Multi-basin and regional calibration based identification of distributed Precipitation-Runoff models for hourly runoff simulation: Calibration and transfer of full and partial parameters

Teklu T. Hailegeorgis[1]* and Knut Alfredsen[2]

[1]Department of Hydraulic and Environmental Engineering, NTNU, NO-7491 Trondheim, Norway.

[2]Department of Hydraulic and Environmental Engineering, NTNU, NO-7491 Trondheim, Norway.

***Corresponding author information:**

Name: Teklu T. Hailegeorgis

Researcher, Department of Hydraulic and Environmental Engineering, NTNU, NO-7491 Trondheim, Norway. Tel. [+47] 73592411, Fax [+47] 735 91298 and e-mail address: tekhi09@gmail.com.

ABSTRACT

Identification of distributed precipitation-runoff models for hourly runoff simulation based on transfer of full parameters (FP) and partial parameters (PP) are lacking for boreal mid-Norway. We evaluated storage-discharge relationships based model (Kirchmod), the Basic-Grid-Model (BGM) and a simplified Hydrologiska Byråns Vattenbalansavdelning (HBV) model for multi-basins (26 catchments). A regional calibration objective function, which uses all streamflow records in the region, was used to optimize local calibration parameters for each catchment and regional parameters yielding maximum regional weighted average (MRWA) performance measures (PM).

Based on regional median Nash-Sutcliffe efficiency (NSE) and NSEln (for log-transformed series) for the calibration and validation periods, the Kirchmod model performed better than the others. Parsimony of the Kirchmod model provided less parameter uncertainty for the FP case but did not guarantee parameter identifiability.

Tradeoffs between parsimony and performance were observed despite advantages of parsimony to reduce parameter correlations for the PP, which requires preliminary sensitivity analysis to identify which parameters to transfer. There are potential advantages of using the MRWA method for parameter transfer in space. However, temporal validation indicated marked deterioration of the PM. The tradeoffs between parameter transfers in space and time substantiate both spatial and temporal validation of the regional calibration methodology.

Key words: Model identification; Hourly runoff; Regional calibration; Parameter uncertainty and identifiability; Parameter transfer; Model validation.

INTRODUCTION

Continuous streamflow simulation by Precipitation-Runoff (P-R) models for prediction purposes are widely employed, for instance to predict streamflow to reservoirs, floods and droughts, and to assess effects of alteration of natural flow regime due to anthropogenic impacts. Moreover, utilization of hydropower reservoirs to satisfy peak energy demands (hydropeaking operation) requires streamflow forecasting at high temporal resolution. The European Water Framework Directive requirements for ecological protection further substantiate the need for better hydrological predictions for ecological impact management in regulated rivers. In addition, prevalence of flood events associated with the issues of land use and climate change require forecasting at high temporal resolution.

The current technology allows for measurements of environmental variables such as rainfall and streamflow with fine temporal resolution and a vast amount of sub-daily data from different sources may be available (see Jones, 2005). However, the majority of previous studies on identification of the P-R models for continuous simulation and prediction purposes in literature are based on a daily time scale, which leaves the potential high information content of available hourly data unexplored. Previous studies (e.g. Kavetski *et al*., 2011; Bastola and Murphy, 2013) illustrated the dependence of optimal model parameters on the temporal resolution of data and substantial drawbacks of parameter transfer from daily calibration to prediction on an hourly time scale. Therefore, there is an interest in hourly calibration and prediction for operational use, which requires comprehensive study relevant to the research gaps on identification of suitable P-R models for the hourly prediction.

Wagener and McIntyre (2005) conducted a study on the identification of lumped conceptual rainfall-runoff models for operational applications based on daily streamflow on three catchments in UK using the ‘split-sample’ and ‘proxy basin’ operational testing schemes of Klemeś (1986), and goodness-of-fit metrics for different flow ranges. Fenicia et al. (2011) used a flexible framework to identify model performance of several model structures for four different catchments in Europe and New Zealand. Smith and Marshal (2009) carried out model selection based on a suite of 30 conceptual, modular structures for snow-dominated, mountainous experimental watershed in USA using 12 hourly data. Orellana *et al*. (2008) applied seven semi-distributed rainfall-runoff model structures using hourly data from four gauging stations in the UK. However, these studies focused on coarse temporal resolutions and/or on a single catchment (with only one or more gauges) or a small number of catchments in a region rather than on fine temporal resolution (e.g. hourly) and multi-basin regional scale modelling based identification of the P-R algorithms.

There are also studies based on both multi-model and multi-basin simulations for both daily and hourly resolutions. Lee *et al*. (2005) conducted a study on the selection of 12 daily conceptual model structures for regionalization for Prediction in Ungauged Basins (PUB) of the rainfall-runoff relationships for 28 UK catchments. Oudin *et al*. (2008, 2010) used two lumped models and daily streamflow records from large number of catchments in France respectively for comparison of regionalization approaches for the PUB and for studying the relationships between physical similarity and hydrological similarity of catchments. Viviroli *et al*. (2009a&b) conducted calibration for 140 mesoscale catchments for hourly flood prediction in ungauged Swiss catchments. However, the majority of the previous studies on multi-model calibration based on multi-basin data mainly focused on regionalization for the PUB rather than on the identification or performance evaluation of the models among alternative hydrological mechanisms as suggested by Jones (2005). An exception is the work by Perrin *et al*. (2001) who conducted a multi-model comparative performance assessment of 19 parsimonious to more complex daily lumped models on 429 catchments mostly located in France.

A thorough study of the identification of P-R models in simulation mode has the potential for improving forecast accuracy. Better performance of the precipitation-runoff models in simulation mode is crucial for forecast modes (see Refsgaard, 1997; Bell and Moore, 1998; Engeland and Steinsland, 2014). In addition, the specific tools used in forecasting for data assimilation and correction affect the performance of a forecast (see Nicolle *et al*., 2014). Therefore, the review indicates that the previous work on hourly identification of P-R models based on multi-basin or regional calibration approach is lacking for boreal snow-dominated catchments. The use of regional scale data and hence data augmentation through the regional calibration is expected to allow more comprehensive performance evaluation than the at-site records based local calibration and 'proxy basin' based model validation.

Identification of the P-R models are dependent on objective functions used for model calibration and performance measures used for model evaluation. For instance, fitting of the P-R models to reproduce the whole hydrograph for scientific research or to a specific flow regime for operational purposes would result in different optimal parameter vectors. For operational applications, it is a common practice to use the P-R models as a 'fit-for-purpose' decision support tools. The commonly used adjustments to make the operational models more right for a 'fit-for-purpose' performance are the error or bias correction parameters for precipitation measurements (e.g. Sevruk, 1983; Yang et al., 1999; Herrnegger *et al*., 2014), but Moine *et al*. (2007) suggested that this practice should be avoided. In addition, an altitudinal gradient parameter for precipitation are considered in some applications but Hingray *et al.* (2010) noted that omitting an altitudinal gradient is a good option to simulate flood events, especially in cases of large precipitation events. Such adjustments for operational settings have the potential to force the models to be 'right for the wrong reasons' (Kirchner, 2006).

