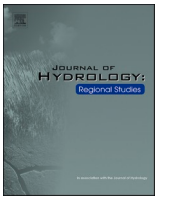
Journal of Hydrology: Regional Studies 41 (2022) 101113

Contents lists available at ScienceDirect

Journal of Hydrology: Regional Studies

journal homepage: www.elsevier.com/locate/ejrh



The effect of weighting hydrological projections based on the robustness of hydrological models under a changing climate

Ernesto Pastén-Zapata [a,b,*], Rafael Pimentel [c,d], Paul Royer-Gaspard [e], Torben O. Sonnenborg [a], Javier Aparicio-Ibañez [c,d], Anthony Lemoine [e], María José Pérez-Palazón [c], Raphael Schneider [a], Christiana Photiadou [f,g], Guillaume Thirel [e], Jens Christian Refsgaard [a]

[a] Department of Hydrology, Geological Survey of Denmark and Greenland, Copenhagen, Denmark
[b] Department of Geographical and Historical Studies, University of Eastern Finland, Joensuu, Finland
[c] Fluvial Dynamics and Hydrology Research Group. Andalusian Institute for Earth System Research, University of Cordoba, Córdoba, Spain
[d] Department of Agronomy, Unit of Excellence María de Maeztu (DAUCO), University of Córdoba, Córdoba, Spain
[e] Université Paris-Saclay, INRAE, HYCAR Research Unit, Antony, France
[f] Hydrology Research and Development, Swedish Meteorological and Hydrological Institute, Norrköping, Sweden
[g] European Environment Agency, Copenhagen, Denmark

ARTICLE INFO

Keywords:
Uncertainty
Climate change impacts
Model weighting
Differential split sampling test, Bayesian model averaging

ABSTRACT

Study region: This study is developed in three catchments located in Denmark, France and Spain, covering different climate and physical conditions in Europe.
Study focus: The simulation skill of hydrological models under contrasting climate conditions is evaluated using a Differential Split Sample Test (DSST). In each catchment, three different hydrological models are given a weight based on their simulation skill according to their robustness considering the DSST results for traditional and purpose-specific metrics. Four weighting approaches are used, each including a different set of evaluation metrics. The weights are applied to obtain reliable future projections of annual mean river discharge and purpose-specific metrics.
New hydrological insights: Projections are found to be sensitive to model weightings in cases where the models show significantly different skills in the DSST. However, when the skills of the models are similar, there is no significant change when applying different weighting schemes. Nevertheless, the methodology proposed here increases the reliability of the purpose-for-fit hydrological projections in a climate change context.

1. Introduction

Water-related impact assessments are often included in climate services across Europe (Soares et al., 2018) and their reliability significantly depends, among other factors, on validated scientific methods (Hewitt et al., 2012). Methods behind the assessment of climate change impacts on hydrology and/or the projected impacts in different regions have been used extensively (e.g. Rojas et al., 2011; Brigode et al., 2013; Prudhomme et al., 2013; van Vliet et al., 2015; Wagner et al., 2017; Teutschbein et al., 2018; Fonseca and Santos, 2019; Pastén-Zapata et al., 2020, Hundecha et al., 2020), and have at the same time highlighted that the contribution of

* Corresponding author at: Department of Hydrology, Geological Survey of Denmark and Greenland, Copenhagen, Denmark.
E-mail address: ernestopasten@gmail.com (E. Pastén-Zapata).

https://doi.org/10.1016/j.ejrh.2022.101113

Available online 26 May 2022

hydrological models to the total uncertainty in the projected hydrological variables can be substantial (Hagemann et al., 2013; Prudhomme et al., 2014). Therefore, it is relevant to evaluate the simulation skill of the hydrological models used for impact assessments. This will eventually improve the usability of the projections by providing access to quality assured information.

Traditionally, the simulation skill of hydrological models is assessed using a split-sample test (SST). This test divides an observation period into two non-overlapping sub-segments of similar length, using one period to calibrate the model and the remaining to evaluate it (Klemeš, 1986). Model parameters are usually calibrated by optimizing a performance metric, the objective function. A model (and its parameters) is deemed transferable in time if it performs well in the evaluation period. Nevertheless, in a climate change context, this approach might be inadequate because it assumes stationary conditions (Refsgaard et al., 2014). Typically, a model with good simulation skill in one period does not necessarily perform as well in subsequent periods where climate is changing (Thirel et al., 2015a). Hence, in climate change impact assessments, evaluating the skill of hydrological models to simulate hydrological processes under contrasting climate conditions should be a standard approach (Seiller et al., 2012). When models satisfy such tests, the confidence in the future projection increases and their contribution to the total uncertainty decreases (Krysanova et al., 2018).

An alternative to a standard SST is the Differential Split Sample Test (DSST) (Klemeš, 1986), where historical non-stationary periods of contrasting climatic regimes are used to calibrate and evaluate the hydrological model. During the last decade, assessments of the skill of hydrological models in contrasting climates based on the DSST have been performed in many contexts on many models. Most of these studies concluded that, in general, hydrological models may lack robustness in a changing climate context (e.g., Refsgaard and Knudsen, 1996; Merz et al., 2011; Coron et al., 2012; Thirel et al., 2015b). The consequences of off-target hydrological projections for water planning is one of the key issues of modern hydrology (Blöschl et al., 2019, Unsolved Problems in Hydrology n°1, 20 and 21). In general, hydrologists stress the need to improve hydrological process descriptions toward better plausibility and extrapolation skills (e.g. Merz et al., 2011; Coron et al., 2014; Fowler et al., 2020).

A few authors have proposed different versions of the DSST, such as the Generalised Split-Sample Test (Coron et al., 2012) or the General Differential Split-Sample Test (Dakhlaoui et al., 2019), but they are generally based on the same idea as DSST. DSST is based on observed changes, however, climate projections indeed suggest that, in some parts of the world, future changes may be completely beyond the range recorded in historical data. Hence, the approach might underestimate model potential failures in hydrological projections (Stephens et al., 2020).

While the DSST has repeatedly been applied to hydrological models (e.g., Li et al., 2012; Thirel et al., 2015b; Broderick et al., 2016), simulation skills are usually evaluated using criteria such as the Nash-Sutcliffe Efficiency (NSE, Nash & Sutcliffe, 1970) or the Kling-Gupta Efficiency (KGE, Gupta et al., 2009). However, such criteria might be inconclusive, especially when dealing with specific water management issues such as water supply, flood impacts, droughts, or stresses on ecosystems. It may be expected that a model performing well on reproducing the main hydrological characteristics of a catchment appears more reliable, compared to a model performing well only on very specific hydrological signatures while showing poor general performance. From an end-user point of view, however, when the possibility to choose from a handful of models with different simulation capabilities is present, the model selection might not always be straightforward. Hence, the inclusion of metrics specifically designed to fit end-user purposes in addition to usual performance metrics could benefit both the usability and the quality of future projections.

Besides, rather than selecting one well-performing hydrological model, it might be preferred to use an ensemble of models weighted based on their simulation skills. The advantage of using ensembles is to compensate for flaws of individual models (Knutti, 2010) such as deficient model structures, non-stationarity of the parameters and biased climate inputs (Broderick et al., 2016; Seiller et al., 2012). Ensemble simulations also allow for uncertainty assessment of the projections. However, acknowledging that not all models are equally skilled, weighting models is useful to increase the reliability of ensemble projections.

From the perspective of climate models, Knutti (2010) points out that weighting procedures should be used as a tool to discard unskillful models rather than to select the best model. Christensen et al. (2010) examined different ways of aggregating performance metrics into single weights for climate models. Although it was shown that the choice of the aggregation method has a rather small influence on the performance of the model ensemble, it was suggested that the model weighting should be considered as an additional level of uncertainty. Nowadays, assigning different weights to climate models of an ensemble according to their simulation skills is frequently used in practice (e.g., Haughton et al., 2015; Knutti et al., 2017; Wang et al., 2019).

Compared to climate models, weighting of hydrological models in climate change impact assessments has rarely been explored. Najafi et al. (2011) assessed, for example, the hydrological model uncertainty with a set of four hydrological models of different complexities in a catchment in the USA and discussed the contribution of the Bayesian Model Averaging method for model weighting. Likewise, Broderick et al. (2016) found that model averaging techniques remained consistent under temporal transferability in several Irish catchments and recommended objective-based weighting methods. Despite the numerous insights in the role of hydrological models within the uncertainty cascade in climate change impacts (e.g., Velázquez et al., 2013; Dams et al., 2015; Karlsson et al., 2016; Lemaitre-Basset et al., 2021) and the long-standing tradition of model comparison in the hydrological community, none has investigated how weighting of the hydrological models affects the projections and their uncertainty when evaluating the model performance in contrasting climate (by applying a DSST) and including metrics that are specifically tailored to specific end-users.

Methods to tailor the choice of climate and hydrological model ensemble would improve the usability, user acceptance and potential uptake of the projections. Therefore, the main objectives of this study are (a) to assess the robustness of hydrological models by evaluating their ability to describe the impact of climate change using a DSST, (b) to estimate weights for hydrological models based on their robustness, (c) to compare the effect of different weighting schemes on the projection of future climate change impacts, (d) to assess the robustness of the models to project changes in purpose-specific metrics that were not used during calibration, and (e) to quantify the uncertainty of the projections considering the different weighting schemes. The study is based on three European case studies, using ensembles of climate and hydrological models to evaluate the future projections. For each case study, three hydrological

models are evaluated with different sets of metrics in a DSST experiment to identify which are more skillful to reproduce changing climate conditions. The sensitivity of future projections to the different sets of weights is then assessed, with a specific focus on the uncertainty.

2. Methodology

2.1. Case studies and available data

Three case studies across Europe are analyzed. These are located in Denmark, France, and Spain (Fig. 1). Each case focuses on providing information for decision-making specific to that particular site. There are significant differences between the catchments; the Danish catchment is mostly flat whereas the French and Spanish sites are topographically complex with snow processes and snowmelt timing playing important roles. In addition, the recent 30-year trends for the annual mean precipitation and streamflow are positive for the Danish catchment and negative for the other catchments (Fig. 2).

2.1.1. Denmark

Climate change has a significant impact in the Storå catchment in western Denmark (Fig. 1). Focus is given to hydrological conditions of importance for agriculture, where dry and wet conditions both have negative impacts. Thus, the target of this case study is to assess the change in the future occurrence of high and low flows in the catchment. Additionally, we look at how the mean groundwater level is projected to change in the future. Such aspects are important in the study catchment as wet conditions in Denmark, primarily occurring in winter (Fig. 2), lead to waterlogged agricultural fields, increasing the drainage needs. On the other hand, dry conditions, commonly in summer, increase the irrigation needs. Additionally, in recent years, there have been river floods in urban areas within the catchment.

2.1.2. France

The hydrological cycle of the Durance catchment (Fig. 1), in the Southern French Alps, is expected to be strongly affected by climate change. Among the potential impacts, reduced snowpack and shifted timing of snowmelt flows may impact the Serre-Ponçon reservoir, which fits many purposes for water management in the region (Maughan, 2015; Branche, 2017). In this case study, we focused on the average water discharge in the reservoir in winter and spring, as well as on the timing of spring peak flows.

2.1.3. Spain

Future climate scenarios suggest a reduction in snow persistence over the Guadalfeo River catchment (Fig. 1), in Sierra Nevada Mountain Range, southern Spain. This catchment constitutes an example of alpine conditions in a semiarid area (i.e., high evaposublimation rates (Herrero and Polo, 2016)), with occurrence of several snowmelt accumulation/ablation cycles within the year (Pimentel et al., 2017b) and a high variability in river discharge linked to the snow dynamics (Pérez-Palazón et al., 2018, Fig. 2). Water resources from the snowpack are key in the area, where different activities compete for allocating water. Therefore, this case study tries to assess the impact of future climate scenarios on water availability in the area. Thus, the projected streamflow volume, and the future occurrence of extreme conditions, wet and dry, are analyzed.

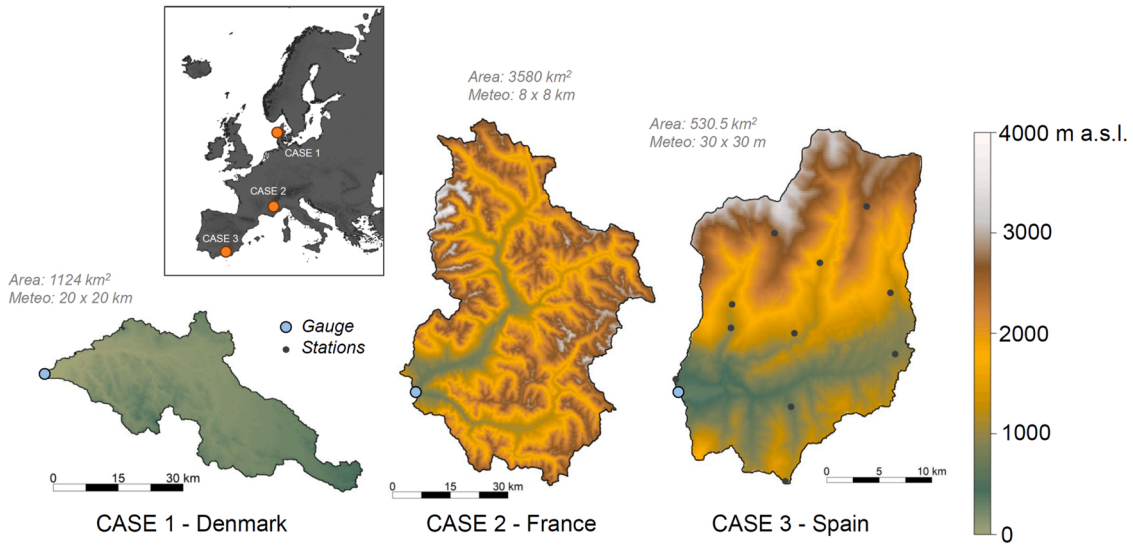


Fig. 1. Location of the three study catchments within Europe along with their elevation.

