Klaus Vormoor | Maik Heistermann | Axel Bronstert | Deborah Lawrence

Hydrological model parameter (in) stability

“crash testing” the HBV model under contrasting flood seasonality conditions

Suggested citation referring to the original publication:
Hydrological Sciences Journal 63 (2018) 7, pp. 991-1007
DOI http://dx.doi.org/10.1080/02626667.2018.1466056
ISSN (print) 0262-6667
ISSN (online) 2150-3435

HYDROLOGICAL SCIENCES JOURNAL
2018, VOL. 63, NO. 7, 991–1007
https://doi.org/10.1080/02626667.2018.1466056

OPEN ACCESS Check for updates

Hydrological model parameter (in)stability – "crash testing" the HBV model under contrasting flood seasonality conditions

Klaus Vormoor[a], Maik Heistermann[a], Axel Bronstert[a] and Deborah Lawrence[b]

[a]Institute of Earth and Environmental Science, University of Potsdam, Potsdam, Germany; [b]Hydrology Department, Norwegian Water Resources and Energy Directorate (NVE), Oslo, Norway

ABSTRACT
This paper investigates the transferability of calibrated HBV model parameters under stable and contrasting conditions in terms of flood seasonality and flood generating processes (FGP) in five Norwegian catchments with mixed snowmelt/rainfall regimes. We apply a series of generalized (differential) split-sample tests using a 6-year moving window over (i) the entire runoff observation periods, and (ii) two subsets of runoff observations distinguished by the seasonal occurrence of annual maximum floods during either spring or autumn. The results indicate a general model performance loss due to the transfer of calibrated parameters to independent validation periods of −5 to −17%, on average. However, there is no indication that contrasting flood seasonality exacerbates performance losses, which contradicts the assumption that optimized parameter sets for snowmelt-dominated floods (during spring) perform particularly poorly on validation periods with rainfall-dominated floods (during autumn) and *vice versa*.

ARTICLE HISTORY
Received 8 September 2017
Accepted 23 March 2018

EDITOR
A. Castellarin

ASSOCIATE EDITOR
C. Perrin

KEYWORDS
hydrological modelling; flood seasonality; differential split-sample test; flood generating processes; Nordic catchments

1 Introduction

Climate change impact assessments are usually based on multi-model/multi-parameter ensembles which lead to a cascade of uncertainty (Wilby and Dessai 2010). The majority of studies that try to attribute uncertainties to the different steps within these ensembles usually indicate a larger contribution from the climate models and the emission scenarios to the overall uncertainty, while the contribution of hydrological model uncertainty tends to be relatively minimal (e.g. Kay *et al.* 2009, Dobler *et al.* 2012, Addor *et al.* 2014). However, there is increasing concern that the hydrological models used in climate change impact assessments are not perfectly suited to dealing with changes in the hydro-meteorological conditions and their related catchment processes due to the conceptual representation and parameterization of the hydrological system (Thirel *et al.* 2015a).

Hydrological model uncertainty emerges both from the model structure and from the parameterization of the model, and from data uncertainty (Refsgaard *et al.* 2006, Matott *et al.* 2009). From the perspective of climate change, we should be particularly interested in the reliability of hydrological model simulations if applied under transient hydro-climatological boundary conditions as potentially imposed by climate change (Bronstert 2004, Blöschl and Montanari 2010). From that perspective, we need to thoroughly verify the transferability of both model structures and calibrated model parameters under transient hydro-climatological conditions. Surely, the performance of process-based models should be robust against changing conditions. However, model structures can become invalid if the dominant processes fundamentally change. Hydrological model parameters related to a specific process may become invalid if the process is not well represented during the calibration period. That is, model calibration against observation data accounts for the specific climate characteristics found in the data period. In turn, different calibration periods showing contrasting hydro-meteorological conditions may already yield different best-fit parameter sets, highlighting a lack of parameter robustness over time (Wagener *et al.* 2003, Merz *et al.* 2011).

In climate change impact analysis, non-stationarity in climate conditions is implicitly considered by the future climate projections, and the time transferability of hydrological parameters is a critical issue that has gained a lot of research interest in recent years, e.g. the Special Issue of *Hydrological Sciences Journal* (Vol. 60, Issue 7–8), Thirel *et al.* (2015b). For testing the robustness of hydrological model simulations under contrasting hydro-meteorological conditions, Klemeš (1986) proposed the differential split-sample test (DSST) in which a

CONTACT Klaus Vormoor kvormoor@uni-potsdam.de

hydrological model is calibrated and validated on two (or more) hydro-meteorologically contrasting periods. Refsgaard *et al.* (2013) recently recommend performing DSSTs to generate further confidence in the hydrological models used for climate change impact projections. Examples of the application of such tests can be found in the studies by Seibert (2003), Vaze *et al.* (2010), Merz *et al.* (2011), Coron *et al.* (2012), and Brigode *et al.* (2013). Most of these authors found a considerable decrease in model performance after transferring calibrated parameter sets between climatologically contrasting periods.

Seibert (2003) calibrated the HBV model in four Swedish catchments for years with lower runoff peaks and tested the calibrated model for years with higher peaks, finding a decrease in model performance. More recently, Vaze *et al.* (2010) applied the DSST for four different conceptual hydrological models in 63 Australian catchments and found that the models calibrated under wetter conditions performed worse for dryer periods than *vice versa*. Coron *et al.* (2012) introduced generalized split-sample tests, which systematically test all possible combinations of calibration–validation periods using a 10-year moving window over the observation time period. Using three hydrological models in 216 catchments in southeast Australia, they also found systematic over- and underestimation of average runoff volumes when transferring calibrated parameters from wetter to drier conditions and *vice versa*. Merz *et al.* (2011) calibrated a conceptual hydrological model for six consecutive 5-year periods for 273 catchments in Austria. They found that the parameters controlling snow dynamics and soil moisture processes depend significantly on the hydro-climatological conditions of the calibration period, which leads to notable biases in high flows especially in snow-affected catchments. Finally, Brigode *et al.* (2013) found that two hydrological models calibrated for 63 catchments in France were sensitive to climatologically contrasted calibration sub-periods (dry *vs* wet) and that this lack of model robustness has a stronger impact on the uncertainty of hydrological projections of future streamflow as compared to the use of several multiple parameter sets. Fowler *et al.* (2016), however, indicate that the reason for failing the DSST is often due to insufficient model calibration techniques, rather than to the models themselves, which can lead to a false negative impression of the capabilities of conceptual hydrological models under changing climate conditions.

Surely, there is no general solution for ensuring the robustness of calibrated parameter sets under transient conditions: transferability always needs to be verified for a specific setting characterized by the region, its scale and its dominant hydrological processes, the transient properties of hydro-climatic and other environmental boundary conditions, and, of course, the hydrological model structure and the observations underlying the calibration (Andréassian *et al.*, 2009). Nevertheless, climate change impact research should aim to characterize parameter transferability for settings characterized by typical combinations of regions, dominant processes, the transient properties and the underlying model type. Until now, the majority of such studies have focused on climatologically contrasting periods in terms of dry *vs* wet and warm *vs* cold conditions. None so far has explicitly studied the robustness of calibrated hydrological model parameters under contrasting conditions in terms of flood seasonality and flood generating processes.

For mountainous and northern regions, where the role of snowmelt *vs* rainfall as the most important flood generating processes is highly relevant for the seasonal flood regimes, the impacts of climate change on runoff and flooding are expected to be more severe than in other regions (Viviroli *et al.* 2011). Climate change impact studies for Scandinavian catchments with mixed snowmelt/rainfall regimes (e.g. Arheimer and Lindström 2015, Vormoor *et al.* 2015, 2016) have indicated a temperature-driven shift in flood seasonality from spring to autumn and early winter, with an increasing relevance of rainfall as a dominant flood generating process. In this sense, it is crucial to investigate the transferability of calibrated hydrological model parameters under contrasting flood seasonality conditions, since this may aid in selecting reasonable calibration periods for the optimization of hydrological model parameter sets that will be used in climate change impact studies for this particular type of setting: Nordic catchments with a complex flood seasonality involving flood generation by both rainfall and seasonal snowmelt.

From this perspective, we have developed a testing protocol that adapts the operational testing schemes for the temporal transferability of hydrological model parameters introduced by Klemeš (1986), extended by the generalization of these tests as proposed by Coron *et al.* (2012). The testing protocol has been applied in five Norwegian catchments with mixed snowmelt/rainfall regimes and flood peaks occurring either during spring or during autumn and early winter. Contrasting flood seasonality has been defined by the seasonal occurrence of the annual maximum floods (AMFs). The testing protocol allows for analysing model performance losses when transferring calibrated hydrological model parameters under stationary and non-stationary flood seasonality conditions, and the following two particular research questions are addressed by this study: (1) How large is the general hydrological model performance loss due to the transfer of calibrated parameter sets to independent validation periods with similar flood seasonality? and (2) Do performance losses increase if we transfer calibrated hydrological model parameters to validation periods with a contrasting flood seasonality reflecting a difference in the role of snowmelt and rainfall as dominant flood generating processes?

2 Material and methods

2.1 Study catchments: hydro-meteorological conditions and flood regimes

The testing protocol (described in Section 2.3) has been applied in five Norwegian catchments: Kråkfoss, Austenå, Bulken, Jogla and Fustvatn. All catchments are characterized by a mixed snowmelt/rainfall regime so that annual flood peaks occur either during spring and early summer or during autumn and early winter. Figure 1 shows the locations of the five catchments and the annual hydrographs for their respective data periods. Detailed catchment characteristics are given in Table 1.