Therefore, comprehensive identification of the P-R models is required for reliable continuous simulation of streamflow (e.g. Wagener, 2003). Hailegeorgis *et al*. (2015b) focused on multi-model based identification of four different types of regionalization methods including the regional calibration method defined by parameter sets yielding maximum regional weighted average (MRWA) performance measures (PM) based on transfer of full set of local calibrated parameters (FP). The authors applied the three P-R models on 26 catchments in mid Norway, which are also used in the present study. Due to similar performance of the regionalization methods based on the MRWA and transferring of regional median parameters (RMedP), the authors suggested that it is worth testing the performance of fixing some of the parameters to regional median values, for instance the snow and runoff routing routines parameters that are common for the three models, and then perform calibration and transfer of partial parameters (PP). Fixing some of the parameters is advantageous since it allows a more parsimonious parameterization while it may have potential disadvantages of reducing the performance of the models. However, studies related to the issues of transferring the full parameter set or partial parameters are necessary to further improve the results of regionalization tasks.

The main objective of the present study is the identification of the three P-R models for hourly runoff simulation based on calibration and transfer of partial parameters (PP) for the 26 catchments in mid Norway compared to a study for the same region using full parameter calibration and transfer (FP) case of Hailegeorgis *et al*. (2015b).

THE STUDY REGION AND DATA

The study region is the boreal mid Norway, which consists of 26 unregulated gauged catchments ranging from 39 to 3090 km^2 in size (Table 1 and Fig. 1). Streamflow and climate records of hourly time resolution (01.09.2008-01.01.2012) were used for model calibration. The climate forcing are precipitation (P), temperature (T), wind speed (W_s), relative humidity (H_R) and global radiation (R_G). Figure 1 shows locations of precipitation and streamflow gauging stations. Table 1 contains some characteristics of the catchments and streamflow stations.

Precipitation occurs in the form of snowfall during winter and rainfall dominates during summer, spring and autumn. The catchments exhibit wide ranges of variations in elevation and terrain slope. There is no systematic relationship between elevation and mean annual precipitation for the region and hence we did not consider altitudinal gradient corrections for the hourly precipitation data. An environmental lapse rate of -0.65^{o}C/100m was used to account for elevation-temperature relationship. The dominant land uses/land covers in the study area are mountainous terrain above timberline and forests. Predominant soil or loose material is glacial tills and the dominant bedrock types for the study catchments are metamorphic and igneous rocks (http://www.ngu.no).

MODELS AND METHODS

We evaluated three different distributed (1x1 km^2 grid) precipitation-runoff models namely the ‘top-down’ water balance model based on Kirchner (2009) or Kirchmod, the Basic-Grid-Model based on Bell and Moore (1998) or BGM and a simple configuration HBV model. Table 2 presents lists of calibrated parameters and their prior ranges or values of fixed parameters for both full parameter transfer (FP) and partial parameter transfer (PP) of the present study. For the PP case, parameters that are common for the three models were fixed to their multi-model regional median or MMRMedP (Eqn. 9) values of the respective parameters obtained from calibration of the FP case. Similarly, parameters in the soil moisture accounting routine of the HBV model and exponent parameter of the subsurface drainage equation (Eqn. 6) of the BGM model were fixed to their regional median or RMedP (Eqn. 7) values. A total of 6, 7 and 9 parameters were calibrated for the FP case for the Kirchmod, BGM and HBV models respectively. A total of 3 parameters were calibrated for the PP case for all models. Therefore, for the PP case a total of 3, 4 and 6 parameters of the Kirchmod, BGM and HBV models respectively were fixed. Brief descriptions of the models are given here. Descriptions of the models that are more detailed are referred to Hailegeorgis et al. (2015b).

Kirchner's runoff response routine (Kirchmod)

The main assumption in the Kirchner's method (Kirchner, 2009) is the discharge Q depends solely on the amount of water stored in the catchment S based on a nonlinear catchment storage-discharge relationship and a water balance equation:

$$\frac{dQ}{dt} = \frac{dQ}{dS}\frac{dS}{dt} = \frac{dQ}{dS}\left(I - AET - Q\right) = g\left(Q\right)\left(I - AET - Q\right), \tag{1}$$

where g(Q) = *dQ*/*dS* is the discharge sensitivity function (Kirchner, 2009). The following linear regression relationship were inferred based on streamflow recession analysis following Kirchner (2009):

$$S\left(Q\right) = \int \frac{1}{g\left(Q\right)} dQ \; ; \; \ln\left(g\left(Q\right)\right) \approx b_0 + b_1 \ln\left(Q\right) \tag{2}$$

The AET was computed from potential evapotranspiration (PET) and discharge:

$$AET = PET\left\{1-\exp\left(-\frac{Q}{EvR}\right)\right\}\left(1-SCA\right) , \quad (3)$$

where the actual evapotranspiration (AET), infiltration (I) = rainfall + snow melt (SM) and Q are in mm/hr, storage (S) is in mm and t is a time variable. The EvR denotes a discharge at which AET equals 0.95*PET. The SCA is snow-covered fraction of grid cell to set the AET to zero for snow-covered areas. A Runge Kutta 4th order method was used to solve the integral (eqn. 2) over the time step. The Q is an instantaneous simulated discharge obtained from the solver while an average Q over the time step is used for calibration against an hourly averaged observed discharge. Observed discharge before the start of model run was used as an initial discharge for the numerical solver. Only the three response routine parameters b_0, b_1 and EvR were calibrated for the PP case.

Basic Grid Model (BGM) Runoff Response Routine

The BGM is a simple distributed model based on Bell and Moore (1998). The infiltration excess runoff, R_{iex}[L] (Horton, 1933), saturation excess runoff, R[L] (Dunne and Black, 1970a&b) and a subsurface drainage (D_{rv}) runoff generation mechanisms are considered:

$$R_{iex} = \max\left[0,(SNOWOUT - I_c)\right] ; TOSOIL = SNOWOUT - R_{iex} \quad (4)$$

$$R = \max\left[0, S(t) + TOSTORAGE - S_{\max}\right] ; S(t+\Delta t) = \max\left[0, S(t) + TOSTORAGE - R\right] \quad (5)$$

$$D_{rv} = k\left[S(t)\right]^n ; AET = PET \times \frac{S}{S_{\max}} ; TOSTORAGE = TOSOIL - AET - D_{rv} , \quad (6)$$

where SNOWOUT[L] is the rainfall and snowmelt outflow from snow routine, TOSOIL[L] is the infiltration into the soil, TOSTORAGE[L] is the net input to the subsurface storage (S[L]), PET[L] and AET[L] are as defined earlier, Drv[L] is the subsurface flow or drainage per unit area, and L and T denote length and time dimensions. The I_c[L/T] or an infiltration capacity, the coefficient k[$L^{1-n/T}$] and maximum subsurface storage capacity or S_{max}[L] were calibrated parameters. Since marked correlation between the k and n[-] parameters was observed for the FP case in Hailegeorgis *et al*. (2015b), in the present study the parameter n was fixed to its calibrated RMedP (eqn. 7) value of the FP case to reduce the correlation and non-identifiability between the two parameters:

$$RMedP = Median \;\; P_1, P_2, P_3, ..., P_{N_C} \;\; , \tag{7}$$

where RmedP denotes regional median parameter, P_1 to P_{NC} denotes calibrated values of the parameter for each catchment and N_C is the total number of catchments calibrated.