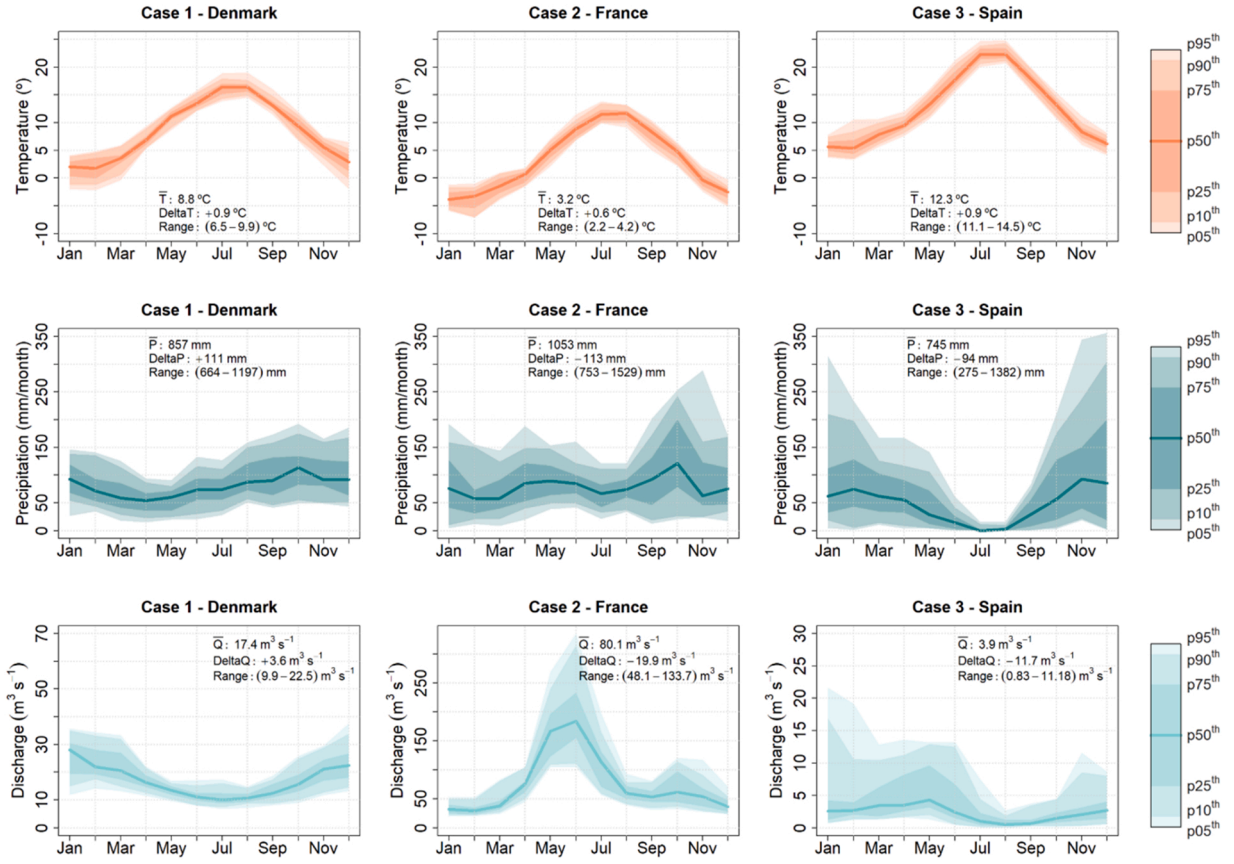


Fig. 2. Observed monthly mean air temperature, monthly accumulated precipitation and monthly average discharge regimes for the Danish (first column), French (second column) and Spanish (third column) sites. The darker line represents the observed mean, and the shaded areas represent different percentiles. Please note the difference in the y-axis for the river discharge. Delta = change over 30 years, following the linear trend of the available observations (1990–2017 for the Danish site, 1976–2005 for the French case and 1961–2015 for the Spanish case) Range = Maximum mean annual temperature – Minimum mean annual temperature, or, maximum accumulated annual precipitation – minimum accumulated annual precipitation.

2.2. Hydrological modelling

Three different hydrological models are set up for each site to account for the hydrological model uncertainty (Fig. 3, a). The selected hydrological models are frequently employed in each of the regions to assess the impacts of climate change. Therefore, a specific hydrological model ensemble is used for each catchment.

A potential benefit of the different modelling setups, spanning from conceptual rainfall-runoff models to distributed hydrological models at the three case studies, is that a broader range of model behaviors may be observed. By setting up different frameworks for the experiment, doubts to the applicability of the conclusions to other contexts are partly avoided. The main drawback of this approach is, however, a more limited understanding of the results, since the different components of the setups (e.g., regional climate characteristics, model hypotheses, etc.) cannot be easily isolated for comparison between the sites. The models used in each case are described next.

2.2.1. Denmark

The MIKE SHE model is the base model of the Danish national water resources model (Henriksen et al., 2003; Højberg et al., 2013; Stisen et al., 2019a; 2019b). MIKE SHE is a physically-based, integrated and fully-distributed groundwater-surface water model (Abbott et al., 1986; Graham and Butts, 2005). Here, the model is set up at a 250 m x 250 m resolution. Input climate data come from datasets developed by the Danish Meteorological Institute gridded at a 20 km x 20 km resolution for temperature and potential evapotranspiration and 10 km x 10 km for precipitation (Scharling, 2012). Daily streamflow data come from National Environmental Monitoring Programme (NOVANA) from the Danish Ministry of Environment and Food (DCE), described by Stisen et al. (2019a); (2019b) and made available by Koch and Schneider (2022). Three different conceptualizations (models) of the unsaturated zone are used:

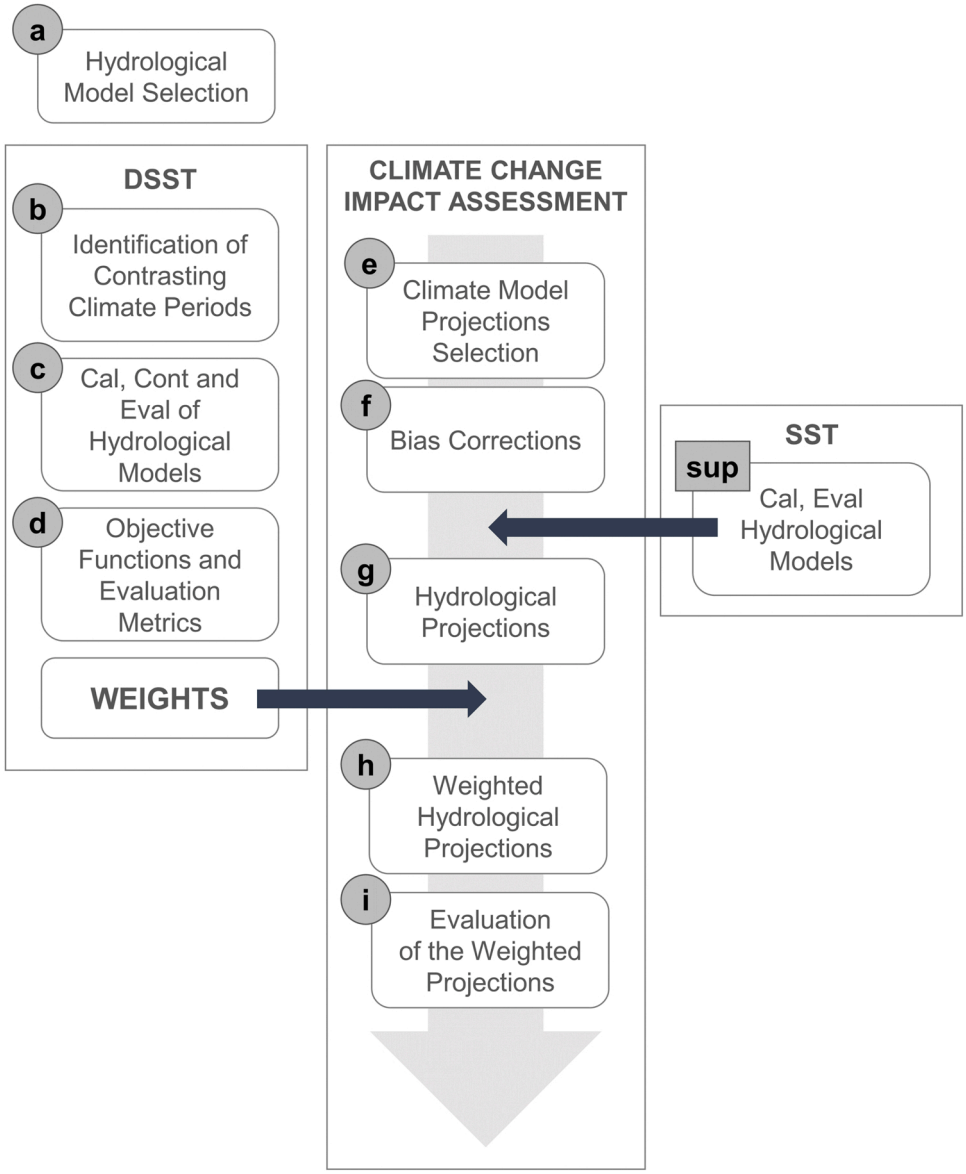


Fig. 3. Flow chart of the analysis performed in this study.

a) Richards' equation with the soil profiles discretized vertically. It is based on the continuity equation and Darcy's law (Graham and Butts, 2005). Vertical flow is a function of gravity and capillary forces. The actual evapotranspiration (AE) is estimated as a function of the vegetation status and soil moisture (Kristensen and Jensen, 1975).
b) Gravity Flow with a vertical discretization of the soil profiles. It is a simple form of the Richards' equation, where the vertical flow depends entirely on the gravity (Graham and Butts, 2005). AE is estimated as a function of the vegetation status and soil moisture (Kristensen and Jensen, 1975).
c) Two-layer model of the unsaturated zone. The unsaturated zone is divided into a root zone and a zone below the root zone. Water infiltrating from the upper unsaturated zone layer flows directly to the saturated zone whenever the water content in the lower unsaturated zone layer equals field capacity (Graham and Butts, 2005). AE depends on the potential evapotranspiration (PE), vegetation status and soil moisture content (Yan and Smith, 1994).

The three versions of the MIKE SHE model do not represent three totally different models. However, the unsaturated zone is a central component of the model that controls water balance and a number of flow components. Hence, it is assumed that working with the three versions of unsaturated zone of MIKE SHE is sufficient to obtain results that are significantly different.

The models are calibrated using a multi-criteria objective function (Demirel et al., 2018) that includes river discharge (NSE and KGE), groundwater head (RMSE) and estimated irrigation, simulated in a demand-driven manner (RMSE). Automatic calibration is

done using PEST (Doherty et al., 1994).

2.2.2. France

Three lumped daily rainfall-runoff models (GR4J, HBV0 and TOPMO) associated with a snow module (CemaNeige) are used to simulate the hydrology of the French catchment. These models have been widely used in the country. Data come from the Hydro-SAFRAN daily data set (Delaigue et al., 2020). Daily streamflow measurements at the outlet of the catchments were retrieved from Banque HYDRO (Leleu et al., 2014). Daily meteorological data were supplied by the SAFRAN dataset, an 8 km x 8 km atmospheric reanalysis (Vidal et al., 2010) aggregated at catchment scale. Potential evaporation (PE) was computed using the temperature- and radiation-based formula proposed by Oudin et al. (2005). A short description of the models is provided here:

a) GR4J (Perrin et al., 2003) is a parsimonious four-parameter lumped conceptual model. AE depends on PE and soil moisture content.
b) HBV0 (Bergström and Forsman, 1973; Bergstrom, 1995) is a nine-parameter lumped model designed for snow-covered catchments. Here, its native snow module is replaced by the CemaNeige snow module. AE depends on PE and soil moisture content.
c) TOPMO (Michel et al., 2003) is a seven-parameter lumped adaptation of TOPMODEL (Beven and Kirkby, 1979) where the distribution of the topographical index is approximated by a calibrated two-parameter probability distribution function instead of being derived from catchment topography. AE depends on PE and soil moisture content.

CemaNeige (Valéry et al., 2014) is a two-parameter degree-day snow module. It divides the catchment into five altitude layers of equal area, where the snow cover on each layer is represented as a conceptual reservoir filled by solid precipitation. The daily amount of melted water is used as input to the rainfall-runoff models in addition to liquid precipitation. The parameters of CemaNeige were calibrated simultaneously with the parameters of the hydrological models. GR4J and CemaNeige are used within the airGR R package (Coron et al., 2017, 2018). HBV0 and TOPMO are used within the airGRplus R package (in development, Coron and Perrin, 2018).

The models are automatically calibrated by optimizing the KGE computed on the square-root of streamflow. The optimization algorithm is based on a prior global screening of the parameter space followed by a local search from the identified best parameter set (see Edijatno et al., 1999; Mathevet, 2005).

