldsymbol{\varphi}),\, y(t)\right] = 1 - \frac{\sum_{t=1}^{N}\left[x(t|\boldsymbol{\varphi}) - y(t)\right]^2}{\sum_{t=1}^{N}\left[y(t) - \mathrm{E}\left[y(t)\right]\right]^2} \tag{15}$$

where $y(t)$ is the observed discharge at time step t, $x(t|\boldsymbol{\varphi})$ is the simulated discharge given parameter set $\boldsymbol{\varphi}$ and N the number of observed and simulated time steps.

We define a Nash-based performance measure to be minimized as follows

$$R_N\left[\boldsymbol{\varphi}|x(t|\boldsymbol{\varphi}),\, y(t)\right] = 1 - L_N\left[\boldsymbol{\varphi}|x(t|\boldsymbol{\varphi}),\, y(t)\right] \tag{16}$$

For the synthetic case studies, where the (exact) best model parameter set exists, we also use a Fourier-domain performance measure based on the Whittle likelihood, computed according to Montanari and Toth (2007) as:

$$L_F\left[\boldsymbol{\varphi}|J_x(\lambda|\boldsymbol{\varphi}),\, J_y(\lambda)\right] = \exp\left[-\sum_{j=1}^{N/2}\left\{\log\left[J_x(\lambda_j|\boldsymbol{\varphi}) + f_e(\lambda_j|\boldsymbol{\varphi})\right] + \frac{J_y(\lambda_j)}{J_x(\lambda_j|\boldsymbol{\varphi}) + f_e(\lambda_j|\boldsymbol{\varphi})}\right\}\right] \tag{17}$$

where $\lambda_j = 2\pi j/N$ are the Fourier frequencies. J_x is the periodogram of the simulated discharge time series and J_y the periodogram of the observed discharge time series. f_e is the Fourier-power spectrum of the modeling error (for details, refer to Montanari and Toth, 2007). We define the performance measure R_F as

$$R_F\left[\boldsymbol{\varphi}|J_x(\lambda|\boldsymbol{\varphi}),\, J_y(\lambda)\right] = -\log\left(L_F\left[\boldsymbol{\varphi}|J_x(\lambda|\varphi),\, J_y(\lambda)\right]\right) \tag{18}$$

which has to be minimized.

4.3.2 Search algorithm

We use a global optimization algorithm for model calibration. The range of possible parameter values is fixed based on a priori information. The used optimizer is a multi-objective evolutionary algorithm called Queueing Multi-Objective Optimiser (QMOO) developed by Leyland (2002) for energy system design. For an application of this optimizer to hydrology, see (Schaefli et al., 2004) and (Schaefli, 2005).

The algorithm has been designed to identify difficult-to-find optima and to solve far more complex problems than the ones presented here, involving much more decision variables (parameter to identify) (see Leyland, 2002). We, therefore, assume that all identified parameter sets correspond to the best identifiable solutions of the optimization problem. The stopping criterion for the search algorithm is fixed as follows: we assume that that the algorithm has converged to the optimum solution if the objective function value of the best found solution does not vary more than 5‰ between two successively identified best solutions.

4.3.3 Penalization

As discussed in Sect. 3.2, we penalize solutions (parameter sets) that lead to a large bias between the observed and the simulated time series. The penalization is completed based on

$$R'_k = \begin{cases} R_k & \text{if} \quad B < 0.1 \\ R_k + B & \text{if} \quad 0.4 > B > 0.1 \\ R_k + 10 \cdot B & \text{if} \quad B \geq 0.4 \end{cases} \tag{19}$$

where R_k, k={W,F,N} is the objective function value (to be minimized) and B is the relative bias between the observed and the simulated time series computed as

$$B\left[\boldsymbol{\varphi}|x(t|\boldsymbol{\varphi}),\, y(t)\right] = \frac{1}{N}\sum_{k=1}^{N}\left[\frac{|x(k,\boldsymbol{\varphi}) - y(k)|}{y(k)}\right] \tag{20}$$

This penalization has been chosen because for all used performance criteria, B and R_k have the same order of magnitude for good solutions. The penalization should not be too strong for low biases because this would hinder the optimization algorithm to explore the parameter space.

5 Results

5.1 Synthetic case studies

5.1.1 Experiment 1

The parameter ranges used as search space for model calibration as well as the identified best parameter sets under each performance criterion are given in Table 1.

For the perturbed reference series for which the results are reported here, none of the performance criteria leads to an exact recovery of the ARMAX parameters. For the specific realization of white noise, there is a parameter set that fits the signal better in a least square sense (Table 1). As expected, for this theoretic Gaussian case with uncorrelated error, the solution in the time-domain and in the Fourier-domain is equivalent. The best parameter set under R_W is different, b_2 even has a wrong sign. In fact, adding Gaussian white noise adds power to all scales (recall that the Fourier power-density spectrum of Gaussian white noise is constant and equals its variance (see, e.g. Priestley, 1981). This induces, thus, an offset between the wavelet-power spectrum of the perturbed reference series and the exact series. As a result, for the model of Eq. (14), there is a parameter set with a closer match to the wavelet-power spectrum of the perturbed reference series. This effect becomes even more

Table 1. Exact parameter values of the ARMAX process, intervals delimiting the search space for parameter estimation and the identified best parameter sets under R_W, R_F and R_N (columns 5–7). For each parameter set, the values of the performance criteria are given (instead of R_N, the more familiar $L_N = 1 - R_N$ is given). The criteria values listed under "exact" are calculated between the unperturbed ("unpert") original series and the perturbed ("pert") series; other abbreviations: corr: linear correlation; min; best possible criterion value; max: worst possible criterion value; inf: no absolute reference value; in bold: the best performance of each row.