The HBV Runoff Response Routines

The HBV runoff response routine used in the present study consists of two linear reservoirs i.e. upper and lower reservoirs:

$$Q_{UZ} = k_1 \times UZ \;\; ; Q_{LZ} = k_0 \times LZ \;\; , \tag{8}$$

where Q_{UZ} and Q_{LZ} respectively are outflows from the upper and lower reservoirs. Percolation from the upper to the lower reservoir in the runoff response routine is controlled by percolation parameter (*PERC*). The soil moisture accounting routine was based on a non-linear partitioning curve for infiltration into change in soil moisture storage (ΔSM) and recharge (R) to the upper zone (Bergström, 1976). Only the three parameters of the runoff response routine namely, recession coefficients in the upper reservoir (k_1), base flow recession coefficient (k_0) and the percolation rate (PERC) to the lower zone were calibrated for the PP case. Two of the soil moisture accounting parameters namely, shape parameter of the partitioning curve (β) and field capacity (FC) were fixed to RMedP (eqn.7) values calibrated for the FP case (Table 2). The ‘limit for potential evaporation’ (LP) was set to a constant value of 0.90, which is a default value of HBV-96 (Booij, 2005).

Snow Accounting Routine

The snow routine uses a mass balance approach to simulate the melt water release (snowmelt runoff) from saturated snow (Qs) and the remaining unmelted snow storage or the snow water equivalent (SWE) based on the Gamma distributed snow depletion curve (SDC). The SDC uses radiation for surface layer energy and phase change calculations (Kolberg and Gottschalk, 2006) as implemented in ENKI hydrological modelling platform (Kolberg and Bruland, 2012). The parameters in this routine are common for the three models and include rainfall-snowfall threshold temperature (TX) and snowmelt sensitivity to wind speed (WS). These parameters were fixed to MMRMedP values (Eqn. 9) of respective parameters calibrated for the FP case:

$$MMRMedP = Median \quad RMedP \quad M_1, M_2, M_3 \quad , \tag{9}$$

where M_1, M_2 and M_3 denotes Kirchmod, BGM and HBV models respectively.

Potential Evapotranspiration Routine

In the present study, we used the PriestleyTaylor method (Priestley and Taylor, 1972) for the calculation of potential evapotranspiration, PET (mm/h):

$$PET = \alpha \frac{\Delta}{\Delta + \gamma} \; R_n \; \left(\frac{\Delta t}{L_v} \right), \tag{10}$$

where α is the Priestley Taylor constant, Δ is the slope of saturation vapor pressure curve at air temperature at 2m (kPa/^{o}C), γ is the psychrometric constant (0.066 kPa/^{o}C), R_n (W/m^2) is net radiation, L_v (kJ/m^3) is volumetric latent heat of vaporization and Δt (s) is the simulation time step in seconds. The net radiation is the sum of net shortwave radiation and net longwave radiation. We computed the net shortwave radiation from the global radiation (R_G) and land albedo, and the net longwave radiation based on Sicart *et al*. (2006). Following Teuling *et al*. (2010), α = 1.26 was used to reduce the number of calibrated parameters.

Runoff Routing

Hailegeorgis *et al*. (2015a) applied a source-to-sink routing with effective velocity of flow for mountainous catchments in mid-Norway. Li *et al*. (2014) applied cell-to-cell routing and source-to-sink routing with spatially distributed velocity of flow for mountainous catchments in central southern Norway. Following Hailegeorgis et al. (2015b) a simple translation based on a 1-hr travel time isochrones was used to translate the runoff response from the hillslope (1x1 km^2 grid cells) to the catchment outlet. Routed simulated streamflow at the outlet is the sum of contributions from each grid cell:

$$Qsim_t = \sum_{i=1}^{N} qsim_{t-T_i}^{i} \; ; \; \mathrm{T_i} = \frac{L_i}{V}, \tag{11}$$

where *t* and *i* represent time and grid cells, *N* is the number of grid cells in the catchments, *Qsim* [LT^{-3}] is streamflow at the outlet, *qsim* [LT^{-3}] is runoff generated at each grid cell, T_i [T] is flow travel time lag to the outlet for each grid, L_i [L] is flow travel path length computed from 25m Digital Elevation model (DEM). The V [LT^{-1}] is velocity of flow, which is a parameter common to the three models and was fixed to MMRMedP (eqn. 9) of calibrated values for the FP case.

Model calibration and evaluation

For the regional calibration, the Differential Evolution Adaptive Metropolis (DREAM) algorithm (Vrugt *et al*., 2009) was used with residuals based log-likelihood (*L-L*) objective function, which was implemented in ENKI hydrological modelling platform (Kolberg and Bruland, 2012):

$$L-L\left(\delta / \sigma_i^2, \sum_{i=1}^{N_C}\sum_{t=1}^{n_i} \left(Qsim_{t,i}^{(\theta)} - Qobs_{t,i}^{(\theta)}\right)^2\right) = \left\{\sum_{i=1}^{N_C}\left(\frac{-n_i}{2}\log\left(2\pi\right) - \frac{n_i}{2}\log\left(\sigma_i^2\right) - \frac{\sum_{t=1}^{n_i}\left(Qsim_{t,i}^{(\theta)} - Qobs_{t,i}^{(\theta)}\right)^2}{2\sigma_i^2}\right)\right\} \times f, \tag{12}$$

where δ denotes model parameter, σ_i^2 and n_i respectively are error variance and the length of non-missing records of streamflow for catchment *i*, N_C is the total numbers of catchments in the region, $Qsim^{(\theta)}$ and $Qobs^{(\theta)}$ respectively are Box-Cox (Box and Cox, 1964) transformed observed and simulated streamflow time series, θ is the Box-Cox transformation parameter and *f* represents a fraction of effectively independent observations which can be estimated from the autoregressive (AR1) model of error covariance (Zięba, 2010). We used the Box-Cox transformation to approximate Normality and homoscedasticity of the residuals. Values of θ between 0.25 and 0.30 are common in literature (e.g. Willems, 2009). We used θ = 0.3 and f = 0.001 for the sake of consistency among the catchments. The DREAM calibration algorithm converges as the Gelman-Rubin convergence (Gelman and Rubin, 1992) comes below 1.2. Details of the DREAM algorithm can be found from Vrugt *et al*. (2009).