2.2.3. Spain

Previous experience in the Spanish case study suggests the use of process-oriented, semi-distributed or physically-based models. Gridded daily meteorological data (30 m x 30 m) generated from in situ weather stations, using specific interpolation algorithms for each meteorological variable (, 2010; Herrero et al., 2007), were used in the historical runs. This information is aggregated at the catchment scale for HYPE and SWAT and used as distributed maps for WiMMed. The daily streamflow data come from the Water Authority in the Region. The selected models have been previously used in the area.

a) HYPE (Hydrological Prediction for the Environment; Lindström et al., 2010) is a daily-timestep semi-distributed process-based hydrological model, using Hydrological Response Units (HRUs) as calculation units. HYPE uses a snow melting routine that combines three decay factors linked to air temperature, radiation and fractional snow cover (Samuelsson et al., 2011). AE is estimated using the modified Hargreaves-Samani PE formulation (Hargreaves and Samani, 1982) whereas infiltration into three soil layers is calculated using a water table discrimination model (Lindström et al., 2010).
b) SWAT (Soil and Water Assessment Tool; Arnold et al., 1998) is a daily-timestep conceptual semi-distributed hydrological model, in which the spatial resolution is fixed by HRUs. Snowmelt is calculated using a degree-day method and AE is estimated using the Penman-Monteith PE formulation (Penman, 1948; Monteith et al., 1964). Infiltration is calculated using Green and Ampt (1911) in a single soil layer.
c) WiMMed (Watershed Integrated Model for Mediterranean Environments; Polo et al., 2010), is a hourly-timestep distributed physically-based model. The snow module (SnowMed) uses a punctual mass and energy balance extended to a distributed scale using depletion curves (Herrero et al., 2009; Pimentel et al., 2017a). AE is calculated using the Penman-Monteith PE formulation (Penman, 1948; Monteith et al., 1964) whereas infiltration is estimated using the Green and Ampt (1911) approach in a two-layer soil discretization.

Calibration is done through a process-oriented stepwise approach based on expert knowledge, minimizing mean absolute error in daily water volumes.

2.3. Differential Split Sample Test (DSST)

2.3.1. Principles of the DSST

The DSST is a scheme used to evaluate the extent at which the hydrological models are capable of skillfully simulating hydrological processes in a climate change context. Three different historical periods are identified – the calibration, control and evaluation periods. The calibration and control periods are comparable (although they concern different years) whereas the evaluation period is defined to obtain the largest contrast in simulation results. Hence, the method aims at reproducing a change between calibration conditions and evaluation conditions. The method allows flexibility in the selection of the periods (e.g., length and continuity or discontinuity) as shown in previous studies that applied it (e.g., Broderick et al., 2016; Li et al., 2012; Seiller et al., 2012; Brigode et al., 2013; Gelfan and Millionshchikova, 2018). In this study, the DSST results are used to rank the models based on their performance and define a weight for

each model within the ensemble.

2.3.2. Setup of the DSST in the case studies

The aim of this study is to evaluate the performance of hydrological models on periods with contrasting climate conditions. Thus, to enhance the contrast among the periods, the lengths of the calibration, control and evaluation periods are set to three (possibly non-consecutive) years (Table 1). The climate variable used for classifying the periods is case dependent, based on the needs of each case study: the annual accumulated precipitation is used in the Danish and Spanish cases, whereas the annual mean temperature is used for the French case. All sites rank the annual values of the variable of interest from the maximum to the minimum. Using this approach, the wettest and driest (Denmark and Spain) and the warmest and coldest (France) years are identified (Fig. 3, b). Based on this rank, a calibration period is defined (used for adjustment of the parameters of the model) along with a control period (to assess the calibrated model skill under similar climate conditions but outside the calibration period) and an evaluation period (to assess transferability of the model to contrasting climate conditions). Thus, for the Danish and Spanish sites, the calibration period is set to comprise the 1st, 3rd and 5th driest years, the control period includes the 2nd, 4th and 6th driest years and the evaluation period is composed by the 1st, 3rd and 5th wettest years. Likewise, for the French case, calibration and control take place during the coldest years, while the evaluation on contrasting conditions is carried out in the warmest three years (Fig. 3, c).

Next, the entire observation period is simulated, but the parameters of the models are calibrated only for the calibration period. The objective functions used during calibration are those defined in Section 2.2.

2.3.3. Evaluation metrics

Widely used evaluation metrics are applied to assess the simulation skill of the models. In addition, as the goal is to provide useful information for the end-users, purpose-specific metrics are included in the assessment. Overall, eight evaluation metrics are used to assess the skill of each hydrological model (Table 2). Metric f6 involves three purpose-specific metrics.

The aim of the proposed purpose-specific metrics for the Danish case (Table 3) is to evaluate the frequency of high (f6.1) and low flows (f6.2), and the importance of groundwater head for the shallow wells (f6.3). In the French case, the metrics focus on mean flows during the fall-winter season (f6.1) and during the spring-summer season (f6.2), and on the date at which half of the total water volume flowing during the six first months of the year is reached (f6.3). Finally, for the Spanish case, total water volume (f6.1) and the occurrence of drought conditions (f6.2) and extreme wet conditions (f6.3) are assessed.

2.3.4. Hydrological model ranking and assignment of weights

Evaluation of the climate model performance in the present climate at local scale is controversial (Knutti, 2010). This is mainly because climate models are intended to simulate large-scale climate processes instead of local climate. Thus, climate models are not submitted to a SST calibration procedure to fine-tune the models. Hence, although different studies can be found where climate models are evaluated based on their performance and assigned weights accordingly, in this study, climate models are given the same probability of occurrence.

Hydrological models, however, are ranked based on their simulation skill with respect to the metrics described in Tables 2 and 3 (Fig. 3d). According to their relative skill for each of the metrics, a score of 3 is assigned to the best performing hydrological model, 2 to the model with the intermediate skill and 1 to the hydrological model with the lowest skill.

Once the scores for each metric are defined, they are combined (multiplicatively) in four different sets, termed as 'final weights' (W1 to W4). These weights are normalized to sum 1. Weight 1 (W1) includes all metrics except the purpose-specific metrics. Weight 2 (W2) only includes the purpose-specific metrics. Weight 3 (W3) includes all the metrics. Finally, weight 4 (W4) represents the 'model-democracy' approach, where all hydrological models have the same weight (one third). W4 is used as reference for comparison because it is the approach that is commonly used (i.e., no weights are applied to the hydrological models). The four weights are defined as:

$$w_{1,i} = r_{1,i} \cdot r_{2,i} \cdot r_{3,i} \cdot r_{4,i} \cdot r_{5,i} \Big/ \sum_i w_{1,i} \tag{1}$$

Table 1
Selected years for the calibration, control and evaluation periods in each case study. For the French and Spanish cases, hydrological years are selected instead of natural years (October to September in France, and September to August in Spain).

	Calibration	Control	Evaluation
Denmark	1st, 3rd, and 5th driest years in record:1996 (600 mm)2003 (700 mm)2005 (750 mm)	2nd, 4th, and 6th driest years in record:1997 (700 mm)1991 (730 mm)2013 (760 mm)	1st, 3rd, and 5th wettest years in record:2015 (1000 mm)1999 (1000 mm)2007 (990 mm)
France	1st, 3rd, and 5th coldest years in record:1983–1984 (2.2 °C)1976–1977 (2.5 °C)1979–1980 (2.7 °C)	2nd, 4th, and 6th coldest years in record:1977–1978 (2.4 °C)1995–1996 (2.6 °C)1990–1991 (2.7 °C)	1st, 3rd, and 5th warmest years in record:1989–1990 (4.2 °C)1982–1983 (4.1 °C) 1988–1989 (3.8 °C)
Spain	1st, 3rd, and 5th driest years in record:1994–1995 (275 mm)1993–1994 (452 mm)2006–2007 (524 mm)	2nd, 4th, and 6th driest years in record:2004–2005 (388 mm)2005–2006 (503 mm)1992–1993 (552 mm)	1st, 3rd, and 5th wettest years in record:2009–2010 (1382 mm)1995–1996 (1296 mm)2010–2011 (1062 mm)

Table 2
Metrics used to evaluate the simulation skill of the hydrological models. Q_i represents the observed daily streamflow of day i, $\widehat{Q}_i$ denotes the simulated daily streamflow of day i. $Q_{m,j}$ ($\widehat{Q}_{m,j}$) denotes the average observed (simulated) monthly streamflow of the j-th month. r represents the correlation coefficient, μ and σ are the symbols for the mean and standard deviation of daily streamflow, respectively. N is the number of days included in the assessment.

ID	Name (units)	Equation	Result range
f1	Mean absolute bias in the daily river discharge (m^3/s)	$daily\ bias = \frac{1}{N}\sum_{i=1}^{N}\lvert Q_i - \widehat{Q_i}\rvert$	[0;+ ∞]
f2	Mean absolute bias in the monthly river discharge (m^3/s)	$monthly\ bias = \frac{1}{12}\sum_{j=1}^{12}\lvert Q_{m,j} - \widehat{Q}_{m,j}\rvert$	[0;+ ∞]
f3	Correlation coefficient of the daily river discharges (-)	$r = \frac{\sum_{i=1}^{N}(Q_i-\mu)*(\widehat{Q_i}-\widehat{\mu})}{\sigma * \widehat{\sigma}}$	[0;1]
f4	NSE of the daily river discharges (-)	$NSE = 1 - \frac{\sum_{i=1}^{N}(Q_i-\widehat{Q_i})^2}{\sigma^2}$	[-∞ ;1]
f5	KGE of the daily river discharges (-)	$KGE = 1 - \sqrt{(1-r)^2 + \left(1-\frac{\widehat{\sigma}}{\sigma}\right)^2 + \left(1-\frac{\widehat{\mu}}{\mu}\right)^2}$	[-∞ ;1]
f6	Purpose-specific metrics	(SeeTable 3)	

Table 3
Purpose-specific metrics assessed in each case study. The unit of each metric is stated in parenthesis. Q10 refers to the flow that is reached or exceeded 10% of the period and Q90 refers to the flow that is reached or exceeded 90% of the period. These thresholds are estimated using the historical timeseries of the reference period.

Case study	Purpose specific metrics		
	f6.1	f6.2	f6.3
Denmark	Absolute bias in the annual number of days with discharge above or equal to the Q10 (d)	Absolute bias in the annual number of days with discharge below or equal to the Q90 (d)	Absolute bias in the mean groundwater head in shallow wells, not more than 5 m deep (m)
France	Relative bias in the mean winter streamflow (-)	Relative bias in the mean spring streamflow (-)	Error in the winter-spring center volume day (d) (Boyer et al., 2010)
Spain	Absolute bias in the annual streamflow volume (hm^3)	Absolute bias in the annual number of days with discharge below or equal to the Q90 (d)	Absolute bias in the annual number of days with discharge above or equal to the Q10 (d)

$$w_{2,i} = r_{6.1,i} \cdot r_{6.2,i} \cdot r_{6.3,i} \Big/ \sum_i w_{2,i} \tag{2}$$

$$w_{3,i} = r_{1,i} \cdot r_{2,i} \cdot r_{3,i} \cdot r_{4,i} \cdot r_{5,i} \cdot r_{6.1,i} \cdot r_{6.2,i} \cdot r_{6.3,i} \Big/ \sum_i w_{3,i} \tag{3}$$

$$w_{4,i} = \frac{1}{3} \quad (model \quad democracy) \tag{4}$$

where $w_{j,i}$ is the weight of model i (1−3) in the j-th weighting scheme (1−4) after normalization, and $r_{f,i}$ is the rank of model i with the f-th metric (1, 2, 3, 4, 5, 6.1, 6.2 or 6.3). In the following, $w_{j,i}$ will denote weights after normalization, so that $\sum_i w_{j,i} = 1$. Weighted projections are obtained by combining the weights and the hydrological projections (see Fig. 3, h).

2.4. Climate change impact assessment

2.4.1. Climate models selection

Five GCM-RCM combinations from the Euro-CORDEX initiative and specifically EUR-11 (~12 km spatial resolution grid) (Jacob et al., 2014) are used (see Table 4). This is a subset of the EUR-11 ensemble, since here, a climate model ensemble of opportunity (ensemble based on model availability - Annan and Hargreaves, 2010) is used at each study site as driving climate for the hydrological

Table 4
GCM-RCM combinations used in this study.

ID	GCM	RCM
CM1	ICHEC-EC-EARTH	SMHI-RCA4
CM2	ICHEC-EC-EARTH	KNMI-RACMO22E
CM3	MPI-M-MPI-ESM-LR	SMHI-RCA4
CM4	MOHC-HadGEM2	KNMI-RACMO22E
CM5	MOHC-HadGEM2	SMHI-RCA4

models. The combinations include two different RCMs driven by three different GCMs (Fig. 3, e). The Representative Concentration Pathway 8.5 (RCP 8.5) is used to assess the projected changes. RCP 8.5 represents the high-emission scenario, and it reproduces the current emission trend more accurately than other pathways (Hayhoe et al., 2017).

2.4.2. Bias correction

Daily temperature and precipitation were bias-corrected to match the observed climate data (Fig. 3,f) using methodologies previously tested in each region. For the Danish case, the distribution-based scaling method (Seaby et al., 2013) is used. For precipitation, a double Gamma distribution is used with the cutoff at the 90th percentile and a zero-precipitation threshold to match the number of observed and simulated dry days. Similarly, the distribution of the temperature simulations is fitted to a normal distribution (Pastén-Zapata et al., 2019).

For the French case, biases were corrected with the CDF-t statistical method (Michelangeli et al., 2009). The method assumes that the cumulative distribution function (CDF) of a local climate variable can be derived from a translation T of the CDF of the associated GCM variable in the historical period, and that this transformation is also valid in the projection period. The CDF-t technique has been applied for the French case study along with the Singularity Stochastic Removal to better deal with dry days, as done by Vrac et al. (2016).

Finally, an empirical quantile mapping is used for bias adjustment in the Spanish case (i.e., Boé et al., 2007, Sun et al., 2011). Climate model CDFs of daily precipitation and temperature for the whole reference period are matched with the observed CDFs. For precipitation, the CDFs are truncated to maintain the same number of dry days, fixing a threshold of 1 mm (Hay and Clark, 2003).

