A bootstrap-based differential split-sample test to assess the transferability of conceptual rainfall-runoff models under past and future climate variability

Hamouda Dakhlaoui, Denis Ruelland, Yves Tramblay

A bootstrap-based differential split-sample test to assess the transferability of conceptual rainfall-runoff models under past and future climate variability

Hamouda DAKHLAOUI [1, 2], Denis RUELLAND [3], Yves TRAMBLAY [4]

[1] *LMHE, Ecole Nationale des Ingénieurs de Tunis, University of Tunis El Manar, BP 37, 1002 Tunis le Belvedère, Tunisia*

[2] *Ecole Nationale d'Architecture et d'Urbanisme, University of Carthage, Rue El Quods, 2026, Sidi Bou Said, Tunisia*

[3] *CNRS, Laboratory HydroSciences Montpellier, University of Montpellier, Place Bataillon, 34395 Montpellier, France*

[4] *IRD, Laboratory HydroSciences Montpellier, University of Montpellier, Place Bataillon, 34395 Montpellier, France*

Correspondence to denis.ruelland@umontpellier.fr

Abstract This study proposes a general differential split-sample test (GDSST) based on an oriented bootstrap to assess the transferability of conceptual rainfall-runoff models to climatically contrasting periods. Compared to existing benchmark techniques, the GDSST allows a larger number of climatically contrasted discontinuous periods to be sampled, and is computationally more effective than the basic bootstrap to identify the most contrasted periods. When applied to three hydrological models (GR4J, HBV and IHACRES) in five catchments in northern Tunisia, the GDSST provided clear limits of the transferability of the models under changing precipitation (P) and temperature (T) conditions towards drier and hotter conditions. According to the criteria and thresholds retained, approximate limits of model transferability are drawn. The models are roughly transferable for relative changes in precipitation $\Delta P < (0.08\ \Delta T - 0.18)$ with $\Delta P \in [-30\%, +80\%]$, and changes in temperature $\Delta T \in [-2\ °C, +2\ °C]$. These transferability limits suggest selecting a past sub-period as close as possible to the future climate to identify calibration parameters, which can be used for hydrological projections. The limits of transferability were then compared to climate projections by eight high-resolution Regional Climate Model (RCM) simulations resulting from the EURO-CORDEX initiative. The RCMs' precipitation and temperature simulations of the historical period 1970–2000 were first assessed to select the most realistic ones for future projections. A delta-change monthly correction was used to perturb the observed climate series according to climate simulations under two Radiative Concentration Pathway (RCP) scenarios (RCP4.5 and RCP8.5) for one medium-term horizon (2040–2070) and one long-term horizon (2070–2100). The effects of the selected past calibration period on the hydrological projections were then analysed. The RCP 8.5 climate projections fall outside the limits of transferability of all rainfall-runoff models tested. Models calibrated on the whole observed period were found

to underestimate the impacts of climate change on runoff by 5% to 20% in comparison with models calibrated on sub-periods with mean annual P and T closer to projected climate conditions.

Key words Rainfall-runoff modelling; Split-sample tests; Parameter transferability; Climate scenarios; CORDEX; Tunisia.

1. INTRODUCTION

1.1. Simulating the impact of climate change on runoff

Several studies have shown that climate change has already affected available water resources worldwide (e.g. Haddeland et al., 2014) and is expected to have even more severe impacts in the future (e.g. Hageman et al., 2013; IPCC, 2013). However, assessing the potential impact of climate change on runoff precisely requires both high-resolution climate projections and models capable of reliably representing the hydrological processes in recent decades and at the scale of basins that are sufficiently representative of water management issues, i.e. several hundred to several thousand square kilometres (Fabre et al., 2016). In developing countries, the use of physically-based distributed models is generally hampered by insufficient data to force and control the models. Consequently, model validation is usually based on the streamflow at the outlet, which does not guarantee that the water redistribution processes and their interactions are well represented within the catchment. Application is even harder when it comes to testing the ability of models to reproduce multi-decadal hydrological variability which is a prerequisite for studying the impact of climate change on water resources (Ruelland et al., 2012). Therefore, conceptual models representing the functioning of the basins using a small number of empirical equations whose few parameters can be calibrated with a minimum of data are often preferred, particularly in a context of data scarcity (e.g. Bastola et al., 2011; Chen et al., 2011; Hublart et al., 2016; Ruelland et al., 2012; 2015). On the other hand, their relative simplicity and the need to calibrate their parameters on current data do not always plead for their use when conditions shift beyond the range of prior experience (Hublart et al., 2015) as it can be the case in a framework of future change (see e.g. Vaze et al., 2010). The use of conceptual models is therefore conditional on the estimation of the uncertainty associated with the modelling process itself, which is a combination of uncertainties evolving from input and control data, model structures, parameterization and parameter transferability under non-stationary conditions, such as climate change or variability.

1.2. Review of the literature on the problem of model parameter transferability to climate change

Parameter transferability can be defined as the ability of a model to perform with the same level of accuracy under conditions that differ from those used for its calibration (Seiller et al., 2012). It requires particular attention in studies on the impact of climate change, because it can be an important source of uncertainty in the hydrological, modelling chain as already pointed out by many authors (e.g. Brigode et al., 2013; Coron et al., 2012; Poulin et al., 2011; Fowler et al. 2016; Melsen et al., 2018). Many techniques have thus been proposed to assess potential parameter transferability under different climate conditions. For instance, the differential split-sample test (DSST, Klemeš, 1986) is a powerful procedure to evaluate transferability under climate variability. It consists of calibration and validation exercises for hydrological models using sub-periods with contrasted climate conditions, which make it possible to evaluate model transferability from one climate condition to another. The idea behind performing a DSST is that the errors made when extrapolating from one set of observed climate conditions to another different set could correspond to the errors made when reference data is used for calibration and extrapolation to future climate conditions (Seibert, 2003). The main variables used in the DSST to discretize sub-periods of different climate conditions are precipitation and/or temperature (e.g. Refsgaard and Knudsen 1996; Vaze et al., 2010; Seiller et al., 2012; Tramblay et al. 2013; Ruelland et al., 2015; Hartmann and Bárdossy, 2005; Dakhlaoui et al. 2017), potential evapotranspiration (Coron et al., 2012), and runoff (Seibert, 2003; Vormoor et al., 2018).

The use of DSST is generally based on clustering reference periods on two climate contrasted sub-periods, and generally according to only one climate variable (e.g. Refsgaard and Knudsen, 1996; Seibert, 2003; Wu and Johnston, 2007; Vaze et al., 2010; Ruelland et al., 2015). As a result, it provides a limited number of calibration and validation samples directly linked to the available data, which does not help fully understand the behaviour of the rainfall-runoff models (RRM) under climate variability. Moreover, the implications of RRM robustness under a changing climate remains unknown, since potential changes in climate may go beyond the observed variability (see Ruelland et al., 2012; Guo et al. 2018; Zheng et al., 2018). These limits led several authors to increase the number of calibration-validation exercises and to expand the range of hydro-climatic changes between these periods, to better explore model robustness under climate variability. For example, Hartmann and Bárdossy (2005) divided a 30-year observation period into three sub-periods, first in terms of mean annual temperature (warm, normal and cold), and second, in terms of annual precipitation (wet, normal, and dry years). This made it possible to increase the number of

validation exercises for each climate variable and catchment to six compared to the two DSST periods usually used, and to sample more contrasted sub-periods. Tolson and Shoemaker (2007) divided the observation period into three sub-periods: a 6-year period for calibration and two independent validation series of three years and one year respectively, which allowed them to check the model's performance under contrasted hydrological conditions. Dakhlaoui et al. (2017) simultaneously used mean annual temperature and precipitation values to generate four climate contrasted sub-periods (hot/wet, cold/wet, hot/dry, and cold/wet), which allowed a bigger number of validation exercises (12 per catchment) to be considered and the use of contrasted sub-periods in terms of temperature and precipitation. Coron et al. (2012) developed a generalized version of SST (so-called general split-sample test: GSST, and hereafter called sliding-window SST), which allows a large number of calibration-validation exercises by sampling sub-periods based on a sliding window over the reference period. This technique allowed even more validation exercises than the previous ones and better exploration of observed climate variability in 216 catchments in southeast Australia. Coron (2013) proposed another version of GSST (hereafter called random bootstrap SST) by generating sub-periods according to a bootstrap where a large number of randomly selected combinations of years are used for calibration and for validation. This technique was revisited recently by Arsenault et al. (2018) to evaluate how the length of the calibration period impacted the transferability of two hydrological models in three North American catchments. To explore a larger continuum of model behaviour, Guo et al. (2018) used a stochastic weather generator to generate synthetic climate data to represent future climate conditions, which made it possible to assess the transferability of three RRMs (GR4J, AWBM and CMD) beyond conditions in existing records. However, one limitation of the proposed methodology is that it is based on only one climate variable (precipitation). Furthermore, the authors reported that the stochastic weather generator had difficulty representing natural variability.

1.3. Research needs revealed by the review of literature

Even though the GSST (sliding-window SST or random bootstrap SST) seems to offer the most complete SST, its use presents some limitations. Indeed, the GSST was not specifically designed to identify contrasted periods only, but rather to create an ensemble of conditions, ranging from similar to contrasted. As a result, unlike DSST, it does not explicitly ensure that the most climatically contrasted periods are selected. When applied to continuous years through a sliding-window, the GSST provides contrasted climatic conditions,

which result mainly from the climate trends over the reference period and/or smoothed climatic variability over the selected continuous periods. When applied to discontinuous years through a random bootstrap, the GSST can theoretically select the most contrasted periods if all combinations are sampled, which is computationally unrealistic when considering multi-decadal periods. On the contrary, the DSST applications found in the literature were intended to explicitly identify a climatic contrast between the periods based on a statistical climate analysis, for example by grouping the wettest (driest) years in the same period. On the other hand, the GSST has the major advantage of generating a large number of time periods (composed of a continuum of climatic conditions) for a more complete assessment of the transferability of conceptual hydrological models. This calls for a SST technique which could sample a large number sub-periods composed of discontinuous years and gathering similar to contrasted conditions in terms of precipitation and temperature, while ensuring that the most climatically contrasted sub-periods are sampled.

Although many authors have used DSST to evaluate RRM transferability under climate variability (e.g. Refsgaard and Knudsen 1996; Vaze et al., 2010; Seiller et al., 2012; Tramblay et al. 2013; Ruelland et al., 2015; Hartmann and Bárdossy, 2005; Seibert, 2003), only a few quantified changes in climatic variables which allow acceptable transferability of model results, and generally, only precipitation was used to define the limits of transferability. For example Vaze et al. (2010) tested four rainfall-runoff models in 61 catchments in southwest Australia and suggested that calibration periods of at least 20 years were needed for robust models under climate variability but only if the difference in mean rainfall between calibration and validation period was greater than -15% (for drier climates) and less than +20% (for wetter climates). Similarly, Bastola et al. (2011) found for two Irish catchments that model transferability was less affected when the difference in rainfall between calibration and validation periods was less than 10%. Singh et al. (2011) identified an acceptable range of changes in precipitation (-10% to +20%) with no marked effect on model transferability for five catchments across continental USA. Coron et al. (2012) evaluated the robustness of three RRMs (GR4J, MORDOR6 and SIMHYD) under simultaneous changes in precipitation and potential evapotranspiration (PET) in southeast Australia, and reported that, on average, a 20% absolute bias was observed with a 10%–20% change in precipitation and a 1%–2% change in PET between the calibration and validation periods. However, as mentioned above, the sliding-window SST used in Coron et al. (2012) was not intended to specifically focus on contrasted periods but rather to create a continuum of conditions, from similar climatic conditions to contrasted ones, in order to better evaluate the evolution of

model behaviour with increasing contrasts between different time periods. In a recent study (Dakhlaoui et al. 2017) we evaluated the transferability of three RRMs (GR4J, HBV and IHACRES) under simultaneous precipitation and temperature variability in catchments representative of hydro-climatic conditions in northern Tunisia. We showed that the difference in climate conditions between calibration and validation periods progressively affected the performance of hydrological models. We also showed that the models tested were transferable to wetter and/or colder conditions. However, the model robustness became unacceptable when climate conditions involved a decrease of more than 25% in annual precipitation and an increase in annual mean temperatures of more than +1.75 °C. However the DSST we used only generated small number of calibration and validation exercises (four calibration exercises and 12 validation exercises for each catchment) and provided little information on RRMs transferability under moderate climate changes. In addition, only a few studies have put the RRM transferability limits into perspective in the context of climate projections (see e.g. Singh et al., 2011; Guo et al., 2018). As a result, there is a need for a more precise definition of the limits of transferability of conceptual models in terms of ΔT and ΔP and for these limits to be put into perspective with respect to available high-resolution climate projections. This is of particular importance in the Mediterranean region which is known to be a hot spot of climate change, notably the southern rim (Cramer et al., 2018). Indeed recent climate change scenarios in the Mediterranean region predicted a potential 20% decrease in total precipitation and a +1 °C to +3 °C increase in mean annual temperature by the 2050 horizon compared with the 1971–1990 period (Milano et al., 2012; Schilling et al., 2012; Terink et al., 2013; Tramblay et al. 2018). These climate changes would have a considerable effect on the surface water resources of southern Mediterranean countries which already suffer from water paucity (Blinda and Thivet, 2009). On the other hand, the few studies which evaluated the hydrological impacts of climate change in the Mediterranean region (e.g. Milano et al., 2012; Drooger et al. 2012; Sellami et al., 2016) did not take into account the limits of transferability of the hydrological models they used. This calls for an evaluation of the effect of the expected loss in performance of RRMs under future climate change and for more reliable hydrological projections to enable better climate change adaptation strategies.

1.4. Objectives

This paper proposes an improved SST technique to test the robustness of hydrological models by selecting time periods with contrasted conditions in terms of temperature and precipitation. The proposed technique

171 was compared to three other split-sampled methods to demonstrate its efficiency. It was then used to assess
172 the transferability of conceptual rainfall-runoff models under multi-decadal climate variability with a view to
173 simulating hydrological scenarios based on high-resolution climate projections in northern Tunisia. We
174 compared the limits of transferability with future climate scenarios obtained from high resolution EURO-
175 CORDEX regional climate models. Finally we analysed the effect of the selection of the calibration period
176 on the limits of RRM robustness on the hydrological projections.