Parameter	Exact	Min	Max	R_W	R_F	R_N
a	−0.850	−0.999	−0.001	−0.847	−0.860	−0.851
b_1	0.080	−2.000	2.000	0.134	0.089	0.084
b_2	0.018	−2.000	2.000	−0.040	0.012	0.013
b_3	0.029	−2.000	2.000	0.032	0.017	0.028
corr pert	0.98	−1	1	0.96	0.98	**0.98**
R_W pert	0.15	1	0	**0.12**	0.13	0.14
R_F pert	−1.69	+ inf	−inf	−1.58	**−1.70**	−1.70
L_N pert	0.95	−inf	1	0.92	0.95	**0.95**
corr unpert	1	−1	1	0.98	1.00	**1.00**
R_W unpert	0	1	0	0.01	0.02	**0.01**
R_F unpert	NaN	+inf	−inf	−2.23	−3.08	**−3.18**
L_N unpert	1	−inf	1	0.96	1.00	**1.00**

important if we apply a stronger error (results not shown). In the unperturbed case, R_W enables an exact recovery of the true parameter set.

The convergence criterion was reached for all ARMAX experiments between 3500 and 4000 model evaluations. There is no significant difference between the different performance criteria. Another interesting question is whether one criterion needs a longer time series to converge efficiently. For all criteria, the convergence is slowed down if the length of the calibration time series is reduced; for R_W this slowdown is more important because the data length limits the number of resolvable scales. For this case study, below 50 data points, the convergence time becomes prohibitive (more than 10 000 evaluations).

5.1.2 Experiment 2

The parameter set used for the generation of the synthetic reference discharge series set is given in Table 3 and a zoom on the time series is shown in Fig. 1b. This exact series is perturbed with a Gaussian white noise having zero mean and a standard deviation of 0.44 (corresponding to 25% of the standard deviation of the exact series). The parameter ranges used as search space for calibration are given in Table 2 and the identified best parameter sets under each performance criterion are given in Table 3.

For the case where the perfect model exists but the series is perturbed, R_W as well as the other criteria recover the exact value of the most sensitive parameter, the degree-day factor for snowmelt (for details about parameter sensitivity, see Schaefli et al., 2005). For the 3 least sensitive parameter values, the identified values are less close to the real values than for a calibration under R_N or R_F. The performance difference of the identified best simulations under all three calibration criteria is, however, hardly detectable. There is, nonetheless, an interesting difference: the optimum parameter values under the two frequency-domain criteria are clearly much better defined than under R_N (Fig. 3a and b). This holds in particular for the least sensitive parameter, the nonlinear direct runoff parameter β. It is noteworthy, however, that this does not indicate a better identifiability in the wavelet-domain in general but depends on the chosen formulation of the time-domain objective function (for a discussion of the shape of time-domain objective functions, refer to Beven and Freer, 2001).

5.1.3 Experiment 3

If we try to estimate the model parameters on a reference series that was generated with a different model structure, i.e. with a variable degree-day factor for snowmelt, the performances under R_N and R_W are very close but each of the criteria lead to distinct solutions for the constant degree-day factor (Table 4), both of which lead to good simulations. The two solutions are hardly distinguishable based on the used performance measures (Table 4) and are very close to the generated reference series (see Fig. 4). A look on the average wavelet-power over certain ranges of scales, however, clearly shows that the simulations having a constant degree-day factor do not reproduce the true dynamics (Fig. 4), neither for the best parameter set in the time-domain nor for the best parameter set in the wavelet-domain.

5.2 Real-world case study

5.2.1 Parameter estimation in the wavelet-domain versus in the time-domain

There is certain trade-off between the time-domain and the wavelet-domain objective function. The optimum for R_N does not correspond to the optimum for R_W (Table 5). It is noteworthy that for this case study, the apparently high Nash values (0.94 for the best simulation under R_N, 0.91 under R_W) do not necessarily mean that the hydrological model does a particularly good job, high Nash values are easy to achieve for times series with a strong annual cycle (see Schaefli and Gupta, 2007).

At a first glance, the optimal parameter sets do not seem to be fundamentally different under R_N and under R_W (see Table 5). A closer look shows however some notable differences. Even if physically meaningless parameter sets are penalized, the optimization under R_N leads to a global optimal solution where the degree-day factor for ice is smaller than for snow. This is physically highly questionable (e.g., Hock, 2003; Schaefli et al., 2005). The global optimal solution in the wavelet-domain respects this physical constraint.