We evaluated the local and regional calibration based on the Nash-Sutcliff efficiency or NSE (Nash and Sutcliffe, 1970) and Nash-Sutcliffe efficiency for log-transferred series (NSEln) performance measures (PM). The NSE gives greater weight to high flows and the NSEln gives greater weight to low flows.

The regional calibration used in the present study can be regarded as an 'importance sampling' strategy for each catchment, where we sample according to an ‘importance surface’ reflecting where we believe the optimum is likely to be (Hailegeorgis et al., 2015b). The objective function in eqn. (12) uses streamflow data from all stations in the region rather than using at-site streamflow records from only a particular site. Therefore, parameter sets among the DREAM samples which provide maximum performance measures (PM) for each catchment are taken as optimized parameters for local calibration (LC) for a specific catchment. Optimal parameter sets for the regional calibration are parameter sets among the DREAM samples that provided maximum regional weighted average (MRWA) performance measures. In the present study, the term regional calibration and the MRWA are used interchangeably. Hailegeorgis et al. (2015b) reported nearly equivalent performance of the MRWA method to more advanced regionalization methods like the physical similarity and spatial proximity methods. In the present study, the MRWA is used to evaluate the regional performance and hence performance of the models for prediction in ungauged basins. We allocated the weight for each catchment based on their length of non-missing streamflow records during the calibration period:

$$NSE_{MRWA} = \frac{1}{N_C}\sum_{i=1}^{N_C}\left(\frac{n_i}{N_{TS}}\right)NSE_i \ ; NSE\ln_{MRWA} = \frac{1}{N_C}\sum_{i=1}^{N_C}\left(\frac{n_i}{N_{TS}}\right)NSE\ln_i \ , \qquad (13)$$

where N_{TS} is the total length of time series for the calibration period. The weights for each catchment are the term in the parenthesis, which are assigned based on the length of their non-missing streamflow records.

The classical split-sample test (Klemeś, 1986) was used for validation of the models (for FP and PP cases) outside the period used for calibration based on NSE for both local and regional calibration, and NSEln for regional calibration. Due to lack of long records, a validation period of only one year (01.01.2006-01.01. 2007) was used. The regional calibration used in the present study is similar to regional calibration works, among others, (Fernandez et al., 2000, Beldring et al., 2003 and Engeland et al., 2006) except the fact that weighted average performance measures are used than arithmetic averages for model evaluations. Model validation for this type of regional calibration is not common in literature. However, Beldring et al. (2003) used a hierarchical scheme for model validation (Klemeś, 1986) which distinguishes between simulations performed for the catchment used for calibration and for a different catchment by noting that the scheme is more adequate than the split-sample scheme using streamflow data from the same catchment during both calibration and validation.

We used histograms or distribution fits (e.g. Schoups and Vrugt, 2010) and linear correlation coefficient matrix of the posterior parameters (e.g. Moreda *et al*., 2006; Blasone *et al*., 2007; Schoups and Vrugt, 2010) to show parameter uncertainty and identifiability. The last 50 % of the posterior parameters accepted by the DREAM algorithm after the burn-in iterations (Vrugt et al., 2009) were used to construct the histograms of posterior parameters and to calculate the correlation coefficients among the posterior parameters. Burn-in iteration refers to discarding an initial portion of the samples to minimize the effects of initial conditions (Hailegeorgis and

Alfredsen, 2014). Hailegeorgis and Alfredsen (2014) provided more details of the DREAM algorithm used in the present study.

RESULTS

Figure 2a-c and Figure 3a-c display performance of the LC and MRWA of the models for the NSE and NSEln respectively. For many catchments, the performance of the three models seems to be close but for some catchments (e.g. catchment 15), the HBV model performed markedly better than the others did. There are tradeoffs of reduction in performance due to the parsimony by fixing some of the parameters to their RMedP and MMRMedP values for the PP case as the large number of free parameters favors for calibration performance for the FP case. The LC performance of FP is better than that of the PP for all catchments for the three models. For the MRWA, the NSE values of the FP are higher than that of the PP for the majority of the catchments except for catchments 2, 12 and 17 for the Kirchmod and BGM models, and catchments 2, 13 and 19 for the HBV model (Figure 2). This may be related to different levels of model performance sensitivity to the fixed parameters among the catchments. Generally, the MRWA for the FP case performed better than the PP case in terms of performance for individual catchment.

Similarly, the NSEln values of the FP is higher than that of the PP except slightly higher NSEln values for some catchments, for instance catchment 2 for the Kirchmod and BGM models. Table 3 shows the regional median values of the PM or the regional performance of the models. In terms of the regional median of the NSE corresponding to the LC and MRWA, the Kirchmod model followed by the BGM model performed better than the HBV model (Table 3). However, the NSE for the Kirchmod and BGM are nearly similar for the FP case. In terms of the regional median of the NSEln corresponding to the LC and MRWA, the Kirchmod model followed by the HBV model performed better than the BGM model except for the FP case for

MRWA (Table 3). However, performance of the HBV model and BGM model are nearly similar.

Figure 4a-c present the NSE values for the validation period for both LC and MRWA. For the validation period, only 12 catchments exhibited NSE ≥ 0.50 for both FP and PP cases for the LC of the Kirchmod model. Only 8 and 6 catchments exhibited NSE ≥ 0.50 for FP and PP cases respectively for the MRWA of Kirchmod model. Only 9 and 8 catchments for FP and PP cases respectively exhibited NSE ≥ 0.50 for both local calibration and MRWA for the BGM model. For the HBV model, only 8 catchments exhibited NSE ≥ 0.50 for both FP and PP cases for the local calibration while only 6 catchments exhibited NSE ≥ 0.50 for both FP and PP cases for the MRWA. However, for the calibration period up to 23 and 16 catchments respectively exhibited NSE ≥ 0.50 for the LC and MRWA. Therefore, the results of split-sample validation indicated marked deterioration of the NSE for both the FP and PP cases for the three models.

Table 4 presents the regional median NSE for validation period for both the LC and MRWA, and regional median NSEln for the MRWA. In terms of the regional median of the NSE corresponding to the LC, the Kirchmod and HBV models exhibited equally better performance followed by the BGM model for the FP case while the Kirchmod model performed better followed by the BGM model and HBV model for the PP case (Table 4). For the MRWA, the Kirchmod model performed better followed by the HBV model and BGM model for both FP and PP cases. In terms of the regional median of the NSEln corresponding to the MRWA, the Kirchmod model performed better while the BGM and HBV models performed equally for the FP case. However, for the PP case, the Kirchmod model performed better followed by the BGM model while the HBV model exhibited the worst performance. The marked deterioration in performance of the HBV model for the PP case is most probably attributable to fixing the three parameters of the soil moisture accounting routines namely FC, LP and β to their RMedP values (Table 2) in addition to parameters that are common to the three models. Therefore, the validation results also show that the Kirchmod model performed relatively better than the BGM and HBV models.