177 **2. DESCRIPTION OF THE SPLIT-SAMPLE TECHNIQUES**

178 **2.1. Three benchmark SST techniques**

179 The SST methods selected to be tested on the study catchments were: (i) a sliding-window SST (Coron et
180 al., 2012); (ii) a random bootstrap SST (Coron, 2013; Arsenault et al., 2018); and (iii) a 4-sub-period DSST
181 (Dakhlaoui et al., 2017). These three techniques were selected because they enable simultaneous
182 investigation of the effect of T and P on model transferability under climate variability. The first technique
183 was adopted in several studies (Coron et al., 2012; 2014; Guo et al., 2018; Vormoor et al., 2018), the second
184 one inspired our proposed DSST to randomly combine independent years in the sample and the third one
185 was recently successfully used in northern Tunisia to sample very climatically contrasted sub-periods. These
186 three techniques are described in Figure 1 and Table 1.
187
188 Figure 1 to be inserted near here (colour).
189
190 The sliding-window SST technique (Coron et al., 2012) consists in using calibration-validation tests on
191 independent sub-periods of equal length, considering all possible pairs of sub-periods. The sampling method
192 used to generate sub-periods is based on sliding windows applied over the reference period. The technique
193 enables the identification of l-n+1 calibration sub-periods, where n is the number of years composing each
194 sub-periods and l is the total number of years of the reference period.
195 The random bootstrap SST technique (Coron, 2013; Arsenault et al., 2018) relies on a sub-period
196 sampling technique which is based on a random combination of discontinuous years (bootstrap). This
197 sampling technique is time consuming since the possible number of calibration sub-periods is equal to C_n^l .
198 For example the random bootstrap SST technique results in around six million possible 8-year sub-periods if

applied to a 30-year reference period. Its application then requires a priori selection of the number of permitted calibration exercises, due to limited time budget for model calibration and validation.

The implementation of the 4-sub-period DSST (Dakhlaoui et al., 2017) requires the calculation of the annual precipitation and mean temperature for each hydrological year of the reference period. The sub-periods are thus made up of clusters of climatically contrasted years. To create these clusters, the hydrological years are first distributed into two equal groups of hydrological years (dry years and wet years) according to the annual precipitation median for the reference period (Fig. 1c). Dry and wet years are defined as years with respectively less or more total precipitation than the median of the reference period. For each group, the median of the mean annual temperature is then calculated, which serves to distinguish hot and cold years. The four final groups of hydrological years are: hot/dry (HD), hot/wet (HW), cold/dry (CD) and cold/wet (CW) years (see Figure 1).

Using the three above techniques makes it possible to identify different numbers of calibration sub-periods of n years (see Table 1). All n-year periods which do not have any year in common with a given n-year calibration period can thus be considered as independent validation exercises. As a result, the number of validation exercises may not be the same for all calibration periods selected with the sliding-window and random bootstrap SST. For the 4-sub-period SST, there are three possible validation exercises for each of the 4 calibration sub-period.

Table 1 Overview of the split-sample test techniques. n is the number of years composing each sub-periods, l is the total number of years of the reference period, SST stands for split-sample test and DSST for differential split-sample test.

Method	Sliding-window SST	Random bootstrap SST	4 –sub-period SST	GDSST
Reference	Coron et al., 2012	Coron, 2013 Arsenault et al., 2018	Dakhlaoui et al., 2017	Current paper
Test type	SST	SST	DSST	DSST
Sub-period	Continuous years	Discontinuous years	Discontinuous years	Discontinuous years
Sub-period generation technique	Sliding window	Random bootstrap	hot/dry, hot/wet, cold/dry and cold/wet years	Oriented random bootstrap
Number of sub-periods	l-n+1	C_n^l as maximum, need to be defined a priori	4	C_n^l as maximum, need to be defined a priori
Number of validation exercises	All n-year periods which do not have any year in common with the n-year calibration period	All n-year periods which do not have any year in common with the n-year calibration period	12	All n-year periods which do not have any year in common with the n-year calibration period

220 **2.2. Proposal for a general differential split-sample test (GDSST)**

221 Based on the existing SST methods, we developed a technique which can take benefit from the random
222 bootstrap SST technique to provide a large number of validation exercises while accounting for the much
223 contrasted ΔT and ΔP detected with the 4-sub-period DSST. In other words, the idea was to design a method
224 which uses the sampling of the random bootstrap SST technique, but which is oriented so as to obtain the
225 extreme climate contrast provided by the 4-sub-period DSST. The proposed method was called general
226 differential split-sample test (GDSST) and is described in Figure 2.

227 The procedure used to generate *k* *n*-year sub-periods from the *l* hydrological years (from the 1st of
228 September to the 31st of August) of reference period, is as follows. The first year of the *n*-year sub-period to
229 be sampled is randomly selected from the *l* years of the reference period (step 1 in Fig. 2). The *l*-1 remaining
230 years of the reference period are then sorted based on the order of increasing distance of Mahalanobis (1936)
231 to the first selected year in the space of mean annual temperature (T) and total annual precipitation (P) (step
232 2 in Fig. 2). The Mahalanobis distance is computed according to the following expression:

233

234 $$d\left(\overrightarrow{C_y}, \overrightarrow{C_1}\right) = \sqrt{(\overrightarrow{C_y} - \overrightarrow{C_1})^T \Sigma^{-1} (\overrightarrow{C_y} - \overrightarrow{C_1})} \quad (1)$$

235

236 where $d\left(\overrightarrow{C_y}, \overrightarrow{C_1}\right)$ is the Mahalanobis distance between a year *y* from the *l*-1 remaining years of the reference
237 period and the first selected year in the space of mean annual temperature (T) and total annual precipitation
238 (P); $\overrightarrow{C_y}$ is a vector representing a year *y* from the *l*-1 remaining years of the reference period in the T and P;
239 $\overrightarrow{C_1}$ is a vector representing the first selected year in the T and P space; Σ is the covariance matrix between T
240 and P of the *l* years of the reference period.

241 Using the Mahalanobis distance aims at rescaling the T and P axes in order to account for the
242 correlations between the two variables and to calculate standard Euclidean distance in a transformed space
243 having unit variance. In other words, it aims to reduce the dominance of one climatic variable over the other
244 when computing "climatic" distance between years. A trapezoidal probability is then assigned to the *l*-1
245 remaining years of the reference period, as follows (step 3 in Fig. 2):

246

247 $$P(i) = 2(m + 1 - i)/m(m + 1) \ , i = 1, \dots, m \quad (2)$$

$$P(i) = 0\,, i = m+1, \ldots, l-1 \quad (3)$$

where P(i) is the probability assigned to the year with rank i; i is the rank of the remaining years of the reference period sorted in order of increasing Mahalanobis distance to the originally selected year; m is a number selected randomly at each sub-period selection from the interval [n-1, l-1]. The year closest to the year originally retained has the highest probability ($P(1) = 2/m+1$) and the farthest years has the lowest probability ($P(m) = 2/m(m+1)$ and $P(i) = 0$ for $i > m$).

The n-1 remaining years of the sub-period are then selected from the l-1 remaining years of the reference period according to the trapezoidal probability distribution giving more chance to be selected to the years which are the closest to the initial year retained according to the Mahalanobis distance defined in the T and P space (step 4 in Fig. 2). The trapezoidal distribution allows only the m years closest to the initial year retained, to be selected in the sub-period. This gives more chance to years with similar climatic conditions to be selected in order to generate more climatically contrasted sub-periods. However, varying randomly m for each sub-period generation also allows years with different climatic conditions to be selected. This aims at creating a continuum of climatic conditions, from similar to contrasted, between the sampled sub-periods in view of evaluating the model transferability under increasing climate contrasts. In case the new created sub-period was already sampled, it is not retained (step 5 in Fig. 2). The procedure (steps 1 to 5 in Fig. 2) is repeated until the required number of sub-periods is reached (step 7 in Fig. 2).

The random selection of years in the proposed procedure allows a larger number of sub-periods to be selected than with a deterministic procedure (where the closest years to the originally retained year are selected). In fact, in the best case, the deterministic procedure provides a number of sub-periods equal to the number of observed years (e.g. 30 sub-periods for a 30-year reference period). The number of calibration sub-periods which can be generated by the proposed technique is similar to the random bootstrap SST technique (C_n^l). That is why its application requires a priori selection of the number of permitted calibration exercises. Similarly to the three benchmark SST (section 2.1), all n-year periods which do not have any year in common with a given n-year calibration period can be considered as independent validation exercises with the GDSST (see Table 1).

277 Figure 2 to be inserted near here (colour).

278

279 **3. EVALUATION PROTOCOL**

280 In this section, we present successively the study basins, the hydro-climatic data and the methods to assess
281 the transferability of three hydrological models under climate-contrasted conditions based on the proposed
282 GDSST.

283 **3.1. Study basins**

284 Five catchments located in Northern Tunisia (Fig. 3) were used for the evaluation protocol. These basins
285 were selected based on the following criteria: (i) their streamflow regime can be considered as natural since
286 they are located upstream from major hydraulic installations, such as dams and water transfers; (ii) the
287 availability of hydro-climatic series for the same 30-year period to enable sufficiently climate contrasted
288 sub-periods to be sampled for the split-sample tests; and (iii) the availability of good quality hydrological
289 data according to hydrological reports, with the aim of reducing the impact of data errors on the results. The
290 basins are located in the region that produces most of the surface water in Tunisia (Baouab and Cherif,
291 2015). The study catchments are situated within a semi-arid to humid Mediterranean climate with a hot
292 season (Henia, 2008). The hydro-climatic characteristics are given in Figure 3. More details on the basins
293 can be found in Dakhlaoui et al. (2017).

294

295 Figure 3 to be inserted near here (colour).

296

297 **3.2. Hydro-climatic data**

298 3.2.1. In-situ meteorological data

299 A total of 123 daily rain-gauges located in the study region were used (see Fig. 3). These stations were
300 selected because they had less than 30% of daily gaps in the period 1970–2000, thus providing a stable,
301 coherent network of measurements for the spatial interpolation of precipitation forcing. Eight meteorological
302 stations with monthly mean series of daily minimum and maximum temperatures (Fig. 3) were used to

compute mean monthly air temperatures, which were considered as daily values. Climate forcing was interpolated on a 2-km grid with the inverse distance weighting technique. Temperature (T) was interpolated by accounting for a lapse rate of -0.65 °C/100 m (see details on the method used in Ruelland et al., 2014) while precipitation (P) was interpolated by accounting for altitude via a 4.10^{-4} corrective factor in the exponential function proposed by Valéry et al. (2010). The formula of Oudin et al. (2005) was chosen to estimate PET. This formula is based on mean daily air temperature and on estimated clear daily sky solar radiation depending on the latitude of the grid cells.

3.2.2. High-resolution climate simulations

The climate model data used (Table 2) are simulations and projections of daily mean temperature and precipitation from eight pairs of RCMs (Regional Climate Models) forced by different Global Circulation Models), from the EURO-CORDEX initiative, the most recent climate simulations for the Euro-Mediterranean region with 0.11° resolution (~12x12 km). The model data for the period 1951–2005 correspond to the historical simulation, and 2006–2100 corresponds to future projections. Two Radiative Concentration Pathway (RCP) scenarios were used for the future projections: RCP 4.5 and RCP 8.5, which represent respectively moderate and strong greenhouse gas emission scenarios. The climate projections were based on two 30-year future periods: one medium-term (2040–2070) and one long-term (2070–2100) horizon. They were compared to the past reference period (1970–2000).

Table 2 RCMs selected from the EURO-CORDEX initiative.

Model acronym	Name of RCM	Modelling center (country)	Name of GCM	Modelling center (country)	Reference
CLM-HAD	CLM11	CLM community (USA)	HadGEM2-ES	Met Office Hadley Centre (UK)	Rockel et al. (2008)
CLM-MPI	CLM11	CLM community (USA)	MPI-ESM-MR	Max-Planck-Institut für Meteorologie (Germany)	Stevens et al. (2013)
CNR-CNR	ALADIN 5.3	Centre National de Recherches	CNRM-CM5	Centre National de Recherches	Voldoire et al. (2013)

		Météorologiques (France)		Météorologiques (France)	
KNM-ECE	KNM11	Royal Netherlands Meteorological Institute	EC-EARTH	EC-EARTH consortium (Europe)	van Meijgaard et al. (2012)
SMH-CNR	SMH11	Swedish Meteorological and Hydrological Institute	CNRM-CM5	Centre National de Recherches Météorologiques (France)	Voldoire et al. (2013)
SMH-ECE	SMH11	Swedish Meteorological and Hydrological Institute	EC-EARTH	EC-EARTH consortium (Europe)	Samuelsson et al. (2011)
SMH-HAD	SMH11	Swedish Meteorological and Hydrological Institute	HadGEM2-ES	Met Office Hadley Centre (UK)	Samuelsson et al. (2011)
SMH-MPI	SMH11	Swedish Meteorological and Hydrological Institute	MPI-ESM-MR	Max-Planck-Institut für Meteorologie (Germany)	Stevens et al. (2013)

3.3. Model's transferability evaluation under climate-contrasted conditions

3.3.1. Hydrological models

Three conceptual hydrological models, running at a daily time step, were used: GR4J (Perrin et al., 2003), HBV (Bergström, 1976; Bergström and Lindström, 2015) and IHACRES (Jakeman et al., 1990; Croke and Jakeman, 2004). They were selected because they are parsimonious models and all have to be calibrated based on precipitation, PET and runoff data, but differ in the way they conceptualise the hydrological processes and in their complexity (4 to 8 free parameters, 2 to 3 conceptual reservoirs, see Table 3). This makes inter-comparison of the model simulations possible. All three models have recently been applied in Tunisia (see Dakhlaoui et al., 2009; 2012; 2017). Further description on the model versions used can be found in Dakhlaoui et al. (2017).

Table 3 Overview of the characteristics of the three tested models (modified from Dakhlaoui et al., 2017).