The catchments considered are included in the Norwegian benchmark dataset for climate change studies and are tested for their suitability for daily analysis of flood discharge (Fleig *et al.* 2013). The five catchments vary in size from 31 km^2 (Jogla) to 1092 km^2 (Bulken) (Table 1). Dominant land cover types are either exposed (crystalline) bedrock with sparse vegetation above the tree line (i.e. the highest elevated catchments, Bulken: 54%, Jogla: 92%) or boreal forest (Kråkfoss: 76%, Austenå: 62%, Fustvatn: 38%). Continuous daily streamflow measurements are provided by the Norwegian Water Resources and Energy Directorate covering 41 years for Jogla (1973–2014), 47 years for Kråkfoss (1967–2014) and 53 years for Austanå, Bulken and Fustvatn (1961–2014). Daily

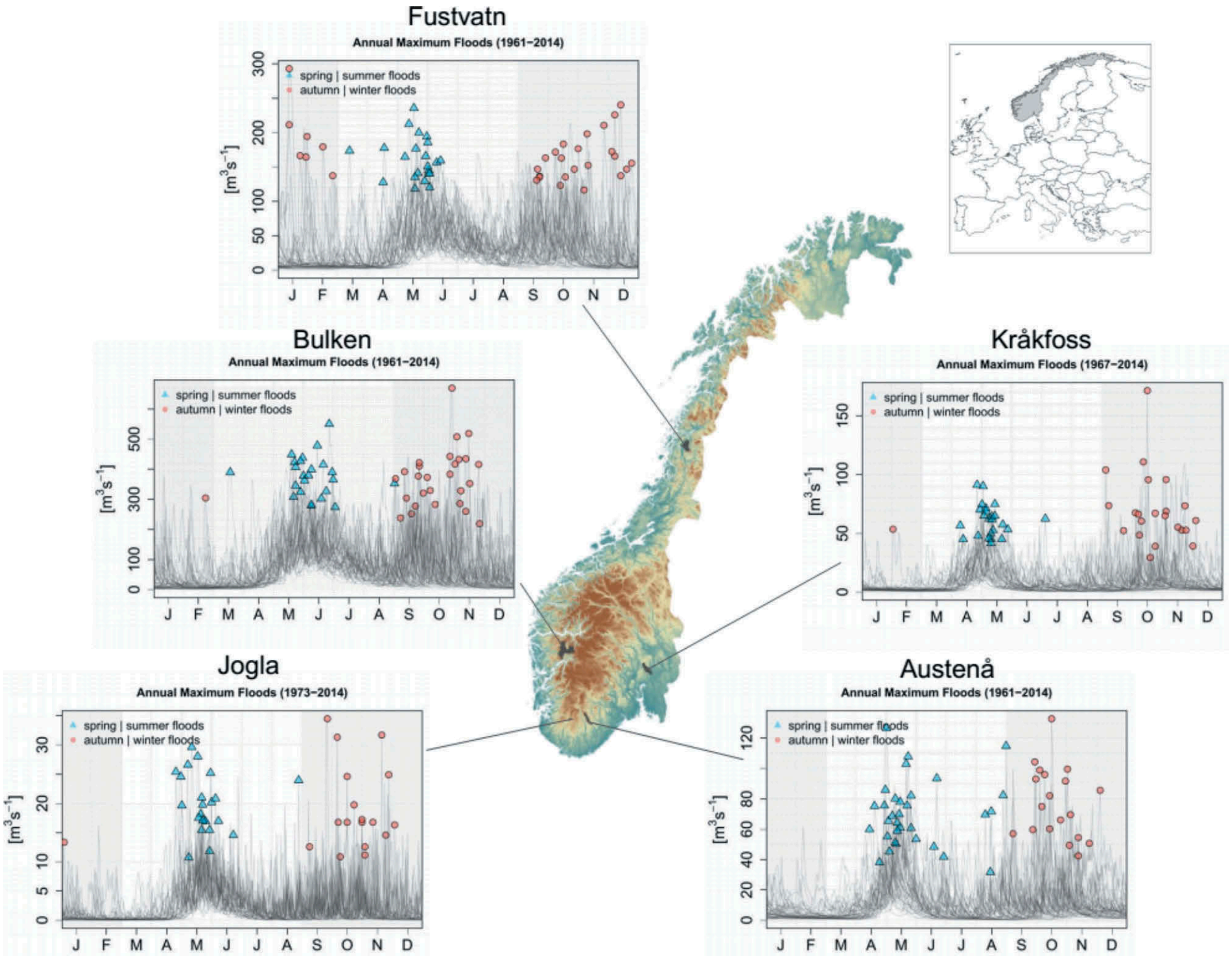

Figure 1. Location of the five study catchments and summary plots of annual hydrographs for the catchment-specific length of runoff observation time series (length given in the header of each plot). The hydrographs also show the seasonal occurrence of the annual maximum floods: snowmelt-dominated spring/summer floods between March and August; rainfall-dominated autumn/ winter floods between September and February.

Table 1. Main characteristics of the five study catchments.

Catchment property	Kråkfoss	Austenå	Bulken	Jogla	Fustvatn
Area (km^2)	433	276	1092	31	526
Median elevation (m a.s.l.)	445	738	867	1002	436
Elevation range (m a.s.l.)	105–803	228–1146	47–1602	612–1194	39–1530
Mean annual *T* (°C) [1981–2010]	3.3	2.9	2.3	1.5	2.2
Average annual *P* (mm) [1981–2010]	1022	1897	2819	2965	2363
Average annual *Q* (mm) [1981–2010]	605	1137	2044	2057	1945
Frequency of spring *vs* autumn events (count)	24/23	32/24	24/29	23/19	22/30
Median magnitude of spring *vs* autumn events (m^3 s^{-1})	62/65	39/73	379/369	20/17	158/163
Land cover, % lake	4	12	4	3	6
Land cover, % glacier	0	0	< 1	0	< 1
Land cover, % forest	76	62	32	3	38
Land cover, % marsh and bog	5	6	2	1	5
Land cover, % sparse vegetation above treeline	0	20	54	92	37
Anthropogenic land use (%)	11.2	< 1	4	0	0

temperature and precipitation data are inferred from nationwide gridded maps with a 1 km × 1 km spatial resolution and a temporal coverage from 1 September 1957 until the present (e.g. Mohr and Tveito 2008; seNorge maps provided to the public at www.senorge.no). For the period 1981–2010, mean annual precipitation and mean annual runoff varied between 1022 and 605 mm (Kråkfoss), and 2965 and 2057 mm (Jogla), respectively. Both variables show gradients in west–east and altitudinal directions, while precipitation and runoff depths increase towards the west coast and with higher altitude. Mean annual temperature varied between 1.5°C (Jogla) and 3.3°C (Kråkfoss) and runoff coefficients generally tend to be high due to low evapotranspiration.

The annual hydrographs shown in Figure 1 illustrate the mixed snowmelt/rainfall regimes of the catchments with spring and summer floods occurring during March–August and autumn and early winter floods occurring during September–February. Snowmelt plays a remarkable role for the temporally clustered peak flows during March–May (Kråkfoss), April–June (Austenå, Jogla, Fustvatn) and May–July (Bulken). Rainfall is the dominant flood generating process for the events occurring during autumn and early winter. Note, however, that some events do not immediately reflect flood generating processes that are associated with their seasonal occurrence. We cannot exclude, for instance, that a phase of early snowmelt has contributed to peak discharge for events in late winter (e.g. January–February at Fustvatn, Bulken, Kråkfoss). At the same time, we can exclude that snowmelt has contributed to event discharge in late summer (e.g. July–August at Bulken, Jogla, Kråkfoss, Austenå). Therefore, we have excluded the years with these events from the analyses (see Section 2.3).

The annual hydrographs further illustrate differences between the catchments regarding the magnitude and frequency of spring/summer floods and autumn/winter floods, which are summarized in Table 1 (rows 7–8). While the frequency of AMFs during spring is (slightly) larger than those during autumn at Kråkfoss (24 *vs* 23 events), Austenå (32 *vs* 24) and Jogla (22 *vs* 19), it is the other way around at Bulken (23 *vs* 28) and Fustvatn (22 *vs* 30). Regarding the median magnitude of AMF seasonality, spring events are slightly larger at Bulken (379 *vs* 369 m^3 s^{-1}) and Jogla (20 *vs* 17 m^3 s^{-1}), whereas autumn events are slightly larger at Kråkfoss (65 *vs* 62 m^3 s^{-1}), Austenå (73 *vs* 39 m^3 s^{-1}), and Fustvatn (163 *vs* 158 m^3 s^{-1}).

2.2 The HBV model and its calibration

The hydrological model that is tested for its transferability under contrasting flood seasonality is the "Nordic" version (Sælthun 1996) of the HBV model (Bergström 1976, 1995). The HBV model is a conceptual lumped precipitation–runoff model that simulates streamflow using temperature and precipitation as inputs. The model has been applied widely in the Nordic countries (e.g. Lindström *et al.* 1997) and it provides a suitable conceptual representation of the dominant runoff generating processes, while it does not impose excessive data requirements. Evapotranspiration is estimated by the model using the temperature index method, rather than using monthly values as model input. The model is applied on a daily time step and consists of three basic subroutines: (a) a snow routine, (b) a soil moisture routine, and (c) a runoff response routine. Table 2 lists all model parameters including a short description and the parameter ranges as they are considered for model calibration. More detailed descriptions of the model structure can be found in Sælthun (1996) (for the model version used in this paper), Bergström (1995), Lindström *et al.* (1997) and Seibert

Table 2. HBV parameter ranges used in the DDS optimization.

HBV parameter	Description	Range considered	Unit
Snow routine			
CX	Degree day correction factor	1.0–5.0	mm d^{-1} $°C^{-1}$
PGRAD	Precipitation lapse rate	0.0–0.1	100 m^{-1}
PKORR	Precipitation correction factor	0.8–3.0	-
SKORR	Snowfall correction factor	1.0–3.0	-
TS	Threshold temperature for snowmelt	−1.0 to 2.0	°C
TX	Threshold temperature for rain/snow	−1.0 to 2.0	°C
TTGRAD	Temperature lapse rate – clear days	−1.0 to −0.5	°C 100 m^{-1}
TVGRAD	Temperature lapse rate during precipitation	−0.7 to −0.3	°C 100 m^{-1}
Soil moisture routine			
BETA	Soil moisture parameter – shape coefficient	1.0–4.0	-
FC	Field capacity – maximum storage in soil box	50.0–500.0	mm
Response routine			
KLZ	Recession constant – lower zone	0.001–0.1	d^{-1}
KUZ1	Recession constant – upper zone 1	0.01–1.0	d^{-1}
KUZ2	Recession constant – upper zone 2	0.1–1.0	d^{-1}
PERC	Percolation – upper to lower zone	0.5–2.0	mm d^{-1}
UZ1	Threshold for quick runoff	10.0–100.0	mm

and Vis (2012) (for general information about the HBV model structure).

