



Effect of model calibration strategy on climate projections of hydrological indicators at a continental scale

Yeshewatesfa Hundecha¹ · Berit Arheimer¹ · Peter Berg¹ · René Capell¹ · Jude Musuza¹ · Ilias Pechlivanidis¹ · Christiana Photiadou¹

Received: 4 February 2020 / Accepted: 30 August 2020 / Published online: 02 October 2020
© The Author(s) 2020

Abstract

The effect of model calibration on the projection of climate change impact on hydrological indicators was assessed by employing variants of a pan-European hydrological model driven by forcing data from an ensemble of climate models. The hydrological model was calibrated using three approaches: calibration at the outlets of major river basins, regionalization through calibration of smaller scale catchments with unique catchment characteristics, and building a model ensemble by sampling model parameters from the regionalized model. The large-scale patterns of the change signals projected by all model variants were found to be similar for the different indicators. Catchment scale differences were observed between the projections of the model calibrated for the major river basins and the other two model variants. The distributions of the median change signals projected by the ensemble model were found to be similar to the distributions of the change signals projected by the regionalized model for all hydrological indicators. The study highlights that the spatial detail to which model calibration is performed can highly influence the catchment scale detail in the projection of climate change impact on hydrological indicators, with an absolute difference in the projections of the locally calibrated model and the model calibrated for the major river basins ranging between 0 and 55% for mean annual discharge, while it has little effect on the large-scale pattern of the projection.

Keywords Model calibration · HYPE · Climate change · Europe · Ensemble modeling

This article is part of a Special Issue on “How evaluation of hydrological models influences results of climate impact assessment”, edited by Valentina Krysanova, Fred Hattermann and Zbigniew Kundzewicz.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s10584-020-02874-4>) contains supplementary material, which is available to authorized users.

✉ Yeshewatesfa Hundecha
yeshewatesfa.hundecha@smhi.se

Extended author information available on the last page of the article

1 Introduction

In the recent past, a large number of hydrological studies have focused on assessing the impact of climate change on hydrological variables that are considered crucial for different water sectors (e.g., Arheimer et al. 2017; Forzieri et al. 2014; Pechlivanidis et al. 2017). This has typically been done by employing hydrological models that operate at a global or continental scale (e.g., Dankers et al. 2014; Donnelly et al. 2017; Hagemann et al. 2013; Prudhomme et al. 2014), large river systems (Arheimer and Lindström 2015; Krysanova et al. 2017), or catchment scales (Vormoor et al. 2015; Pechlivanidis et al. 2018). Despite their convenience of application at very large scales, continental or global scale models are often coarse in their spatial resolution and may not represent local details of the hydrological system adequately (Dankers et al. 2014). Regional hydrological models with higher resolution, on the contrary, may offer an attractive alternative to accounting for local details in the processes relevant to the study at hand (Olsson et al. 2016). However, this can only be achieved if the model is parameterized through appropriate parameter estimation schemes.

Regional hydrological models employed for climate impact studies are often calibrated against observed catchment response, often river discharge. This may be performed in different details depending on the intended purpose of the model's application, availability of data for model calibration, and the involved computational effort. In many applications of regional hydrologic models, calibration is often performed for a river basin as a whole. Although this approach may lead to acceptable model performance in reproducing the aggregated model response from the basin, small-scale variability of certain processes may not be well represented, and the local climate change signal projected by the model may not be robust (Krysanova et al. 2018). In some applications, however, calibration is performed with more detail by allowing all or a subset of the model parameters, such as soil and land use dependent parameters, to vary locally while keeping general parameters constant throughout, and tuning the model against observed catchment response simultaneously at several locations within a basin. The model parameters that vary locally are often linked to catchment physiographic characteristics in order to ensure their physical consistency across the basin (e.g., Hundecha et al. 2016). This may increase the confidence one attributes to the climate change projections made by the model at local scales (Krysanova et al. 2018).