	GR4J	HBV	IHACRES
Number of parameters	4	8	5
Production module	Precip. interception by PE,	PE extracted from stored soil moisture	Precip. interception by PE,

	Non-linear soil moisture (SM) accounting store,	The level of soil moisture accounting store determines the quantity of precipitation intercepted by the soil	Non-linear soil moisture accounting store,
	Actual evapotranspiration parabolic function (of SM), Yes	Actual evapotranspiration piecewise linear function (of SM)	Actual evapotranspiration exponential function (of SM)
Water exchange	Inter-catchment groundwater flow	No	No
Routing module	Two unit hydrographs, Non-linear routing store	Two conceptual reservoirs, the upper is non-linear (direct runoff) and the lower is linear (subsurface runoff) One unit hydrograph	Two unit hydrographs in parallel equivalent to two linear routing stores
Source of first publication	Perrin et al. (2003)	Bergström (1976)	Jakeman et al. (1990)

345

346 3.3.2. Calibration and validation methods

347 Model robustness was evaluated through a series of calibration and validation exercises under contrasted
348 precipitation/temperature conditions according to the proposed GDSST. The three hydrological models
349 (GR4J, HBV and IHACRES) were calibrated for the five study catchments over the 100 sub-periods
350 generated by the oriented bootstrap technique. One optimal parameter set per sub-period was thus obtained
351 by calibration. The parameter sets obtained were then used to perform all possible independent validation
352 exercises (i.e. any sub-periods which do not have any year in common with the calibration sub-period). The
353 proposed GDSST method thus enabled testing of the models under different conditions from those used for
354 calibration. Although discontinuous sub-periods were used for both model calibration and validation, the
355 models were run in a continuous way for the whole reference period, while only the years that corresponded
356 to calibration or validation periods were taken into account to compute the efficiency criteria. A 3-year
357 warm-up period (September 1967 to August 1970) was considered before the whole reference period to limit
358 the effect of the storage initialization.

359

360 The model parameters were calibrated using the Kling-Gupta Efficiency index (KGE, Gupta et al., 2009),
361 which represents a compromise between three evaluation criteria (correlation coefficient, bias error and
362 standard deviation ratio) expressed as follows:

363

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \quad (4)$$

where r is the linear correlation coefficient between observed (Q_{obs}) and simulated (Q_{sim}) flows, α is a measure to compare the variability of the observed and predicted data (equal to the standard deviation of Q_{sim} over the standard deviation of Q_{obs}), and β is a measure of bias (equal to the mean of Q_{sim} over the mean of Q_{obs}).

Calibration was performed in a 4D, 8D and 5D parameter space for GR4J, HBV and IHACRES respectively, by searching for the maximum value of KGE. To perform this optimization exercise efficiently, we used the Shuffle Complex Evolution (SCE) algorithm (Duan et al., 1992). The algorithmic parameters of SCE-UA were set to the values recommended by Duan et al. (1994). We also followed the recommendation of Kuczera (1997), who suggested a number of complexes equal to the number of parameters to be optimised and assumed that with these security measures, the risk that SCE-UA falls in local optimal solutions is considerably reduced. The parameter feasible ranges were set to usual parameter limits (see Dakhlaoui et al., 2017). Although the hydrological models were run at a daily time step, their calibration and validation were performed at 10-day time scale (i.e. based on 10-day mean values), since preliminary tests showed that day-to-day variability was difficult to reproduce accurately due to the limited quality of data in space and over time.

Although models were calibrated by KGE Criteria, their performance (transferability) during validation was evaluated based on the analysis of the Nash and Sutcliffe Efficiency index (NSE, Nash and Sutcliffe, 1970) and the cumulated volume error (VE). We selected these two efficiency criteria because, they are commonly used by hydrologists, thus making easier their interpretation notably when defining the model transferability limits. The NSE criterion is as well-known form of the normalized least squares objective function. It represents the overall agreement of the shape of the hydrograph, while placing more emphasis on high flows. Perfect agreement between the observed and simulated values yields an efficiency of 1, while a negative efficiency represents a lack of agreement worse than if the simulated values were replaced with the observed mean values. VE represents the agreement of cumulated runoff volume during the simulation period and is expressed through the proportional difference to observed values. Its optimal value is zero. These criteria are computed according to the following equations:

$$\text{NSE}=1-\left(\sum_{i=1}^{n} (Q_{sim} - Q_{obs})^2 \Big/ \sum_{i=1}^{n} (Q_{obs} - \overline{Q}_{obs})^2 \right) \quad (5)$$

where:

$\overline{Q}_{obs}$: mean observed runoff.

$$VE = \left(\sum_{i=1}^{n} Q_{sim} - \sum_{i=1}^{n} Q_{obs} \right) \Big/ \sum_{i=1}^{n} Q_{obs} \quad (6)$$

To underline model instability between periods with contrasted climate conditions, parameter transferability was assessed as a function of the climate change variables (ΔT and ΔP) between the validation and calibration periods. Since NSE is based on a ratio of the squared model error to the variance of observed flows, any changes in variance or volumes between climatically contrasted periods (dry/wet) can affect result of the comparison. We thus chose to evaluate model transferability by calculating the differences between NSE resulting from the calibration period (RR, receiver) and the NSE calculated in the same period but with parameters provided by model calibration on other sub-periods (DR, donor). This made it possible to represent the results obtained from the GDSST classified in a grid of ΔT and ΔP, with a step of 0.2 °C and 5%, respectively (see Fig. 5).

4. RESULTS AND DISCUSSION

4.1. Comparison of the different SST techniques

The three benchmark SST techniques and the proposed GDSST were applied to the five study catchments in northern Tunisia. The techniques were compared as regards to the number of validation exercises and to the precipitation-temperature differences they provided. The 30-year reference period over each catchment was based on the hydrological years (from the 1st of September to the 31st of August). The length of the sub-periods was set to 8 years for the sliding-window SST, the random bootstrap SST and the GDSST. However, for the 4-sub-period DSST, the 30-year reference period was spread over 7-8 years hot/dry, hot/wet, cold/dry and cold/wet sub-periods (since 30 is not a multiple of four). A ~8-year time span was judged to be suitable for model calibration. Indeed, several studies have shown that three to eight years are generally sufficient for

418 model calibration with validation using a classical split-sample test (e.g. Yapo et al., 1996; Anctil et al.,
419 2004). Moreover, using 8-year periods still provided significant differences in mean climate due to the
420 length of available records (30 years) and the high variability of the climate conditions under study. As a
421 result, we assumed it was an acceptable compromise between the length of the period needed for calibration
422 and the number of possible combinations between calibration-validation periods.

423 The random bootstrap SST results in a large number of possible sub-periods if fully applied to a 30-
424 year period (around six millions 8-year sub-periods). Due to limited time budget for model calibration and
425 validation, we decided to use only 100 randomly selected sub-periods for each catchment. For sake of fair
426 comparison, the same number of randomly selected sub-periods was set with the GDSST. Note that the
427 number of sub-periods for the two other techniques is already limited by their design: 23 sub-periods for the
428 sliding-window SST and four for the 4-sub-period DSST.

429 Figure 4 shows the number of validation exercises obtained with the different sampling techniques
430 according to a grid of the differences in mean annual temperature and precipitation between the validation
431 and calibration sub-periods (ΔT and ΔP with a step of 0.2 °C and 5%, respectively). When a given ΔT and
432 ΔP did not exist in the classification sample, the corresponding square in the figure was coloured in grey.
433 The figure allows the spread of the sample provided by each sampling technique to be evaluated in terms of
434 ΔT and ΔP. Additionally, Table 4 summarizes the number of calibration-validation exercises and the ranges
435 in ΔT and ΔP provided by the different sampling techniques applied over the 1970–2000 period in the five
436 studied basins.

437 Figure 4a shows the sample offered by the sliding-window SST technique when applied to the five
438 study catchments. It provided 1 495 possible validation exercises for 115 (23 x 5 basins) calibration
439 exercises. The differences between the different sub-periods in mean precipitation ranged from -20% to
440 +25% and the differences in temperature ranged from -1.8 °C to +1.8 °C. When looking at the random
441 bootstrap SST technique (Fig. 4b), it provided 5 800 possible validation exercises for a total of 500
442 calibration (100 x 5 basins) exercises. The differences between the different sub-periods in mean
443 precipitation ranged from -35% to +50%, and in temperature from -1.4 °C to +1.4 °C. The 4-sub-period
444 DSST (Fig. 4c) provided 60 possible validation exercises from 20 (4 x 5 basins) calibration sub-periods. The
445 differences in mean precipitation obtained ranged from -40% to +60%, and in temperature from -2 °C to
446 +2 °C. Like the random bootstrap SST, the proposed GDSST (Fig. 4d) provided many possible validation

exercises (9 320) from a total of 500 calibration (100 x 5 basins) exercises. However, the differences in mean precipitation obtained ranged from -45% to +80%, and in temperature from -2 °C to +2 °C.

Table 4 Number of calibration-validation exercises and ranges in ΔT and ΔP provided by the four split-sample methods applied over a 30-year reference period (1970–2000) in the five studied basins. ΔT and ΔP represent respectively the differences in mean annual temperature and the relative difference in annual precipitation between the calibration and validation sub-periods.

Method	Sliding-window SST	Random bootstrap SST	4 –sub-period SST	GDSST
Number of calibration exercises	115 (23 x 5 basins)	500 (100 x 5 basins)	20 (4 x 5 basins)	500 (100 x 5 basins)
Number of validation exercises	1 495 (299 x 5 basins)	5 800 (~1160 x 5 basins)	60 (12 x 5 basins)	9 320 (~1 864 x 5 basins)
Range of ΔT	[-1.8 °C; +1.8 °C]	[-1.4 °C; +1.4 °C]	[-2.0 °C; +2.0 °C]	[-2.0 °C; +2.0 °C]
Range of ΔP	[-20%; +25%]	[-35%; +50%]	[-40%; +60%]	[-45%; +80%]

Although the sliding-window SST technique provided numerous validation exercises, the differences in P (ΔP) were less contrasted than those offered by the three other techniques. The sliding-window technique thus appears to depend too much on the historical climate trends to detect extremely contrasted sub-periods for calibration. Using this method, Coron et al. (2012) found well contrasted precipitation in southeast Australia. However, the authors reported precipitation trends that contributed to obtain a significant contrast in precipitation characteristics between different periods. In northern Tunisia, continuous sliding periods were unable to provide sufficiently contrasted periods because there was no trend in precipitation during the historical study period, as shown by Dakhlaoui et al. (2017). In addition, the study area presents high inter-annual precipitation variability (see also Dakhlaoui et al., 2017). Using continuous sub-periods thus smooths the average precipitation in the sub-periods, thereby reducing the climate contrast between them. However this is not the case for temperature, for which the sliding-window SST technique provided significant differences in T (ΔT) due to the increasing temperature trends in northern Tunisia over 1970–2000 (Dakhlaoui et al., 2017). The random bootstrap SST technique provided an important number of validation exercises (5 800). However it led to limited differences in T (ΔT) and a poor distribution of the sample with high concentration in the centre of the figure, where there is the least significant contrast to test model parameter transferability. The 4-sub-period DSST provided more contrasted ΔT and ΔP than the sliding-window and random bootstrap SST. Indeed it is based on a sampling technique generating highly climate-contrasted sub-periods. However, although it explored contrasted climatic conditions in the historical period,

473 the technique provides very few insights into moderate ΔT and ΔP compared to the other techniques. The
474 oriented bootstrap of the GDSST provided more validation exercises than the random bootstrap SST,
475 although both techniques were based on the same number of calibration exercises (500). This can be
476 explained by the fact that the oriented bootstrap favours the selection of independent sub-periods by
477 reducing overlap between them. In addition, the GDSST provided a better spread of validation periods.
478 Indeed, contrary to the random bootstrap technique in which the validation exercises were concentrated in
479 the zone of ΔT and ΔP near 0, the sample provided by the GDSST technique was more concentrated at the
480 extremes ΔT and ΔP, which are the most contrasted sub-periods to test the parameter transferability. Hence,
481 the differences in mean precipitation and temperature between the different sub-periods ranged respectively
482 from -45% to +80%, and from -2 °C to +2 °C, thus providing a more marked climatic contrast between the
483 calibration and validation periods compared with the previous techniques (see Fig. 4).
484
485 Figure 4 to be inserted near here (colour).
486
487 It should be noted that the random bootstrap technique theoretically includes all the spread of ΔT and ΔP
488 provided by the other techniques tested. In other words, the theoretical limits of the tested combinations (if
489 all possible combinations were sampled) should be as large as the largest limits provided by all the other
490 techniques. However the problem is that the application of a bootstrap on all combinations would require
491 excessive computation time and would lead to a very large number (~6 million of 8-year sub-periods) of
492 combinations that could obviously not be tested through cross-validation with hydrological models. The
493 proposed GDSST has the advantage to be more effective: with only a limited number of calibration exercises
494 (100), it provides a large number of sub-periods from similar to contrasted conditions in terms of
495 precipitation and temperature, while ensuring that the most climatically contrasted sub-periods are sampled.

496 **4.2. Model transferability under climate-contrasted conditions using GDSST**

497 These encouraging results led us to use the GDSST to assess the model's transferability under climate-
498 contrasted conditions (see protocol in section 3.3.).

499 Based on the NSE criterion, the model's transferability decreased with an increase in temperature and a
500 decrease in precipitation (Fig. 5b). The simultaneous increase in T and decrease in P resulted in a significant

reduction in model performance. In contrast, an increase in precipitation and/or a decrease in temperatures had a much moderate impact on model efficiency (sometimes even leading to a slight improvement in performance), providing evidence for better parameter transferability under wetter and/or colder conditions. The cases where the model performance was improved can be explained by the fact that the evaluation criterion (NSE) is different from the calibration objective function (KGE). By the way, there were no cases where the KGE value in validation on a given period was larger than the KGE value in calibration on the same period (see Fig 5a), which shows that the method and the optimization algorithm are consistent and robust. Figure 5c shows that VE increases with an increase in temperature and a decrease in precipitation. The water balance decreases when temperature decreases and precipitation increases. This means that runoff is overestimated when moving to hotter and drier climate conditions and underestimated in the reverse case. The transferability as regards to the VE criterion was more affected by changes in precipitation (P) than by changes in temperature (T). An increase in T and/or a decrease in P between periods weakened the model robustness. On the opposite, the transferability is very satisfactory when ΔT and ΔP are low, i.e. when climate conditions are rather similar between the receiver and donor periods (including periods with extreme conditions). This means that the models are robust when applied on periods with similarly climate conditions, whether extreme or not.