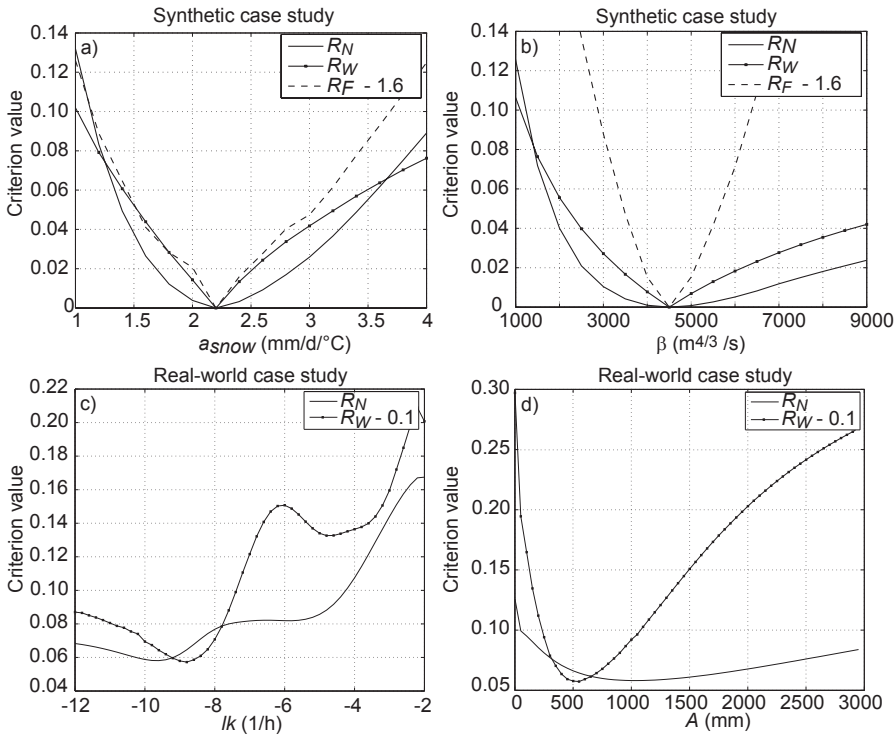

Fig. 3. Parameter sensitivity around the optimum value identified under the different calibration criteria; top: experiment 2, the most sensitive parameter a_{snow} **(a)**, and the least sensitive parameter β **(b)**, the other parameters are kept constant to the values of Table 3; bottom: real-world case study, the two least sensitive parameters lk **(c)** and A **(d)**, the other parameters are kept constant to the values of Table 5.

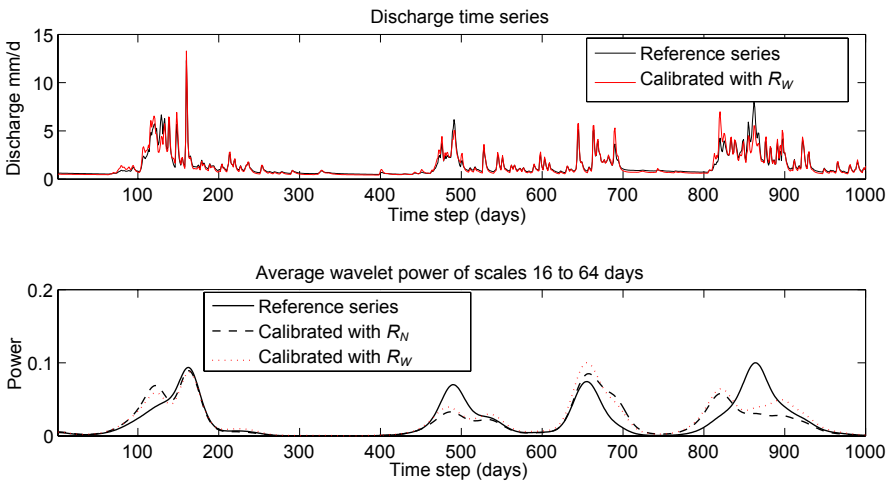

Fig. 4. Experiment 3, top: reference and calibrated discharge under R_W (calibrated discharge under R_N is almost identical and, thus, not plotted); bottom: estimated mean wavelet-power between the scales of 16 days and 64 days for the reference discharge and for the simulated discharge calibrated under R_W, respectively under R_N.

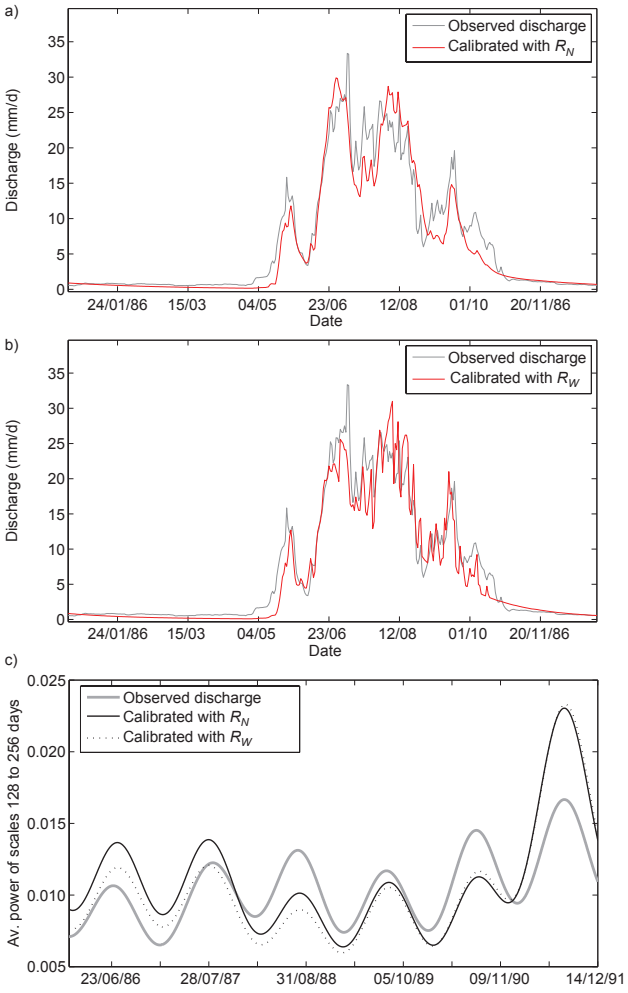

Fig. 5. Real-world case study: **(a)** zoom on the observed discharge for the year 1986 (calibration period) and the simulation calibrated under R_N; **(b)** as (a) but simulation calibrated under R_W; **(c)** zoom on the average estimated wavelet-power between the scales of 128 days and 256 days for the observed discharge and the simulations calibrated under R_W, respectively under R_N.