Regional hydrological models are also employed for continental studies (e.g., Donnelly et al. 2017). The models are typically parameterized through calibration at selected large basins as the computational effort can be prohibitively high to allow calibration at all locations where observations are available. Krysanova et al. (2018) compared climate change impact projections made by a model parameterized in such way and a model calibrated with more local detail for a specific catchment. Their results showed that the model calibrated for continental applications missed some important local processes that were captured by the locally calibrated model.

The objective of this paper is to assess the effect of model calibration strategy of a continental regional hydrological model on the projected climate change impact on hydrological indicators. We specifically aim to assess the differences in the climate change projections of the hydrological indicators when the model is calibrated at a set of major basins only, compared with a detailed calibration that accounts for the local variability of the model parameters. We employ a semi-distributed hydrological model setup at a pan-European scale by subdividing the model domain into tens of thousands of subcatchments and perform different model calibrations. We assess the climate change signals projected by the differently

calibrated models by forcing the models with an ensemble of climate model outputs. Furthermore, we assess the sensitivity of the climate change signal when an ensemble of hydrological models is used. This ensemble was generated through sampling of specific parameters of the model calibrated in a detailed way using the Generalized Likelihood Uncertainty Estimation (GLUE) approach (Beven and Binley 1992).

2 Data and method

2.1 The E-HYPE model

The continuous process-based hydrological model HYPE (Lindström et al. 2010), which simulates components of the catchment water cycle and water quality at a daily or hourly time step, is employed. The model is semi-distributed, in which a river basin may be subdivided into multiple subcatchments, which are further subdivided into hydrological response units (HRUs) based on soil type and land use classes. It has conceptual routines for most of the major land surface and subsurface processes (e.g., snow/ice accumulation and melting, evapotranspiration, surface and macropore flow, soil moisture, discharge generation, groundwater fluctuation, aquifer recharge/discharge, irrigation, abstractions and routing through rivers, lakes and reservoirs, solid-matter and dissolved nutrient pools) that are controlled by a number of parameters that are often linked to physiography to account for spatial variability and estimated through calibration.

The model parameters can generally be categorized into HRU and general parameters. The HRU parameters are either soil or land use type dependent. Unique parameter values are estimated for each soil or land use type and are applied throughout the model domain. The general parameters, such as the routing parameters, are subcatchment scale parameters and may be assigned constant values throughout the model domain or estimated separately for different parameter regions. They can also be estimated using a regionalization approach as functions of catchment descriptors (see “A classification-based stepwise calibration” section).

The model was set up for the pan-European domain and is referred to as E-HYPE (Donnelly et al. 2016). The version used in the present study covers an area of 8.8 million km² and is subdivided into 35,408 subcatchments with an average size of 248 km².

2.2 Data

A range of physiographic and land management data, including reservoirs and irrigated areas, were used in setting up the model (Donnelly et al. 2016; Hundecha et al. 2016), and detailed information can be found at <https://hypeweb.smhi.se/explore-water/>. Originally, daily discharge data at more than 3000 stations were obtained from different sources. A subset of the stations was used in this work for model calibration and validation (see Section 2.3.2). Another difference to the previous E-HYPE applications is that we used the observational meteorological forcing from EFAS-Meteo (Ntegeka et al. 2013). The version used here is from May 2019 and was a pre-release for the version that is being ingested into the Copernicus Climate Change Service and its catalog (C3S). Only daily mean surface air temperature and precipitation are used for the current study. The dataset covers most of Europe, with a resolution of 5 km at daily time steps, and the temporal coverage is from 1990 until present.

Compared with other standard open datasets, such as E-OBS (Cornes et al. 2018), EFAS-Meteo was chosen as it is used as a reference dataset for the hydrological models in the operational service for the water sector in C3S, from which model results are openly distributed. E-OBS (version 20.0e) was used to evaluate the quality of EFAS-Meteo.