Figure 5 to be inserted near here (colour).

Figure 5d shows the limits of transferability of the hydrological models as a function of ΔT and ΔP. These limits were defined according to a decrease in NSE of more than 0.2 and a variation in VE of more than +/- 25%. We acknowledge that these thresholds are somewhat subjective and should be adapted depending on the hydrological conditions and according to the user need, for example to match a sustainable level of uncertainties for water resources management in a given context. For the current study, NSE values were always greater than 0.8 while VE values were always around 0 for all calibration periods and catchments. We thus assumed that beyond a 0.2 decrease in NSE criterion and a 25% increase in VE, the simulations were no longer efficient. The red grid shows non-transferable areas. These results show that transferring parameters to different climate conditions resulted in significant uncertainties when the shift is to a hotter drier climate. Compared to the other hydrological models, GR4J appeared to be the least affected by changes

in temperature and the most affected by changes in precipitation. The limited effect of changes in temperature on the VE criterion can be partly explained by the fact that the changes in temperature mainly occur in the dry season, which contributes little to runoff.

The limits of transferability of the models show clear interdependence of precipitation and temperature. Given the criteria and the thresholds retained, they can be approximated according to an acceptability line (Fig. 5d) computed from a linear relation between ΔT and ΔP: the models are thus roughly transferable for changes in precipitation $\Delta P < (0.08*\Delta T - 0.18)$, with $\Delta P \in$ [-30%, +80%] and changes in temperature $\Delta T \in$ [-2 °C, +2 °C]. These limits are more precise than those presented in previous studies, which showed squared transferability limits (see Coron et al., 2012 and Dakhlaoui et al., 2017). However, it should be noted the presence of outliers above the acceptability line. These outliers correspond to low density of validation samples (see Fig. 4d) and it is difficult to give general conclusion about them. For instance another limit towards wetter and colder conditions may also exist, but it is difficult to identify within the obtained results. Moreover, although Figure 5 provides evidence of transferability problems when moving to drier and hotter conditions, the interdependence of temperature and precipitation partly hides the fact that the models are probably less transferable towards drier conditions than towards hotter conditions.

4.3. Analysis of high resolution climate simulations

4.3.1. Efficiency of the RCMs over the control period

Figure 6 shows the raw RCM outputs (P and T) versus climate observations over the reference period 1970–2000. Figures 6a and 6b compare the mean seasonal precipitation (temperature) observed with the one simulated by each of the eight RCMs over the reference period, while Figure 6c shows the relative errors between mean simulated and observed annual precipitation and mean errors between mean simulated and observed annual temperature.

The RCMs did not correctly reproduce precipitation. NSE values between mean seasonal simulated precipitation and mean seasonal observed precipitation range from negative values to 0.7 depending on the climate models and the catchments (Fig. 6b). The RCMs tested were thus not able to accurately reproduce the average seasonality of precipitation, especially during the wet season. In addition, many models were not even able to represent average annual precipitation, which leads to significant over- or under-estimation of rainfall (see Fig. 6c). The RCM simulations of temperature show a better agreement with observations. The

NSE values were generally above 0.85 (Fig. 6b) showing the good performance of RCM in reproducing the temperature seasonality. The difference between mean annual simulated and observed temperature was very small for CLM-HAD, CLM-MPI, SMH-HAD and SMH-MPI (Fig. 6c). However this difference was greater than 2 °C for CNR-CNR and KNM-ECERCM.

The limited efficiency of RCM (notably in reproducing observed precipitation over the reference period) hampered the direct use of climate model raw outputs for building climate scenarios. Moreover, since CNR-CNR RCM was the least efficient in reproducing the past observed climate in the studied basins, it was excluded from the ensemble. While this did not guarantee better future projections, we nevertheless considered it an essential step to obtain the most reliable and relevant simulations for future projections.

4.3.2. Climate scenarios for the medium- and long-term horizon

The RCM bias in reproducing reference climate (notably precipitation volume and seasonal patterns) led us to apply a simple delta-change method in order to produce a range of climate scenarios from the RCM outputs. All simulations of climate change were thus based on the historical Representative Concentration Pathway (RCP) over the reference period (1970–2000) and scenarios RCP 4.5 and 8.5 for a medium-term horizon (2040–2070) and a long-term horizon (2070–2100). High-resolution climate change forcing was thus obtained by a monthly perturbation method, which assumes that climate models reproduce the relative change in climatic variables better than their absolute values. The method consists in producing future climate scenarios by modifying the observed climatic series so as to reproduce the mean monthly variations obtained between the reference and future climatic simulations produced by climate models. For more details on the method used, see Ruelland et al. (2012).

Figure 6 to be inserted near here (colour).

Figure 7a compares the mean seasonal precipitation observed over the reference period and projected precipitation according to the four combinations of horizons and RCPs, for each of the seven selected RCMs. Figure 7b shows the same comparison for temperature. The climate change signal is very different from one RCM to another especially for precipitation. However all RCMs predict a warmer climate and almost all climate models predict dryer conditions in the future. At the medium-term horizon, a change of

586 +10% to -23% (+6.3% to -35%) in total precipitation is projected under the RCP4.5 (RCP8.5) scenario. At
587 the long-term horizon, a change of +12% to -30% (+15% to -52%) in total precipitation is projected under
588 the RCP4.5 (RCP8.5) scenario. These changes mainly occur in the wet season (November to April). For
589 temperature (Fig. 7b) at the medium-term horizon, an increase of +1.0 to +1.5 °C (+1.8 to +3.5 °C) is
590 projected under the RCP4.5 (RCP8.5) scenario. At the long-term horizon, an increase of +1.7 to +3.3 °C
591 (+3.2 to +5.7 °C) is expected under the RCP4.5 (RCP8.5) scenario. In contrast to precipitation, these
592 changes are expected to occur mainly during summer. The changes in temperature are likely to have little
593 impact on discharge, since there is almost no runoff in summer, but the decrease in winter precipitation may
594 have a critical impact on water resources.
595
596 Figure 7 to be inserted near here (colour).
597

598 **4.4. Comparing climate projections with model parameter transferability**

599 4.4.1. Analysis of the transferability of the models to climate-contrasted periods

600 This section exploits one of the main results obtained in section 4.2 which showed that the difference in
601 climate conditions between calibration and validation periods progressively affects the performances of
602 hydrological models (see Fig. 5).

603 For this purpose, three cases were used for the hydrological projections with each model. The first case
604 uses the parameter sets calibrated over the whole (WHO) reference period (1970–2000). The second case
605 uses the parameter sets calibrated over the 12-year sub-period the closest to the future climate (MSP).The
606 third case uses the parameter sets calibrated over the 12-year sub-period the most different from the future
607 climate (MDP). The median of the seven RCM projections was considered for each RCP, horizon, and
608 catchment. Three hundred sub-periods were generated according to the proposed GDSST to increase the
609 chance of finding the MSP and MDP sub-periods. The closest (the farthest) sub-period to the future climate
610 (according to Mahalanobis distance as regards to the pluviometric and temperature conditions) were
611 considered as MSP (MDP). It should be noted that given the time span (30 years) of the projection periods,
612 we decided to increase the calibration period length from eight (see section 4.2. for the tests regarding the
613 GDSST) to 12 years. The rationale behind this increasing length of calibration periods was to reduce the

ratio of number of years between the calibration period (12) and the future periods (30) under study, thus limiting over-fitted calibration when applying the models to the future 30-year periods.

Figure 8 shows the behaviour of VE and the transferability limits obtained by the GDSST applied to the reference period (1970–2000) as described in section 4.2. For each RCM, RCP, horizon and catchment, three points are shown: one for the future climate seen from the whole period, one for the future climate seen from the MSP, and one for the MDP. The cross indicates the mean reference climate conditions over the 30-year past period or over the most similar or most different 12-year period, i.e. with 0 coordinates as a reference. This makes it possible to position future climate conditions (RCM) in relation to the transferability limits of the models. In the figure, the past climate conditions (WHO, MSP or MDP) are thus considered as calibration periods with respect to the climate conditions predicted by the RCM.

We found that under RCP 4.5, whatever the horizon, the transfer of parameter sets calibrated over the whole 30-year period to the future climate conditions would be acceptable with respect to the defined transferability limits. However, this was not the case under RCP 8.5 at the medium-term horizon, whereas choosing parameter sets calibrated over the MSP would be acceptable. Given the long-term climate predicted under RCP8.5, the parameter sets calibrated over the whole 30-year period or over the most similar 12-year sub-period both fell outside the transferability limits.

Figure 8 to be inserted near here (colour).

4.4.2. Sensitivity of the hydrological projections to the selected calibration period

Figure 9 shows the relative changes in runoff volume in the hydrological projections performed by the three rainfall-runoff models forced by the median of the climate projections of the seven RCMs retained. Three parameter sets (WHO, MSP and MDP) were used for each case. The predicted changes in precipitation and temperature are transformed into hydrological projections by a change of +0.14% to -6.2% in mean annual runoff under RCP4.5 and of -13% to -31% under RCP8.5 at the medium-term horizon, and respectively -16% to -29% and -37% to -57% at the long-term horizon.

The different cases of model parameterization had different impacts on the hydrological projections. MSP generally predicted the largest change in volume compared to the two other cases. In fact, when using parameters calibrated over the whole period, the hydrological impact of climate change was underestimated

by 5% to 20% compared to when the parameters were calibrated on the MSP. Using parameters calibrated on the MDP (MSP) generally led to a smaller (larger) change in volume compared to the two other cases. Additionally, the change in volume simulated when using parameters calibrated over the WHO period was generally closer to the change simulated when they were calibrated over the MDP period rather than the MSP period.

It is clear here that the behaviour concerning the change in volume simulated by the models via the GDSST experiment (Fig. 5) was transferred to the hydrological projections. When moving to drier and hotter conditions (future climate conditions seen relative to past reference conditions over the WHO period or the MDP), the hydrological models tended to overestimate runoff and to generate less change in volume. The overestimation of runoff was reduced in the case of MSP where the decrease in precipitation and increase in temperature were less marked which, according to the GDSST results (section 4.2), could be translated into less overestimation of runoff compared to the WHO period, causing a bigger decrease in runoff. The limited difference between MDP and whole period could be explained by RRM transferability of VE, which is more affected by changes in precipitation than by changes in temperature, as discussed in section 3.3.3. In fact the climate conditions over the whole 30-year period and over the MDP 12-year period are not too different in terms of precipitation, in contrast to temperature. As found in further experiments, a similar behaviour of different rainfall-runoff models can be observed.

Figure 9 to be inserted near here (colour).

5. SUMMARY AND CONCLUSION

We developed a bootstrap-based differential split-sample test to assess the transferability of conceptual rainfall-runoff models under past and future climate variability. The proposed general differential split-sample test (GDSST) aims to sample sub-periods of discontinuous years gathering similar to different conditions in terms of differences in precipitation and temperature. The GDSST was compared to three other existing techniques to select sub-periods over a 30-year past period on a set of five basins under semi-arid

conditions in northern Tunisia. We showed that the GDSST outperformed the other split-sample techniques by providing a larger number of sub-periods from similar to contrasted conditions in terms of precipitation and temperature, while ensuring that the most climatically contrasted sub-periods are sampled.

The GDSST was then used to evaluate the transferability of three hydrological models under various past climate conditions in the five basins. Our results showed that the difference in climate between calibration and validation progressively affects model performance. The models tested showed acceptable transferability to wetter and/or colder conditions. However, their efficiency was significantly affected under a decrease in precipitation and an increase in temperature. According to the criteria and the thresholds retained, the models were found roughly transferable for relative changes in precipitation $\Delta P < (0.08*\Delta T - 0.18)$, with $\Delta P \in [-30\%, 80\%]$ and changes in temperature $\Delta T \in [-2\ °C, 2\ °C]$. These transferability limits showed clear interdependence between precipitation and temperature and are more accurate than those presented in previous studies which revealed more squared limits (Coron et al., 2012; Dakhlaoui et al., 2017). The models tend to overestimate runoff with an increase in temperature and a decrease in precipitation, and conversely.

The transferability limits were then compared to the future climate projections in seven high-resolution regional climate simulations under two radiative concentration pathway (RCP) scenarios (RCP4.5 and RCP8.5) for one medium-term horizon (2040–2070) and one long-term horizon (2070–2100). At the medium-term horizon, RCMs project a change in mean annual precipitation of +6.3% to -35%. At the long-term horizon, the change in precipitation is expected to reach +12% to -52%. The RCMs foresee an increase in temperature ranging from +1.0 °C to +3.5 °C by the medium-term horizon and from +1.7 °C to +5.7 °C at the long-term horizon. The differences in precipitation and temperature between past and future climate are generally within the limits of modelling transferability under RCP 4.5 regardless of the horizon. However this was not the case under RCP 8.5, regardless of the horizon. Our results showed that it was possible to find a calibration sub-period within the limits of transferability for the medium-term horizon. The projected change in precipitation and temperature are translated into hydrological projections by a +0.14% to -6.2% change in mean annual runoff under RCP4.5 and a -13% to -31% change in runoff under RCP8.5 at the medium-term horizon, and respectively -16% to -29% and -37% to -57% by the long-term horizon.

Finally, the effects of the selected past calibration period on the hydrological projections were analysed. We found that models calibrated on the whole past period underestimated the impact of climate change on

700 mean annual runoff by 5% to 20% in comparison to their calibration on sub-periods with mean annual
701 precipitation and temperature that are closer to future climate conditions. Another key finding was similar
702 transferability between the different hydrological models tested.

703 This paper thus assessed the robustness of hydrological models under climate variability and drew the
704 limits of their parameter transferability in terms of ΔT and ΔP. We proposed to reduce the uncertainty
705 caused by parameter instability through a better strategy of calibration. Understanding the sources of the
706 limited transferability of the models beyond the limits identified in the present study is a complex task.
707 However we showed that the climate conditions of the period used to calibrate hydrological models can have
708 a significant impact on hydrological projections. Using the whole historical period for model calibration can
709 result in systematic underestimation of the impact of climate change on surface water resources. Based on
710 our findings, we recommend selecting a past sub-period in which the climate conditions are as close as
711 possible to those of the future periods to be simulated in order to identify calibration parameters that can be
712 used for hydrological projections, which could significantly reduce uncertainty.