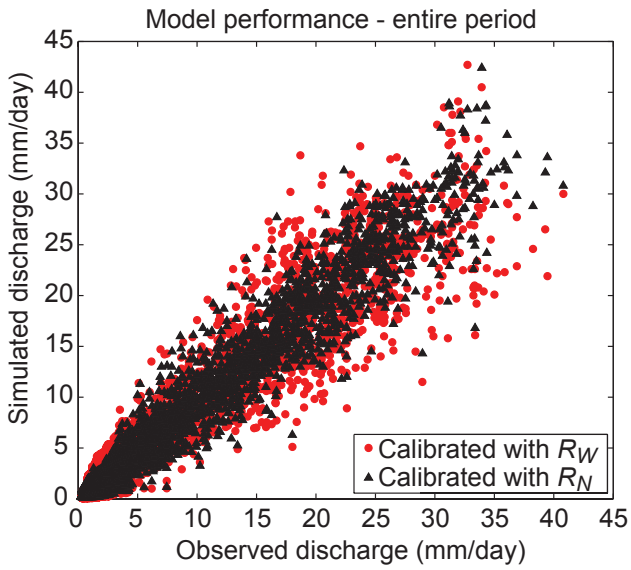

Fig. 6. Real-world case study: scatter-plot of observed discharge (during calibration period) against best simulation under each of the two calibration criteria.

Table 2. Parameter intervals delimiting the search space for the real-world case study and for the synthetic experiments 2 and 3; parameter sets that do not respect the imposed physical conditions are penalized during parameter set evaluation (for more details about these parameters, see Schaefli et al., 2005).

Parameter	Unit	Min	Max	Significance	Condition/penalty
a_{ice}	mm/d/°C	1.0	16.0	Degree factors for ice resp. snow	$a_{ice}>a_{snow}$ penalty = $a_{snow}-a_{ice}$
a_{snow}	mm/d/°C	0.5	12.0		
k_{ice}	d	0.5	45.0	Linear reservoir coefficient for ice resp. snow melt	$k_{ice}<k_{snow}$ penalty=$(k_{ice}-k_{snow})/2$
k_{snow}	d	1.0	45.0		
A	mm	1	3000	Max. storage for linear slow response reservoir	
lk	log(1/h)	−12.0	−0.1	Coeff. for linear slow response	
β	$m^{4/3}/s$	1	30 000	Coeff. for nonlinear, fast response	
T_{crit}	°C	1.0	1.0	Threshold for snowfall	Fixed

In addition, the parameters have (as for the synthetic case study) a better identified optimum under R_W than under R_N, especially for the parameters with the lowest sensitivity, the soil transfer parameters (Fig. 3c and d).

Close inspection of the discharge simulations based on the best parameters obtained in the time-domain and in the wavelet-domain, respectively, shows that both parameter sets yield rather different discharge dynamics (see zooms on the simulations in Fig. 5 and scatter-plots of observed against simulated discharge in Fig. 6). Without further cross-validation data (e.g. observed ice melt data), it is difficult to judge which parameter set captures the observed dynamics better. An interesting hint is, however, given by the following analysis: we build a prediction interval based on 90 of the 100 best random simulations. Under R_N, this interval includes around 80% of all observations for the calibration period. This success rate decreases constantly if we evaluate it for discharges above a certain threshold (Fig. 7). For R_W, the success rate, which is overall slightly lower than under R_N, remains constant for all discharge thresholds. This clearly suggests that for this case study, good simulations in the wavelet-domain capture equally well all discharge ranges. Good simulations under R_N, however, capture particularly well small discharges, i.e. the long periods of easy to predict low flows, which, due to their temporal dominance, tend to have a strong influence on the time-domain objective function for this case study.

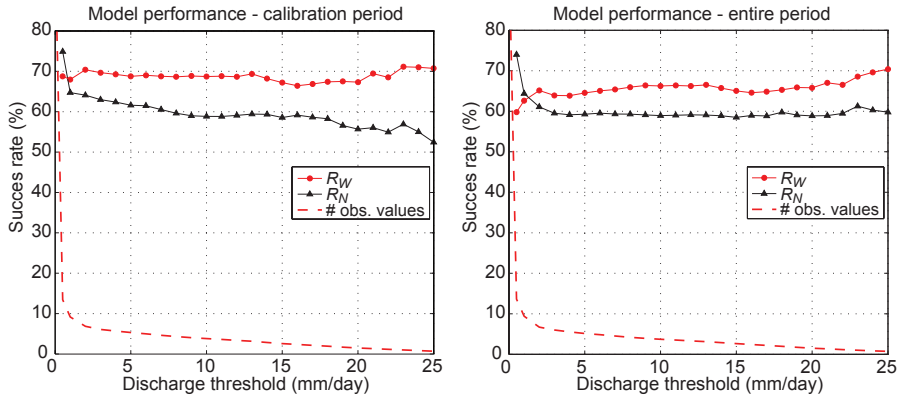

Fig. 7. Real-world case study: model performance for the best 100 random simulations (of 20 000 random parameter sets) under R_W, respectively R_N; the success rate measures the relative number of observed daily discharges above a certain threshold that fall within the 90% prediction range of the retained 100 simulations; left calibration period, right entire period.