Figure 5a-c present values of calibrated parameters for the FP and PP, and RMedP or MMRMedP values of the fixed parameters for the PP case. The values of the calibrated parameters for the FP and PP are different, which show the sensitivity of calibrated parameters to fixing some of the parameters i.e. the calibrated parameters compensate for the fact that some parameters were fixed to their RMedP (Eqn. 7) or MMRMedP (Eqn. 9) values. Figure 6a-f present the histograms and 'best-fit' distributions fitted using the Statistics Toolbox 9.0 in matlab for the posterior parameters obtained from the DREAM algorithm. The calibration resulted in different types of 'best-fit' posterior distributions of the parameters while the uniform prior distribution (Table 2) was used for all.

For the FP case, the three parameters of the Kirchmod model exhibit narrow posterior distributions (Figure 6a) indicating less parameter uncertainty compared to the parameters for the BGM model (Figure 6c) and HBV model (Figure 6e). In addition, some parameters like the coefficient for the storage-discharge relationship (k) of the BGM model and the slow flow recession coefficient (k_0) of the HBV model exhibit narrow posterior distributions. Even though there are equal numbers of calibrated parameters in the Kirchmod and HBV response routines, wider posterior distributions (hence large uncertainty) for the HBV response routine parameters for the FP case probably indicate less sensitivity of the response routine parameters and interactions between the soil moisture accounting routine and the response routine parameters for the HBV model. For the PP case, posterior distributions of calibrated parameters are wider than the FP cases (i.e. large uncertainty) for the Kirchmod (Figure 6b), BGM (Figure 6d) and HBV (Figure 6f) models.

Table 5 shows correlation matrices of posterior parameters as a measure of identifiability of the parameters. The correlation matrices showed considerable interactions among some parameters manifested by large positive or negative correlations. Positive correlation coefficients greater than 0.60 were observed between the regression parameters b_0 and b_1 for the Kirchmod model for both FP and PP case. The two parameters support each other to influence the discharge sensitivity for the change in storage (g(Q)) based on eq. (2), which shows challenges of parameter non-identifiability even for parsimonious parameterization.

For the BGM model for the FP case, there is a positive correlation greater than 0.6 between S_{max} and the coefficient k and there is a large negative correlation ($r < -0.6$) between the exponent parameter n and k, which show that the S_{max} and k support each other while n and k compensate each other according to eq. (6) for computation of the subsurface drainage. For the PP, there is no case of $r > 0.6$ or $r < -0.6$ for the BGM model that shows parameterization by fixing the n in the subsurface drainage equation resulted in reduction of parameter correlations, which fulfilled the intention of fixing the parameter n.

For the HBV model, there is a positive correlation greater than 0.60 between the quick flow recession coefficient (k_1) and percolation to the lower zone (PERC) for the FP case. This shows that an increase in k_1 for the discharge from the upper zone (Q_{UZ}) compensates the decrease in the upper zone storage due to an increase in the PERC. However, there is less correlation between k_1 and PERC for the PP case most probably due to fixing the soil moisture accounting parameters. There is a large negative correlation ($r < -0.6$) between k_0 and PERC for the HBV for the FP case, which shows that the two parameters compensate each other for the baseflow contribution from the lower reservoir (Q_{LZ}). For the HBV model, there is a large negative correlation ($r < -0.6$) between the response routine parameters k_0 and k_1 for both FP and PP cases. This compensation between the discharge from the upper and the lower reservoirs in the response routine regardless of the parsimony obtained by fixing the parameters of the soil moisture accounting routine indicates higher challenges of parameter non-identifiability in the multiple storage HBV model.

DISCUSSION

Performance of calibration and validation

Local and regional calibration based model performance for the NSE (Figure 2 and Figure 3) indicate that the Kirchmod and BGM models provided better performance for the majority of the catchments. However, the HBV model provided best NSE and NSEln performance for local calibration for some catchments, e.g. catchments no. 15, 17 and 19. Generally, the best performing model varies among the catchments and performance measures and hence it is not possible to identify a unique model structure for the region. This complies with the uniqueness of place (Beven, 2000) and previous findings that one cannot expect similar calibration performance for a model across different ranges of magnitudes of streamflow series (Gupta et al., 1998; Wagener et al., 2001, Madsen, 2003). Lee *et al*. (2005) investigated if it is justifiable to use one model structure to cover a range of catchment types and found that there is no evidence of relationships between catchment type and preferred model structure. The authors found the results based on classification of 28 catchments over a range of hydrological types and wide geographical extent in the UK based on different combinations of three catchment characteristics namely catchment area, a baseflow index from the hydrology of soil types classification and annual average rainfall for the period 1941–1970.

Due to higher values of regional median NSE, the Kirchmod and the BGM models are more suitable than the HBV model for the MRWA, which has a potential for prediction of high flows in ungauged basins. For the NSEln, the Kirchmod model provided higher performance than the HBV and BGM models; however, the HBV model provided slightly higher NSEln than the BGM model probably due to separate simulation of baseflow from the lower reservoir for the HBV model. Hailegeorgis *et al*. (2015b) found similar performance of the MRWA to other more advanced regionalization methods and hence selection of the models based on their MRWA performance for the PUB is valid for the region. The 'top-down' Kirchmod model, which is based on a single catchment storage-discharge relationships and does not consider an infiltration excess overland flow, performed better in terms of regional NSE and NSEln than the BGM model that considers both the infiltration excess and saturation excess runoff generation mechanisms and the HBV routines with multiple storage reservoirs. However, the general trends in performance of the three models are very close to each other for the majority of the catchments except for some catchment e.g. catchment 15.

Deterioration of the NSE and NSEln from their values obtained for the LC were observed for the MRWA for nearly all of the catchments (Figure 2 and Figure 3). The NSE and NSEln values for both the LC and MRWA are lower for the PP case than the FP case (Figure 2 and Figure 3) for the majority of the catchments. These show that despite parsimony could be achieved by fixing some of the parameters to their RMedP or MMRMedP values, there are tradeoffs of noticeable deterioration in performance. The catchments with poor NSE and NSEln are of different sizes and located in different parts of the study region. However, the majority of these catchments are located far from precipitation gauging stations and hence the less representativeness of the precipitation stations probably affected the performance for these catchments.