713 Future studies could focus on a better understanding of parameter instability to improve RRMs
714 robustness. For instance, the choice in the calibration period length could be further explored. As mentioned
715 by Coron at al. (2012), choosing the sub-period length used in the sampling methodology is a difficult task:
716 the calibration period should be long enough to allow for correct parameter determination. Several studies
717 (see e.g. Guo et al 2018; Vaze et al., 2010) thus claimed that longer calibration periods lead to more robust
718 RRM under climate variability, since they represent more diversified climate conditions. At the same time,
719 using overly long periods may play against the study's objectives as it would reduce the contrast between
720 periods. Also, the number of independent test periods per catchment decreases when the sub-period length
721 increases. In the present study dealing with a 30-year reference period, we have considered 8-12 year
722 calibration sub-periods, which appeared as an acceptable compromise between the length of the period
723 needed for calibration and the number of possible combinations between calibration-validation periods.
724 However, we acknowledge that shorter (longer) sub-periods could provide more (less) different contrasts in
725 terms of ΔT and ΔP while possibly increasing (reducing) model robustness. Future work could also use
726 physically-based models to test whether the more detailed processes they attempt to represent make them
727 less climate dependent than the conceptual models in realistically representing the multi-decadal flow.
728 Finally, the proposed GDSST was developed to sample sub-periods of discontinuous years, which is suitable

for a semi-arid climate with a long dry summer like in Mediterranean environments. However, it may not be suitable for other climates under which the hydrological processes are strongly influenced by the preceding years and with discharge sustained by groundwater flows during dry periods, thus leading to very various initial conditions. Future studies could thus focus on adapting the GDSST to such climates.

Acknowledgements This work is part of the Postdoc research of the first author at the HSM laboratory, Montpellier France, which was supported by an Erasmus Mundus Alyssa mobility scholarship. It was carried out as part of the ENVI-Med CLIHMag (*Changement cLimatique et Impacts Hydrologiques au Maghreb*) project funded by the program INSU-MISTRALS (2014–2015). The first author has benefited also of Erasmus+ MIC short stay support at HSM in 2018. The authors thank INM (*Institut National de la Météorologie*) and DGRE (*Direction Générale des Ressources en Eau*) in Tunisia for providing the necessary hydro-climatic data for the study. They are sincerely grateful to the two anonymous reviewers for their careful reading of the original manuscript and their many insightful comments and constructive suggestions for improvements.

REFERENCES

Anctil, F., Perrin, C., Andréassian, V., 2004. Impact of the length of observed records on the performance of ANN and of conceptual parsimonious rainfall-runoff forecasting models. Env. Model. Soft., 19, 357–368. doi: 10.1016/S1364-8152(03)00135-X

Arsenault, R., Brissette, F., Martel, J.-L., 2018. The hazards of split-sample validation in hydrological model calibration. J. Hydrol., 566, 346–362. doi:10.1016/j.jhydrol.2018.09.027

Bastola, S., Murphy, C., Sweeney, J., 2011. The role of hydrological modelling uncertainties in climate change impact assessments of Irish river catchments. Adv. in Water Res., 34, 562–576. doi:10.1016/j.advwatres.2011.01.008

Baouab M. A. and Cherif S., 2015. Changement climatique et ressources en eau : tendances, fluctuations et projections pour un cas d'étude de l'eau potable en Tunisie. La Houille Blanche, 5, 99–107. doi: 10.1051/lhb/20150061

Bergström, S., 1976. Development and application of a conceptual rainfall-runoff model for the Scandinavian catchments. SMHI RH07, Norrköping. doi:10.2166/nh.1973.0012

Bergström, S., Lindström, G., 2015. Interpretation of runoff processes in hydrological modelling: experience from the HBV approach. Hydrol. Proc., 29, 3535–3545. doi:10.1002/hyp.10510

Blinda, M., Thivet, G., 2009. Ressources et demandes en eau en Méditerranée : situation et perspectives. Sécheresse, 20, 9–16. doi:10.1684/sec.2009.0162

757 Brigode, P., Oudin, L., Perrin, C., 2013. Hydrological model parameter instability: A source of additional uncertainty in estimating the
758 hydrological impacts of climate change? J. Hydrol., 476, 410–425. doi:10.1016/j.jhydrol.2012.11.012

759 Chen, J., Brissette, F. P., Poulin, A., Leconte, R., 2011. Overall uncertainty study of the hydrological impacts of climate change for a
760 Canadian watershed. Water Resour. Res., 47, W12509. doi:10.1029/2011WR010602

761 Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., Hendrickx, F., 2012. Crash testing hydrological models in
762 contrasted climate conditions: an experiment on 216 Australian catchments. Water Resour. Res., 48, W05552.
763 doi :10.1029/2011WR011721

764 Coron, L., 2013. Les modèles hydrologiques conceptuels sont-ils robustes face à un climat en évolution ? ISIVE, AgroParisTech,
765 364 p.

766 Coron, L., Andréassian, V., Perrin, C., Bourqui, M., Hendrickx, F., 2014. On the lack of robustness of hydrologic models regarding
767 water balance simulation: a diagnostic approach applied to three models of increasing complexity on 20 mountainous catchments,
768 Hydrol. Earth Syst. Sci., 18, 727–746. doi:10.5194/hess-18-727-2014

769 Cramer, W., Guiot, J., Fader, M., Garrabou, J., Gattuso, J.-P., Iglesias, A., Lange, M. A., Lionello, P., Llasat, M. C., Paz, S., Peñuelas,
770 J., Snoussi, M., Toreti, A., Tsimplis, M. N., Xoplaki, E., 2018. Climate change and interconnected risks to sustainable
771 development in the Mediterranean. Nature Climate Change online. doi:10.1038/s41558-018-0299-2.

772 Croke, B. F. W., Jakeman, A. J. 2004. A catchment moisture deficit module for the IHACRES rainfall-runoff model. Environ. Model.
773 Softw., 19, 1–5. doi:10.1016/j.envsoft.2003.09.001

774 Dakhlaoui, H., Bargaoui, Z, Bárdossy, A., 2009. Comparaison de trois méthodes d'usage de la technique des voisins les plus proches en
775 vue d'amélioration de la performance de l'algorithme SCE-UA appliqué pour le calage du modèle pluie-débit HBV. In:
776 Hydroinformatics in Hydrology, Hydrogeology and Water Resources, IAHS Publ., 331, 139–153.

777 Dakhlaoui, H., Bargaoui, Z., Bárdossy, A., 2012. Toward a more efficient Calibration Schema for HBV Rainfall-Runoff Model.
778 J. Hydrol., 444–445, 161–179. doi:10.1016/j.jhydrol.2012.04.015

779 Dakhlaoui, H., Ruelland, D., Tramblay, Y., Bargaoui, Z., 2017. Evaluating robustness of conceptual rainfall-runoff models under
780 climate variability in northern Tunisia. J. Hydrol., 550, 201–217. doi:10.1016/j.jhydrol.2017.04.032

781 Droogers, P., Immerzeel, W. W., Terink, W., Hoogeveen, J., Bierkens, M. F. P., van Beek, L. P. H., Debele, B., 2012. Water resources
782 trends in Middle East and North Africa towards 2050. Hydrol. Earth Syst. Sci., 16, 3101–3114. doi:10.5194/hess-16-3101-2012

783 Duan, Q. Y., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water
784 Resour. Res., 28, 1015–1031. doi:10.1029/91WR02985

785 Duan, Q., Sorooshian, S., Gupta, V., 1994. Optimal Use of the SCE-UA global optimization method for calibrating watershed models.
786 J. Hydrol., 158, 265–284. doi:10.1016/0022-1694(94)90057-4

787 Fabre, J., Ruelland, D., Dezetter, A., Grouillet, B., 2016. Sustainability of water uses in managed hydrosystems: human- and climate-
788 induced changes for the mid-21st century. Hydrol. & Earth Syst. Sci., 20, 3129–3147. doi: 10.5194/hess-20-3129-2016

789 Fowler, K. J., Peel, M. C., Western, A. W., Zhang, L., Peterson, T. J., 2016. Simulating runoff under changing climatic conditions:
790 Revisiting an apparent deficiency of conceptual rainfall-runoff models. Water Resour. Res., 52, 1820–1846.
791 doi:10.1002/2015WR018068

792 Guo, D., Johnson, F., Marshall, L., 2018. Assessing the potential robustness of conceptual Rainfall-Runoff Models under a changing
793 climate. Water Resour. Res., 54, 5030–5049. doi:10.1029/2018WR022636

794 Gupta, H. V., Kling, H., Yilmaz, K. K., Martinez, G. F., 2009. Decomposition of the mean squared error and NSE performance criteria
795 Implications for improving hydrological modelling. J. Hydrol., 377, 80–91. doi:10.1016/j.jhydrol.2009.08.003

796 Haddelanda,I., Heinkeb,J., Biemansd, H., Eisnere,S., Flörkee, M., Hanasakif, N., Konzmannb, M., Ludwigd, E., Masakif, Y., Scheweb,
797 J., Stackeg, T., Tesslerh, Z.D., Wadai, Y., and Wisseri, D. 2014. Global water resources affected by human interventions and
798 climate change. Proceedings of the National Academy of Sciences · March 2014 . DOI: 10.1073/pnas.1222475110.

799 Hagemann, S., Chen, C., Clark, D. B., Folwell, S., Gosling, S. N., Haddeland, I., Hanasaki, N., Heinke, J., Ludwig, F., Voss, F.,
800 Wiltshire, A. J., 2013. Climate change impact on available water resources obtained using multiple global climate and hydrology
801 models. Earth Syst. Dynam., 4, 129–144. doi:10.5194/esd-4-129-2013.

802 Hartmann, G., Bárdossy, A., 2005. Investigation of the transferability of hydrological models and a method to improve model
803 calibration. Adv. in Geosciences, 5, 83–87.

804 Hublart, P., Ruelland, D., Dezetter, A., Jourde, H., 2015. Reducing structural uncertainty in conceptual hydrological modeling in the
805 semi-arid Andes. Hydrol. Earth Syst. Sci., 19, 2295–2314. doi:10.5194/hess-19-2295-2015

806 Hublart, P., Ruelland, D., Garcia de Cortázar-Atauri, I., Gascoin, S., Lhermitte, S., Ibacache, A., 2016. Reliability of lumped
807 hydrological modelling in a semi-arid mountainous catchment facing water-use changes. Hydrol. & Earth Syst. Sci., 20, 3691–
808 3717. doi: 10.5194/hess-20-3691-2016

809 IPCC – Intergovernmental Panel on Climate Change, 2013. Climate Change 2013: The Physical Science Basis. Contribution of
810 Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T. F., D., Qin, G.-
811 K., Plattner, M., Tignor, S. K., Allen, J., Boschung, A., Nauels, Y., Xia, V., Bex and P. M., Midgley (eds.)]. Cambridge
812 University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp. doi:10.1017/CBO9781107415324

813 Jakeman, A. I., Littlewood, I. G., Withehead, P. G., 1990. Computation of the instantaneous unit hydrograph and identifiable
814 component flows with application to two small upland catchments. J. Hydrol., 117, 275–300. doi:10.1016/0022-1694(90)90097

815 Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrol. Sci. J., 31, 13–24. doi:10.1080/02626668609491024

816 Kuczera, G., 1997. Efficient subspace probabilistic parameter optimization for catchment models. Water Resour. Res., 33, 177–185.
817 doi:10.1029/96WR02671

818 Mahalanobis, P. C., 1936. On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India, 2(1), 49–
819 55.

820 Melsen, L. A., Addor, N., Mizukami, N., Newman, A.J., Torfs, P. J. J. F., Clark, M.P., Uijlenhoet, R., and Teuling, A.J., 2018 .
821 Mapping (dis)agreement in hydrologic projections. Hydrol. Earth Syst. Sci., 22, 1775–1791. https://doi.org/10.5194/hess-22-
822 1775-2018

Milano, M., Ruelland, D., Fernandez, S., Dezetter, A., Fabre, J., Servat, E., 2012. Facing global changes in the Mediterranean basin: How could the current water stress evolve by the medium-term? C. R. Geoscience, 344, 432–440. doi:10.1016/j.crte.2012.07.006

Nash, J. E., Sutcliffe, J. V., 1970. River flow forecasting through conceptual models – Part I: A discussion of principles. J. Hydrol., 10, 282–290. doi:10.1016/0022-1694(70)90255-6

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., Loumagne, C., 2005. Which potential evapotranspiration input for a lumped rainfall-runoff model? Part 2: towards a simple and efficient potential evapotranspiration model for rainfall-runoff modelling. J. Hydrol., 303, 290–306. doi:10.1016/j.jhydrol.2004.08.025

Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. J. Hydrol., 279, 275–289. doi:10.1016/S0022-1694(03)00225-7

Poulin, A., Brissette, F., Leconte, R., Arsenault, R., Malo, J. S., 2011. Uncertainty of hydrological modelling in climate change impact studies in a Canadian, snow-dominated river basin. J. Hydrol., 409, 626–636. doi:10.1016/j.jhydrol.2011.08.057

Refsgaard, J. C., Knudsen, J., 1996. Operational validation and intercomparison of different types of hydrological models. Water Resour. Res., 32, 2189–2202. doi:10.1029/96WR00896

Rockel, B., Will, A., Hense, A., 2008. Special issue regional climate modelling with COSMO-CLM (CCLM). Meteorol., 17, 347–348. doi:10.1127/0941-2948/2008/0309.

Ruelland, D., Ardoin-Bardin, S., Collet, L., Roucou, P., 2012. Simulating future trends in hydrological regime of a large Sudano-Sahelian catchment under climate change. J. Hydrol., 424–425, 207–216. doi:10.1016/j.jhydrol.2012.01.002

Ruelland, D., Dezetter, A., Hublart, P., 2014. Sensitivity analysis of hydrological modelling to climate forcing in a semi-arid mountainous catchment. In: Hydrology in a changing world: environmental and human dimensions, IAHS Publ., 363, 145–150.