The state-of-the-art in regional climate projections for Europe is currently the Euro-CORDEX ensemble at 0.11 degree (about 12.5 km) horizontal resolution. A sub-selection of models was done due to computational expenses. The sub-ensemble was chosen to sample different driving global climate models and regional climate models: EC-Earth (r12i1p1) with RACMO22E, EC-Earth (r12i1p1) with RCA4, HadGEM2-ES with RACMO22E (v2), HadGEM2-ES with RCA4, and MPI-ESM-LR (r1i1p1) with RCA4 (v1a). Each model was bias-adjusted towards the EFAS-Meteo reference dataset, separately for temperature and precipitation, using a method described in Text S1 in the supplementary material. The analysis is restricted to the RCP8.5 greenhouse gas concentration scenario, and simulations cover the period 1970–2100.

Monthly actual evapotranspiration data were obtained for the period 2000–2012 from the Moderate Resolution Imaging Spectroradiometer (MODIS) dataset (Mu et al. 2011). The product is based on a Penman-Monteith approach, and the dataset covers the entire domain at a spatial resolution of 1 km, which is later spatially aggregated to the E-HYPE model resolution.

2.3 Methods

2.3.1 Data quality assessment

Observed discharge time series used for calibration and evaluation of E-HYPE were quality-assessed during the model setup phase. The assessment consisted of evaluation of station metadata and time series quality. The former included cross-checking station positions, upstream areas, and other available metadata such as river names between source databases and E-HYPE setup and manual adjustments of the model setup where necessary. Time series were visually assessed for erroneous data, e.g., interpolated values in observation gaps, and for the influence of upstream regulation, e.g., plateaus in hydrographs or breaks in flow duration curves. Stations with evidence of upstream regulation as well as with less than 3 years of data during the calibration period were excluded.

The reference forcing data (EFAS-Meteo) was evaluated against the E-OBS gridded dataset, which is the current standard open dataset for pan-European applications. The evaluation consisted of comparing different aspects of the temperature and precipitation data using the following indicators: annual mean temperature and precipitation, the number of precipitation events above 10 mm/day, and number of consecutive dry day periods of at least 5 days, with a dry day defined as a day with precipitation less than 1 mm/day.

2.3.2 Calibration and spatiotemporal evaluation of E-HYPE

Three approaches of model calibration were employed in this work, which are described in the following subsections. Calibration was performed for a set of parameters identified in Hundecha et al. (2016) as important (See Table S1).

Calibration using downstream stations in major river basins Model calibration was performed at a set of river flow gauging stations from the major river basins with at least 3 years of data during 1991 to 2010 and located as close to the sea as possible, i.e., not draining into other rivers. The station must have drained at least 70% of the basin as well. The basins had a minimum area of 5000 km², which ensures preservation of sub-grid variability in the forcing data. Fifty-seven stations were selected, but some were excluded because they were very close to others or were outside the EFAS-Meteo domain, which left 37 stations, each from an independent river basin (Fig. 1a, Table S2). Every subcatchment located upstream and downstream of each of the stations was assigned the same parameter set estimated at the station. Subcatchments outside the 37 major basins were assigned the parameter set of the closest calibrated basin.

The SAFE toolbox (Pianosi et al. 2015) was used to generate 15,000 uniformly distributed samples of parameter set (see Table S1) following the Latin hypercube strategy. Identical HRU parameters were applied to all basins, but the general parameters were allowed to vary between basins. Since the HRU parameters are the same in all basins, simultaneous calibration was performed at all 37 stations. The sum of the Nash-Sutcliffe efficiency (NSE; Nash and Sutcliffe 1970) at all stations was employed as a basis of the objective function. Since a poor performance at some stations could be offset by a good performance at others, this could lead to an unbalanced model performance between the stations. The objective function was, therefore, modified in such a way that more emphasis is given to the station where the performance is the poorest (Hundecha and Bárdossy 2004):