2.4.3. Uncertainty of the hydrological projections

Hydrological projections are produced using simulations from the hydrological models that were calibrated and evaluated using the standard SST (a common approach, in hydrology, which is described in Supplementary material S1). Bias-corrected climate data from the five GCM-RCM combinations are used as input to produce hydrological projections (divided into different periods 2010–2039, 2040–2069 and 2070–2099). The projected changes in the annual mean streamflow regime and purpose-specific metrics are assessed considering the different weights (Fig. 3, i). A Bayesian Model Averaging approach (Hoeting et al., 1999; Neuman, 2003) is used to assess the uncertainty associated with the weighted outputs using the variance of the projection ($\mathrm{Var}_j[x]$) for a (hydrological) variable (x):

$$Var_j[x] = \sum_{i=1}^{n_{HM}} w_{j,i} S_i + w_{j,i} \left(E_j[x] - \overline{x_i} \right)^2 \tag{5}$$

$$E_j[x] = \frac{1}{n_{CM}} \sum_{k=1}^{n_{CM}} \sum_{i=1}^{n_{HM}} w_{j,i} x_{k,i} \tag{6}$$

$$\overline{x_i} = \frac{1}{n_{CM}} \sum_{k=1}^{n_{CM}} x_{k,i} \tag{7}$$

$$S_i = \frac{1}{n_{CM} - 1} \sum_{k=1}^{n_{CM}} \left(x_{k,i} - \overline{x_i} \right)^2 \tag{8}$$

where $w_{j,i}$ refers to the computed weight for the hydrological model i for the *j-th* weighting scheme (1−4); S_i is the variance of hydrological variable for the hydrological model i; $E_j[x]$ is the mean of the ensemble of all climate and hydrological models; $\overline{x_i}$ is the mean of the ensemble of climate models for hydrological model i; $x_{k,i}$ is the hydrological variable value for the hydrological model i, when driven by climate model k; n_{HM} is the number of hydrological models (3); n_{CM} is number of climate models (5).

The uncertainty can also be shown as the standard deviation of the projection, defined as the square root of the variance. Bayesian Model Averaging has been used in previous studies to assess the uncertainty of hydrological projections (e.g., Najafi and Moradkhani, 2015; Refsgaard et al., 2012, Najafi et al., 2011).

3. Results

Focus in the first part of the analysis is given to the evaluation of the hydrological models using DSST and the metrics presented in Section 2.3.3 (Fig. 3d). This approach provides an assessment of the overall simulation skill of each hydrological model relative to the other models included in the ensemble, with emphasis on their ability to generate reliable results for periods with contrasting climate conditions (3 dry years vs 3 wet years). The weights (Eqs. 1–4) are subsequently combined with results from the projections of the three hydrological models to obtain a single future streamflow record in the beginning (2010–2039), middle (2040–2069) and end (2070–2099) of the century. The results generated in these periods are compared to the results from the reference period.

3.1. Simulation skill of the monthly streamflow regime under contrasting conditions

The simulation skill with respect to the monthly streamflow regime is important as decision-making is usually done at the monthly

scale for each case study. For the Danish case, shifts in the streamflow regime affects the agricultural management (e.g., water availability, drought periods, flood periods). For the French and Spanish cases, changing snow conditions impact the monthly streamflow regime.

Fig. 4 shows the observed and simulated discharge from each case during the calibration, control, and evaluation periods. In each period, the monthly average for three years is presented. In the Danish case, all three models simulate the monthly streamflow in the evaluation period accurately. The two-layer model underestimates the streamflow in late spring and summer months. In contrast, the winter flow is overestimated by the gravity flow and Richards' equation model. The best performing model is the gravity flow model, with the highest values for correlation coefficient (0.92), Nash-Sutcliffe coefficient (0.76) and the Kling-Gupta Efficiency (0.76), see Table S2.

For the French case, the three models tend to underestimate the spring peak flow resulting from snowmelt during the evaluation period. In summer, TOPMO and GR4J to a lesser extent, display some problems to adequately represent low flows compared to HBV0. The overall better fit between HBV0 simulations and the observations reflects the higher skill of the model considering the f2 metric (Table 5). It should nonetheless be noted that the differences between the models are rather small (around 5 m^3/s) compared to the average streamflow in the catchment (80 m^3/s).

The models from the Spanish case show large differences in the performance of the calibration, control and evaluation periods. Biases are larger during the evaluation period, where WiMMed has the smallest relative biases, followed by SWAT and finally SN-HYPE. WiMMed and especially SWAT tend to overestimate the monthly streamflow whereas the SH-HYPE model largely underestimates it. Despite this underestimation, SN-HYPE, like WiMMed, is able to capture the timing of the flow, which is not the case for SWAT.

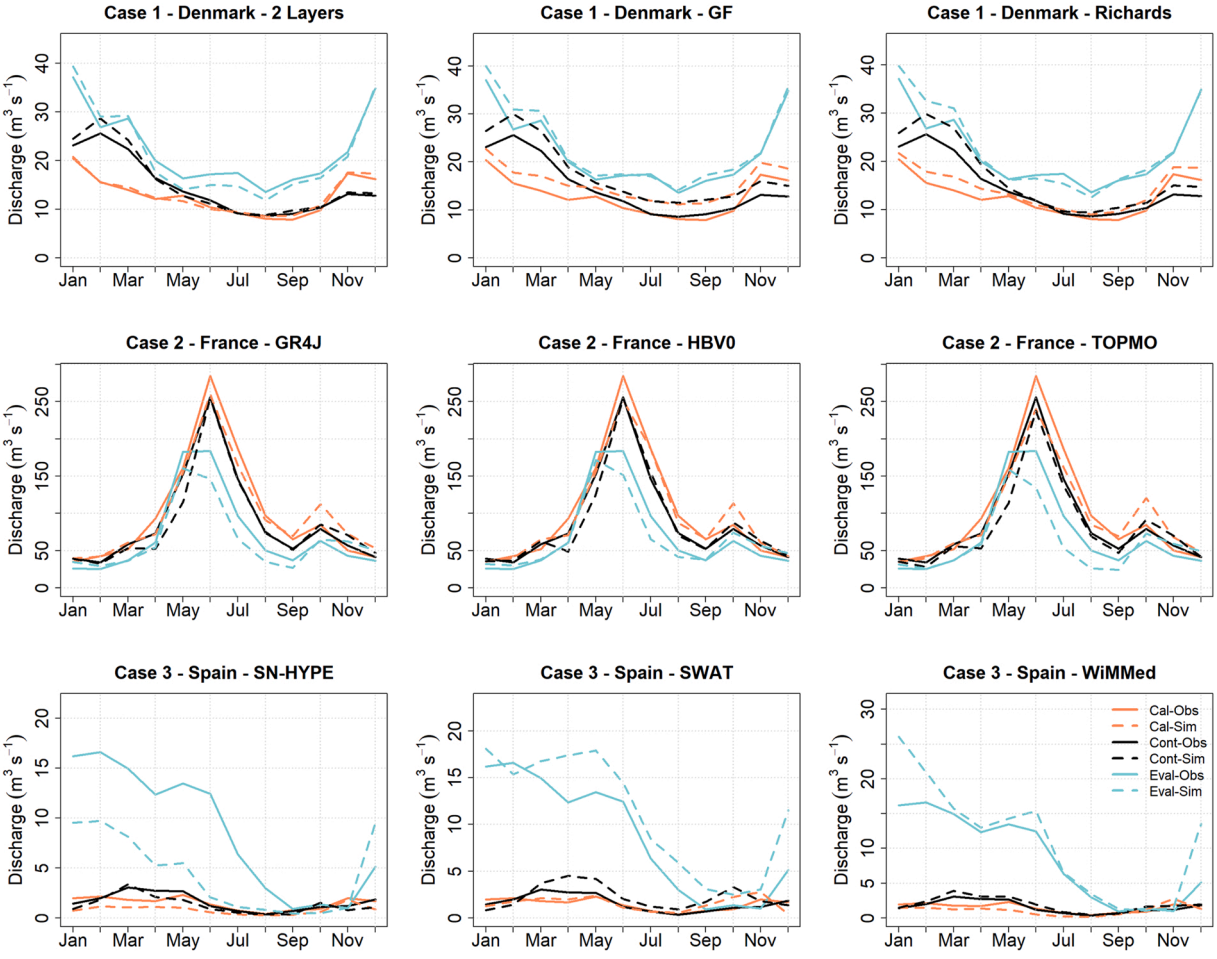


Fig. 4. Simulated and observed monthly streamflow regimes for the three hydrological models (columns) in each case study (rows) for the calibration, control and evaluation periods (DSST).

Table 5
DSST results of the metrics for each hydrological model in the evaluation period. Underlined and bold scores denote the result with the poorest and the best skill, respectively. General metrics are listed first (f1-f5), while purpose-specific metrics are listed at the bottom (f6.1- f6.3).

	CASE 1 – Denmark			Case 2 - France			Case 3 - Spain		
	Two-Layer vv	Gravity flow	Richards' Equation	GR4J	HBV0	TOPMO	SN-HYPE	SWAT	WiMMed
General metrics									
f1 (m^3/s)	3.49	**2.97**	3.70	20.4	**17.5**	23.3	5.53	3.78	**3.74**
f2 (m^3/s)	1.63	**1.22**	1.38	14.3	**10.8**	16.7	5.05	3.22	**2.94**
f3 (-)	**0.92**	**0.92**	0.88	0.76	**0.79**	0.73	0.69	0.78	**0.79**
f4 (-)	0.72	**0.76**	0.65	0.75	**0.78**	0.71	0.18	0.46	**0.49**
f5 (-)	**0.76**	**0.76**	0.69	0.79	**0.82**	0.74	0.23	0.65	**0.70**
Purpose-specific metrics									
f6.1	**1 (d)**	10	12	1.24 (-)	**1.23**	1.24	144 (hm^3)	68.5	**63.2**
f6.2	3 (d)	**0**	2	0.83 (-)	0.87	**0.79**	25.3 (d)	**18.0**	23.7
f6.3	2.53 (m)	2.64	**2.45**	1.3 (d)	**0.7**	2.3	99.7 (d)	47.7	**27.3**

3.2. Simulation skill for the general metrics under contrasting conditions

Results from the metrics indicate that no hydrological model gives the best results for all metrics (Table 5). Nevertheless, there are models that show a higher simulation skill than others or models whose simulation skill is low for most of the metrics (Table 5).

For instance, in the Danish case, the Richards' equation model has the lowest skill for half of the metrics, and it is the best performing model only for one metric, f6.3. The model with the best simulation skill varies between the two-layer and gravity flow models, indicating that the skill of these two models is similar for this set of metrics.

For the French case, the results indicate that the HBV0 model is the best performing model for seven out of eight metrics. It can also be observed that the TOPMO model has the worst performance for seven out of eight metrics. The simulation skill of the GR4J model is intermediate.

In the Spanish case, WiMMed shows the best results for seven out of the eight defined metrics. SN-HYPE has the worst results for all the metrics. From the three models used in the Spanish case, the simulation skill of SWAT falls in between the other two. It has the highest skill for one of the metrics, but its performance is similar to the WiMMed model.

3.3. Simulation skill for the purpose-specific metrics

Here, it is shown how the hydrological models simulate each of the purpose-specific metrics in the evaluation period (Table 5, metric f6's).

In the Danish case, the annual mean days with river flow equal or above the Q10 (f6.1) are generally simulated better during calibration (Table S2). For this metric, biases are of similar magnitude during control and evaluation, except for the two-layer model that has a better skill during control. All models seem to adapt well to the simulation of days with low flows (f6.2) outside the calibration period as the biases during evaluation are reduced compared to the other periods. The mean absolute error of the groundwater level in shallow wells (f6.3) is larger during control than evaluation, except for the Richards' equation model which has a similar skill in all periods. However, during calibration and control, the error is significantly higher for Richards' equation compared to two-layer with the Gravity formulation in between. During control, the best performing model varies for each purpose-specific metric (all models are the best for one metric).

For the French case, the HBV0 model generally outperforms the other models in all periods. Interestingly, while large differences in the ability of the models to reproduce winter volumes during the calibration period are found, very similar performances are observed during the evaluation period. Considering all periods and models, the accumulated winter streamflow is overestimated, whereas accumulated spring streamflow is systematically underestimated. The error in the accumulated spring streamflow is lower for HBV0 during all periods, whereas TOPMO has the highest error for all periods. The biases in the winter-spring center volume (WSCV) day increase in the control and evaluation periods compared to calibration. Errors in timing of the three models are small. All simulate peak flows approximately 3 days too early during the cold control period but display a short delay in the evaluation period.

For the Spanish case, the biases in the annual streamflow are relatively small for all models during the calibration and control but increase at least by a factor of four during the evaluation for all models because of the high variability between dry and wet conditions. The model with the lowest bias varies in each of the periods, while WiMMed never shows the highest errors. The bias in the annual mean number of days with river flow equal or below the Q90 are lower for the SN-HYPE model during the control and evaluation periods. In this case, SWAT largely overestimates the occurrence of low flows. Finally, for the annual mean number of days with river flow equal to or above the Q10 the biases are larger for the calibration and control compared to the evaluation period, due to the small values of the observed streamflow during the calibration and control. SN-HYPE and WiMMed always show the same signal in their biases whereas changes in SWAT go in the opposite direction.

3.4. Assigned weights based on the model simulation skill

For the French and Spanish cases, the three weighting schemes based on the DSST results (all except W4) give a clear preference to

one model out of the three (Table 6). These models are HBV0 for the French case and WiMMed for the Spanish case. Whatever the weighting scheme, these models have the highest weight and thus have the highest influence on the projection of impacts. For the Danish case, W1 and W3 produce the same weights while W2 results in 'model democracy'.