For model calibration, we apply the dynamically dimensioned search (DDS) (Tolson and Shoemaker 2007), which is a global optimization algorithm for the calibration of multi-parameter models. DDS is based on a neighbourhood search, and within a user-specified maximum number of model evaluations it automatically scales the search to find best-possible solutions. In this study, we set the maximum number of model evaluations to 800, which provided a good balance of calibration performance and computational cost. A modified version of the Nash-Sutcliffe efficiency (NSE_w) was used as the objective function so as to put even more emphasis on high-flow events than the regular NSE would do (e.g. Ott *et al.* 2013, Vormoor *et al.* 2015):

$$NSE_w = 1 - \frac{\sum_{i=1}^{n}\left[Q_{obs}^{i}\left(Q_{obs}^{i} - Q_{sim}^{i}\right)^2\right]}{\sum_{i=1}^{n}\left[Q_{obs}^{i}\left(Q_{obs}^{i} - \overline{Q_{obs}}\right)^2\right]} \qquad (1)$$

where Q_{obs} represents the observed discharges, Q_{sim} represents the modelled discharges, and n is the number of daily time steps. The squared errors in the numerator and denominator are weighted by the observed discharge. A mismatch between high observed and simulated discharges is, therefore, penalized proportionally to the observed discharge value. That way, we train the HBV model parameters to particularly represent the processes that lead to high-flow events, be it during spring and early summer or during autumn and early winter.

For one of the study catchments (Kråkfoss), we compared the differences in matching the magnitude of AMFs between model simulations that are based on calibrated parameter sets using the modified NSE_w criterion and the regular NSE criterion, respectively, as the objective function. For both NSE variants, we calibrated and evaluated the model on 6-year periods over a moving window moved forward by 1 year over the entire runoff observation time period (see Section 2.3). This comparison shows that simulations using calibrated parameter sets based on the NSE_w outperform those based on the regular NSE (mean absolute errors on median, NSE_w: 1.56 m^3 s^{-1} *vs* NSE: 2.63 m^3 s^{-1}).

Furthermore, we tested whether the calibration of the HBV model is sensitive to the AMF using the DDS optimization algorithm and the NSE_w as objective function. To that end, we calibrated the model on one 6-year period (1986–1991) preceded by a 5-year spin-up period in two different ways: (a) on the entire period including all AMFs, and (b) on the entire period after removing the AMFs (peak discharge plus one day concentration time and an event-specific number of days of runoff recession below the 75th streamflow percentile – 2–3 days in this case). This test shows that the removal of AMFs modifies the calibration considerably (NSE_w 0.80 *vs* 0.62) and leads to differences in the calibrated parameter sets.

2.3 Testing the transferability of hydrological model parameters under similar and contrasting flood seasonality conditions

Klemeš (1986) introduced four hierarchical testing methods to describe how well a hydrological model can be transferred in space and time. Since we are interested in the temporal transferability of calibrated hydrological model parameters, the split-sample test (SST) and the differential split-sample test (DSST) are relevant for our purpose. The SST is a well-established calibration–validation procedure which assumes stationary hydro-meteorological conditions between calibration and validation periods. The DSST, in contrast, assumes non-stationary conditions for testing model simulations under potentially transient conditions. In this study, we apply both principles: (a) SSTs for quantifying the general validity of calibrated hydrological model parameters for the simulations of high flows for independent time periods with similar flood seasonality; and (b) DSSTs for testing the ability of the hydrological models to predict high flows under contrasting conditions in terms of flood seasonality.

Since the number of possible transfer tests is usually small due to the limited temporal coverage of runoff observation data, Coron *et al.* (2012) proposed a generalization of the standard SST and DSST to fully utilize the available runoff observations, and referred to this procedure as the generalized split-sample test (GSST). The generalization is achieved by creating sub-periods of equal length using a moving window which is moved forward by 1 year over the entire runoff observation time period. For each sub-period, the hydrological model is calibrated, and the optimized parameter sets are validated on all possible (independent) sub-periods. In addition to fully utilizing the available observation data, the GSST has the advantage that no prior knowledge regarding transient boundary conditions is needed since the approach tests the hydrological model in as many varied climate configurations as covered by the data. Analyses of the contributions of, e.g., hydro-meteorological factors to performance losses can be undertaken posterior to running the GSST.

We are, however, able to determine – *a priori* and purely data-driven – the contrasting conditions (or processes) that have caused the AMFs if we expect that snowmelt is the dominant flood generating process in years with AMFs occurring during spring, and rainfall is the dominant flood generating process in years with AMFs occurring during autumn and early winter (Vormoor *et al.* 2015). Since the test calibration has proven to be sensitive to the AMF (see Section 2.1), we can assume that the calibration will account for the certain process dominance that has caused annual peak flow discharge either during spring or autumn. Therefore, we can apply the generalized scheme of the SST and DSST both to the entire time series and to two contrasting blocks that cover the years in which AMFs occur either during spring or autumn and early winter to distinguish these two types of process dominance. Note that the determination of the two blocks with contrasting flood seasonality is based purely on the seasonal occurrence of annual floods. This excludes neither that a second prominent high-flow event occurs during the contrasted season, nor that the basic hydro-meteorological conditions fundamentally differ between these years. We further elaborate this in Section 3.1. In the following, we describe the individual steps of the testing protocol as it is applied in this study with reference to the example catchment Kråkfoss. The individual steps are illustrated in Figure 2.

Step 1: Creation of sub-periods for the calibration and validation of the hydrological model using a moving window as proposed by Coron *et al.* (2012). A time window of 6 years is moved forward stepwise by 1 year over the entire runoff observations time series, which results in 43 sub-periods. We choose a window length of 6 years to ensure that (i) in each sub-period and for each catchment both types of AMFs are present, and (ii) a sufficient number of independent periods are left for model validation. For each of these sub-periods, the HBV model is calibrated using the DDS global optimization algorithm. In contrast to the procedure described by Coron *et al.* (2012), the resulting best-fit parameter sets are not validated on every independent sub-period only, but on all other 42 sub-periods including those with overlapping years. Therefore, we are able to estimate systematic model performance losses with a decreasing overlap in the years of the calibration sub-periods. Note that for both the calibration and validation, a 5-year spin-up period, which is made up of five times repeating the year prior to the first year of each sub-period, is applied to estimate the system states at the beginning of the simulations.

Steps 2–3: The same procedure is then applied for the two subsets of the runoff observation time series that have been separated based on the seasonal occurrence of AMFs during spring and autumn, respectively. For each subset, this results in 18 sub-periods used for the calibration and validation of the HBV model. We are, therefore, able to estimate (i) the differences in the calibration results both between the two subsets and between the two subsets as compared to the results of Step 1, and (ii) the general validity of the HBV model under similar flood seasonality conditions with similar flood generating processes. As in Step 1, 5-year spin-up periods are applied to estimate the initial system states for each sub-period. To estimate the initial states for each year within the sequence of discontinuous years per sub-period (as indicated in Fig. 2), the model is run over all continuous years covered by a sub-period but only the relevant years are used in the evaluation.

Step 4: Having estimated 18 best-fit parameter sets for the subsets with dominant spring floods and autumn floods, respectively, we are able to test these parameter

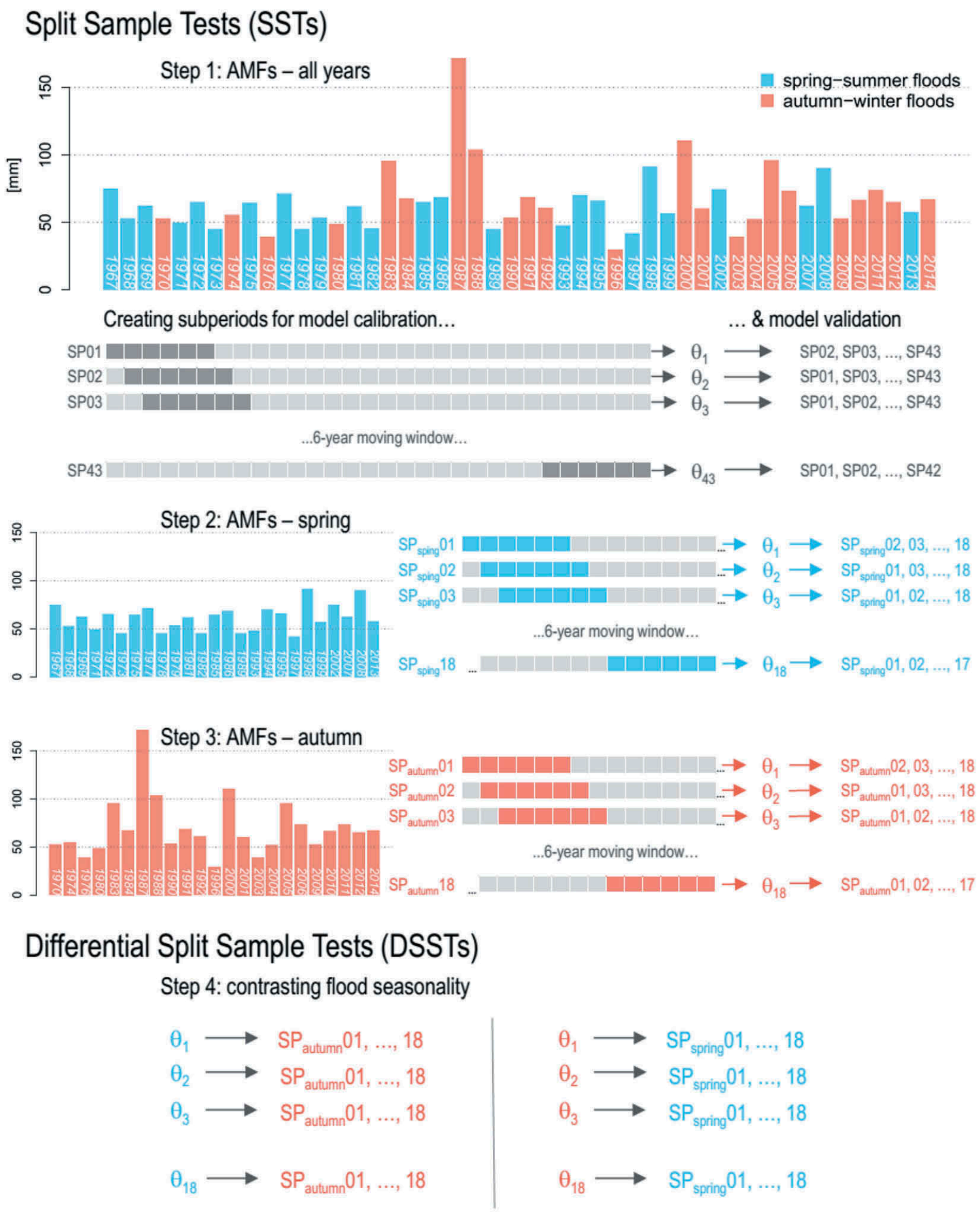
Split Sample Tests (SSTs)
Step 1: AMFs – all years
spring–summer floods
autumn–winter floods
[mm]
0
50
100
150
1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
Creating subperiods for model calibration…
… & model validation
SP01 → θ1 → SP02, SP03, ..., SP43
SP02 → θ2 → SP01, SP03, ..., SP43
SP03 → θ3 → SP01, SP02, ..., SP43
...6-year moving window...
SP43 → θ43 → SP01, SP02, ..., SP42
Step 2: AMFs – spring
SPsping01 → θ1 → SPspring02, 03, ..., 18
SPsping02 → θ2 → SPspring01, 03, ..., 18
SPsping03 → θ3 → SPspring01, 02, ..., 18
...6-year moving window...
SPsping18 → θ18 → SPspring01, 02, ..., 17
Step 3: AMFs – autumn
SPautumn01 → θ1 → SPautumn02, 03, ..., 18
SPautumn02 → θ2 → SPautumn01, 03, ..., 18
SPautumn03 → θ3 → SPautumn01, 02, ..., 18
...6-year moving window...
SPautumn18 → θ18 → SPautumn01, 02, ..., 17
Differential Split Sample Tests (DSSTs)
Step 4: contrasting flood seasonality
θ1 → SPautumn01, ..., 18
θ2 → SPautumn01, ..., 18
θ3 → SPautumn01, ..., 18
θ18 → SPautumn01, ..., 18
θ1 → SPspring01, ..., 18
θ2 → SPspring01, ..., 18
θ3 → SPspring01, ..., 18
θ18 → SPspring01, ..., 18