The model validation using the split-sample test showed that the NSE for both LC and MRWA deteriorate for outside calibration period (Figure 4 and Table 4). Similarly, the NSEln of the MRWA deteriorate for the validation period. For instance, the NSE values for validation period for LC of the BGM model for catchment 6 are 0.40 and 0.39 for the FP and PP respectively (Figure 4b) compared to NSE values of 0.83 and 0.81 for the FP and PP respectively for LC for the calibration period (Figure 3b). Hailegeorgis et al. (2015a) obtained NSE values of 0.84 for both calibration and validation periods by calibrating catchment 6 by using only streamflow records for the catchment. The NSE values for validation period for the LC of the HBV model for catchment 6 are 0.35 and 0.49 for the FP and PP respectively (Figure

4c) compared to NSE values of 0.74 for both the FP and PP for the calibration period for the LC (Figure 3c). Hailegeorgis and Alfredsen (2014) based on calibration of the HBV model for catchment 6 by using a streamflow data only from the catchment obtained NSE values of 0.75 and 0.71 respectively for calibration and validation periods. The results demonstrated that performance of calibration using only a streamflow data for a particular catchment would probably result in optimal parameter that has better transferability in time. Split-sample test for validation of the regional calibration methodology used in the present study, which uses all available streamflow records from all catchments, is not common in literature. However, the results of the present study comply with the study by Beldring et al. (2003), who found that regional calibration of a model failed to model the dynamics of hydrological processes for several catchments based on a hierarchical scheme for model validation.

However, there are merits of the multi-basin regional calibration to derive regional parameters, which yields the MRWA PM to transfer these parameters in space for prediction in ungauged basins in the region. The multi-basin and regional calibration approach would provide an opportunity for a more comprehensive evaluation of models better than the proxy basin (Klemeś, 1986; Wrede et al., 2013) approach. Fenicia *et al*. (2011) proposed a flexible framework for conceptual hydrological modelling, SUPERFLEX, with one of the objectives towards a more robust and reliable performance in operational contexts. For operational purposes, combined flexible models and multi-basin based identification of robust and reliable model structures, parameterizations and modelling paradigms (e.g. 'bottom-up' process models and 'top-down' inferences from observations) among a pool of plausible competing options are advisable. Currently, fixed model and catchment scale modelling are more common due to their simplicity and less computational demand.

The model calibration based on continuous time series and model evaluations based on different performance measures (e.g. NSE and NSEln) could not necessarily yield optimal parameter sets, which can simultaneously simulate floods associated to high rainfall and snowmelt events, and low flows especially when extrapolated to the streamflow magnitude outside the calibration conditions. Wagener and McIntyre *et al*. (2005) on identification of rainfall-runoff models for operational applications suggested that a more empirical approach to identification of models for specific forecasting problems are preferable to trying to achieve a good all-round representation of the rainfall-runoff processes. Calibration for a specific modelling objective or reproducing a specific runoff signature may provide reliable prediction for the specific purpose.

Parameter uncertainty and identifiability

Uhlenbrook *et al*. (1999) found considerable implications of parameter uncertainty and identifiability on the predictive uncertainty, and noted that parameter and model structure uncertainties should be considered for operational (practical) predictions. Wider posterior distributions (i.e. large uncertainty) of calibrated parameters for the PP case than the FP case for the Kirchmod (Figure 6b), BGM (Figure 6d) and HBV models (Figure 6f) show that parsimony in the number of parameters and longer data series for calibration do not necessarily provide less parameter uncertainty. However, while comparing the models for the FP case, narrow posterior parameter distributions of the Kirchmod (Figure 6a) compared to the other runoff response routines (Figure 6c&e) indicate that a small number of free parameters exhibits least parameter uncertainty. In addition to the parsimony, the model structure based on the 'top-down' modelling paradigm and relationship between catchment storage and discharge inferred from streamflow recession analysis might have contributed to the reduction in parameter uncertainty. For a given model structure, there is a likelihood of less predictive uncertainty from less parameter uncertainty, but uncertainties due to input data also contribute to the predictive uncertainty.

Few pairs of response routine parameters exhibit correlation coefficients (r) with either $r > 0.60$ or $r < -0.60$ (Table 5). The parsimony for the PP case reduced correlations in runoff response routine parameters of the BGM and HBV models than the Kirchmod model. In terms of parameter correlations, the BGM model benefited much better from the parameterization of the subsurface drainage equation based on fixing the exponent parameter. Correlation of parameters results in lack of identifiability because a change in one parameter compensated by a change in another, such that multiple parameter sets give the same output according to some quantity of interest (Libelli *et al.*, 2014). The existence of either positive or negative correlations is an indication of non-identifiability of parameters and hence the potential for non-identifiability of the performance of the models, which is one of the main challenges in precipitation-runoff modelling. Hailegeorgis and Alfredsen (2014) found that compensation between the discharge from the upper reservoir and baseflow from the lower reservoir in the different HBV configurations resulted in indistinguishable streamflow hydrographs but less reliable baseflow simulation by some of the configurations.

The differences in the values of the calibrated parameters for the FP and PP cases (Figure 5a-c) show the sensitivity of runoff simulation to the fixed parameters, and compensations and correlations among the parameters. Parameterization issues have potential impacts on regionalization based on transferring of parameters for PUB. Therefore, regionalization of precipitation-runoff models should be augmented by preliminary parameter sensitivity analysis to determine which parameters to transfer. The quality of input (both climate and streamflow data) should also be able to constrain the model parameters during calibration.

Data quality

The expected conditions for the model calibration is that there is no considerable error in the observed streamflow data and uncertainty in estimation of precipitation fields is low. Errors in the observed streamflow and errors in estimation of precipitation fields have the potential to affect the reliability of calibrated (optimized) parameters. However, the discrepancies in the data potentially affect the reliability of modelling inferences and predictions, which is one of the challenges in hydrological modelling. The density and representativeness of precipitation gauging stations are crucial to capture the spatial variability of precipitation, for instance, localized intense precipitation events to reproduce the flood events. Sparse gauging networks for the hourly precipitation input, which may yield less accurate spatially interpolated precipitation fields on the 1x1 km^2 grids, seems to be a major factor for the low NSE or poor estimation of peak flows. Engeland and Steinsland (2014) mentioned that they applied a hydrological forecasting model at daily time-step for small size catchments (with time of concentration less than one day) in southwestern Norway due to the availability of most input data at daily resolution, which matches the current daily hydropower scheduling models. In addition to the density of precipitation data, the density of streamflow data is also important for the regional modelling. Pokhrel and Gupta (2011) noted the importance of multiple (high-density) streamflow gauging stations at interior catchments and exploiting the spatial information on soil moisture and evapotranspiration to infer the spatial catchment variability from streamflow hydrographs and for better identification of models.

CONCLUSIONS

We conducted identification of three spatially distributed precipitation-runoff response models based on multi-basin local and regional calibration based on calibration and transfer of both full parameter (FP) and partial parameter (PP) for hourly runoff simulation in mid-Norway. The best performing model structure varies among the catchments, which may be related to uniqueness of catchments. Different best performing models for a catchment were observed for different PM, which is attributed to different sensitivities of the PM to various parts of the hydrograph and different quality of streamflow records on various parts of the hydrograph.