For the Danish case, considering W1 and W3, the weight of the two-layer model is almost three times higher than the model with the second highest weight. The difference in the French case is even larger, considering the weighting schemes W1, W2 and W3. For W1 and W3, the TOPMO model has a weight lower than 0.1. In all cases, the weights defined by W2 (purpose-specific metrics) show less dispersion than the weights of W1 and W3. This is the result of the more even distribution of model rankings according to the simulation skill of purpose-specific metrics and/or of the lower number of metrics that are assessed. Therefore, no model is completely discarded by this scheme.

3.5. Impact assessment

In this section the impact on the hydrological projections obtained from climate models is quantified (corresponding to Fig. 3, step g).

3.5.1. Projected changes

Changes in hydrological factors are based on climate model results. Below, the impact on different quantities and metrics are described.

3.5.1.1. Annual mean streamflow. The future streamflow is projected using the weights for each model (Table 6) at the beginning (2010–2039), middle (2040–2069) and end (2070–2099) of the century. The results from these periods are compared to those from the reference period (1976–2005 for the Danish and French cases, and 1971–2000 for the Spanish case) (Fig. 5). Additionally, the effect of the weighting method can be observed when compared to the standard 'model-democracy' approach, W4.

For the Danish case (Fig. 5, left panel), all weighting schemes indicate a median slight increase (approximately less than 5%) in the annual mean streamflow during the near and mid future. For the far future, the projections indicate a median increase of approximately 15% in the annual mean streamflow. However, the difference between the weighting scheme and hence the choice of hydrological model is small. In the French case (Fig. 5, central panel), annual mean streamflow is projected to increase in the near future while it is expected to decrease in the mid and far future. In the far future period, most of the models project a 10% decline compared to the historical period. However, projected values exhibit a rather large inter model uncertainty for most future periods. The impact of the weighting method on the change in discharge is in some cases significant. In the far future, changes between 0% and 10% are found where W4 consistently yields smaller values than the other weighting schemes.

Finally, in the Spanish case the annual mean streamflow is projected to decrease for all weighting methods and time horizons. The median decrease ranges from about 10% in the near future to almost 20% in the mid-future and between 20% and 30% in the far future. By the end of the century, all projections agree on the direction of the change signal for the Danish case. However, for the French and Spanish cases different change directions are projected for different models.

For the Danish case, weighting schemes W1 and W3 project a larger increase in streamflow than W2 and W4. However, the change throughout the century follows a similar pattern. For the French case, all four weighting approaches project rather similar changes across the century, although the discrepancy increases with time. In the far future period, the largest difference is found between W3 and W4, where the first simulates the largest future discharges, whereas the second simulates the lowest future discharge. For the Spanish case, a clear decreasing pattern in discharge is found for all weights, however, W2 always projects the driest scenario with a slightly higher value of change when compared to the other weights. On the contrary, W4 projects the lowest changes for all periods except for the mid future. For the mid future, W1 and W3, which always have similar patterns, project a scenario that is slightly wetter.

3.5.1.2. Purpose-specific metrics. Projections of the purpose-specific metrics are assessed by the end of the century, as the difference between the projections are larger here. For the Danish case, all weighting methods project an increase between 119% and 122%, compared to the reference period (Table 7). The impact of the choice of weights (or hydrological model) is larger for the number of days where streamflow is below or equal to the Q90 (f6.2), which decreases between 1% and 13%, compared to the reference period. The changes in the depth to groundwater (f6.3) are similar for all weighting schemes. Overall, the projections indicate a higher occurrence of wet extremes and an increase in river discharge along with a relatively small increase in groundwater levels. In all cases the impact of weighting is insignificant, which is no surprise considering the small differences in evaluation results being the basis for

Table 6
Weights assigned to each model for the four weighing methods (W1 - All but purpose-specific, W2 - Only purpose-specific, W3 - All metrics and W4 - Model democracy).

	Case 1 - Denmark			Case 2 - France			Case 3 - Spain		
	Two-Layer	Gravity	Richards	GR4J	HBV0	TOPMO	HYPE	SWAT	WIMMed
W1	0.128	0.862	0.011	0.116	0.880	0.004	0.022	0.113	0.865
W2	0.333	0.333	0.333	0.174	0.783	0.043	0.032	0.387	0.581
W3	0.128	0.862	0.011	0.028	0.971	0.001	0	0.08	0.92
W4	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333	0.333

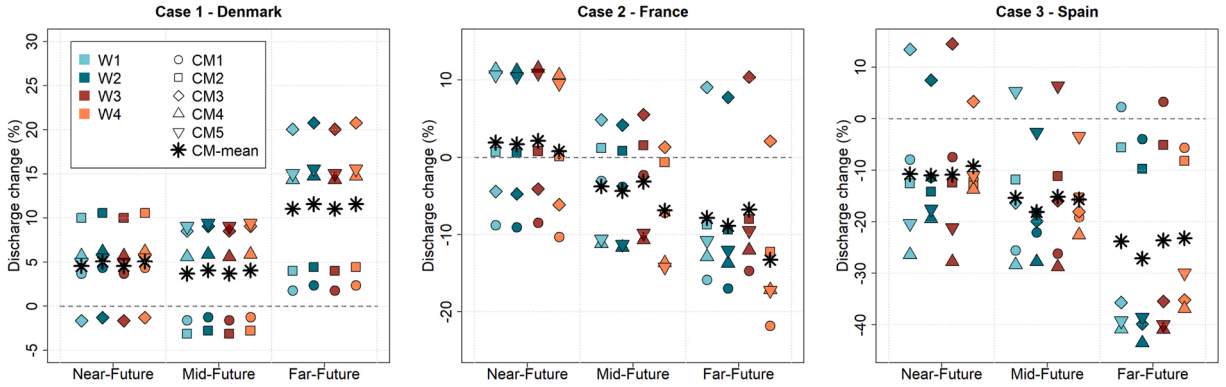


Fig. 5. Projected streamflow changes (%) in the near (2010–2039), mid (2040–2069) and far future (2070–2099) periods, compared to the reference period used for each site. Different colors represent the different weighting schemes, e.g., W1 applied to the hydrological models and different symbols represent the projection of the different climate models (e.g., CM1), see Table 4 for climate models.

Table 7
Projected relative change for the purpose-specific metrics for the different weights in each case study, for the RCP8.5 far-future (2070–2099) compared to the reference period. Please note that all changes are in percentage, except for f6.3 in the French case, where the change is in days.

Weights	Case study								
	Denmark			France			Spain		
	f6.1 (%)	f6.2 (%)	f6.3 (%)	f6.1 (%)	f6.2 (%)	f6.3 (d)	f6.1 (%)	f6.2 (%)	f6.3 (%)
W1	119	-1	11	15	-15	-24	-21	-43	118
W2	122	-13	13	14	-15	-24	-19	-40	98
W3	119	-1	11	17	-15	-24	-18	-43	113
W4	122	-13	13	5	-18	-24	-21	-35	72

weighting.

For the French case, the weighting schemes project very similar results apart from the change in winter average streamflow (f6.1). For this metric, relative changes range from + 5% to + 17% with the highest increases being obtained with the scheme W3. The average spring-summer streamflow (f6.2) is consistently projected to decrease by around −15%. Finally, all weights project a change of −24% for the WSCV day (f6.3), which corresponds to a 32-day shift earlier in the year of the winter spring half-flow day. These results indicate a shift from spring snowmelt to winter flow, the decline of spring streamflow being partially balanced by the increase in winter streamflow. Weighting (choice of hydrological model) is however relative unimportant for the projections of the purpose specific metrics.

In the Spanish case (Table 7), the annual streamflow volume (f6.1) is projected to decrease, ranging from −18% to −21% for the different weighting methods, compared to the reference period. The annual mean number of days below or equal to the Q90 (f6.2) is projected to decrease by −35% to −43%. Finally, the annual mean number of days above or equal to the Q10 are projected to increase by 72–118%, which is the largest difference between the weighting schemes for this variable. Summarizing, the general pattern shows a decrease in streamflow volume, a reduction in the number of days with low flows and a higher number of days with high values of the streamflow. The choice of hydrological model (weighting) is significant in a few cases.

3.6. Projection uncertainty

The computed standard deviations show how the use of the different weights impacts the uncertainty of the projection (Table 8).

Table 8
Uncertainty (shown as the standard deviation) of the weighted projections by the end of the century for the mean annual discharge (Q) and purpose specific metrics (f's).

	Case study											
	Denmark				France				Spain			
	Q (m^3/s)	f6.1 (d)	f6.2 (d)	f6.3 (m)	Q (m^3/s)	f6.1 (-)	f6.2 (-)	f6.3 (d)	Q (m^3/s)	f6.1 (hm^3)	f6.2 (d)	f6.3 (d)
W1	2.06	16.78	17.60	0.41	7.03	7.87	10.55	21.28	0.65	0.65	20.63	5.50
W2	2.11	16.41	20.27	0.56	7.24	7.86	10.72	21.45	0.81	0.81	28.93	6.05
W3	2.06	16.78	17.60	0.41	6.76	7.81	10.39	21.14	0.61	0.61	17.65	5.33
W4	2.11	16.41	20.27	0.56	7.23	7.44	10.88	22.20	0.75	0.75	39.33	6.56

Changes in standard deviation by the end of the century (2077–2099) are presented because it is the period with the largest simulation spread. The impact of weights is generally relatively small with the largest response found in the Spanish case.

For the Danish site, W2 and W4 give the same weights to all models. Therefore, no differences are found on the uncertainty. Similarly, W1 and W3 produced the same weighting scheme, and therefore no differences are found among them. For the projected change in the annual mean river discharge (Q), there is no large difference between the variance of W1 and W4. Compared to W4, the uncertainty of the projected occurrence of high flows increases (f6.1) when the W1 weights are used. In contrast, when using the weights of W1 the variance of the projection decreases for the occurrence of low flows (f6.2) and mean groundwater head (f6.3). From the analyzed metrics, the largest change in the variance is found for the projections of low flows.

Standard deviation values show a limited sensitivity to the weighting of hydrological models for the French case study. In general, model democracy (W4) is associated with the highest uncertainty, except for the winter accumulated flow metric (f6.1). Discarding the TOPMO model, and to a lesser extent, the GR4J model, marginally reduces the uncertainty. The projected average annual flow is the metric where the impact of hydrological model weighting is largest, with a decrease up to 5% from W4 to W3. In general, the weighting schemes with the most significant differences between the weights (W3) yields the highest reduction of the uncertainty.

For the Spanish case, the standard deviation for all analyzed variables follows a similar pattern. Higher values are always found for W4, followed by W2, with slightly lower values, and then W1 and W3, respectively, with significantly smaller variance values. For the projected change in annual mean river streamflow (Q and f6.1), the standard deviation can vary significantly depending on the selection of the weighting scheme, for instance from 0.61 to 0.81 for W2 and W3, respectively. This proportion drastically increases for the number of days with low flows (f6.2). In this case, the W4 is more than twice the value of W3, with standard deviation of 17.65 days versus 39.33 days. The ratio is lower for the number of days with high flows. In this case, the standard deviation ranges from 5.33 days to 6.56 days, for W3 and W4 respectively. Therefore, in the Spanish case, the uncertainty is clearly conditioned by the weighting scheme. Lower uncertainty relates to those weights with a clear preference for a specific model (W1 and W3).

4. Discussion

This study investigated the applicability of the DSST as a tool for quantifying the robustness of three hydrological models to describe the impact of climate change and to assess the impact of climate change as a function of the choice of hydrological model. Four weighting schemes were applied to three case studies in Europe to assess the sensitivity of the metrics and the corresponding uncertainty of the projections.

4.1. Discussion of DSST results

4.1.1. Do the hydrological models used in each case have the same transferability skills?

The DSST methodology highlighted different capabilities among hydrological models to describe changing climatic conditions, especially in the mountainous snow-dominated catchments (France and Spain) where one of the hydrological models clearly had better transferability skills despite comparable calibration results.

In the Spanish case, the different snow modules used by the hydrological models might explain the different transferability capabilities. The physically-based conceptualizations of WiMMed were able to capture the high variability in snow dynamics found in these catchments (especially the role of evaposublimation (Herrero and Polo, 2016)), that the degree-day approaches were not able to reproduce. Additionally, the higher spatiotemporal resolution used by WiMMed and the other models (30 x 30 m and hourly compared to 100 km^2 and daily, respectively) can also explain the differences in skill.

In the French case, the better performance of the HBV0 model is most certainly explained by a better structural description of the catchment processes, although snow processes dominate the hydrological regime in winter and spring and all three models have the same snow module. The interactions between the snow module and the parameters of the rainfall-runoff models could yield equifinality and affect model calibration. For instance, the parameter governing snowpack inertia in TOPMO was significantly different from the other models and might have caused an inappropriate scaling of the snow processes in warm conditions. Better representation of the catchment processes in the models might avoid undesirable compensations in calibration between hydrological model parameters and the snow module parameter, and explain the better performance obtained with HBV0.

In the Danish case, the DSST approach does not yield as straightforward results as for the other case studies. The relatively small differences among the alternative versions of the MIKE SHE model might explain these smaller contrasts compared to the other cases. Overall, the performance of the Gravity Flow model was found to be the best. The most complex representation of the unsaturated zone (Richard's equation) does not improve the simulation skill. This might be a result of problems to estimate optimal parameters because of the highly non-linear description of the soil water physics described by Richard's equation (retention curve, unsaturated hydraulic conductivity curve, both presented by van Genuchten (1980)). Additional explanations include the uncertainty related to the calibration of the required parameters or because the assessed metrics might not be sensitive to such parameters. However, the Richard's equation was able to describe the groundwater (purpose-specific) metric slightly better than the other models despite a relatively poor calibration result.