Table 3. Experiment 2: parameter values used to generate the synthetic discharge time series and the identified optimal parameters sets under R_W, R_F and R_N; the glacier surface is set to 0, which eliminates the parameters a_{ice}, k_{ice} and k_{snow}; calib: calibration period; valid: validation period; for other abbreviations, see Table 1.

		Calibration R_W		Calibration R_N		Calibration R_F	
Parameter/criterion	Exact	Calib	Valid	Calib	Valid	Calib	Valid
a_{snow}	2.2	2.2		2.2		2.2	
A	550	691		544		557	
$\log(k)$	−9.1	−9.6		−9.1		−9.1	
β	4500	4748		4546		4478	
corr pert	0.97	0.97	0.96	0.97	0.97	0.97	0.97
R_W pert	0.15	0.14	0.15	0.15	0.15	0.15	0.15
R_F pert	−1.01	−0.84	−0.37	−0.85	−0.38	−0.85	−0.38
L_N pert	0.94	0.93	0.92	0.94	0.93	0.94	0.93
corr unpert	1	1.00	1.00	1.00	1.00	1.00	1.00
R_W unpert	0	0.04	0.05	0.00	0.00	0.00	0.00
R_F unpert	NaN	−1.44	−0.95	−1.51	−0.98	−1.51	−0.98
L_N unpert	1	0.99	0.99	1.00	1.00	1.00	1.00

Table 4. Experiment 3: parameters used to generate the synthetic reference discharge time series and the identified optimal parameters values using the R_W and R_N performance criteria; the reference time series has been generated using a variable a_{snow} parameter throughout the year; the a_{snow} parameter values for each month are {1.2, 1.4, 1.4, 2.0, 4.0, 5.0, 6.0, 7.0, 5.0, 2.0, 1.2 , 1.1}; for abbreviations see Table 1.

Parameter	Exact	R_W	R_N
a_{snow}	variable	2.2	2.4
A	550	584	438
lk	−9.1	−9.5	−8.7
β	4500	4180	4731
k_{snow}	15.6	29.1	24.8
corr	1	0.94	0.95
R_W	0	0.08	0.09
L_N	1	0.88	0.90

The above results suggest an important difference between the best parameter sets in the time-domain and the best parameter sets in the wavelet-domain. A look on the parameter space of the most sensitive parameters, the degree-day factors for snowmelt (a_{snow}) and for ice melt (a_{ice}) illustrates this difference: Fig. 8 shows a scatter-plot of a_{snow} against a_{ice} for all "physically feasible" parameter sets, i.e. parameter sets that lead to a bias smaller than 10% (the visible dependance between physically feasible snow and ice melt factors is a common result for this type of models, see Schaefli et al., 2005). The 100 parameter sets that, among all physically feasible parameter sets, have the lowest R_W values are highlighted in red; the parameter sets with the lowest R_N values are highlighted as triangles. These two groups show that the best parameter sets under R_N correspond to another area of the physically feasible parameter space than the best parameter sets under R_W. Retaining good solutions under the quadratic error-based criteria R_N further reduces the physically feasible parameter space. In contrast, the group of the best parameter sets under R_W appears to show the same dependance between a_{snow} and a_{ice} as the overall group of physically feasible parameter sets. This result suggests that:

1. The bias criterion could be sufficient to ensure solutions that reproduce the dominant frequencies (which is a very interesting result for snow- and ice melt modeling).

2. The R_N could be too restrictive and exclude solutions that can reproduce the frequency content of the observed time series. This means that for this particular case study, a parameter estimation based on the R_N could be misleading.

This last hypothesis is supported by the fact that R_N only leads to unbiased parameter estimates if the model residuals are Gaussian with a constant variance (for a discussion, see Kavetski et al., 2006a). For the present case study, the residuals clearly have a different variance during the summer and the winter season (see Schaefli et al., 2006).

5.2.2 Model diagnostics

A visual inspection of the average wavelet-power at bands of scales ranging from weeks to a few months shows that even the best models do not well capture the observed dynamics (see an example in Fig. 5c): both best models (under R_W and R_N) do not well reproduce the frequency content of the observed series. The models have a somewhat different behavior but the plot of the average power at high scales suggests that given the observed input and the current model structure, the model cannot produce a closer fit to the estimated wavelet-power of the observed series. The power content tends to be either largely over- or underestimated.

The model's inability to reproduce the frequency content at high scales (months) suggests that either a storage term is missing or is not well parameterized in the current model structure. In fact, the model does not simulate separately the melt and the transfer of firn, i.e. of old snow that lasts more than one season and that is a transition state between ice and snow. Firn has a degree-day factor and melt water transfer times that are between the ones of ice and of snow (e.g. Klok et al., 2001) and induces a further partitioning of the discharge during the melt season. Its contribution to the total discharge depends on the annual snowfall.

The mismatch of the calibrated and observed wavelet periodograms at intermediate scales (weeks) could be a hint that the model does not fully capture the relationship between temperature and snow- and ice melt. This relationship is constant in the current model parameterization but it is known to be variable throughout the melt season (for example due to changes in the albedo, see, e.g. Rango and Martinec, 1995). But the models' inability to reproduce these scales could also indicate the need to improve the meltwater transfer through the glacier and the overlaying snowpack. In the model, this transfer is encoded by a linear transfer function having a constant parameter. In reality, the transfer of melt water through a glacier is highly variable in time since the glacier drainage system evolves throughout the melt season (see, e.g. Willis, 2005). In warm years, it develops and evacuates water much faster than during cold years. A detailed analysis of the wavelet-power at different bands of the observed and the simulated time series during years with high snowfall and low snowfall, respectively during cold and warm years could help gaining further insights into which of the above model structural deficiencies are more important.