Figure 2. Illustration of the methodology used for testing the transferability of calibrated hydrological model parameters based on the principles of the split-sample test (Step 1) and the differential split-sample test (Step 2) (Klemeš 1986). The generalization of those schemes, i.e. the 6-year moving window, is adapted from Coron *et al.* (2012). The numbers in this illustration stem from the study catchment Kråkfoss. Note that a 5-year spin-up period is applied for each sub-period in both calibration and validation modes. Moreover, for Steps 2–4, the model is run over continuous time periods and only the relevant years are used for model evaluation.

sets on the sub-periods of their contrasting groups. This enables us to study the performance loss of the HBV model due to the transfer of calibrated model parameters under contrasting AMF seasonality. If the contrasting conditions in AMF seasonality have an impact on the robustness of the hydrological model parameters, the performance loss should be larger than the performance loss within the individual subsets (Steps 2–3). The initial system states are estimated as described in Steps 2–3.

As mentioned in Section 2.1, years with AMFs occurring during late summer and late winter have been excluded from the analysis since these AMFs do not necessarily reflect the typical flood generating processes that are associated with their season of occurrence (i.e. snowmelt *vs* rainfall). This leads to the fact that, for each of the catchments Kråkfoss, Bulken and Jogla 2 years, for Austenå 5 years, and for Fustvatn 7 years are excluded from the analysis. Another possible concern is the relatively small length of the sub-periods used for model calibration (i.e. 6 years). However, Brigode *et al.* (2013), for instance, have shown that model calibration on even shorter time periods (3 years) can provide reasonable results.

2.4 Estimating the transferability of calibrated parameter sets

Performances losses caused by the transfer of optimized parameter sets to (in)dependent validation periods are estimated by the model robustness criterion (MRC) proposed by Coron *et al.* (2012):

$$\mathrm{MRC}_{D\rightarrow R}=\frac{\varepsilon_{D\rightarrow R}}{\varepsilon_{R\rightarrow R}}-1 \tag{2}$$

where ε is the performance criterion (in this case the NSE_w) to be maximized during the calibration. The idea behind the MRC is that a parameter set that has been calibrated for a certain period acts as a "donor" parameter set (D) for a model application on a validation ("receiver", R) period. The quality of this donor parameter set is assessed relative to the quality of the parameter set that has been optimized for the receiver period. Consequently, MRC varies depending on the ability of the parameter set optimized on the period D to simulate discharge, and particularly high flows, on the period R. That is, the MRC is zero if the parameter set optimized on period D performs as well as the parameters calibrated on the period R, while negative values indicate a decrease in the suitability of the parameter sets for the period R. The more negative the MRC is, the less transferable is the parameter set, and a MRC-value of, say −0.1 corresponds to a 10% performance loss. A positive MRC estimate, on the other hand, would mean that the parameter set from the period D performs better on the period R than that which has been optimized for R. Such cases may indicate problems with the calibration of hydrological model parameters on the receiver period, i.e. the global optimum could not be found properly.

3 Results

3.1 The dominance of annual maximum floods per calibration period

As indicated previously, the determination of the two blocks with contrasting flood seasonality is based purely on the seasonal occurrence of the AMFs, which does not exclude that a second prominent high-flow event occurs during the contrasting season. Thus, the level of dominance of AMFs during spring and autumn, respectively, varies across the years and the different periods used for calibration and validation. In this regard, Figure 3 shows the ranked levels of dominance of AMFs for the 6-year calibration periods with AMFs occurring during either spring or autumn. The level of dominance indicates to what degree high-flow discharge from the contrasting season reduces the dominance of AMFs during spring or autumn for a certain calibration period. That is, a dominance of AMF seasonality of 100% would mean that there are no second prominent high-flow events during the contrasting season, and a dominance of AMF seasonality of 0% would mean that the second prominent high-flow events during the contrasting season are as large as the seasonal AMFs.

Kråkfoss and Austenå catchments show the largest dominance of AMF seasonality per calibration period and, thus, the most pronounced contrasting conditions in terms of flood seasonality. For Kråkfoss, the level of dominance for 18 calibration sub-periods with AMFs during spring and autumn ranges from 28 to 61% and from 26 to 46%, respectively. For Austenå, we have 19 calibration periods covering the years with AMFs during spring, and the level of dominance ranges from 27 to 49%. For 16 calibration periods that cover the years with AMFs during autumn, the level of dominance ranges from 25 to 51%, with eight periods showing larger than 40% dominant autumn floods. Compared to both previous catchments, the range of the dominance of seasonal AMFs at Bulken is smaller: for 18 periods with AMFs during spring the level of dominance varies between 21 and 33%; and for 23 periods the dominance of AMFs during autumn ranges from 20 to 35%. Due to the comparatively short runoff observation time series for Jogla (1973–2014), only 30 periods are available for the calibration and validation (17 with AMFs during spring, 13 with AMFs during autumn). Here, the range of dominance of periods with spring and autumn floods is 21–42% and 19–41%, respectively. For Fustvatn, the seasonal AMFs show the comparatively lowest level of dominance per calibration period. For 17 periods with dominant AMFs during spring, the level of dominance ranges from 15 to 38%, and for 27 periods with AMFs during autumn the level of dominance ranges from 18 to 37%.

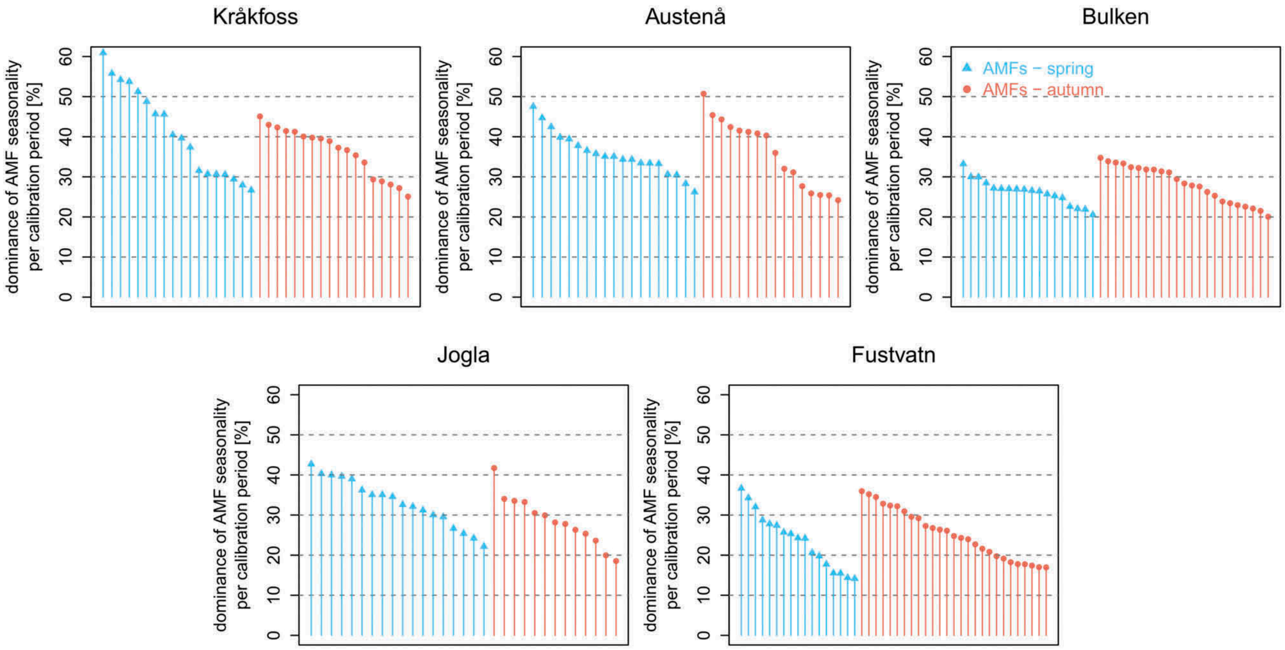

Figure 3. The dominance of AMFs during spring and autumn, respectively, per calibration period as a measure of the level of contrasting flood seasonality conditions between the periods used for SSTs and DSSTs. A dominance level of 0% would mean that the sum of second-order peak-flow discharge during the respective contrasting season is as high as the sum of all AMFs during a certain calibration period.