4.1.2. Which are the main benefits and drawbacks of the DSST approach for weighting hydrological models?

The main strength of the DSST approach is to provide a better evaluation of the sensitivity of the hydrological model to changes in forcing conditions. Here, this advantage has been used to assess the robustness of models to cope with contrasting climate conditions and assign different weights to hydrological models. Relying on model performance in calibration may lead to inappropriate

assessments because of parameter overfitting (Andréassian et al., 2012). Even in an SST approach, there is a risk that a model's suitability for extrapolation may be overestimated because model robustness to expected climatic changes is not explicitly evaluated. Model rankings and weights would have been different if a traditional SST approach had been used, especially in the Danish and Spanish cases (see Table S2).

However, the DSST approach used in this study also has limitations. When focus is given to a single variable to discriminate between calibration, control and evaluation periods, the hydrological response might be sensitive to the choice of variable. To circumvent this problem, simultaneous changes in air temperature and precipitation could have been used (Dakhlaoui et al., 2017, 2019). Another constraint is the limited amplitude and the differences in character of climatic changes encompassed in past historical data compared to future climate change (Stephens et al., 2020). Nevertheless, possibilities to evaluate models beyond conditions measured in existing records are limited to techniques based on synthetic weather generators, which may have difficulties to represent natural climate variability (e.g., Guo et al., 2018). As noted by Klemeš (1986), passing the DSST may not be a sufficient condition for a model to be valid, yet it is a necessary condition that every model intended to be used in extrapolation should satisfy.

4.1.3. Does the selection of evaluation metrics have an impact on the weighted projections?

Purpose-specific metrics, which were defined from an end-user perspective and considering the adequacy-for-purpose view (Parker, 2020), have proven to modify the ranking of hydrological models when compared with traditional metrics in two of the cases (Denmark and Spain). Indeed, while model rankings derived from general metrics were relatively similar across the considered metrics, they appeared to vary from one purpose-specific metric to another. This result stresses the interest in using several performance criteria focusing each on different features of the hydrological regime to precisely inform model selection based on case requirements. Although, it should be noted that addressing specific needs of stakeholders requires that performance criteria are defined with caution. Otherwise, noise can be introduced in the analysis of projections (Weichselgartner and Arheimer, 2019). This fact highlights the need of a co-development approach, connecting user needs and models strengths (Photiadou et al., 2021). Overall, the subjectivity in the selection of metrics used in this type of analyses is something that cannot be avoided and should be considered when interpreting the resulting projections (Christensen et al., 2010).

4.2. Opportunities and challenges for hydrological model weighting

4.2.1. How does the weighting scheme affect the analysis?

Interestingly, when as many metrics as possible were included in the computation of the weights, at least one model out of the three was discarded (almost two in the Spanish case). This is the combined effect of the multiplicative rather than additive approach in the weight's computation, as well as the choice to derive metric-wise weights from model ranking. The sensitivity of the adopted scheme could be reduced if weights were attributed according to performance rather than ranks. However, this alternative method is difficult to apply in practice. Indeed, the variations of metric values with model skill are often non-linear and thus difficult to aggregate into a single inclusive score. An alternative to our method for model rankings would be to define thresholds to evaluate performance and set scores to the models (e.g. 1 if below the threshold, 2 if above). However, the choice of a threshold is subjective as well as context-dependent since the characteristics of the observed streamflow timeseries might significantly influence model scores (e.g., Schaefli and Gupta, 2007; Criss and Winston, 2008; Knoben et al., 2019). Hence, even if a usual performance threshold may apply in some catchments, it may simply be unobtainable in other catchments for most of the current hydrological models. Hence, we found that ranking models based on the ranking of their scores was more appropriate given the diversity of our case studies than defining performance thresholds. Similarly, the multiplicative scoring of the weights used here is subjective. The impact of using another type of aggregating scoring, such as additive, can be explored in future studies. Although, for different experimental setups, the selection of different scoring aggregation approaches has a small influence on the ensemble (Christensen et al., 2010).

Further, the number of metrics included in the weighting computations also influences the discrepancies between model weights. In our approach, since metrics are not completely independent from each other, poor models may rank last for many metrics and tend to be more severely discarded by the weighting schemes with more metrics in the computation.

Discarding the worst models can reduce ensemble uncertainty and avoid very poor models to reduce the effectiveness of the multi-modelling approach (Najafi et al., 2011). Nonetheless, it is recommended to include as many hydrological models as possible when implementing such methodology. As pointed out by Seiller et al. (2012), an ensemble of models generally improves temporal and spatial transposability compared to individual models.

4.2.2. Approaches for the selection of hydrological and climatological models

In hydrological impact analyses, the selection of climate models represents an important source of uncertainty as it affects changes in precipitation, which is the main driving variable of hydrological models (Refsgaard et al., 2016). This is often the case for catchments where streamflow change is mainly driven by climate variables, such as precipitation. Nevertheless, there are instances where the hydrological model has a large impact on the uncertainty of the projection, such as low flows, because streamflow generation can be strongly driven by hydrological model states (Najafi et al., 2011). Thus, careful selection of climate and hydrological models is important to improve the reliability of the projection (Her et al., 2019).

Assigning weights to climate models in impact studies based on their performance has been suggested before. For instance, some studies focused on the reproduction of the present local climate (e.g., Wang et al., 2019) or on the reproduction of large-scale climate processes (e.g. Haughton et al., 2015; Knutti et al., 2017). Nevertheless, these approaches are not entirely supported by the climate modeling community as, traditionally, every climate model is considered as having the same probability of occurrence and being only

useful when combined with other climate models (Collins, 2017). Climate models are intended to reproduce large-scale climate processes, and it is not trivial to reduce this objective to streamflow metrics. Nevertheless, it should be noted that selecting a specific ensemble of climate models (without weighting) is common and supported by selection methods (e.g., Pechlivanidis et al., 2018; Lutz et al., 2016). In contrast, hydrological models are commonly developed for specific catchments and undergo calibration efforts to better simulate the hydrologic processes within the catchment. In addition, hydrological model weighting based on simulation skill can be a mean to increase the reliability of the simulation outputs (Krysanova et al., 2018; Her et al., 2019). This highlights the potential of assigning weights to hydrological models for local impact studies.

4.2.3. Is the uncertainty of the projections impacted by the weighting schemes?

Our results show that weighting hydrological models affects the uncertainty of the projection for some metrics. For most of the metrics in Table 8, there is no reduction in uncertainty when using the weighted simulations in the Danish and French cases. Considering that the involved purpose-specific metrics in the French case are related to snow processes, the choice of using the same snow module for all hydrological models leads to the apparent very low sensitivity of the uncertainty to the hydrological models. In the Danish case, where the only difference is found on the conceptualization of the processes in the unsaturated zone, the high degree of similarity in model structures could explain this result. Nevertheless, the reduction of the uncertainty associated with the purpose-specific metric focusing on low flows (Tables 8, f6.2) with non-model-democracy weighting schemes (W1 and W3) shows that the contribution of hydrological model uncertainties to the total uncertainty cannot be neglected on the whole hydrological cycle. In the Spanish case, the decrease in uncertainty is relatively larger. This shows the importance of using a reliable hydrological model or weighting approach for such an area, where uncertainty can be reduced to half of its initial value for most metrics.

Our experiment demonstrates that considering the performance of the hydrological models is likely to provide more reliable simulations in impact studies. However, it is crucial to understand that reducing uncertainties is not a goal that justifies any means. It is likely that in the presented case studies the total uncertainty is underestimated. Some modelling choices made in our case studies for the sake of simplicity probably lead to neglection of complexity of the represented hydrological processes. Therefore, it should be noticed that the methodology presented in this article is focused on improving the reliability of hydrological projections and should be applied on an appropriately representative ensemble of hydrological models.

An appropriate weighting scheme that increases the reliability of the projection requires a careful analysis of the performance of the models. The main contributor (namely climate model or hydrological model) to the uncertainty varies for each process. To identify the main contributor, it is important to understand the relevant hydrologic processes within the catchment and how these are simulated. As the proposed approach focuses on hydrological model weighting, it would have larger benefits for processes and variables which are more influenced by the simulation of hydrological models (Her et al., 2019; Najafi et al., 2011; Vidal et al., 2016).

5. Conclusions

This study aimed to assess a framework for assigning weights to hydrological models based on their robustness under changing climate conditions and evaluate the effect of weights on the uncertainty of the projected impact. Three case studies were selected in a North-South gradient across Europe, capturing not only the hydrological differences between the cases, but also the different water management issues to be assessed in a future climate context. The main findings of this study can be summarized as:

- The proposed methodology covers different aspects relevant to climate change impact assessments, such as: (1) integrating purpose-specific metrics addressing relevant hydrological issues for water managers, (2) assessing the robustness of the hydrological models in a climate change context, and (3) model weighting based on the robustness of the hydrological models under changing climate conditions. By integrating these aspects, the adequacy-for-purpose and reliability of the projections increases.
- Model weighting can help to better understand the uncertainty of the projections and increase the reliability of their assessment, particularly regarding water management metrics. We recommend using this methodology with a large ensemble of climate and hydrological models to account for different conceptions and parameterizations of the climate and hydrological systems and by including different purpose-specific metrics in the analysis. For the purpose-specific metrics, our results indicate that the change in uncertainty was not much when hydrological models had similar simulation skill (robust), but it was impacted by the weighting scheme when the performance among the hydrological models differed significantly.
- The use of purpose-specific metrics highlights the importance of evaluating the quality of a model based on its purpose, helping to prioritize hydrological models under certain hypotheses. This fact gains importance for decision-making where information needs to be adapted to potential user's requirements in favour of usability.

CRediT authorship contribution statement

EPZ – Development of the research idea, development of the methodology, leadership of the Danish case, bias-correction of the climate model outputs, calibration and weighting of the hydrological models, analysis of the results and development of the manuscript. RP – Development of the methodology, leadership of the Spanish case, bias-correction of the climate model outputs, calibration and weighting of the hydrological models, analysis of the results, development of the manuscript. PRG – Development of the methodology, leadership of the French case, calibration and weighting of the hydrological models, analysis of the results and development of the manuscript. TSO – Development of the methodology and analysis for the Danish case. JAI – Calibration and weighting of the hydrological models for the Spanish case. MJP – Calibration and weighting of the hydrological models for the Spanish case. AL –

Production of the bias corrected climate outputs for the French case. RS – Model setup and calibration for the Danish case. CP – Development of the methodology and climate model selection assessment. GT – Development of the methodology and analysis of the results. JCR – Uncertainty assessment of the projections. All coauthors reviewed and commented on the original draft of the manuscript.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgements

This work was funded by the project AQUACLEW, which is part of ERA4CS, an ERA-NET initiated by JPI Climate, and funded by FORMAS (SE), DLR (DE), BMWFW (AT), IFD (DK), MINECO (ES), ANR (FR) with co-funding by the European Commission [Grant 690462]. R. Pimentel acknowledges fundings by the modality 5.2 of the Programa Propio-2018 of the University of Cordoba and the *Juan de la Cierva Incorporación Programme of the Spanish Ministry of Science and Innovation (IJC2018–038093-I). J. Aparicio acknowledges fundings by the Programme Ayudas para contratos predoctorales para la formación de doctores of the Spanish Ministry of Science and Innovation (PRE2019–090493). R. Pimentel and J. Aparicio are members of DAUCO, Unit of Excellence ref. CEX2019–000968-M, with financial support from the Spanish Ministry of Science and Innovation, the Spanish State Research Agency, through the Severo Ochoa and María de Maeztu Program for Centers and Units of Excellence in R&D.*

The authors would like to thank the Euro-CORDEX initiative for providing access to the raw climate projections.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ejrh.2022.101113.

References

Abbott, M.B., Bathurst, J.C., Cunge, J.A., O'Connell, P.E., Rasmussen, J., 1986. An introduction to the european hydrological system - systeme hydrologique Europeen, "SHE", 1: history and philosophy of a physically-based, distributed modelling system. J. Hydrol. https://doi.org/10.1016/0022-1694(86)90114-9.

Andréassian, V., Le Moine, N., Perrin, C., Ramos, M.H., Oudin, L., Mathevet, T., Lerat, J., Berthet, L., 2012. All that glitters is not gold: the case of calibrating hydrological models. Hydrol. Process. 26, 2206–2210. https://doi.org/10.1002/hyp.9264.

Annan, J.D., Hargreaves, J.C., 2010. Reliability of the CMIP3 ensemble. Geophys. Res. Lett. 37 (2) https://doi.org/10.1029/2009GL041994.

Arnold, J.G., Srinivasan, R., Muttiah, R.S., Williams, J.R., 1998. Large area hydrologic modeling and assessment part I: Model development. J. Am. Water Resour. Assoc. 34 (1), 73–89. https://doi.org/10.1111/j.1752-1688.1998.tb05961.x.

Bergstrom, S., 1995. The HBV model. In: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, Highlands Ranch, CO, pp. 443–476.

Bergström, S., Forsman, A., 1973. Development of a conceptual deterministic rainfall-runoff model. Hydrol. Res. 4 (3), 147–170. https://doi.org/10.2166/nh.1973.0012.

Beven, K.J., Kirkby, M.J., 1979. A physically based, variable contributing area model of basin hydrology. Hydrol. Sci. J. 24 (1), 43–69. https://doi.org/10.1080/02626667909491834.

Blöschl, G., Bierkens, M.F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Stumpp, C., 2019. Twenty-three unsolved problems in hydrology (UPH)–a community perspective. Hydrol. Sci. J. 64 (10), 1141–1158. https://doi.org/10.1080/02626667.2019.1620507.

Boé, J., Terray, L., Habets, F., Martin, E., 2007. Statistical and dynamical downscaling of the Seine basin climate for hydro-meteorological studies. Int. J. Climatol. 27 (12), 1643–1655. https://doi.org/10.1002/joc.1602.