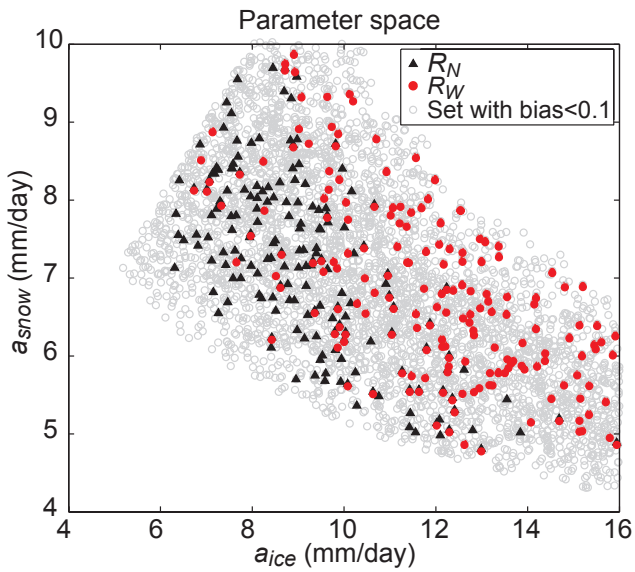

Fig. 8. Real-world case study: relationship between a_{snow} and a_{ice} for the 100 best parameters sets (of 20 000 random parameter sets) under R_N, respectively under R_W.

Table 5. Real-world case study: optimal parameter values identified using global optimization and R_W and R_N as objective-functions; calib=calibration period, valid=validation period, var. bias=relative difference between variance of observed and simulated time series.

	R_W	R_N
a_{ice}	8.5	7.8
a_{snow}	6.8	8.2
lk	−8.8	−9.7
A	539	1014
β	6279	274
k_{ice}	1.5	2.1
k_{snow}	37.0	26.0
R_W calib	0.086	0.088
R_W valid	0.087	0.090
L_N calib	0.91	0.94
L_N valid	0.91	0.93
B	−0.06	−0.05
Var. bias	−0.09	−0.04

The next step would now be to improve the model structure. Monitoring the model performance in the wavelet-domain will help to verify that a model modification really acts on the dynamics at the assumed ranges of temporal scales or whether an achieved performance increase is just "pure chance", due for example to compensations between structural errors. This step requires a considerable reformulation of the model and is left for future research.

5.3 Computational costs

The computation of the wavelet-domain performance criterion involves first of all the computation of the wavelet periodogram of the analyzed time series, which requires convolving the signal (observed or simulated times series) with the wavelet at each scale. This implies a number of inverse Fast Fourier Transforms that equals the number of analyzed scales. This "pre-treatment" of the time series before the computation of the performance criterion increases the computational cost by a factor at least equal to the number of scales. The Kolmogorov -Smirnov distance-based performance criterion also involves more calculations than the computation of a squared error-based distance measure. In our case, using a Matlab® code on a laptop with a Intel®Pentium® M 1.5 GHz processor, the computation of the inverse Fast Fourier Transform for one scale is roughly twice longer than the computation of a Nash criterion over the entire time series. For a time series with 6939 time steps, computing the Nash efficiency takes typically 0.01 s whereas computing the wavelet periodogram for 122 scales takes 1.9 s and computing the Kolmogorov-Smirnov distance takes another 0.6 s.

6 Conclusions and outlook

The present paper discusses and illustrates parameter estimation and model performance analysis of rainfall-runoff models in the wavelet-domain with the main purpose to show how this could contribute to hydrological model diagnostics and to model structure improvement.

As discussed based on theoretical considerations and based on the presented examples, parameter estimation for at least partly misspecified models in the wavelet-domain can yield different results than parameter estimation in the time-domain. Especially for observed time series having a strongly time-varying frequency content, the suggested approach allows estimation of model parameter sets in the wavelet-domain that are internally consistent and allow simulations with more plausible dynamics than a parameter estimation in the time-domain. However, it is at the current stage difficult to determine a priori in which cases a calibration in the wavelet-domain could yield better representations of the true system dynamics. Future case studies and theoretical developments should provide insights into this question. A key hereby will be the detailed study of the behavior of the wavelet-domain performance criterion in presence of errors in the input or output data.

In general, a detailed investigation of the origin of the differences between the best solutions in the wavelet-domain and in the time-domain can offer additional and new pieces to the puzzle of understanding conceptual model behavior and shortcomings. For the real-world case study presented in this paper, the best parameter sets in the wavelet-domain do for example not show the same dependence structure as the best parameter sets in the time-domain. Such a result is a hint that the model has structural deficiencies. These deficiencies can then be further investigated by analyzing in detail how the model performs over relevant ranges of temporal scales, by visually inspecting the power content of the wavelet periodograms or by computing wavelet performance measures over certain scales instead of over the entire range of resolvable scales. As illustrated for the real-world case study, this can give valuable indications on model deficiencies and how to overcome them.

Just as different objective functions can be formulated in the time-domain, the presented wavelet-based criterion corresponds to one possible performance measure in the wavelet-domain. Other formulations (and other wavelets) are possible and would potentially yield other optimal parameter sets. While the statistical properties of different time-domain objective functions are well understood, applications of wavelet spectral analysis to geosciences are still relatively recent and statistical questions have to be further evaluated. We would thus like to emphasize that the potential of parameter estimation in the wavelet-domain lies in the information that it yields for model improvement.