Ruelland, D., Hublart, P., Tramblay, Y., 2015. Assessing uncertainties in climate change impacts on runoff in Western Mediterranean basins. In: Hydrologic non-stationarity and extrapolating models to predict the future, IAHS Publ., 371, 75–81. doi:10.5194/piahs-371-75-2015, 2015

Samuelsson, P., Gollvik, S., Ullerstig, A., 2006. The land-surface scheme of the Rossby Centre regional atmospheric climate model (RCA3). SMHI Rep. Met., 122–125.

Schilling, J., Freier, K. P., Hertig, E., Scheffran, J., 2012. Climate change, vulnerability and adaptation in North Africa with focus on Morocco. Agric. Ecosyst. Environ., 156, 12–26. doi: 10.1016/j.agee.2012.04.021

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

Seiller, G., Anctil, F., Perrin, C., 2012. Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. Hydrol. Earth Syst. Sci., 16, 1171–1189. doi:10.5194/hess-16-1171-2012

Sellami, H., Benabdallah, S., La Jeunesse, I., Vanclooster, M., 2015. Quantifying hydrological responses of small Mediterranean catchments under climate change projections. Sci. Total Env., 543, 924–936. doi:10.1016/j.scitotenv.2015.07.006

Singh, R., Wagener, T., van Werkhoven, K., Mann, M. E., Crane, R., 2011. A trading-space for-time approach to probabilistic continuous streamflow predictions in a changing climate–accounting for changing watershed behavior. Hydrol. Earth Syst. Sci., 15, 3591–3603. doi:10.5194/hess-15-3591-2011

857 Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., Salzmann, M., Schmidt, H., Bader, J., Block, K., Brokopf, R.,
858 Fast, I., Kinne, S., Kornblueh, L., Lohmann, U., Pincus, R., Reichler, T., Roeckner, E., 2013. Atmospheric component of the
859 MPI-M Earth System Model: ECHAM6. J. Adv. Model. Earth. Syst., 5, 146–172. doi:10.1002/jame.20015

860 Terink, W., Immerzeel, W. W., Droogers, P., 2013. Climate change projections of precipitation and reference evapotranspiration for the
861 Middle East and Northern Africa until 2050. Int. J. Climatol., 33, 3055–3072. doi:10.1002/joc.3650

862 Tolson, B. A., Shoemaker, C. A. 2007. Cannonsville Reservoir watershed SWAT2000 model development, calibration and validation.
863 J. Hydrol., 337, 68–86. doi:10.1016/j.jhydrol.2007.01.017

864 Tramblay, Y., Ruelland, D., Somot, S., Bouaicha, R., Servat, E., 2013. High-resolution Med-CORDEX regional climate model
865 simulations for hydrological impact studies: a first evaluation of the ALADIN-Climate model in Morocco. Hydrol. Earth Syst.
866 Sci., 17, 3721–3739. doi:10.5194/hess-17-3721-2013

867 Tramblay, Y., Jarlan, L., Hanich, L., Somot, S., 2018. Future scenarios of surface water resources availability in North African dams.
868 Water Res. Management, 32, 1291–1306. doi:10.1007/s11269-017-1870-8

869 Valéry, A., Andréassian, V., Perrin, C., 2010. Regionalization of precipitation and air temperature over high-altitude catchments:
870 learning from outliers. Hydrol. Sci. J., 55, 928–940. doi:10.1080/02626667.2010.504676

871 Van Meijgaard, E., Van Ulft, L. H., Lenderink, G., de Roode, S. R., Wipfler, L., Boers, R., Timmermans, R. M. A., 2012. Refinement
872 and application of a regional atmospheric model for climate scenario calculations of Western Europe. Climate changes spatial
873 planning publication: KvR 054/12, ISBN/EAN 978-90-8815-046-3, pp. 44.

874 Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J. M., Viney, N. R., Teng, J., 2010. Climate non-stationarity – validity of calibrated
875 rainfall-runoff models for use in climate change studies. J. Hydrol., 394, 447–457. doi:10.1016/j.jhydrol.2010.09.018

876 Voldoire, A., Sanchez-Gomez, E., Salas y Mélia, D., Decharme, B., Cassou, C., Sénési, S., Valcke, S., Beau, I., Alias, A., Chevallier,
877 M., Déqué, M., Deshayes, J., Douville, H., Fernandez, E., Madec, G., Maisonnave, E., Moine, M.-P., Planton, S., Saint-Martin,
878 D., Szopa, S., Tyteca, S., Alkama, R., Belamari, S., Braun, A., Coquart, L., and Chauvin, F., 2013. The CNRM-CM5.1 global
879 climate model: description and basic evaluation. Clim. Dyn., 40, 2091–2121.

880 Vormoor, K., Heistermann, M., Bronstert, A., Lawrence, D., 2018. Hydrological model parameter (in)stability – "crash testing" the
881 HBV model under contrasting flood seasonality conditions. Hydrol. Sci. J., 63, 991–1007. doi:10.1080/02626667.2018

882 Wu, K., Johnston, C. A., 2007. Hydrologic response to climatic variability in a Great Lakes Watershed: A case study with the SWAT
883 model. J. Hydrol., 337, 187–199. doi:10.1016/j.jhydrol.2007.01.030

884 Yapo, P. O., Gupta, H. V., Sorooshian, S., 1996. Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration
885 data. J. Hydrol., 181, 23–48. doi:10.1016/0022-1694(95)02918-4

886 Zheng, F., Maier, H. R., Wu, W., Dandy, G. C., Gupta, H. V., Zhang, T., 2018. On lack of robustness in hydrological model
887 development due to absence of guidelines for selecting calibration and evaluation data: demonstration for data-driven models.
888 Water Res. Research, 54, 1013–1030. doi:10.1002/2017WR021470

FIGURE CAPTIONS

Fig. 1 Split-sample methods according to (a) a sliding-window SST technique, (b) a random bootstrap SST technique, and (c) a 4-sub-period DSST technique.

Fig. 2 Processing steps to sample climate contrasted sub-periods in the proposed GDSST. Each point represents a hydrological year from the reference period. The years circled are those selected

Fig. 3 Location of study basins in Tunisia and of the precipitation, temperature and streamflow stations. The main hydro-climatic characteristics are averaged over the period 1970–2000. (Dakhlaoui et al., 2017).

Fig. 4 The number of validation exercises classified in a grid of ΔT and ΔP according to four split-sample methods applied over a 30-year reference period (1970–2000) in the five studied basins: (a) sliding-window SST; (b) random bootstrap SST; (c) 4-sub-period DSST; and (d) the proposed General DSST. ΔT and ΔP represent respectively the differences in mean annual temperature and the relative difference in annual precipitation between the calibration and validation sub-periods. When a given ΔT and ΔP did not exist in the sampled sub-periods, the corresponding square in the figure is coloured grey (No available information).

Fig. 5 Evaluation of model transferability as a function of changes in the mean climate variables (ΔT and ΔP) between the validation and calibration sub-periods, according to differences in (a) KGE, (b) NSE and (c) VE between the receiver (RR, i.e. validation) and the donor (DR, i.e. calibration) periods, and (d) transferability limits defined by a decrease in NSE of more than 0.2 and a variation in VE of more +/-25%. Each coloured square represents the mean results of five catchments obtained with each model (GR4J, HBV and IHACRES). When a given ΔT and ΔP did not exist in the sampled sub-periods, the corresponding square in the figure is coloured grey (no information available). It should be noted that the absence of information in the top-right and bottom-left parts of the figures reflects the effect of anti-correlation between annual total precipitation and mean annual temperature for the Mediterranean semi-arid climate of northern Tunisia.

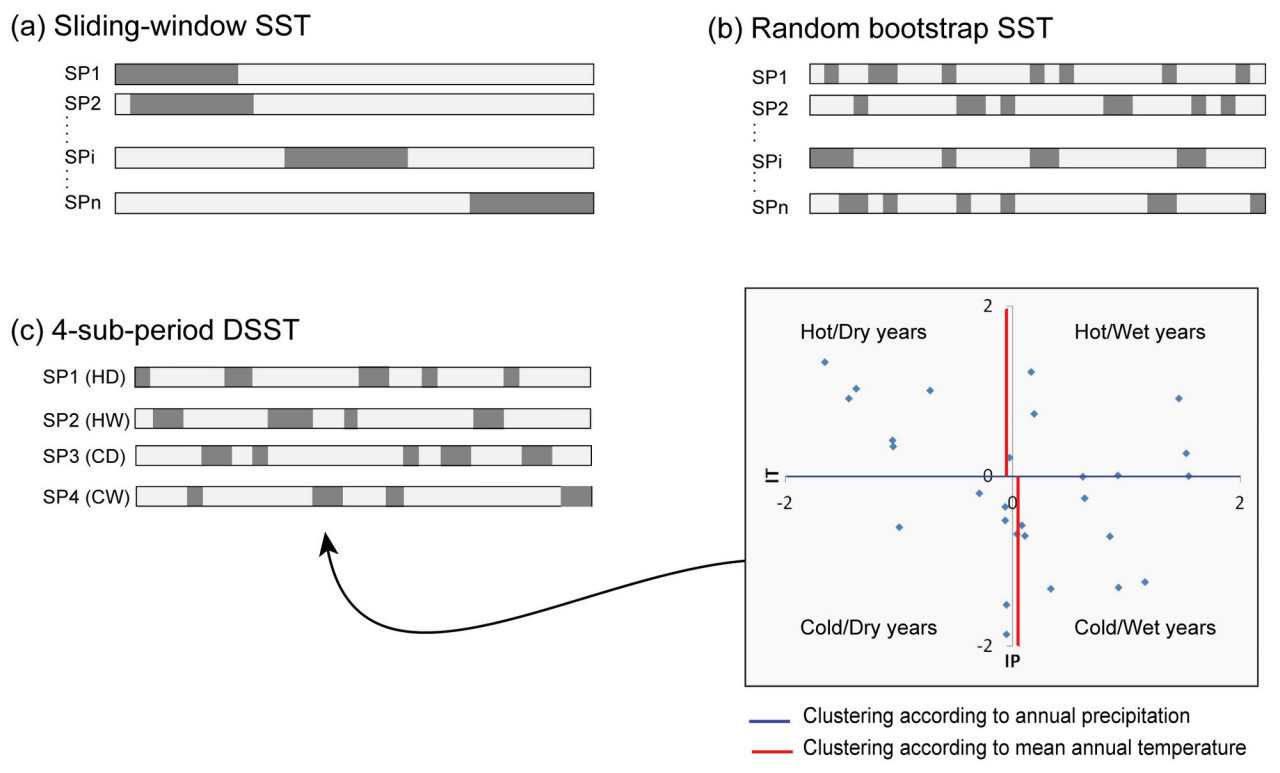
Fig. 6 Raw historical RCM outputs (P and T) versus climate observations in the five study catchments over the reference period 1970–2000: (a) mean seasonal precipitation and temperature in the RCM simulations; (b) performance of the RCM simulations according to the NSE criterion in reproducing observed seasonal precipitation and temperature; and (c) relative errors between mean simulated and observed annual precipitation (expressed in %) and mean errors between mean simulated and observed annual temperature (expressed in °C).

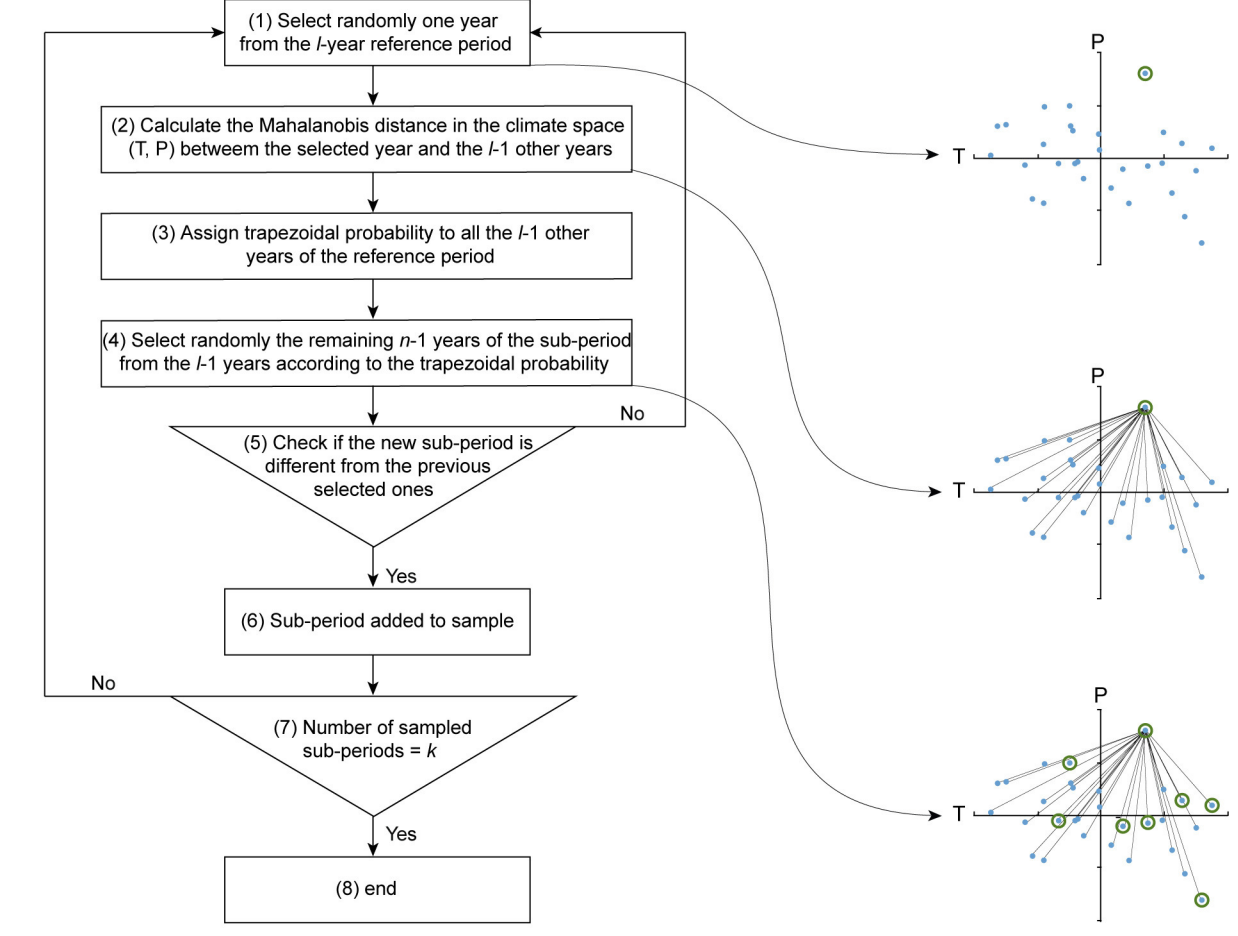
Fig. 7 Changes in (a) mean annual precipitation and (b) mean annual temperature predicted by the seven RCMs for the medium-term horizon (2040–2070) and long-term horizon (2070–2100) under RCP 4.5 and RCP 8.5.