Boyer, C., Chaumont, D., Chartier, I., Roy, A.G., 2010. Impact of climate change on the hydrology of St. Lawrence tributaries. J. Hydrol. 384, 65–83. https://doi.org/10.1016/j.jhydrol.2010.01.011 (1-2).

Branche, E., 2017. The multipurpose water uses of hydropower reservoir: the SHARE concept. Comptes Rendus Phys. 18 (7–8), 469–478. https://doi.org/10.1016/j.crhy.2017.06.001.

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

Broderick, C., Matthews, T., Wilby, R.L., Bastola, S., Murphy, C., 2016. Transferability of hydrological models and ensemble averaging methods between contrasting climatic periods. Water Resour. Res. 52 (10), 8343–8373. https://doi.org/10.1002/2016WR018850.

Christensen, J.H., Kjellström, E., Giorgi, F., Lenderink, G., Rummukainen, M., 2010. Weight assignment in regional climate models. Clim. Res. 44 (2–3), 179–194. https://doi.org/10.3354/cr00916.

Collins, M., 2017. Still weighting to break the model democracy. Geophys. Res. Lett. 44 (7), 3328–3329. https://doi.org/10.1002/2017GL073370.

Coron, L., and Perrin, C. (2018). airGRplus: Additional Hydrological Models to the 'airGR' Package. R package version 0.8.1.2.

Coron, L., Andreassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., Hendrickx, F., 2012. Crash testing hydrological models in contrasted climate conditions: an experiment on 216 Australian catchments. Water Resour. Res. 48 (5) https://doi.org/10.1029/2011WR011721.

Coron, L., Andréassian, V., Perrin, C., Bourqui, M., Hendrickx, F., 2014. On the lack of robustness of hydrologic models regarding water balance simulation: a diagnostic approach applied to three models of increasing complexity on 20 mountainous catchments. Hydrol. Earth Syst. Sci. 18 (2), 727–746. https://doi.org/10.5194/hess-18-727-2014.

Coron, L., Thirel, G., Delaigue, O., Perrin, C., Andréassian, V., 2017. The suite of lumped GR hydrological models in an R package. Environ. Model. Softw. 94, 166–171. https://doi.org/10.1016/j.envsoft.2017.05.002.

Coron L., Perrin, C., Delaigue, O., Thirel, G. and Michel, C. (2018). Suite of GR hydrological models for precipitation-runoff modelling, R package version 1.0.12.3.2. ⟨https://webgr.inrae.fr/en/airGR/⟩.

Criss, R.E., Winston, W.E., 2008. Do Nash values have value? discussion and alternate proposals. Hydrol. Process.: Int. J. 22 (14), 2723–2725. https://doi.org/10.1002/hyp.7072.

Dakhlaoui, H., Ruelland, D., Tramblay, Y., Bargaoui, Z., 2017. Evaluating the robustness of conceptual rainfall-runoff models under climate variability in northern Tunisia. J. Hydrol. 550, 201–217. https://doi.org/10.1016/j.jhydrol.2017.04.032.
Dakhlaoui, H., Ruelland, D., Tramblay, Y., 2019. A bootstrap-based differential split-sample test to assess the transferability of conceptual rainfall-runoff models under past and future climate variability. J. Hydrol. 575, 470–486. https://doi.org/10.1016/j.jhydrol.2019.05.056.
Dams, J., Nossent, J., Senbeta, T.B., Willems, P., Batelaan, O., 2015. Multi-model approach to assess the impact of climate change on runoff. J. Hydrol. 529, 1601–1616. https://doi.org/10.1016/j.jhydrol.2015.08.023.
Delaigue, O., Génot, B., Lebecherel, L., Brigode, P., and Bourgin, P.Y. (2020). Database of watershed-scale hydroclimatic observations in France, Université Paris-Saclay, INRAE, HYCAR Research Unit, Hydrology group, Antony, [data set], available at: ⟨https://webgr.inrae.fr/base-de-donnees/⟩, last access: 31 July 2020.
Demirel, M.C., Koch, J., Mendiguren, G., Stisen, S., 2018. Spatial pattern oriented multicriteria sensitivity analysis of a distributed hydrologic model. Water 10 (9), 1188. https://doi.org/10.3390/w10091188.
Doherty, J., Brebber, L., Whyte, P., 1994. PEST: Model-Independent Parameter Estimation, 122. Watermark Computing, Corinda, Australia, p. 336.
Edijatno Nascimento, N.D.O., Yang, X., Makhlouf, Z., Michel, C., 1999. GR3J: a daily watershed model with three free parameters. Hydrol. Sci. J. 44 (2), 263–277. https://doi.org/10.1080/02626669909492221.
Fonseca, A.R., Santos, J.A., 2019. Predicting hydrologic flows under climate change: the Tâmega Basin as an analog for the mediterranean region. Sci. Total Environ. 668, 1013–1024. https://doi.org/10.1016/j.scitotenv.2019.01.435.
Fowler, K., Knoben, W., Peel, M., Peterson, T., Ryu, D., Saft, M., Ki-Weon, S., Western, A., 2020. Many commonly used rainfall-runoff models lack long, slow dynamics: Implications for runoff projections. Water Resour. Res. 56 (5) https://doi.org/10.1029/2019WR025286 e2019WR025286.
Gelfan, A.N., Millionshchikova, T.D., 2018. Validation of a hydrological model intended for impact study: problem statement and solution example for Selenga River basin. Water Resour. 45 (1), 90–101.
Graham, D.N., Butts, M.B., 2005. Flexible, integrated watershed modelling with MIKE SHE. Watershed Models 849336090, 245–272.
Green, W.H., Ampt, G.A., 1911. Studies in soil physics: I. the flow of air and water through soils. J. Agric. Sci. 4, 1–24. https://doi.org/10.1017/S0021859600001441.
Guo, D., Westra, S., Maier, H.R., 2018. An inverse approach to perturb historical rainfall data for scenario-neutral climate impact studies. J. Hydrol. 556, 877–890. https://doi.org/10.1016/j.jhydrol.2016.03.025.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J. Hydrol. 377 (1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.
Hagemann, S., et al., 2013. Climate change impact on available water resources obtained using multiple global climate and hydrology models. Earth Syst. Dyn. 4, 129–144. https://doi.org/10.5194/esd-4-129-2013.
Hargreaves, G.H., Samani, Z.A., 1982. Estimating potential evapotranspiration. J. Irrig. Drain. Div. 108 (3), 225–230.
Haughton, N., Abramowitz, G., Pitman, A., Phipps, S.J., 2015. Weighting climate model ensembles for mean and variance estimates. Clim. Dyn. 45 (11–12), 3169–3181. https://doi.org/10.1007/s00382-015-2531-3.
Hay, L.E., Clark, M.P., 2003. Use of statistically and dynamically downscaled atmospheric model output for hydrologic simulations in three mountainous basins in the western United States. J. Hydrol.. 282, 56–75. https://doi.org/10.1016/S0022-1694(03)00252-X.
Hayhoe, K., Edmonds, J., Kopp, R.E., LeGrande, A.N., Sanderson, B.M., Wehner, M.F., Wuebbles, D.J., 2017. Climate models, scenarios, and projections. In: Wuebbles, D.J., Fahey, D.W., Hibbard, K.A., Dokken, D.J., Stewart, B.C., Maycock, T.K. (Eds.), Climate Science Special Report: Fourth National Climate Assessment, Volume I. U.S. Global Change Research Program, pp. 133–160. https://doi.org/10.7930/J0WH2N54.
Henriksen, H.J., Troldborg, L., Nyegaard, P., Sonnenborg, T.O., Refsgaard, J.C., Madsen, B., 2003. Methodology for construction, calibration and validation of a national hydrological model for Denmark. J. Hydrol. 280 (1–4), 52–71. https://doi.org/10.1016/S0022-1694(03)00186-0.
Her, Y., Yoo, S.H., Cho, J., Hwang, S., Jeong, J., Seong, C., 2019. Uncertainty in hydrological analysis of climate change: multi-parameter vs. multi-GCM ensemble predictions. Sci. Rep. 9 (1), 1–22. https://doi.org/10.1038/s41598-019-41334-7.
Herrero, J., Polo, M.J., 2016. Evaposublimation from the snow in the Mediterranean mountains of Sierra Nevada (Spain). Cryosphere 10, 2981–2998. https://doi.org/10.5194/tc-10-2981-2016, 2016.
Herrero, J., Aguilar, C., Polo, M.J. and M.A.,Losada, 2007: Mapping of meteorological variables for runoff generation forecast in distributed hydrological modelling. Proceeding, Hydraulic Measurements & Experimental Methods Conference, New York, 606–611, 2007.
Herrero, J., Polo, M.J., Moñino, A., Losada, M.A., 2009. An energy balance snowmelt model in a Mediterranean site. J. Hydrol. 371 (1–4), 98–107. https://doi.org/10.1016/j.jhydrol.2009.03.021.
Hewitt, C., Mason, S., Walland, D., 2012. The global framework for climate services. Nat. Clim. Change 2 (12), 831–832. https://doi.org/10.1038/nclimate1745.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 14 (4), 382–417.
Højberg, A.L., Troldborg, L., Stisen, S., Christensen, B.B.S., Henriksen, H.J., 2013. Stakeholder driven update and improvement of a national water resources model. Environ. Model. Softw. 40, 202–213. https://doi.org/10.1016/j.envsoft.2012.09.010.
Hundecha, Y., Arheimer, B., Berg, P., et al., 2020. Effect of model calibration strategy on climate projections of hydrological indicators at a continental scale. Clim. Change 163, 1287–1306. https://doi.org/10.1007/s10584-020-02874-4.
Jacob, D., Petersen, J., Eggert, B., Alias, A., Christensen, O.B., Bouwer, L.M., Yiou, P., 2014. EURO-CORDEX: new high-resolution climate change projections for European impact research. Reg. Environ. Change 14 (2), 563–578. https://doi.org/10.1007/s10113-013-0499-2.
Karlsson, I.B., Sonnenborg, T.O., Refsgaard, J.C., Trolle, D., Børgesen, C.D., Olesen, J.E., Jensen, K.H., 2016. Combined effects of climate models, hydrological model structures and land use scenarios on hydrological impacts of climate change. J. Hydrol. 535, 301–317. https://doi.org/10.1016/j.jhydrol.2016.01.069.
Klemeš, V., 1986. Operational testing of hydrological simulation models. Hydrol. Sci. J. 31 (1), 13–24. https://doi.org/10.1080/02626668609491024.
Knoben, W.J.M., Freer, J.E., Woods, R.A., 2019. Technical note: inherent benchmark or not? comparing Nash–sutcliffe and Kling–Gupta efficiency scores. Hydrol. Earth Syst. Sci. 23 (10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019.
Knutti, R., 2010. The end of model democracy? Clim. Change 102, 395–404. https://doi.org/10.1007/s10584-010-9800-2.
Knutti, R., Sedláček, J., Sanderson, B.M., Lorenz, R., Fischer, E.M., Eyring, V., 2017. A climate model projection weighting scheme accounting for performance and interdependence. Geophys. Res. Lett. 44, 1909–1918. https://doi.org/10.1002/2016GL072012.
Koch, J., Schneider, R., 2022. Long short-term memory networks enhance rainfall-runoff modelling at the national scale of Denmark. GEUS Bull. 49.
Kristensen, K.J., Jensen, S.E., 1975. A model for estimating actual evapotranspiration from potential transpiration. Nord. Hydrol. 6, 70–88.
Krysanova, V., Donnelly, C., Gelfan, A., Gerten, D., Arheimer, B., Hattermann, F., Kundzewicz, Z.W., 2018. How the performance of hydrological models relates to credibility of projections under climate change. Hydrol. Sci. J. 63 (5), 696–720. https://doi.org/10.1080/02626667.2018.1446214.
Leleu, I., Tonnelier, I., Puechberty, R., Gouin, P., Viquendi, I., Cobos, L., Foray, A., Baillon, M., Ndima, P.-O., 2014. La refonte du système d'information national pour la gestion et la mise à disposition des données hydrométriques. Houille Blanc. 1, 25–32. https://doi.org/10.1051/lhb/2014004.
Lemaitre-Basset, T., Collet, L., Thirel, G., Parajka, J., Evin, G., Hingray, B., 2021. Climate change impact and uncertainty analysis on hydrological extremes in a French Mediterranean catchment. Hydrol. Sci. J. 1–16. https://doi.org/10.1080/02626667.2021.1895437.
Li, C.Z., Zhang, L., Wang, H., Zhang, Y.Q., Yu, F.L., Yan, D.H., 2012. The transferability of hydrological models under nonstationary climatic conditions. Hydrol. Earth Syst. Sci. 16 (4), 1239–1254. https://doi.org/10.5194/hess-16-1239-2012.
Lindström, G., Pers, C., Rosberg, J., Strömqvist, J., Arheimer, B., 2010. Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales. Hydrol. Res. 41 (3–4), 295–319. https://doi.org/10.2166/nh.2010.007.
Lutz, A.F., ter Maat, H.W., Biemans, H., Shrestha, A.B., Wester, P., Immerzeel, W.W., 2016. Selecting representative climate models for climate change impact studies: an advanced envelope-based selection approach. Int. J. Climatol. 36 (12), 3988–4005. https://doi.org/10.1002/joc.4608.
Mathevet, T. (2005). Quels modèles pluie-débit globaux au pas de temps horaire? Développements empiriques et comparaison de modèles sur un large échantillon de bassins versants (Which Rainfall-Runoff model at the hourly time-step? Empirical development and intercomparison of rainfall–runoff models on a large sample of watersheds), PhD thesis, ⟨http://webgr.irstea.fr/publications/theses/⟩, ENGREF, Paris, France, 354 pp.
Maughan, N. (2015). The Serre-Ponçon dam and the Durance river: the founding act towards the most regulated French waterway. Arcadia.