For very long time series, the computational cost for the evaluation of the wavelet criterion can become important. This aspect is however counterbalanced by the gained insights. We are confident that future case studies including namely not only discharge data but also other sources of validation data will provide additional evidence for the potential of parameter estimation and model diagnostics in the wavelet-domain.

A Matlab® code for the computation of the presented performance measure can be obtained from the first author.

Acknowledgements. The research of the first author was funded by Fellowships of the Swiss National Science Foundation and of the German National Science Foundation. Our Matlab code is based on a code originally written by C. Torrence and G. Compo that is available at http://paos.colorado.edu/research/wavelets/. The code for the IAFFT algorithm has been written by V. Venema and is available at http://www.meteo.uni-bonn.de/mitarbeiter/venema. The meteorological time series have been provided by the Swiss Meteorological Institute MeteoSwiss and the hydrological time series by the Hydrological Section of the Swiss Federal Office for the Environment. We also would like to thank the anonymous reviewers for their detailed suggestions to improve the paper structure.

Edited by: E. Todini

References

Beven, K. J. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, J. Hydrol., 249, 11–29, 2001.

Blöschl, G. and Zehe, E.: Invited commentary – On hydrological predictability, Hydrol. Process., 19, 3923–3929, 2005.

Boyle, D. P.: Multicriteria calibration of hydrological models, Department of Hydrology and Water Resources, University of Arizona, Tucson, 193 pp., 2000.

Contreras-Cristán, A., Gutiérrez-Peña, E., and Walker, S. G.: A Note on Whittle's Likelihood, Commun. Stat.-Simul. C., 35, 857–875, 2006.

Daubechies, I.: Ten Lectures on Wavelets, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 357 pp., 1992.

Frei, C. and Schär, C.: Detection probability of trends in rare events: Theory and application to heavy precipitation in the Alpine region, J. Climate, 14, 1568–1584, 2001.

Gabor, D.: Theory of communication, J. IEE, 93, 429–457, 1946.

Gaucherel, C.: Use of wavelet transform for temporal characterisation of remote watersheds, J. Hydrol., 269, 101–121, 2002.

Grossmann, A. and Morlet, J.: Decomposition of Hardy functions into square integrable wavelet constant shape, SIAM J. Math. Anal., 15, 723–736, 1984.

Gupta, H. V., Beven, K. J., and Wagener, T.: Model Calibration and Uncertainty Estimation, in: Encyclopedia of Hydrological Sciences, edited by: Anderson, M. G., Wiley, Chichester, UK, 2015–2032, 2005.

Hannan, E. J.: The asymptotic theory of linear time-series models, J. Appl. Prob., 10, 130–145, 1973.

Herbst, M. and Casper, M. C.: Towards model evaluation and identification using Self-Organizing Maps, Hydrol. Earth Syst. Sci., 12, 657–667, 2008, http://www.hydrol-earth-syst-sci.net/12/657/2008/.

Herren, E. R., Bauder, A., Hoelzle, M., and Maisch, M.: The Swiss Glaciers 1999/2000 and 2001/2002, Glaciological Commission of the Swiss Academy of Sciences, Zürich, Glaciological Report 121/122, 73, 2002.

Hock, R.: Temperature index melt modelling in mountain areas, J. Hydrol., 282, 104–115, 2003.

Holschneider, M.: Wavelets: an analysis tool, Oxford University Press, Oxford, UK, 423 pp., 1998.

Kaiser, G.: A friendly Guide to Wavelets, Birkhäuser, New York, USA, 300 pp., 1994.

Kavetski, D., Kuczera, G., and Franks, S. W.: Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, Water Resour. Res., 42, W03407, doi:10.1029/2005WR004368, 2006a.

Kavetski, D., Kuczera, G., and Franks, S. W.: Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts, J. Hydrol., 320, 173–186, 2006b.

Klok, E. J., Jasper, K., Roelofsma, K. P., Gurtz, J., and Badoux, A.: Distributed hydrological modelling of a heavily glaciated Alpine river basin, Hydrol. Sci. J., 46, 553–570, 2001.

Labat, D.: Recent advances in wavelet analyses: Part 1. A review of concepts, J. Hydrol., 314, 275–288, 2005.

Lafrenière, M. and Sharp, M.: Wavelet analysis of inter-annual variability in the runoff regimes of glacial and nival stream catchments, Bow Lake, Alberta, Hydrol. Process., 17, 1093–1118, 2003.

Lane, S. N.: Assessment of rainfall-runoff models based upon wavelet analysis, Hydrol. Processes, 21, 586–607, 2006.

Leyland, G. B.: Multi-Objective Optimisation Applied to Industrial Energy Problems, Laboratoire d'Energétique Industrielle, Ecole Polytechnique Fédérale de Lausanne, Switzerland, available at: http://library.epfl.ch/theses, 188 pp., 2002.

Maraun, D. and Kurths, J.: Cross wavelet analysis: significance testing and pitfalls, Nonlin. Processes Geophys., 11, 505–514, 2004, http://www.nonlin-processes-geophys.net/11/505/2004/.

Maraun, D., Kurths, J., and Holschneider, M.: Non-stationary Gaussian Processes in Wavelet Domain: Definitions, Estimation and Significance Testing, Phys. Rev. E, 75, 016707, doi:10.1103/PhysRevE.75.016707, 2007.

Montanari, A. and Toth, E.: Calibration of hydrological models in the spectral domain: an opportunity for ungauged basins?, Water Resour. Res., 43, W05434, doi:10.1029/2006WR005184, 2007.