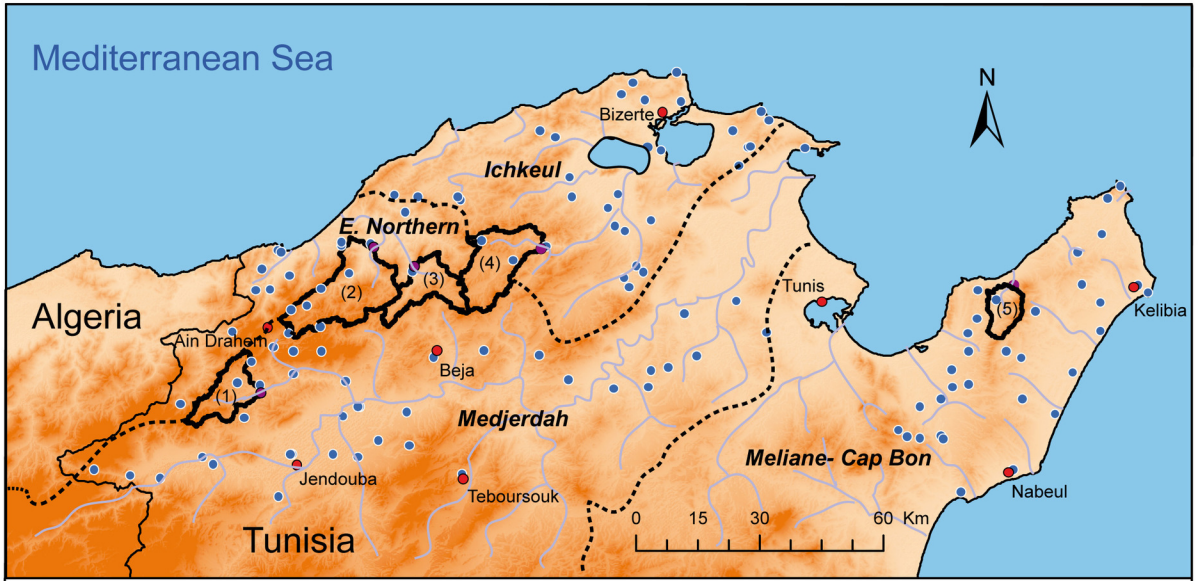
Fig. 8 Comparison of future climate projections (ΔT and ΔP) with limits to model transferability in the Melah catchment. For each RCM, RCP and horizon three points were drawn: the empty circles represent the future climate relative to the whole period, the empty squares represent the future climate relative to the most similar sub-period (MSP), and the empty triangle represent the future climate relative to the most different sub-period (MDP). The crosses indicate the mean reference climate conditions over the 30-year past period

920 (WHO) or over the most similar (MSP) or different (MDP) 12-year period.

921 **Fig. 9** Changes in mean annual runoff (ΔQ) produced by models calibrated over the whole past period (WHO), the most similar period
922 (MSP) to future climate and the most different period (MDP) to future climate.


(a) Sliding-window SST
SP1
SP2
SPi
SPn
(b) Random bootstrap SST
SP1
SP2
SPi
SPn
(c) 4-sub-period DSST
SP1 (HD)
SP2 (HW)
SP3 (CD)
SP4 (CW)
Hot/Dry years
Hot/Wet years
Cold/Dry years
Cold/Wet years
IT
IP
2
0
-2
Clustering according to annual precipitation
Clustering according to mean annual temperature

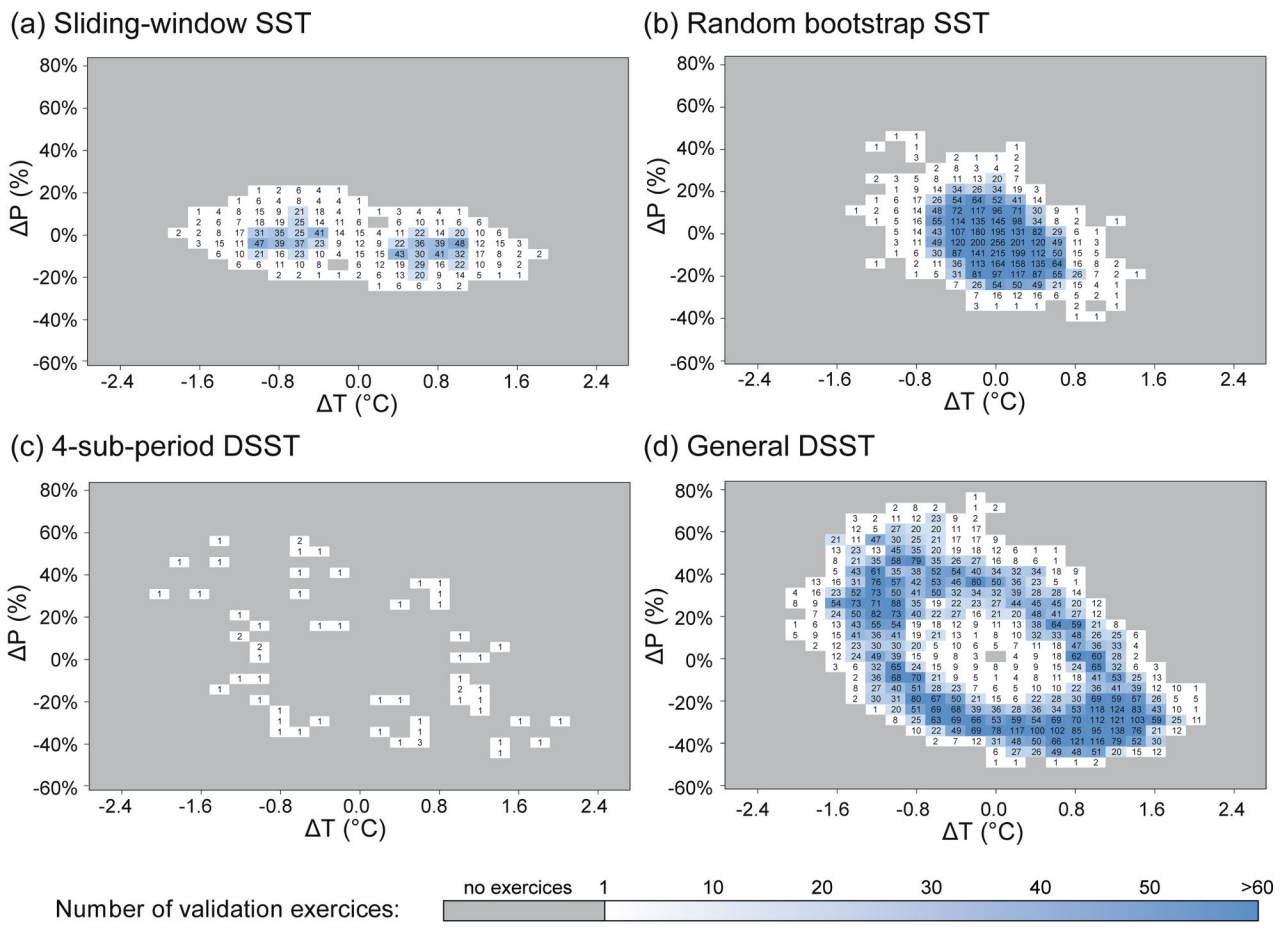

(1) Select randomly one year from the l-year reference period
(2) Calculate the Mahalanobis distance in the climate space (T, P) betweem the selected year and the l-1 other years
(3) Assign trapezoidal probability to all the l-1 other years of the reference period
(4) Select randomly the remaining n-1 years of the sub-period from the l-1 years according to the trapezoidal probability
(5) Check if the new sub-period is different from the previous selected ones
No
Yes
(6) Sub-period added to sample
No
(7) Number of sampled sub-periods = k
Yes
(8) end
P
T
P
T
P
T