However, models were identifiable based on their overall regional performance and the calibration and validation results indicated that the Kirchmod model performed best. Even though it is not possible to identify a single best performing model structure for the whole catchments in the region, a flexible model and multi-basin based regional modelling framework were found to be necessary for comprehensive identification of reliable model structure, parameterizations and modelling paradigms for specific objectives of prediction and for prediction in ungauged basins (PUB).

The parsimonious ‘top-down’ model (Kirchmod) provided the least parameter uncertainty for the full parameter transfer (FP). However, parsimony could not guarantee parameter identifiability due to the considerable correlations among the calibrated parameters. The deterioration of performance due to fixing of some of the parameters to their regional median or multi-model regional median values for the partial parameter transfer (PP) substantiates the need for preliminary assessment of parameter sensitivity to identify which parameters to transfer to minimize the tradeoffs between performance and parsimony. In addition, marked deterioration of performance measures for the validation period for the calibration objective function used in the present study, which uses streamflow records from all catchments in the region, indicate tradeoffs in regional calibration for parameter transfer in space for PUB and parameter transfer in time. Therefore, temporal validation tests for this type of regional calibration algorithm by using the split-sample scheme is indispensable. Performance of local calibration by using only at-site records for each catchment should be evaluated compared to the local calibration results obtained from the regional calibration methodology used in the present study, which use streamflow records from all catchments in the region.

Dense hourly precipitation gauging networks, which can provide more accurate spatially interpolated precipitation on the 1x1 km^2 grids, are required for improved hourly prediction especially for high flows and for improved identification of hourly P-R models for the region.

In addition, streamflow measurements from dense hydrological gauging networks or spatially distributed observations of rainfall have the potential to improve multi-basin local and regional calibration based identification of models for the hourly prediction.

Acknowledgments

The Center for Environmental Design of Renewable Energy or CEDREN funded the project. We obtained the climate data from the Norwegian Meteorological Institute, Nord Trøndelag Elektrisitetsverk (NTE) and bioforsk We found the streamflow and landuse data from the Norwegian Water Resources and Energy Directorate and the stream networks from a 1:50000 map from the Norwegian Mapping Authority. We wish to thank anonymous reviewers for their comments, which improved the manuscript.

References

Bastola, S. & Murphy, C. 2013 Sensitivity of the performance of a conceptual rainfall-runoff model to the temporal sampling of calibration data. *Hydrology Research* **44** (3), 484-494.

Bell, V. A., and Moore, R. J. 1998 A grid-based distributed flood forecasting model for use with weather radar data: Part 1. Formulation. *Hydrol. Earth. Syst. Sc.*, **2** (2-3), 265-281.

Bergström, S. 1976 Development and application of a conceptual runoff model for Scandinavian catchments. *SMHI Reports RHO*, No. **7**, Norrköping.

Beven, K. J. 2000 Uniqueness of place and process representations in hydrological modelling. *Hydrol. Earth Syst. Sci.* **4**(2) 203-213.

Blasone, R.-S., Madsen H. and Rosbjerg, D. 2007 Parameter estimation in distributed hydrological modelling: comparison of global and local optimisation techniques. *Nordic Hydrology* **38** (4-5), 451-476.

Booij, M.J. 2005 Impact of climate change on river flooding assessed with different spatial model resolutions. *J. Hydrol*. **303**, 176-198.

Box, G. E. P., and Cox, D. R. 1964 An analysis of transformations.” *Journal of the Royal Statistical Society,* Series B **26**, 211-252.

Dunne, T., Black, R. D. 1970a An experimental investigation of runoff production in permeable soils.” *Water Resour. Res*. **6** (2), 478-490.

Dunne, T., and Black, R. D. 1970b Partial area contributions to storm runoff in a small New England watershed.” *Water Resour. Res*. **6** (5), 1296-1311.

Engeland, K. and Steinsland, I. 2013 Probabilistic postprocessing models for flow forecasts for a system of catchments and several lead times. *Water Resour. Res*. **50** (1), 182-197.

Fenicia, F., Kavetski, D. & Savenije, H.H.G. 2011 Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resour. Res.* DOI:10.1029/2011WR010748.

Gelman, A. and Rubin, D. B. 1992 Inference from iterative simulation using multiple sequences. *Stat. Sci. 7,* 457-472.

Gupta, H. V., Sorooshian, S., and Yapo, P. O. 1998 Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information.” *Water Resour. Res.*, **34** (4), 751-763.

Hailegeorgis, Teklu T., and Alfredsen, Knut 2014 Comparative evaluation of performances of different conceptualizations of distributed HBV runoff response routines for prediction of hourly streamflow in boreal mountainous catchments. *Hydrology Research,* in press.

Hailegeorgis, T., Alfredsen, K., Abdella, Y., and Kolberg, S. 2015a Evaluation of different parameterizations of the spatial heterogeneity of subsurface storage capacity for hourly runoff simulation in boreal mountainous watershed. *J. Hydrol*., 522, 522-533.

Hailegeorgis, T., Abdella, Y., Alfredsen, K., and Kolberg, S. 2015b Evaluation of regionalization methods for hourly continuous streamflow simulation using distributed models in boreal catchments. *J. Hydrol. Eng*., accepted manuscript.

Herrnegger, M., Senoner, T. and Nachtnebel, H.-P. 2014 Distributed precipitation corrections in Alpine areas for a real-time flood forecasting system. Geophysical Research Abstracts Vol. **16**, EGU2014-9664.

Hingray, B. Schaefli, B., Mezghani A. & Hamdi Y. 2010 Signature-based model calibration for hydrological prediction in mesoscale Alpine catchments. *Hydrolog. Sci. J*. **55** (6), 1002-1016.

Horton, R.E., 1933 The role of infiltration in the hydrologic cycle. Trans. Am. Geophys. Union 14, 446-460.

Jones, J. 2005 Intersite comparisons of rainfall-runoff processes in *Encyclopedia of hydrological sciences* eds Andersen, M.G., 1839-1854.

Kavetski, D., Fenicia, F. & Clark, M. P. 2011 Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: Insights from an experimental catchment. *Water Resour. Res*. **47**, W05501.

Kirchner, J.W. 2006 Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resour. Res*. **42**, W03S04, doi:10.1029/2005WR004362.

Kirchner, J.W. 2009 Catchments as simple dynamical systems: catchment characterization, rainfall–runoff modeling, and doing hydrology backward. *Water Resour. Res*. **45**, W02429.

Klemeś, V. 1986 Operational testing of hydrological simulation models. *Hydrolog. Sci. J.* **31**, 13-24.

Kolberg, S. A., and Gottschalk, L. 2006 Updating of snow depletion curve with remote sensing data. *Hydrol. Process*. **20** (11), 2363-2380.