3.2 Calibration performance

Figure 4 shows the calibration performance in terms of NSE$_w$ for the five study catchments. The distributions were derived by optimizing the HBV model parameters for all 6-year calibration periods of the moving window over the time series without any distinction regarding the seasonal occurrence of AMFs (Fig. 4, left), and over the time series distinguished by the seasonal occurrence of AMFs during spring (middle) and autumn (right). The distribution of the calibration performance per group indicates the quality of the reference parameter sets as they are applied for the estimation of the MRC (see next sections).

The calibration of the model yields fair to good results, with NSE$_w$ values ranging from 0.53 (minimum for Jogla) to 0.97 (maximum for Bulken). Note, however, that NSE$_w$ values tend to be somewhat higher than regular NSE values (e.g. 12–14% for Kråkfoss) due to the weighting of the high flows in the discharge data. Focusing on the distributions of NSE$_w$ values, the comparatively worst and best calibration results tend to emerge for Jogla and Kråkfoss, respectively. The medians of the calibration performance using the entire data series without any distinction regarding the seasonal occurrence of AMFs (Fig. 4, left boxes in each plot) vary between 0.70 (Jogla) and 0.89 (Kråkfoss; Bulken). For Kråkfoss (0.91), Austenå (0.88), and Bulken (0.90) the medians of the calibration performance for the periods that cover the years with AMFs during spring are slightly larger than the medians of the calibration performance for all years (0.89, 0.87, 0.89). For these three catchments, the calibration of the hydrological model on periods with dominant spring floods tends to outperform the model calibration on periods with dominant autumn floods (both regarding the medians and the inter-quartiles of NSE$_w$ values). This may indicate that high flows during spring can be better simulated than high flows during autumn and early winter, which are often associated with more rapid concentration and recession runoff than snowmelt-dominated high-flow events during spring (see e.g. Lawrence and Haddeland 2011). The visual inspection of observed and simulated hydrographs, however, does not always confirm this assumption. Moreover, the other two catchments, Jogla and Fustvatn, do not immediately promote this pattern, since the medians of the calibration performance on periods with AMFs during autumn are slightly larger than the medians of the calibration performance on periods with AMFs during spring. The upper quartiles of the calibration performance on periods with spring floods are, though, slightly larger than the upper quartiles of the calibration performances on periods with dominant autumn floods.

3.3 Split-sample tests (SSTs)

The results of the SSTs (Steps 1–3 of the testing protocol in Fig. 2) in terms of systematic model performance losses estimated by the MRC are illustrated in

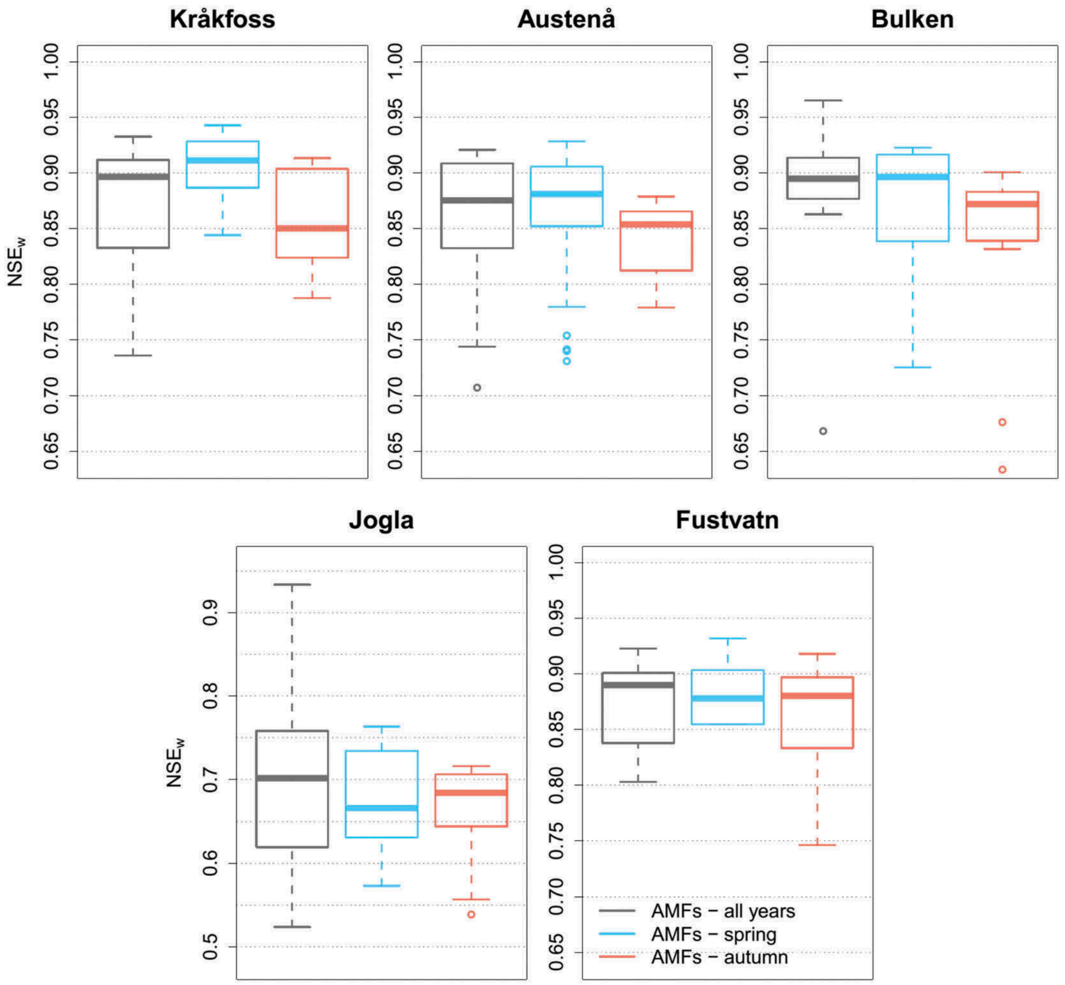

Figure 4. Distribution of the calibration performances in terms of NSE$_w$ for the five study catchments: covering the AMFs of the entire runoff observation time series (left), for periods with AMFs occurring only during spring (middle) and only during autumn (right).

Figure 5 upper panel shows detailed results including those performance losses estimated on validation periods with overlapping years between donor and receiver periods for Kråkfoss; and the lower panel shows model performance losses estimated on only independent validation periods (no overlap) for the four remaining catchments. That is, the boxes and whiskers in the lower panel correspond to the rightmost boxes and whiskers in each plot of the upper panel (i.e. Fig. 5 (a)–(c)).

For Kråkfoss and the entire time series (Fig. 5(a)), model performance decreases with a decreasing level of overlap between donor and receiver periods (about 1.2% of performance loss with each year decrease in overlap). Although expected, the result consistently demonstrates the role of (in)dependency between periods used for calibration and validation. For years with dominant spring floods (Fig. 5(b)), we find a similar decrease in model performance with decreasing overlapping years, though not as distinct as for the entire time series. Note, however, that the sample size of parameter sets is considerably smaller (18 *vs* 43) due to the limited amount of years available to create 6-year sub-periods with spring floods. For the years with dominant autumn floods (Fig. 5(c)), the SSTs show slightly smaller systematic performance losses than for the two previous groups. Moreover, for every set of SSTs, the upper quartile of the distribution shows positive MRC estimates, pointing to a better model performance with the donor parameter sets as compared to the original ones. This indicates, on the one hand, difficulties with estimating best-fit parameter fits for years where autumn and early winter flooding is dominant, and it illustrates, on the other hand, a comparatively large exchangeability of parameter sets that have been calibrated on different periods with AMFs occurring during autumn and early winter. Similar patterns of systematic model performance decrease with a decreasing level of overlapping years are found for the other four catchments (not shown). For Jogla and Fustvatn, however, the upper quartiles of