$$Obj = \sum_{i=1}^N NSE_i + NminNSE \quad (1)$$

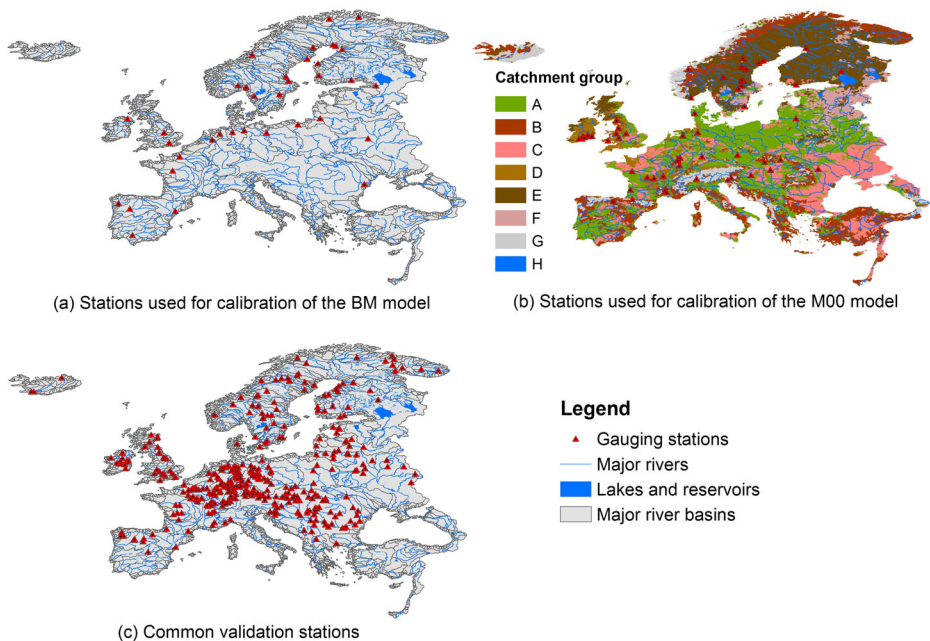


Fig. 1 Stations used for calibration of E-HYPE using different approaches. The catchment groups shown in (b) are those used for regionalization of the M00 model

where N is the number of stations and $\min NSE$ is the minimum of the NSE_i values. The model is herein referred to as the benchmark (BM) model.

A classification-based stepwise calibration This is a more detailed calibration approach, where a catchment classification was performed based on catchment physiographic and climate descriptors and calibration was performed for a set of gauged catchments within each group of catchments. The method is fully described in Hundecha et al. (2016) and is only briefly described here. The meteorological forcing data used in the present study (EFAS-Meteo) does not cover a portion of the eastern part of the E-HYPE domain. Therefore, fewer stations were used for model calibration and validation in this study than in Hundecha et al. (2016).

Catchment classification was first performed by employing a hierarchical minimum-variance clustering method using catchment area, mean catchment elevation and slope, land use classes, soil types, and mean annual catchment precipitation and temperature as catchment physiographic and climate descriptors. This resulted in 16 groups of catchments. Gauging stations whose upstream subcatchments are entirely contained within each group were identified for model calibration. A minimum drainage area of 1000 km² was set as a requirement for the stations since the forcing data were estimated from grids and the sub-grid variability may not be well represented for smaller catchments. Fifty-eight stations were identified, which are different from the 37 stations used to calibrate the BM model (Fig. 1b; Table S2).

The HRU parameters were estimated by calibrating the model simultaneously at the selected gauged catchments in each catchment group at a time and repeating the procedure to the other groups iteratively. Each group of catchments was found to have one or two dominant land use and soil types. For each group, parameters of the dominant soil and land use types were calibrated while setting arbitrary values within the parameter range for the other soil and land use types. The HRU parameters estimated in one group of catchments were then transferred to the next, and parameters of the dominant soil and land use types in the new group were estimated in a similar way as in the previous group. The procedure was repeated until all groups are covered. During this stage of calibration, the general parameters were set constant for all catchments and were manually tuned in such a way that certain features of the simulated hydrograph, such as time shift, which are modeled using the routing parameters, are corrected. The whole procedure was repeated a few times until the overall model performance showed little change.