Merz, R., Parajka, J., Blöschl, G., 2011. Time stability of catchment model parameters: implications for climate impact analyses. Water Resour. Res. 47 (2) https://doi.org/10.1029/2010WR009505.
Michel, C., Perrin, C., Andréassian, V., 2003. The exponential store: a correct formulation for rainfall—runoff modelling. Hydrol. Sci. J. 48 (1), 109–124. https://doi.org/10.1623/hysj.48.1.109.43484.
Michelangeli, P.A., Vrac, M., Loukos, H., 2009. Probabilistic downscaling approaches: application to wind cumulative distribution functions. Geophys. Res. Lett. 36 (11) https://doi.org/10.1029/2009GL038401.
Monteith, J.L., Szeicz, G., Yabuki, K., 1964. Crop photosynthesis and the flux of carbon dioxide below the canopy. J. Appl. Ecol. Vol. 1 (No. 2), 321–337. https://doi.org/10.2307/2401316.
Najafi, M.R., Moradkhani, H., 2015. Multi-model ensemble analysis of runoff extremes for climate change impact assessments. J. Hydrol. 525, 352–361. https://doi.org/10.1016/j.jhydrol.2015.03.045.
Najafi, M.R., Moradkhani, H., Jung, I.W., 2011. Assessing the uncertainties of hydrologic model selection in climate change impact studies. Hydrol. Process. 25 (18), 2814–2826. https://doi.org/10.1002/hyp.8043.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—a discussion of principles. J. Hydrol. 10 (3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.
Neuman, S.P., 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. Stoch. Environ. Res. Risk Assess. 17 (5), 291–305. https://doi.org/10.1007/s00477-003-0151-7.
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 (1–4), 290–306. https://doi.org/10.1016/j.jhydrol.2004.08.026.
Parker, W.S., 2020. Model evaluation: An adequacy-for-purpose view. Philos. Sci. 87 (3), 457–477. https://doi.org/10.1086/708691.
Pastén-Zapata, E., Sonnenborg, T.O., Refsgaard, J.C., 2019. Climate change: sources of uncertainty in precipitation and temperature projections for Denmark. Geol. Surv. Den. Greenl. Bull. https://doi.org/10.34194/GEUSB-201943-01-02.
Pastén-Zapata, E., Jones, J.M., Moggridge, H., Widmann, M., 2020. Evaluation of the performance of Euro-CORDEX regional climate models for assessing hydrological climate change impacts in Great Britain: a comparison of different spatial resolutions and quantile mapping bias correction methods. J. Hydrol. 584, 124653 https://doi.org/10.1016/j.jhydrol.2020.124653.
Pechlivanidis, I.G., Gupta, H., Bosshard, T., 2018. An information theory approach to identifying a representative subset of hydro-climatic simulations for impact modeling studies. Water Resour. Res. 54 (8), 5422–5435. https://doi.org/10.1029/2017WR022035.
Penman, H.L., 1948. Natural evaporation from open water, baresoil and grass. Proc. R. Soc. Lond. 193, 120–145. https://doi.org/10.1098/rspa.1948.0037.
Pérez-Palazón, M.J., Pimentel, R., Polo, M.J., Pérez-Palazón, M.J., Pimentel, R., Polo, M.J., 2018. Climate trends impact on the snowfall regime in Mediterranean mountain areas: future scenario assessment in sierra nevada (Spain). Water 10 (6), 720. https://doi.org/10.3390/w10060720.
Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. J. Hydrol. 279 (1–4), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7.
Photiadou, C., Arheimer, B., Bosshard, T., Capell, R., Elenius, M., Gallo, I., Gyllensvärd, F., Klehmet, K., Little, L., Ribeiro, I., Santos, L., Sjökvist, E., 2021. Designing a climate service for planning climate actions in vulnerable countries. Atmosphere 12, 121. https://doi.org/10.3390/atmos12010121.
Pimentel, R., Herrero, J., Polo, M.J., 2017a. Subgrid parameterization of snow distribution at a Mediterranean site using terrestrial photography. Hydrol. Earth Syst. Sci. 21, 805–820. https://doi.org/10.5194/hess-21-805-2017.
Pimentel, R., Herrero, J., Polo, M.J., 2017b. Quantifying snow cover distribution in semiarid regions combining satellite and terrestrial imagery. Remote Sens 2017 (9), 995. https://doi.org/10.3390/rs9100995.
Polo, M.J., Herrero, J., Aguilar, C., Millares, A., Moñino, A., Nieto, S., Losada, M.A., 2010. WiMMed, a distributed physically-based watershed model (I): description and validation. Environ. Hydraul.: Theor. Exp. Comput. Solut. 225–228.
Prudhomme, C., Haxon, T., Crooks, S., Jackson, C., Barkwith, A., Williamson, J., Kelvin, J., Mackay, J., Wang, L., Young, A., Watts, G., 2013. Future flows hydrology: an ensemble of daily river flow and monthly groundwater levels for use for climate change impact assessment across great Britain. Earth Syst. Sci. Data 5, 101–107 doi:10-5194/essd-5-101-2013.
Prudhomme, C., Giuntoli, I., Robinson, E.L., Clark, D.B., Arnell, N.W., Dankers, R., Hagemann, S., 2014. Hydrological droughts in the 21st century, hotspots and uncertainties from a global multimodel ensemble experiment. Proc. Natl. Acad. Sci. 111 (9), 3262–3267. https://doi.org/10.1073/pnas.1222473110.
Refsgaard, J.C., Knudsen, J., 1996. Operational validation and intercomparison of different types of hydrological models. Water Resour. Res. 32 (7), 2189–2202. https://doi.org/10.1029/96WR00896.
Refsgaard, J.C., Christensen, S., Sonnenborg, T.O., Seifert, D., Højberg, A.L., Troldborg, L., 2012. Review of strategies for handling geological uncertainty in groundwater flow and transport modeling. Adv. Water Resour. 36, 36–50. https://doi.org/10.1016/j.advwatres.2011.04.006.
Refsgaard, J.C., Madsen, H., Andréassian, V., Arnbjerg-Nielsen, K., Davidson, T.A., Drews, M., Hamilton, D.P., Jeppesen, E., Kjellström, E., Olesen, J.E., Sonnenborg, V., 2014. A framework for testing the ability of models to project climate change and its impacts. Clim. Change 122 (1–2), 271–282. https://doi.org/10.1007/s10584-013-0990-2.
Refsgaard, J.C., Sonnenborg, T.O., Butts, M.B., Christensen, J.H., Christensen, S., Drews, M., Jørgensen, L.F., Larsen, M.A.D., Rasmussen, S.H., Seaby, L.P., Seifert, D., Jørgensen, F., Vilhelmsen, T.N., 2016. Climate change impacts on groundwater hydrology–where are the main uncertainties and can they be reduced? Hydrol. Sci. J. 61 (13), 2312–2324. https://doi.org/10.1080/02626667.2015.1131899.
Rojas, R., Feyen, L., Dosio, A., Bavera, D., 2011. Improving pan-European hydrological simulation of extreme events through statistical bias correction of RCM-driven climate simulations. Hydrol. Earth Syst. Sci. https://doi.org/10.5194/hess-15-2599-2011.
Samuelsson, P., Jones, C.G., Willén, U., Ullersting, A., Gollvik, S., Hansson, U., Jansson, C., Kjellström, E., Nikulin, G., Wyser, K., 2011. The rossby centre regional climate model RCA3: model descriptor and performance. Tellus A 63 (1), 4–23. https://doi.org/10.1111/j.1600-0870.2010.00478.x.
Schaefli, B., Gupta, H.V., 2007. Do Nash values have value? Hydrol. Process. 21, 2075–2080. https://doi.org/10.1002/hyp.6825.
Scharling, 2012. Climate Grid Denmark. Dataset for use in research and education. Danish Meteorological Institute Technical Report 12–10. ISSN: 1399–1388.
Seaby, L.P., Refsgaard, J.C., Sonnenborg, T.O., Stisen, S., Christensen, J.H., Jensen, K.H., 2013. Assessment of robustness and significance of climate change signals for an ensemble of distribution-based scaled climate projections. J. Hydrol. 486, 479–493. https://doi.org/10.1016/j.jhydrol.2013.02.015.
Seiller, G., Anctil, F., Perrin, C., 2012. Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. Hydrol. Earth Syst. Sci. 16 (4), 1171–1189. https://doi.org/10.5194/hess-16-1171-2012.
Soares, M.B., Alexander, M., Dessai, S., 2018. Sectoral use of climate information in Europe: a synoptic overview. Clim. Serv. 9, 5–20. https://doi.org/10.1016/j.cliser.2017.06.001.
Stephens, C.M., Marshall, L.A., Johnson, F.M., Lin, L., Band, L.E., Ajami, H., 2020. Is past variability a suitable proxy for future change? a virtual catchment experiment. Water Resour. Res. 56 (2) https://doi.org/10.1029/2019WR026275.
Stisen, S., Ondracek, M., Troldborg, L., Schneider, R.J. M., & van Til, M.J. (2019a). National Vandressource Model. Modelopstilling og kalibrering af DK-model 2019. Danmarks Og Grønlands Geologiske Undersøgelse Rapport, 31.
Stisen S., Ondracek, M., Troldborg, L., Schneider, R.J. M. and van Til, M.J. (2019b). National Vandressource Model, Modelopstilling og kalibrering af DK-model 2019. Geological Survey of Denmark and Greenland, Report 2019/31. In Danish.
Sun, F., Roderick, M.L., Lim, W.H., Farquhar, G.D., 2011. Hydroclimatic projections for the Murray–Darling Basin based on an ensemble derived from intergovernmental panel on climate change AR4 climate models. Water Resour. Res. 47, W00G02. https://doi.org/10.1029/2010WR009829.
Teutschbein, C., Grabs, T., Laudon, H., Karlsen, R.H., Bishop, K., 2018. Simulating streamflow in ungauged basins under a changing climate: the importance of landscape characteristics. J. Hydrol. 561, 160–178. https://doi.org/10.1016/j.jhydrol.2018.03.060.
Thirel, G., Andréassian, V., Perrin, C., 2015a. On the need to test hydrological models under changing conditions. Hydrol. Sci. J. 60 (7–8), 1165–1173. https://doi.org/10.1080/02626667.2015.1050027.

Thirel, G., Andréassian, V., Perrin, C., Audouy, J.N., Berthet, L., Edwards, P., Folton, N., Furusho, C., Kuentz, A., Lerat, J., Lindström, G., 2015b. Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments. Hydrol. Sci. J. 60 (7–8), 1184–1199. https://doi.org/10.1080/02626667.2014.967248.

Valéry, A., Andréassian, V., Perrin, C., 2014. 'As simple as possible but not simpler': what is useful in a temperature-based snow-accounting routine? part 2–sensitivity analysis of the cemaneige snow accounting routine on 380 catchments. J. Hydrol. 517, 1176–1187. https://doi.org/10.1016/j.jhydrol.2014.04.058.

van Vliet, M.T., Donnelly, C., Strömbäck, L., Capell, R., Ludwig, F., 2015. European scale climate information services for water use sectors. J. Hydrol. 528, 503–513. https://doi.org/10.1016/j.jhydrol.2015.06.060.

Van Genuchten, M.T., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil science society of America journal 44 (5), 892–898. https://doi.org/10.2136/sssaj1980.03615995004400050002x.

Velázquez, J.A., Schmid, J., Ricard, S., Muerth, M.J., St-Denis, B.G., Minville, M., Turcotte, R., 2013. An ensemble approach to assess hydrological models' contribution to uncertainties in the analysis of climate change impact on water resources. Hydrol. Earth Syst. Sci. 17 (1) https://doi.org/10.5194/hess-17-565-2013.

Vidal, J.P., Martin, E., Franchistéguy, L., Baillon, M., Soubeyroux, J.M., 2010. A 50-year high-resolution atmospheric reanalysis over France with the safran system. Int. J. Climatol. 30 (11), 1627–1644. https://doi.org/10.1002/joc.2003.

Vidal, J.P., Hingray, B., Magand, C., Sauquet, E., Ducharne, A., 2016. Hierarchy of climate and hydrological uncertainties in transient low-flow projections. Hydrol. Earth Syst. Sci. 20 (9), 3651–3672. https://doi.org/10.5194/hess-20-3651-2016.

Vrac, M., Noël, T., Vautard, R., 2016. Bias correction of precipitation through singularity stochastic removal: because occurrences matter. J. Geophys. Res.: Atmos. 121 (10), 5237–5258. https://doi.org/10.1002/2015JD024511.

Wagner, T., Themeßl, M., Schüppel, A., Gobiet, A., Stigler, H., Birk, S., 2017. Impacts of climate change on stream flow and hydro power generation in the Alpine region. Environ. Earth Sci. 76 (1), 4. https://doi.org/10.1007/s12665-016-6318-6.

Wang, H.M., Chen, J., Xu, C.Y., Chen, H., Guo, S., Xie, P., Li, X., 2019. Does the weighting of climate simulations result in a better quantification of hydrological impacts? Hydrol. Earth Syst. Sci. 23 (10), 4033–4050. https://doi.org/10.5194/hess-23-4033-2019.

Weichselgartner, J., Arheimer, B., 2019. Evolving climate services into knowledge–action systems. Weather, Clim., Soc. 11 (2), 385–399. https://doi.org/10.1175/WCAS-D-18-0087.1.

Yan, J.J., Smith, K.R., 1994. Simulation of integrated surface water and ground water systems – model formulation. Water Resour. Bull. 30 (5), 1–12. https://doi.org/10.1111/j.1752-1688.1994.tb03336.x.