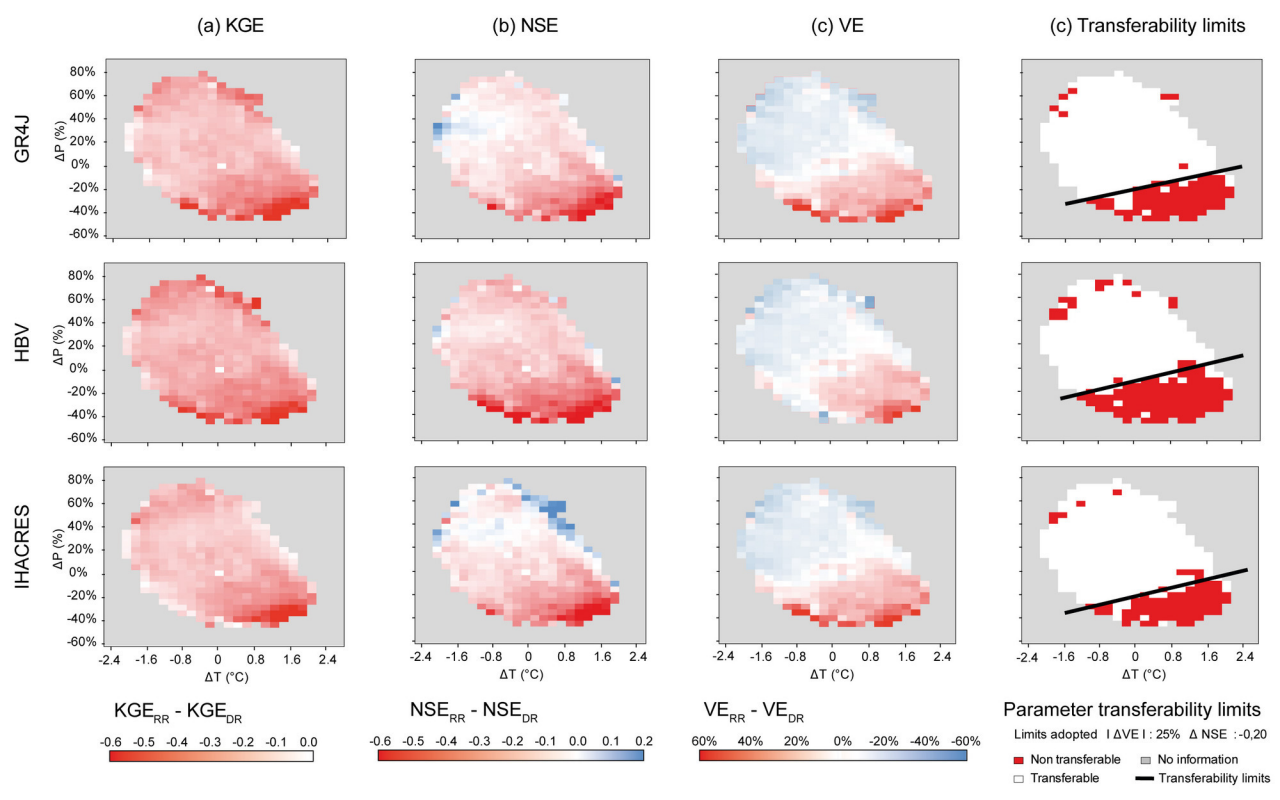


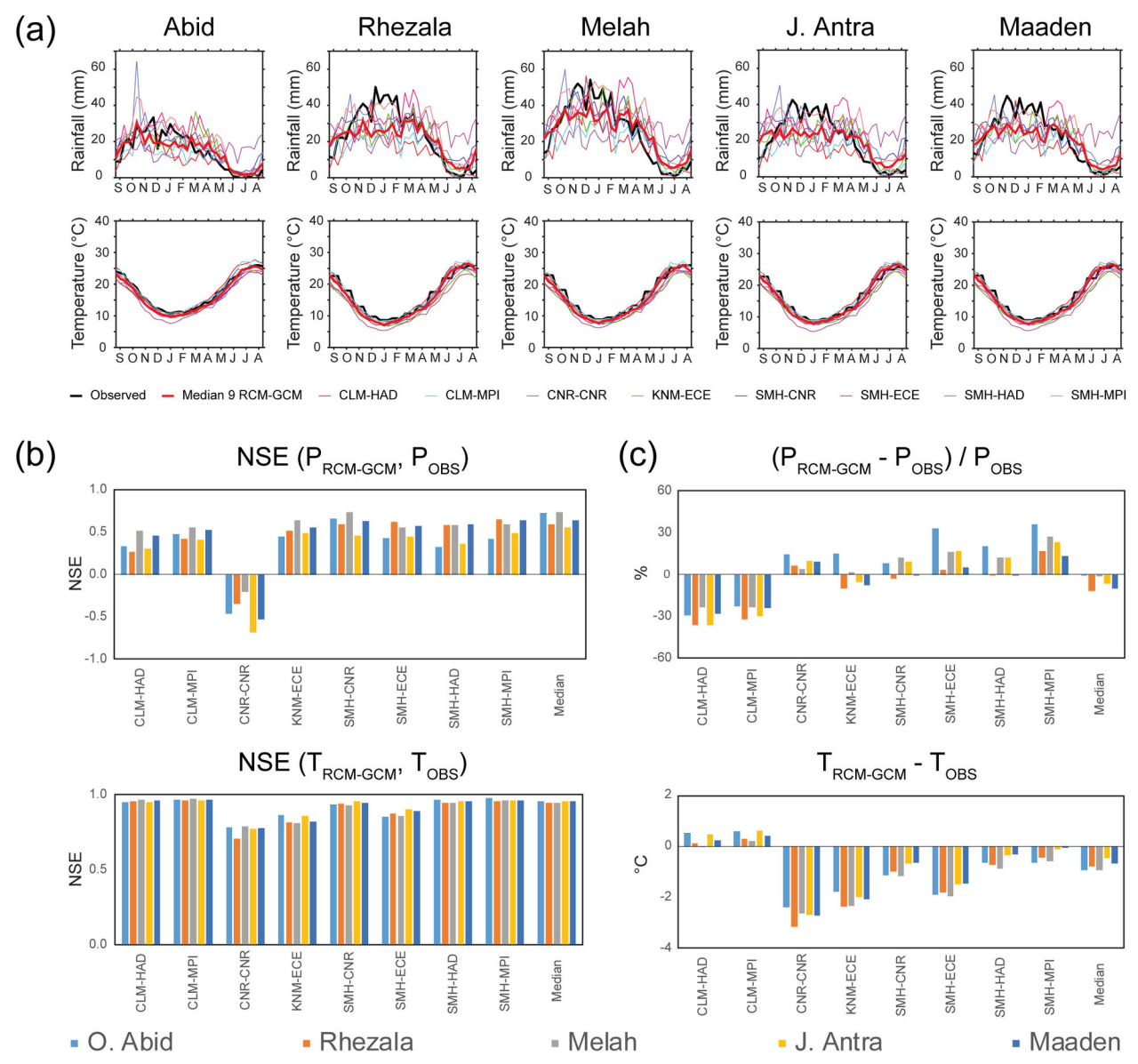
Elevations (m) 1 400 0

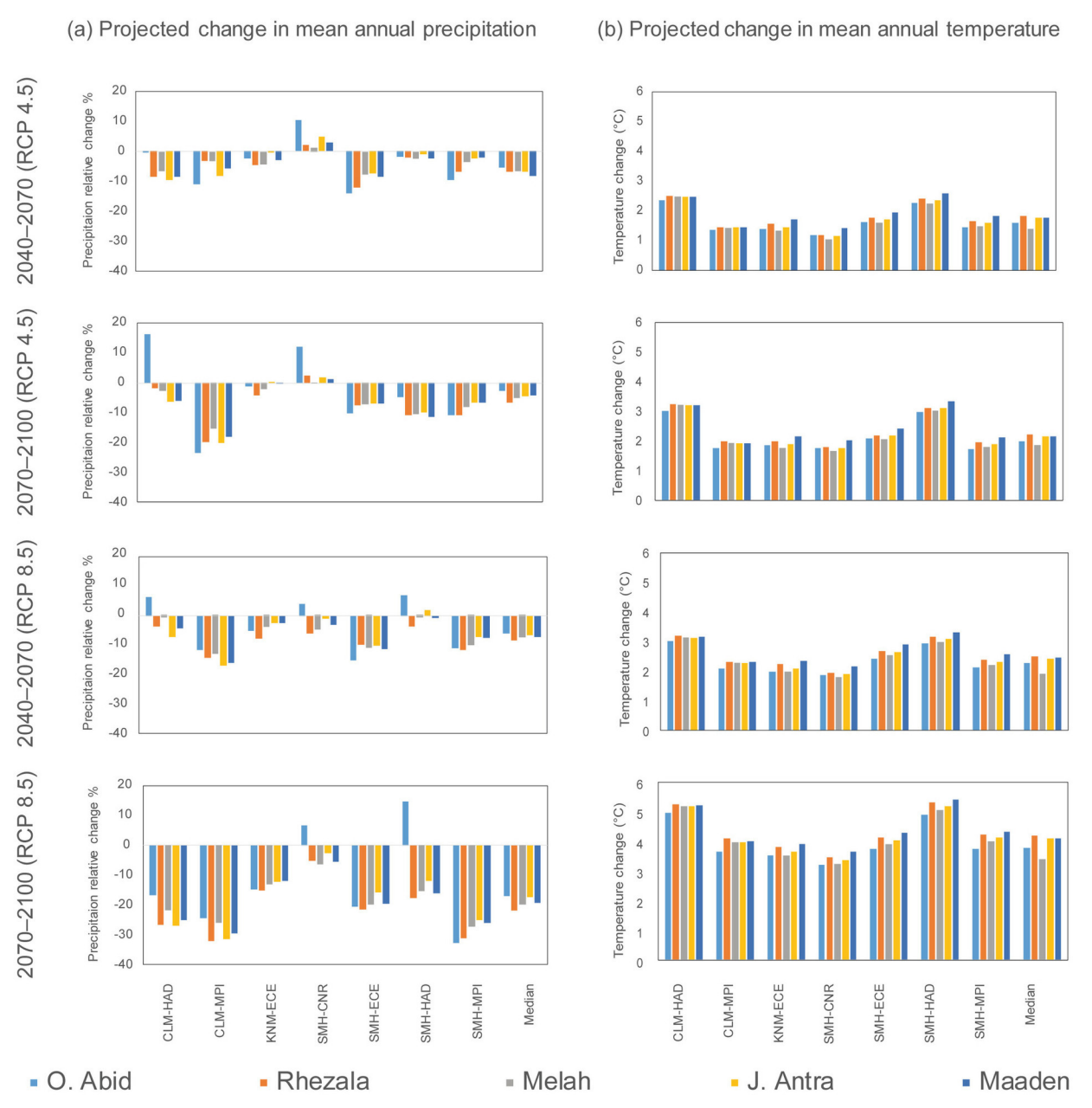
Frontiers
Catchment limits
Hydrographic Basins
Rivers
Precip. gauges
Temp. gauges
Stream flow gauges

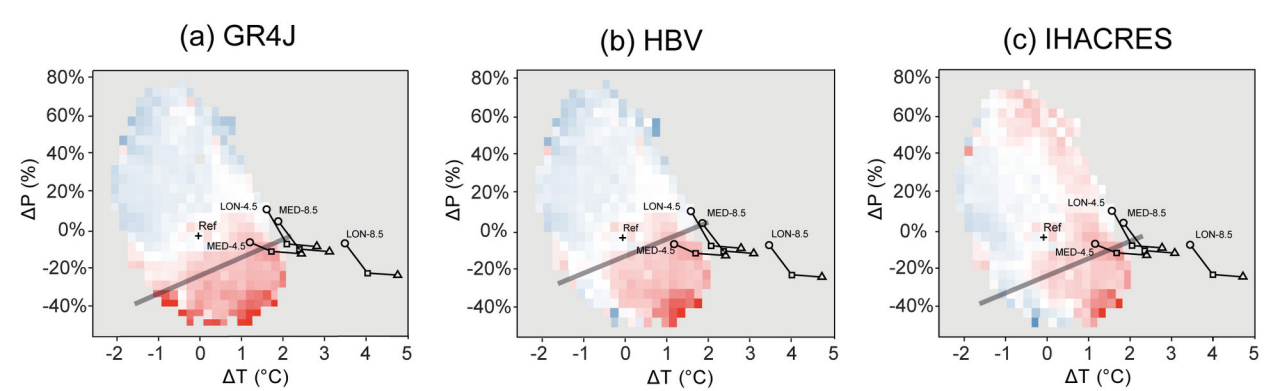
N°	Catchment	Outlet	DGRE* Code	Area (Km²)	Av. Elev. (m)	Precip. (mm)	Temp. (°C)	PET (mm)	Runoff (mm)
(1)	Rhezala	Fernana	5303510	138	470	857	16.3	1020	101
(2)	Melah	Ouchtata	3700250	315	423	928	16.5	1023	179
(3)	Maaden	Bou Brima	3700350	145	337	783	16.6	1029	135
(4)	Joumine	J. Antra	3600110	234	375	742	16.3	1016	116
(5)	El Abid	Pont-route	4300150	81	202	530	17.6	1061	40


(a) Sliding-window SST
(b) Random bootstrap SST
(c) 4-sub-period DSST
(d) General DSST
ΔP (%)
ΔT (°C)
80%
60%
40%
20%
0%
-20%
-40%
-60%
-2.4
-1.6
-0.8
0.0
0.8
1.6
2.4
Number of validation exercices:
no exercices
1
10
20
30
40
50
>60


(a) KGE
(b) NSE
(c) VE
(c) Transferability limits
GR4J
HBV
IHACRES
ΔP (%)
80%
60%
40%
20%
0%
-20%
-40%
-60%
-2.4 -1.6 -0.8 0 0.8 1.6 2.4
ΔT (°C)
KGE_RR - KGE_DR
-0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0
NSE_RR - NSE_DR
-0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2
VE_RR - VE_DR
60% 40% 20% 0% -20% -40% -60%
Parameter transferability limits
Limits adopted I ΔVE I : 25% Δ NSE :-0,20
Non transferable
No information
Transferable
Transferability limits


(a)
Abid
Rhezala
Melah
J. Antra
Maaden
Rainfall (mm)
Temperature (°C)
S O N D J F M A M J J A
Observed
Median 9 RCM-GCM
CLM-HAD
CLM-MPI
CNR-CNR
KNM-ECE
SMH-CNR
SMH-ECE
SMH-HAD
SMH-MPI
(b)
NSE (P RCM-GCM, P OBS)
NSE
NSE (T RCM-GCM, T OBS)
(c)
(P RCM-GCM - P OBS) / P OBS
%
T RCM-GCM - T OBS
°C
Median
O. Abid
Rhezala
Melah
J. Antra
Maaden


(a) Projected change in mean annual precipitation
(b) Projected change in mean annual temperature
2040–2070 (RCP 4.5)
2070–2100 (RCP 4.5)
2040–2070 (RCP 8.5)
2070–2100 (RCP 8.5)
Precipitaion relative change %
Temperature change (°C)
CLM-HAD
CLM-MPI
KNM-ECE
SMH-CNR
SMH-ECE
SMH-HAD
SMH-MPI
Median
O. Abid
Rhezala
Melah
J. Antra
Maaden



Climate difference expressed in ΔT and ΔP:

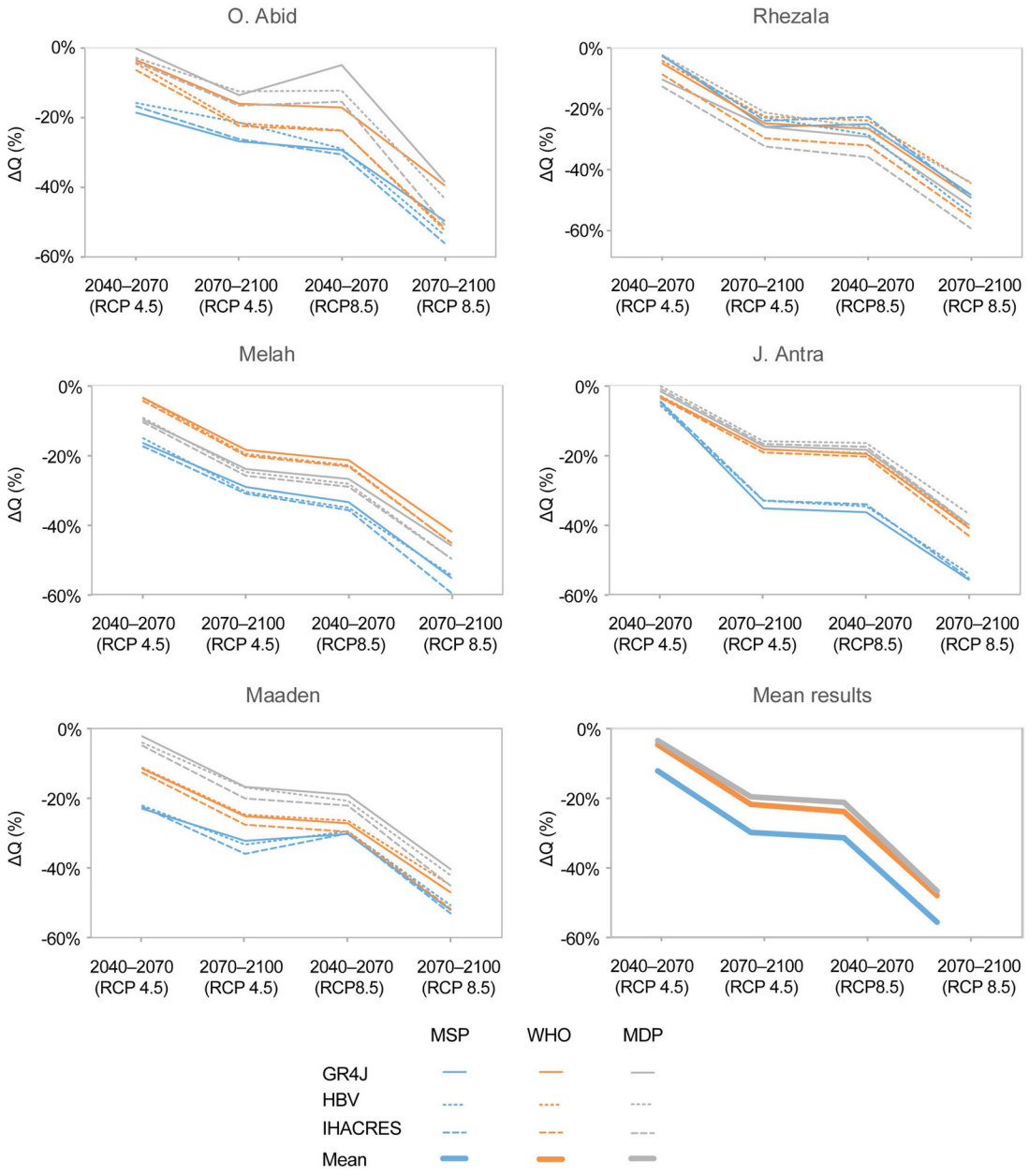
- \+ Past reference climate conditions over the 30-year period or over the most similar or different 12-year period
- ▫ Future climate conditions as compared to the past conditions over the whole 30-year period
- ○ Future climate conditions as compared to the past conditions over the most similar past 12-year period
- △ Future climate conditions as compared to the past conditions over the most different past 12-year period

Models transferability :

VE_{RR} - VE_{DR}

60% 40% 20% 0% -20% -40% -60%

Transferability limits


O. Abid
Rhezala
Melah
J. Antra
Maaden
Mean results
0%
-20%
-40%
-60%
ΔQ (%)
2040–2070 (RCP 4.5)
2070–2100 (RCP 4.5)
2040–2070 (RCP8.5)
2070–2100 (RCP 8.5)
MSP
WHO
MDP
GR4J
HBV
IHACRES
Mean