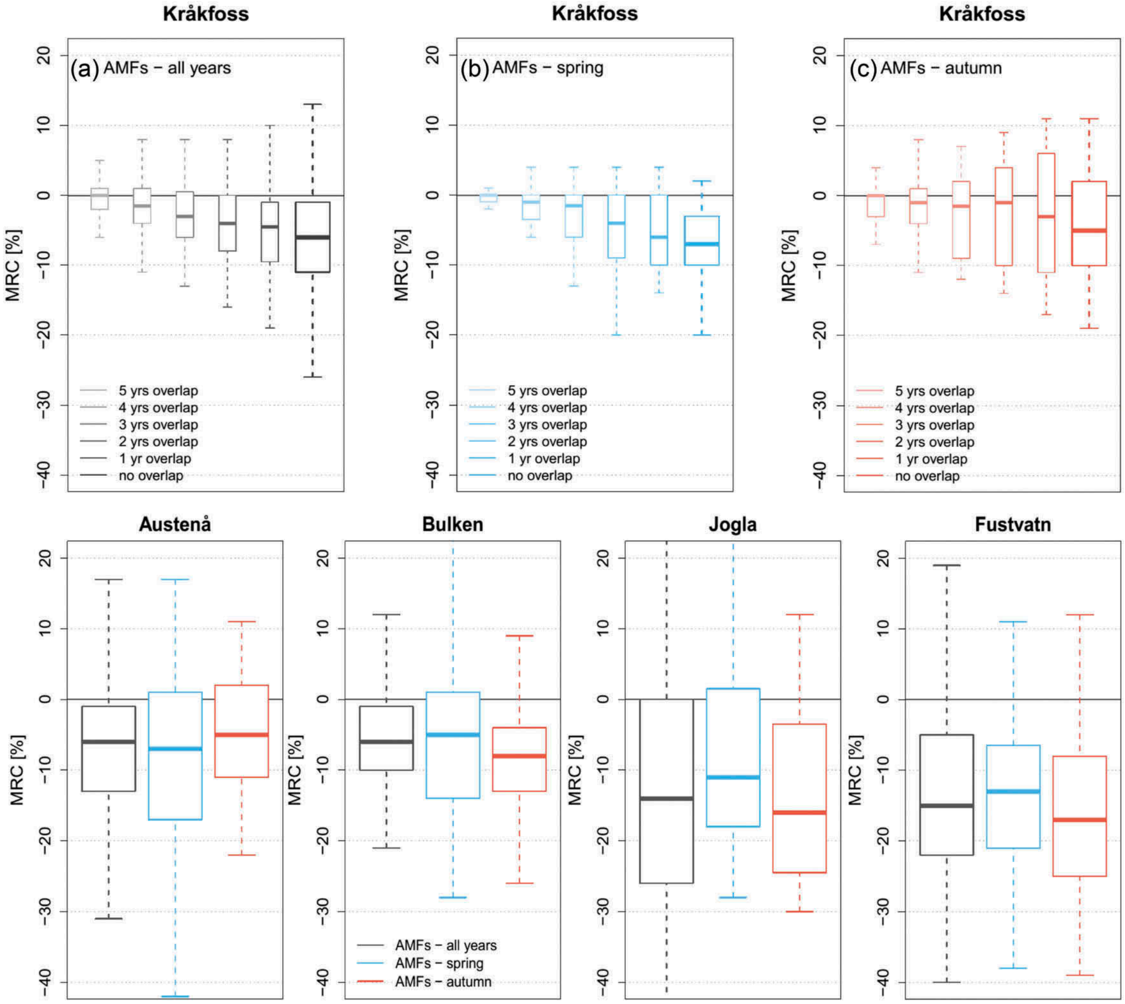

Figure 5. Model performance loss estimated by the MRC using SSTs for 6-year calibration–validation time periods with AMFs of (a) all years (also left boxes in each plot of the lower panel), (b) AMFs occurring during spring (also middle boxes in lower panel plots), and (c) during autumn (also right boxes in lower panel plots). Boxes and whiskers from left to right in each plot of the upper panel indicate a decreasing level of overlap between donor and receiver periods for the study catchment Kråkfoss. The range of the whiskers covers 1.5 times the inter-quartile range represented by the boxes.

all distributions including those for the years with AMFs during autumn show negative MRC estimates. Remember that for the same two catchments the calibration of hydrological model parameters for years with dominant autumn floods worked comparatively well (see Section 3.2).

Focusing on the MRC estimates for the independent validation periods for all catchments over the entire time series (Fig. 5: the largest box to the right (a) and the left boxes in each plot of the lower panel), median performance losses vary between −5% (Austenå) and −15% (Fustvatn). Median performance losses for Kråkfoss, Bulken, and Jogla are −6%, −6% and −14%, respectively. The smallest inter-quartile range of MRC estimates is found for Bulken (−10 to −1%), and the largest inter-quartile range is found for Jogla (−24 to 0%), which has also shown the poorest calibration results (see Fig. 3). For all catchments, some individual donor parameter sets do continue to show a better model performance than the original parameter sets for the independent validation periods.

For the periods with dominant spring floods (Fig. 5(b) and the middle boxes in the lower panel), the median performance losses estimated by the MRC for all independent validation periods within this group are equal or a bit smaller as compared to those of the entire time series (Kråkfoss and Austenå: −7%, Bulken: −5%, Jogla: −11% and Fustvatn: −13%). Again, the largest inter-quartile range is found for Jogla (−18 to +1%), though the inter-quartile range for Austenå is only one percentage point smaller. The smallest inter-quartile range is found for Kråkfoss (−10 to −3%). Again, there are individual donor parameter sets that show better model performances as compared to the original parameter sets for the validation periods, and for Austenå, Bulken, and Jogla even the upper quartile of the distributions of MRC estimates are slightly positive (<+1 to +1.5%).

For the periods with dominant autumn floods (Fig. 5(c) and boxes to the right in the lower panel), the SSTs for the independent validation periods show slightly larger median performance losses for Bulken (–8%), Jogla (–16%) and Fustvatn (–17%) as compared to the two previous groups. For Kråkfoss (–5%) and Austenå (–4.5%), however, the SSTs show slightly smaller systematic performance losses than both previous groups. The pattern regarding the largest inter-quartile range is similar to both previous groups of SSTs (largest for Jogla: – 24 to – 3.5%), though, the smallest inter-quartile range is found for Bulken in this case (–13 to – 4%).

To sum up: although there are considerable differences regarding the magnitude of model performance losses between the catchments, the results do not show large differences in the medians of model performance losses estimated by the SSTs for each individual catchment, irrespective of the group considered. This points to a robust model behaviour when transferring HBV model parameters under similar conditions in terms of flood seasonality.

3.4 Differential split-sample tests (DSSTs)

The question which now arises is whether or not model performance loss increases if we transfer best-fit parameter sets that have been optimized for years with AMFs during spring and autumn, respectively, to their contrasting group of receiver periods. Figure 6 shows the results of the DSSTs (Step 4 of the testing protocol in Fig. 2). For each plot in Figure 6, the boxes and whiskers in front show the distributions of model performance losses in terms of MRC estimates using the donor parameter sets from periods with AMFs occurring during spring (left) and autumn (right) to their contrasting groups of receiver periods. They allow general conclusions to be made regarding the transferability of hydrological model parameters for periods with contrasting flood seasonality. To ease the comparison, boxes and whiskers in the background are adopted from Figure 4 and show the results of the SSTs for the independent validation periods with dominant spring and autumn floods, respectively.

Transferring calibrated parameters from periods with prominent spring floods to periods with prominent autumn floods results in median performance losses of −11% (Jogla) to −3% (Bulken), i.e. not so different from the results of transferring parameters the other way around (range of median performance losses from −12% (Fustvatn) to −4% (Austenå). More interestingly, even, it becomes obvious that the model performance losses estimated by the DSSTs (Fig. 6, thick boxes in front) do not fundamentally differ from the performance losses as estimated by the SSTs (Fig. 6, thin boxes in the background). Moreover, they are not necessarily larger. Only for the donor parameter sets that have been calibrated on periods with dominant autumn floods at Kråkfoss and on periods with prominent spring floods at Austenå is the median model performance loss estimated by the DSSTs slightly larger than that estimated by the SSTs (three and one percentage points, respectively). For all other catchments the performance loss is equal or even smaller when transferring parameters to periods with contrasted AMF seasonality. The largest differences between the median model performance losses estimated by the DSSTs and SSTs are found for Fustvatn (five percentage points for both directions of transfer) and Jogla (six percentage points; transfer from autumn donor periods to spring receiver periods). The most remarkable difference between the results of both tests is that the range of the distributions of the DSSTs tends to be smaller than those of the SSTs for most but not all catchments and directions of transfer. Note that the differences between the distributions of MRC values estimated by the DSSTs are statistically significant (with 95% confidence) for all catchments except for Jogla, as none of the notches of the boxplots in front are overlapping. The differences between the distributions of MRC estimates of the SSTs and DSSTs are statistically significant only for Kråkfoss, Jogla (for autumn donor periods), and Fustvatn (both donor periods).

Altogether, the transfer of parameters between periods of contrasting flood seasonality (and associated flood generating processes), as determined within this study, does not imply more pronounced performance losses as compared to the performance losses that have been estimated for periods with similar flood seasonality conditions.

3.5 On the role of AMF seasonality dominance and model performance loss

To further investigate the role of contrasting flood seasonality conditions on the model performance losses, we show in Figure 7 the relationship between the dominance of AMF seasonality per calibration period (adopted from Fig. 3) and the maximum performance loss in terms of MRC values estimated by the DSSTs.

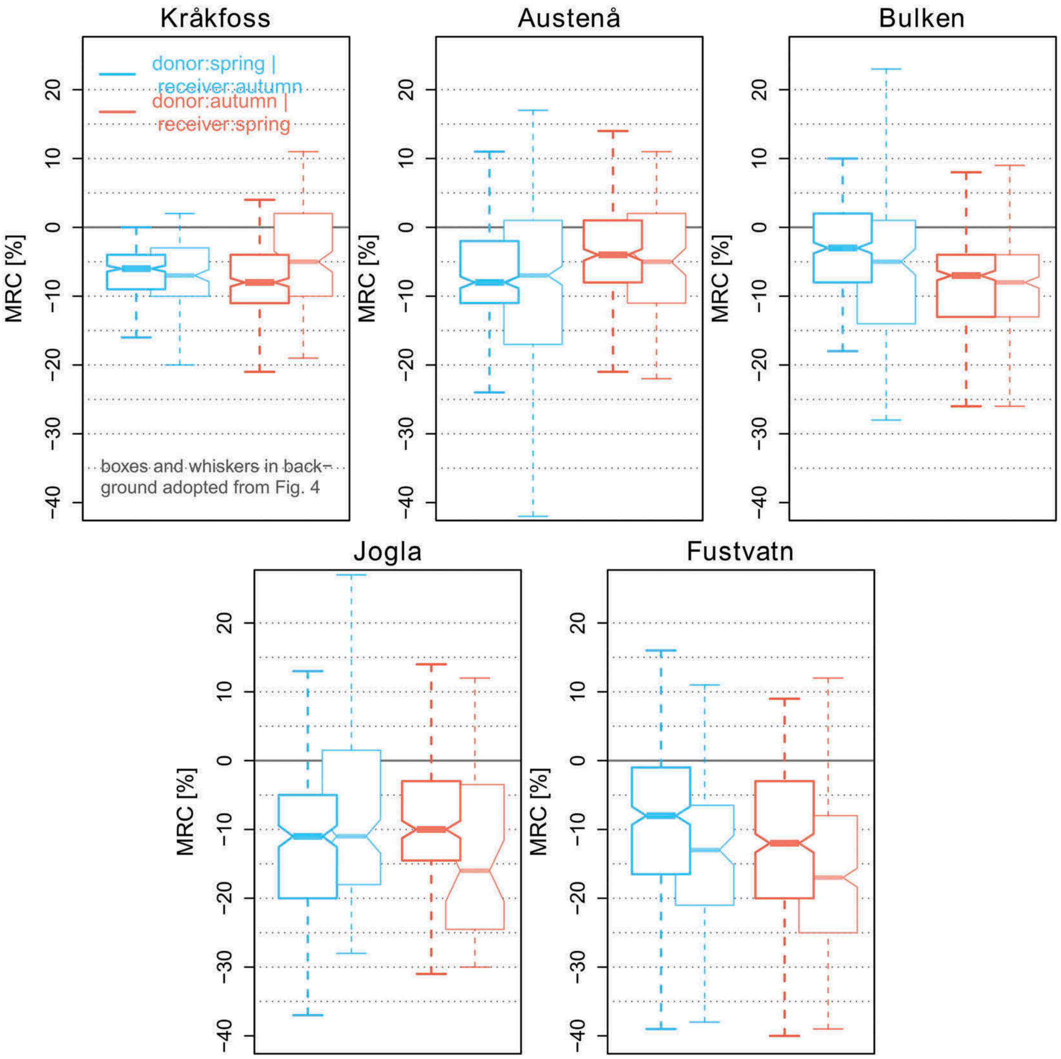

Figure 6. Model performance losses in terms of the MRC estimated by the DSSTs using parameter sets optimized on the 6-year periods with AMFs occurring during spring and applied on the 6-year receiver periods with AMFs occurring during autumn (left boxes in front) and *vice versa* (right boxes in front). Boxes and whiskers in the background show the results of the SSTs on independent validation periods (no overlapping years) for each catchment to ease the comparison. The range of the whiskers covers 1.5 times the inter-quartile range represented by the boxes.