Moulines, E., Roueff, F., and Taqqu, M. S.: A wavelet Whittle estimator of the memory parameter of a non-stationary Gaussian time series, Ann. Stat., 36, 1925–1956, 2008.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models. Part I, a discussion of principles, J. Hydrol., 10, 282–290, 1970.

Nicótina, L., Alessi Celegon, E., Rinaldo, A., and Marani, M.: On the impact of rainfall patterns on the hydrologic response, Water Resour. Res., 44, W12401, doi:10.1029/2007WR006654, 2008.

Priestley, M.: Spectral Analysis and Time Series, Academic Press, London, UK, 884 pp., 1981.

Rango, A. and Martinec, J.: Revisiting the degree-day method for snowmelt computations, Water Resour. Bull., 31, 657–669, 1995.

Reusser, D. E., Blume, T., Schaefli, B., and Zehe, E.: Analysing the temporal dynamics of model performance for hydrological models, Hydrol. Earth Syst. Sci. Discuss., 5, 3169–3211, 2008, http://www.hydrol-earth-syst-sci-discuss.net/5/3169/2008/.

Schaefli, B., Hingray, B., and Musy, A.: Improved calibration of hydrological models: use of a multi-objective evolutionary algorithm for parameter and model structure uncertainty estimation, Hydrology: Science and Practice for the 21st Century, London, UK, 362–371, 2004.

Schaefli, B.: Quantification of modelling uncertainties in climate change impact studies on water resources: Application to a glacier-fed hydropower production system in the Swiss Alps, Ecole Polytechnique Fédérale de Lausanne, Switzerland, available at: http://library.epfl.ch/theses, 209 pp., 2005.

Schaefli, B., Hingray, B., Niggli, M., and Musy, A.: A conceptual glacio-hydrological model for high mountainous catchments, Hydrol. Earth Syst. Sci., 9, 95–109, 2005, http://www.hydrol-earth-syst-sci.net/9/95/2005/.

Schaefli, B., Balin Talamba, D., and Musy, A.: Quantifying hydrological modeling errors through a mixture of normal distributions, J. Hydrol., 332, 303–315, 2006.

Schaefli, B. and Gupta, H.: Do Nash values have value?, Hydrol. Process., 21, 2075–2080, 2007.

Schaefli, B., Maraun, D., and Holschneider, M.: What drives high flow events in the Swiss Alps? Recent developments in wavelet spectral analysis and their application to hydrology, Adv. Water Resour., 30(12), 2511–2525, doi:10.1016/j.advwatres.2007.06.004, 2008

Schaefli, B. and Zehe, E.: Hydrological model performance and parameter estimation in the wavelet-domain, Hydrol. Earth Syst. Sci. Disc., 6, 2451–2498, 2009.

Schmidli, J. and Frei, C.: Trends of heavy precipitation and wet and dry spells in Swizterland during the 20th century, Intern. J.

Climatol., 25, 753–771, 2005.

Schreiber, T. and Schmitz, A.: Surrogate time series, Physica D, 142, 346–382, 2000.

Shumway, R. H. and Stoffer, D. S.: Time Series Analysis and Its Applications, With R Examples, 2nd edn., Springer, New York, USA, 576 pp., 2006.

Si, B. C. and Zeleke, T. B.: Wavelet coherency analysis to relate saturated hydraulic properties to soil physical properties, Water Resour. Res., 41, W11424, doi:10.1029/2005WR004118, 2005.

Torrence, C. and Compo, G. P.: A practical guide to wavelet analysis, B. Am. Meteorol. Soc., 79, 61–78, 1998.

Velasco, C.: Gaussian semiparametric estimation of non-stationary time series, J. Time Ser. Anal., 20, 87–127, 1999.

Vrugt, J. A., Gupta, H. V., Bouten, W., and Sorooshian, S.: A shuffled complex evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic models, Water Resour. Res., 39, 1201, doi:10.1029/2002WR001642, 2003.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., and Sorooshian, S.: A framework for development and application of hydrological models, Hydrol. Earth Syst. Sci., 5, 13–26, 2001, http://www.hydrol-earth-syst-sci.net/5/13/2001/.

Weisstein, E. W.: Metric, From MathWorld-A Wolfram Web Resource, http://mathworld.wolfram.com/Metric.html, last access: 19 March 2009, 2009.

Whittle, P.: Estimation and information in stationary time series, Ark. Mat., 2, 423–434, 1953.

Willis, I.: Hydrology of glacierized basins, in: Encyclopedia of Hydrological Sciences, edited by: Anderson, M. G., Wiley, Chichester, UK, 2601–2631, 2005.

Winsemius, H., Schaefli, B., Montanari, A., and Savenije, H. H. G.: On the calibration of hydrological models in ungauged basins: a framework for integrating hard and soft hydrological information, Water Resour. Res., doi:10.1029/2009WR007706, in press, 2009.

Yao, Q. and Brockwell, P. J.: Gaussian Maximum Likelihood Estimation For ARMA Models. I. Time Series, J. Time Ser. Anal., 27, 857–875, 2006.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resour. Res., 44, W09417, doi:10.1029/2007WR006716, 2008.

Zehe, E., Becker, R., Bardossy, A., and Plate, E.: Uncertainty of simulated catchment runoff response in the presence of threshold processes: Role of initial soil moisture and precipitation, J. Hydrol., 315, 183–202, 2005.

Zehe, E., Elsenbeer, H., Lindenmaier, F., Schulz, K., and Blöschl, G.: Patterns of predictability in hydrological threshold systems, Water Resour. Res., 43, W07434, doi:10.1029/2006WR005589, 2007.