Kolberg, S. A., and Bruland, O. 2012 “ENKI-An Open Source environmental modelling platform.” *Geophys. Res. Abstracts* **14**, EGU2012-13630, EGU General Assembly.

Lee, H., McIntyre, N. Wheater, H., and Young, A. 2005 Selection of conceptual models for regionalization of the rainfall-runoff relationship. *J. Hydrol*. **312**, 125-147.

Li, H., Beldring, S., and Xu, C-Y. 2014 Implementation and testing of routing algorithms in the distributed HBV model for mountainous catchments. *Hydrol. Res*. 45(3), 322-333.

Libelli, M. S., Beck, M. B., Brunner, P., Croke, B. Guillaume J., Jakeman, A. Jakeman, J. Keesman, K. Stigter, H. Practical identifiability analysis of environmental models. International Environmental Modelling and Software Society (iEMSs) 7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA, Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.)

Madsen, H. 2003 Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. *Adv. Wat. Res*. **26**, 205-216.

Moine, N. L., Andre´assian, V., Perrin, C. and Michel, C. 2007 How can rainfall-runoff models handle intercatchment groundwater flows? Theoretical study based on 1040 French catchments. *Water Resour. Res*. **43** (6), W06428, doi:10.1029/2006WR005608.

Moreda, F., Koren, V., Zhang, Z., Reed, S. & Smith, M. 2006 Parameterization of distributed hydrological models: learning from the experiences of lumped modeling. *J. Hydrol.* **320**, 218-237.

Nash, J. E., and Sutcliffe, J. V. 1970 River flow forecasting through conceptual models, I. A discussion of principles. *J. Hydrol*. **10**, 228-290.

Nicolle, P. Pushpalatha, R., Perrin, C., François, D. Thiéry, D., Mathevet, T. Lay, M. L., Besson, F., Soubeyroux, J.-M., Viel, C., Regimbeau, F., Andréassian, V., Maugis, P., Augeard, B. and Morice, E. Benchmarking hydrological models for low-flow simulation and forecasting on French catchments. *Hydrol. Earth. Syst. Sci*. **18**, 2829-2857.

Orellana, B., Pechlivanidis, I.G., McIntyre, N., Wheater H.S. and Wagener, T. 2008. A Toolbox for the Identification of Parsimonious Semi-Distributed Rainfall-Runoff Models: Application to the Upper Lee Catchment. Proceedings of iEMSs 2008: International Congress on Environmental Modelling and Software.

Oudin, L., Andre´assian, V., Perrin, C., Michel, C., and Le, M., N. 2008 Spatial proximity, physical similarity, regression and un-gaged catchments: A comparison of regionalization approaches based on 913 French catchments. *Water Resour. Res*. **44**, W03413, Doi: 10.1029/2007WR006240.

Oudin, L., Kay, A., Andre´assian, V., and Perrin, C. 2010 Are seemingly physically similar catchments truly hydrologically similar?. *Water Resour. Res*. **46**, W11558, doi: 10.1029/2009WR008887.

Perrin, C., Michel, C. & Andréassian, V. 2001 Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *J. Hydrol.* **242**, 275-301.

Pokhrel, P. & Gupta, H. V. 2011 On the ability to infer spatial catchment variability using streamflow hydrographs. *Water Resour. Res*. **47**, W08534, doi:10.1029/2010WR009873.

Priestley, C.H.B., and Taylor, R.J. 1972 On the assessment of surface heat flux and evaporation using large-scale parameters. *Mon. Weather Rev.* **100**, 81-82.

Refsgaard, J.C., 1997 Validation and Intercomparison of different updating procedures for real-time forecasting. Nord. Hydrol. **28**, 65-84.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., and Carrillo, G. 2011 Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. *Hydrol. Earth Syst. Sci*, **15**, 2895-2911.

Schoups, G. & Vrugt, J.A. 2010 A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. *Water Resour. Res.* **46**, W10531, doi:10.1029/2009WR008933.

Sevruk, B. 1983 Correction of measured precipitation in the Alps using the water equivalent of new snow. *Nordic hydrology* **1983**, 49-58.

Smith, T.J. and Marshall, L.A. 2009 A conceptual precipitation-runoff modeling suite: Model selection, calibration and predictive uncertainty assessment. 18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July.

Teuling, J., Lehner, I., Kirchner, J. W., and Seneviratne, S. I. 2010 “Catchments as simple dynamical systems: Experience from a Swiss prealpine catchment. *Water Resour. Res*. **46**, W10502, doi:10.1029/2009WR008777.

Uhlenbrook, S., Seibert, J., Leibundgut, C. & Rohde, A. 1999 Prediction uncertainty of conceptual rainfall–runoff models caused by problems in identifying model parameters and structures. *Hydrol. Sci. J*. **44** (5), 779–797.

Viviroli, D., Zappa, M., Schwankbeck, J., Gurtz, J., Weingartner, R. 2009a Continuous simulation for flood estimation in ungauged mesoscale catchments of Switzerland-Part I: Modelling framework and calibration results. *J. Hydrol.* **377**, 191-207.

Viviroli, D., Mittelbach, H., Gurtz, J., Weingartner, R. 2009b Continuous simulation for flood estimation in ungauged mesoscale catchments of Switzerland-Part II: Parameter regionalization and flood estimation results. *J. Hydrol*. **377**, 208-225.

Vrugt J.A., Ter Braak C.J.F., Diks C.G.H., Robinson B.A., Hyman J.M., and Higdon D. 2009 Accelerating Markov Chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *Journal of Nonlinear Sciences and Numerical Simulation*, **10** (3), 273-290.

Wagener, T., Boyle, D. P., Lees M. J., Wheater, H. S., Gupta, H. V. & Sorooshian, S. 2001 A framework for the development and application of hydrological models. Hydrol. Earth Syst. Sci. **5** (1), 13–26.

Wagener, T. 2003 Evaluation of catchment models. *Hydrol. Process.* **17**, 3375-3378.

Wagener, T. & McIntyre N. 2005 Identification of rainfall–runoff models for operational applications Hydrological Sciences, **50** (5), 735-751.

Willems, P. 2009 A time series tool to support the multi-criteria performance evaluation of rainfall-runoff models. *Environmental modelling and software* **24**, 311-321.

Wrede, S., Seibert, J., and Uhlenbrook, S. 2013 Distributed conceptual modelling in a Swedish lowland catchment: a multi-criteria model assessment. *Hydrol. Res*. **44** (2), 318-333.

Yang, D. 1999 Bias correction of daily precipitation measurements for Greenland. *Journal of Geophysical Research* **104**, 6171-6181.

Zięba, A. 2010 Effective number of observations and unbiased estimators of variance for autocorrelated data an overview. *Metrol. Meas. Syst*. **XVII** (1), 3-16.