The largest correlations between the level of contrasting flood seasonality conditions and model performance loss are found for the transfer of donor parameter sets that have been calibrated on periods with dominant autumn floods at Kråkfoss and Austenå to receiver periods with dominant spring floods (Pearson correlation coefficients of 0.72 and 0.61, respectively). Austenå also shows a comparatively large correlation coefficient for transferring parameter sets the other way around (0.41). For Kråkfoss this is only true for calibration periods in which AMFs during spring are dominant with up to 45% (0.77); periods that show a larger dominance of AMF seasonality do not promote this pattern (correlation coefficient for all periods with dominant spring floods: 0.23).

Regarding the three other catchments, only Fustvatn and Jogla show mentionable positive correlations between the level of dominance of AMF seasonality and the maximum model performance losses: Fustvatn for the case when parameter sets are transferred from donor periods with dominant spring floods to receiver periods with dominant autumn floods (0.33), and Jogla for both directions of transfer (spring to autumn periods: 0.30; autumn to spring periods: 0.37). For Bulken (both directions of transfer) and for Fustvatn (transfer from autumn to spring periods) the correlation between the level of contrasting flood seasonality and model performance losses estimated by the DSSTs is almost zero.

In summary, the correlations between the level of contrasting AMF seasonality and the maximum model

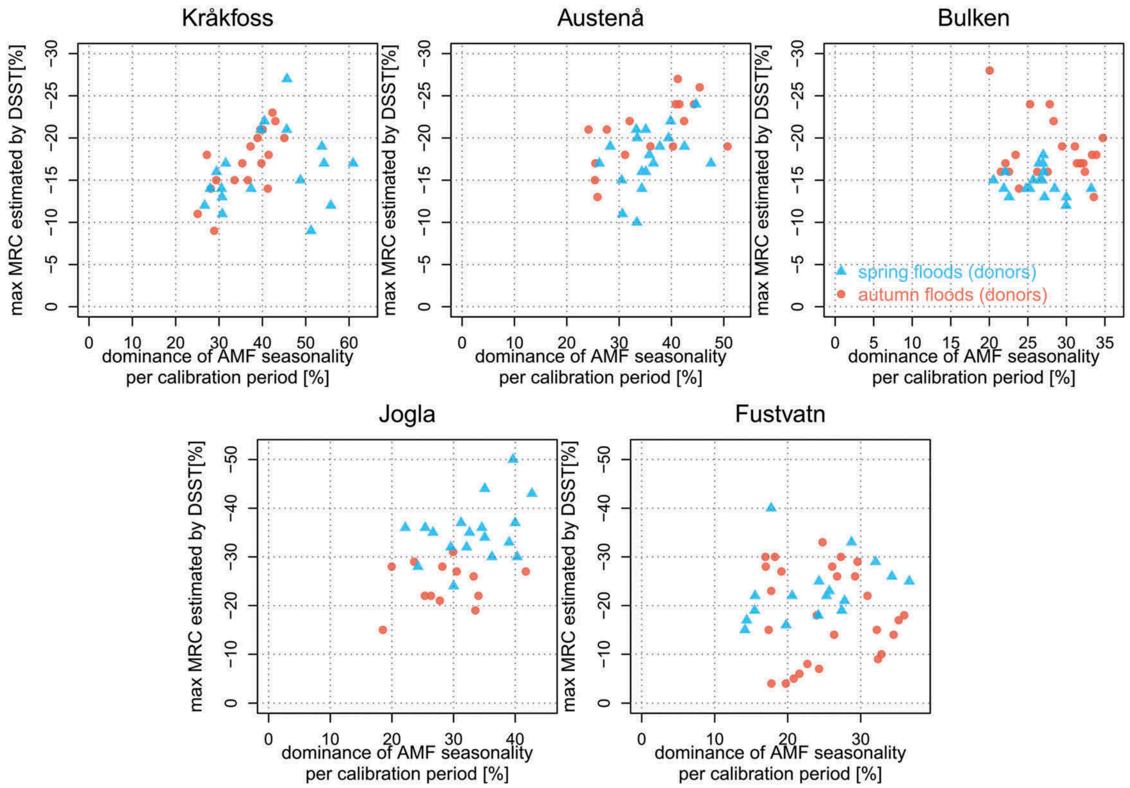

Figure 7. Scatterplots showing the relationship between the level of dominance of AMF seasonality per calibration period (adopted from Fig. 3) and the maximum model performance loss estimated by the DSSTs. The level of dominance of AMF seasonality indicates the level of contrasting conditions in terms of flood seasonality as determined within this study.

performance losses are not as distinct as expected. Only for Kråkfoss and Austenå, which show the most pronounced contrasts in flood seasonality conditions (see Section 3.1), a systematic increase in model performance loss can be correlated with larger levels of contrasting AMF seasonality conditions. In this sense, it is remarkable that these two catchments are the only ones within this study that also show larger median model performance losses estimated by the DSSTs as compared to the SSTs.

4 Critical discussion of the results

4.1 What are the reasons for the unexpected small differences between the model performance losses estimated by the SSTs and the DSSTs?

The classification of the blocks with contrasting flood seasonality is based on seasonal occurrence of AMFs (i.e. one single event during either spring or autumn and early winter). We have shown that the calibration of the HBV model is sensitive to the AMF using the DDS optimization algorithm and the NSE_w as objective function. We have also shown that the dominance of these AMFs per calibration period is reduced by second-order peak flow discharge from the contrasting season by 38% (Kråkfoss) to 85% (Fustvatn). This means that each calibration sub-period for both blocks of dominant AMF seasonality also covers relevant processes from their respective contrasting seasons. Moreover, the determination of the two blocks does not exclude that the basic hydro-meteorological conditions fundamentally differ between these two blocks (see the annual hydrographs in Fig. 1). In consequence, the model calibration seems to take care of all relevant runoff generating processes even though they are not dominant during a certain calibration period. For flood simulations this means that, although a flood generating process might not be dominant in a specific period, it might still be prominent enough to allow for a robust optimization of underlying parameters. This allows for the transferability of calibrated HBV model parameters to periods in which the same process becomes dominant.

From the perspective of *differential* split-sample tests, the results indicate that the contrasting conditions as determined by our classification approach are not contrasting enough to identify systematic model performance losses as intended by the test and as identified in other studies (e.g. Coron *et al.* 2012). So the question arises: which alternative approaches may be more suitable to properly gain contrasting conditions in terms of flood seasonality? One opportunity might be the determination of two contrasting blocks based on the seasonal occurrence of the largest mean monthly discharge per year. In this case we would focus on more aggregated runoff characteristics rather than on single events. We have tested this approach for one example catchment (Kråkfoss) and found similar model performance losses in terms of median MRC values estimated by the DSSTs (spring to autumn: −7%; autumn to spring: −9%, i.e. not so very different from the median MRCs estimated by the DSSTs in this study). Note, however, that the classification of two contrasting blocks in terms of the seasonal occurrence of the largest mean monthly flow results in an unequally distributed number of calibration sub-periods being available for each block: 30 sub-periods with maximum mean flows during spring and early summer, and only six sub-periods during autumn and early winter. In addition, the determination of the two contrasting blocks based on mean monthly flows does not particularly account for *flood* generating processes, as intended by this study.

Another, and probably the most drastic approach for gaining contrasting seasonality conditions between the calibration and validation periods refers to the optimization of hydrological model parameters for each calibration period based on the respective months of seasonal discharge only (i.e. March–August *vs* September–February). We tested this approach (again for Kråkfoss) and applied the generalized scheme of the DSST using a 6-year moving window over the entire runoff time series (1967–2014) to optimize donor parameter sets for spring/summer and autumn/winter conditions, respectively. When we transfer these donor parameter sets to their contrasting group of receiver periods, we identify model performance losses that are considerably larger than those identified within this study (i.e. spring to autumn: −26%; autumn to spring: −15%, on median). However, this approach still does not account for (changing) flood generating processes in particular. Moreover, this approach is somewhat arbitrary since it entirely neglects the seasonal hydro-meteorological regimes of the catchments, which may get altered by climate change but probably not fundamentally modified.

4.2 What are the implications of the results for hydrological flood modelling under climate change?

Our results are promising news for (flood-)hydrological impact modelling under climate change using the HBV model in Nordic catchments with likely shifts in their seasonal high-flow regime as long as the dominant “future processes” are represented to a certain degree in the calibration period (in this study with at least 38% prominence). This conclusion is based on the medians of the MRC estimates. The range of the distributions of the MRC estimates, however, indicates that model performance losses can be quite large for individual parameter sets (up to −27% for Kråkfoss; more than −50% for Austenå, Bulken, and Jogla). This highlights the need for careful selection of calibration periods and for applying a range of calibrated best-fit parameter sets in the multi-model ensembles for assessing the hydrological impacts of climate change. In that perspective, it is reasonable to calibrate hydrological models either for periods of sufficient length to include as many relevant processes as reflected by the observation data, or for periods that reflect likely future conditions most appropriately. The latter case, of course, assumes some prior knowledge about likely future conditions, which may not always be available *a priori*.

Finally, we need to emphasize that our positive conclusions regarding the transferability of calibrated HBV model parameters to periods with changing flood seasonality is based on the particular assumptions and choices of methods we have made in this study: (1) the determination of contrasting conditions based on the AMF seasonality; (2) the selection of the study catchments; (3) the choice of the DDS global optimization algorithm for model calibration; and (4) the application of the NSE_w as objective function. Altering one or several of these decisions may yield different results and conclusions.