In the second step of model calibration, the general parameters, which were set constant throughout the model domain during the first step of calibration, were estimated using a regionalization approach. A method described in Hundecha and Bárdossy (2004), where a linear relationship between a model parameter and a set of catchment descriptors is assumed a priori and the coefficients of the linear function are estimated during model calibration, was employed. The parameter estimation was performed separately for each group of catchments. The number of stations in some of the 16 groups was too few to allow a reasonable estimation of the general parameters as functions of catchment descriptors. Therefore, catchment groups with less than 5 stations were merged into groups with similar land use and soil types that have more than 5 stations. This resulted in 8 groups (see Hundecha et al. 2016). The HRU parameters were not adjusted in this step.

Since the calibration at each step was performed simultaneously at multiple stations, the objective function given in Eq. 1 was employed. An automatic calibration based on the Differential Evolution Markov Chain (DE-MC) algorithm (Ter Braak 2006) was employed, together with manual tuning. This model is referred to as M00.

A multi-model ensemble In order to account for the uncertainty in the estimated model parameters, a multi-model ensemble was generated. This was performed by first identifying the most sensitive parameters for the model calibrated using the procedure described in “A classification-based stepwise calibration” section. The dominant soil and land use parameters corresponding to the eight catchment groups, along with the general parameters (Table S1), were sampled using the SAFE sensitivity and uncertainty analysis toolbox (Pianosi et al. 2015) to generate 10,000 randomly distributed samples with the Latin hypercube strategy. The model performance was assessed in terms of NSE.

Parameter samples whose model performance was above a subjectively defined threshold (top 1% of the 10,000 samples) were accepted based on a GLUE-type parameter identification approach (Beven and Binley 1992). In the next step, the top 1% samples were combined across the eight groups to provide narrower identifiable parameter ranges, and 15,000 sets were sampled within these ranges with the same strategy as before. These new global sets were then evaluated against all calibration stations, and parameter sets were chosen as in the previous step, providing a number of acceptable global parameter sets. Finally, the best 10 sets were selected as the multi-model ensemble, herein referred to as M01-M10.

Model evaluation Temporal and spatial model evaluations were performed for all three model variants. Temporal validation was performed by employing the standard split-sampling technique. Model calibration was performed for the period 1991–2000 and temporal validation was performed for 2001–2013. Spatial validation was performed using a set of independent validation stations that were not used for model calibration. Three hundred eighty-eight common validation stations were selected for evaluation of all models (Fig. 1c). Two model performance measures, NSE and percentage model bias (Pbias), were used for both temporal and spatial model evaluation.

The M00 model and the multi-model ensemble were further evaluated in terms of their ability to reproduce certain features of the daily hydrograph. A set of flow signatures computed from the observed and simulated daily flows for the entire modeling period (1991–2013) were compared. Mean specific daily flow (Qmean) and specific daily flows that exceeded 5% (Q95) and 95% (Q05) of the total number of days were used to characterize the overall daily mean and high and low flows, respectively. The coefficient of variation of the daily flow (CV) was also used to characterize the daily flow variability.

The models’ performances in reproducing trends in the annual values of the mean and extreme flow signatures were also evaluated. The distribution-free rank-based Mann-Kendall trend test (Kendall 1975) was employed. As the maximum length of the annual values is 24 years due to the forcing data period, the distributional assumption of the test statistic may not hold for the assessment of the test significance level. Therefore, a resampling technique based on permutation (Good 1994) was employed to compute the test significance level. Trends were considered significant at 5% level.

Furthermore, model state variables other than discharge were used to evaluate the models’ performance. Water quality variables total nitrogen (TN) and total phosphorus (TP) were used for evaluation. For both variables, observed time series of in-stream concentrations were available at sampling sites across the model domain, typically with biweekly to monthly frequencies. Moreover, nutrient load estimates were calculated using modeled discharge at the sampling sites. Model performance was evaluated for the M00 and ensemble models at 509 (TN) and 718 (TP) sampling sites using Pbias in TN and TP concentrations and long-term average load differences between the ensemble members and the M00 model. Performance

averages and spreads across the model domain were used to accept or reject prior model ensemble members, which were selected based on discharge performance alone. The evaluation was conducted subjectively based on relative performance differences between proposed ensemble members.