5 Conclusions

Using the HBV model in five Norwegian catchments with mixed snowmelt/rainfall regimes, we have systematically analysed the robustness of hydrological model parameters in terms of model performance losses when transferring calibrated parameter sets to validation periods with both similar and contrasting flood seasonality conditions.

On average, the results indicate the expected decrease in model performance when applying calibrated hydrological model parameter sets on independent validation periods. However, there is no indication that contrasting flood seasonality conditions – as they are defined within this study – exacerbate model performance losses. This contradicts our assumptions that calibrated parameter sets, which have been “specialized” for different dominant flood generating processes (i.e. snowmelt *vs* rainfall), perform more poorly under contrasting flood seasonality conditions (spring *vs* autumn) as compared to similar conditions. It appears that the intensity of a flood generating process plays only a minor role for optimizing the corresponding model parameters, as long as the process has some level of prominence in the calibration period (at least 38% in this study).

The results obtained by this study differ from the findings of other studies that have analysed the temporal transferability of hydrological model parameters (e.g. Vaze *et al.* 2010, Coron *et al.* 2012, Brigode *et al.* 2013). All these authors found a considerable decrease in model performance when transferring hydrological model parameters under non-stationary conditions, which indicates a lack of parameter robustness. Note, however, that non-stationarity was usually defined as dry *vs* wet, or warm *vs* cold conditions, and the level of contrast between calibration and validation periods was most probably higher than within this study. Furthermore, mean runoff instead of floods was analysed as the hydrological target variable, and the geographical and hydro-climatological settings of those studies (mid-latitudes and semi-arid areas) differ from the setting represented by the catchments considered in this study.

Therefore, this study underlines that the transferability of hydrological model parameters needs to be scrutinized for specific catchments, models, and cases of non-stationarity. The results presented here establish that, at least for the investigated Nordic catchments with mixed snowmelt/rainfall regimes, changing flood seasonality is not the dominant cause for model performance losses. Thus, the prospects for the transferability of HBV model parameters under contrasting flood seasonality seem good as long as no fundamental changes in the hydro-meteorological regime lead to shifts in the relevance of (flood generating) catchment processes. In such a case, the probability of model failure may increase. Still, detecting the true causes for model performance losses as estimated by both the SSTs and DSSTs is needed and would require a more extensive analytical framework. Detailed analyses of the correlation between model performance and climate characteristics, for instance, may help to identify whether or not some parameters may indeed correspond to changes in climate characteristics, as found by Merz *et al.* (2011) and Osuch *et al.* (2015). This, however, is beyond the scope of this study. In order to substantiate our conclusions, further investigations are required that use several hydrological model structures and more catchments with similar and/or different hydrological regimes in which contrasts can be found (Andréassian *et al.* 2009).

Acknowledgements

We want to thank Charles Perrin for his valuable comments on the latest version of our manuscript. The helpful comments of Benjamin Dewals and an anonymous reviewer on an earlier version are also gratefully acknowledged. All statistical analyses and visualizations (except for the maps) were performed within the R software environment (www-r-project.org).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The first author was part of the graduate research school GeoSim, jointly funded by the Helmholtz Association, the University of Potsdam, and the Free University of Berlin, Germany.

References

Addor, N., *et al.*, 2014. Robust changes and sources of uncertainty in the projected hydrological regimes of Swiss catchments. *Water Resoures Research*, 50, 7541–7562. doi:10.1002/2014WR015549

Andréassian, V., *et al.*, 2009. HESS Opinions " Crash tests for a standardized evaluation of hydrological models". *Hydrology and Earth System Sciences,*13, 1757–1764.

Arheimer, B. and Lindström, G., 2015. Climate impact on floods: changes in high flows in Sweden in the past and the future (1911–2100). *Hydrology and Earth System Sciences*, 19 (2), 771–784. doi:10.5194/hess-19-771-2015

Bergström, S., 1976. Development and application of a conceptual runoff model for Scandinavian catchments. Norrköpping: Swedish Meteorological and Hydrological Institute, Report RHO 7.

Bergström, S., 1995. The HBV model. *In*: V.P. Singh, ed. *Computer models of watershed hydrology*. Highlands Ranch, CO: Water Resources Publications, 443–476.

Blöschl, G. and Montanari, A., 2010. Climate change impacts – throwing the dice? *Hydrological Processes*, 24, 374–381. doi:10.1002/hyp

Brigode, P., Oudin, L., and Perrin, C., 2013. Hydrological model parameter instability: a source of additional uncertainty in estimating the hydrological impacts of climate change? *Journal of Hydrology*, 476, 410–425. doi:10.1016/j.jhydrol.2012.11.012

Bronstert, A., 2004. Rainfall-runoff modelling for assessing impacts of climate and land-use change. *Hydrological Processes*, 18, 567–570. doi:10.1002/hyp.5500

Coron, L., *et al.*, 2012. Crash testing hydrological models in contrasted climate conditions: an experiment on 216 Australian catchments. *Water Resoures Research*, 48, W05552. doi:10.1029/2011WR011721

Dobler, C., *et al.*, 2012. Quantifying different sources of uncertainty in hydrological projections in an Alpine watershed. *Hydrology and Earth System Sciences*, 16, 4343–4360. doi:10.5194/hess-16-4343-2012

Fleig, A.K., *et al.*, 2013. Norwegian hydrological reference dataset for climate change studies. Oslo: Norwegian Water Resources and Energy Directorate (NVE). NVE rapport. No.2.

Fowler, K.J.A., *et al.*, 2016. Simulating runoff under changing climatic conditions: revisiting an apparent deficiency of conceptual rainfall–runoff models. *Water Resources Research*, 52, 1820–1846. doi:10.1002/2015WR018068

Kay, A.L., *et al.*, 2009. Comparison of uncertainty sources for climate change impacts: flood frequency in England. *Climatic Change*, 92, 41–63. doi:10.1007/s10584-008-9471-4

Klemeš, V., 1986. Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, 31, 13–24. doi:10.1080/02626668609491024

Lawrence, D. and Haddeland, I., 2011. Uncertainty in hydrological modelling of climate change impacts in four Norwegian catchments. *Hydrology Research*, 42, 457. doi:10.2166/nh.2011.010

Lindström, G., *et al.*, 1997. Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, 201, 272–288. doi:10.1016/S0022-1694(97)00041-3

Matott, L.S., Babendreier, J.E., and Purucker, S.T., 2009. Evaluating uncertainty in integrated environmental models: a review of concepts and tools. *Water Resources Research*, 45, W06421. doi:10.1029/2008WR007301

Merz, R., Parajka, J., and Blöschl, G., 2011. Time stability of catchment model parameters: implications for climate impact analyses. *Water Resources Research*, 47, W02531. doi:10.1029/2010WR009505

Mohr, C. and Tveito, O.E., 2008. Daily temperature and precipitation maps with 1 km resolution derived from Norwegian weather observations. Preprints, 17th Conf. on Applied Climatology. Whistler, BC: American Meteorological Society, 6.3. Available from: https://ams.confex.com/ams/pdfpapers/141069.pdf. [Accessed 3 May 2018].

Osuch, M., Romanowicz, R.J., and Booij, M.J., 2015. The influence of parametric uncertainty on the relationships between HBV model parameters and climatic characteristics. *Hydrological Sciences Journal*, 60, 1299–1316. doi:10.1080/02626667.2014.967694

Ott, I., *et al.*, 2013. High-resolution climate change impact analysis on medium-sized river catchments in Germany: an ensemble assessment. *Journal of Hydrometeorology*, 14, 1175–1193. doi:10.1175/JHM-D-12-091.1

Refsgaard, J.C., *et al.*, 2013. A framework for testing the ability of models to project climate change and its impacts. *Climatic Change*, 122, 271–282. doi:10.1007/s10584-013-0990-2

Refsgaard, J.C., *et al.*, 2006. A framework for dealing with uncertainty due to model structure error. *Advances in Water Resources*, 29, 1586–1597. doi:10.1016/j.advwatres.2005.11.013

Sælthun, N., 1996. The Nordic HBV model. NVE Publication no. 7. Oslo: Norwegian Water Resources and Energy Directorate (NVE).

Seibert, J., 2003. Reliability of model predictions outside calibration conditions. *Nordic Hydrology*, 34, 477–492.

Seibert, J. and Vis, M.J.P., 2012. Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences*, 16, 3315–3325. doi:10.5194/hess-16-3315-2012

Thirel, G., Andréassian, V., and Perrin, C., 2015b. On the need to test hydrological models under changing conditions. *Hydrological Sciences Journal*, 60, 1165–1173. doi:10.1080/02626667.2015.1050027

Thirel, G., *et al.*, 2015a. Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments. *Hydrological Sciences Journal*, 60, 1184–1199. doi:10.1080/02626667.2014.967248

Tolson, B.A. and Shoemaker, C.A., 2007. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, 43, W01413. doi:10.1029/2005WR004723

Vaze, J., *et al.*, 2010. Climate non-stationarity – validity of calibrated rainfall–runoff models for use in climate change studies. *Journal of Hydrology*, 394, 447–457. doi:10.1016/j.jhydrol.2010.09.018

Viviroli, D., *et al.*, 2011. Climate change and mountain water resources: overview and recommendations for research, management and policy. *Hydrology and Earth System Sciences*, 15, 471–504. doi:10.5194/hess-15-471-2011

Vormoor, K., *et al.*, 2015. Climate change impacts on the seasonality and generation processes of floods – projections and uncertainties for catchments with mixed snowmelt/rainfall regimes. *Hydrology and Earth System Sciences*, 19, 913–931. doi:10.5194/hess-19-913-2015

Vormoor, K., *et al.*, 2016. Evidence for changes in the magnitude and frequency of observed rainfall vs. snowmelt driven floods in Norway. *Journal of Hydrology*, 538, 33–48. doi:10.1016/j.jhydrol.2016.03.066

Wagener, T., *et al.*, 2003. Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis. *Hydrological Processes*, 17, 455–476. doi:10.1002/hyp.1135

Wilby, R.L. and Dessai, S., 2010. Robust adaptation to climate change. *Weather*, 65, 180–185. doi:10.1002/wea.543