Similarly, modeled monthly actual evapotranspiration was evaluated against the corresponding MODIS dataset in the period 2000–2012 using Pbias and correlation coefficient (CC). We hypothesize that the MODIS evapotranspiration dataset is of adequate quality both in space and time and can therefore be used as a reference to evaluate the hydrological models.

Krysanova et al. (2018) suggest that regional hydrological models used for climate change impact assessment be evaluated using a 5-step procedure: assessing observational data quality; performing model calibration and validation over periods of contrasting climate; validating at multiple sites within a basin and for multiple variables; validating in terms of reproducing hydrological indicators of interest; and in terms of reproducing trends or lack thereof in hydrological indicators. The model evaluation strategy presented above and the data quality assessment presented in Section 2.3.1 were partly designed based on this suggestion. We followed all steps except model calibration and validation over periods of contrasting climate. As running E-HYPE for different periods of contrasting climate identified for a large number of calibration and validation stations is computationally cumbersome, we did calibration and validation over a common period for all stations.

2.3.3 Climate change impact assessment

All model variants were run using forcing data from a sub-ensemble of the bias adjusted Euro-CORDEX (see Section 2.2) for the period 1970–2100. The following hydrological indicators were computed from the model outputs separately for the periods 1971–2000 and 2071–2100:

- Mean river discharge (m^3/s): defined as the annual or seasonal simulated outflow from a subcatchment
- Maximum river discharge (m^3/s): defined as the annual daily maximum discharge
- Mean runoff (mm/month): defined as the annual or seasonal mean daily runoff
- Mean annual or seasonal soil moisture: calculated as the root zone soil moisture as fraction of the maximum water content volume and depends on the soil type (land use)
- Mean annual or seasonal actual aridity: defined as the ratio between actual evapotranspiration and precipitation)

The assessment of climate change impact was performed by calculating first the ensemble mean of the entire climate model-based simulations and then taking the difference between the future and historical time periods in relative terms, i.e., $(\text{future} - \text{historical}) / \text{historical} * 100$.

3 Results and discussion

3.1 Assessment of data quality

As shown in Fig. S1, the EFAS-Meteo temperature is very similar to that of E-OBS, except for some differences with strong topography. This is largely a resolution effect, since

remapping to a common E-OBS grid did not account for height differences. The precipitation mean shows larger deviations, which is related to the strong spatial heterogeneity of this variable which requires high station network density for reliable maps. Deviations are therefore likely related to both the underlying station network and the interpolation method. The most notable deviations are in north-west UK-Ireland and Scandinavia and generally in regions of strong topographic variations. It is difficult to determine which dataset is more correct, and more in-depth analysis comparing with gauge data would be required. However, for large parts of the continent, the differences are within $\pm 5\%$, which is a comforting result. In Eastern Europe, one can see “spots,” which are traced back to EFAS-Meteo, which seems not to be able to spread the sparse gauge information there. This is, however, outside the hydrological simulation domain and is therefore not accounted for in the quality assessment. Extreme precipitation has similar characteristics as the mean precipitation, and here one can also see political borders (e.g., Germany), which indicate large gradients in the underlying network density.

Consecutive dry day periods longer than 5 days are quite similar within mostly 5% but with some very strong deviations in Turkey and Greece. It looks like E-OBS has some issues here, with clear extrapolations in the southern and eastern Mediterranean countries. In conclusion, we can determine that EFAS-Meteo is a fair alternative to the state-of-the-art dataset E-OBS, however, with some local large differences, for better or worse.

3.2 Model calibration and evaluation

3.2.1 Temporal and spatial model evaluation

Generally, the difference in the model performance at the stations used for model calibration in the calibration period (1991–2000) between all models is not very big (Fig. 2). However, one can see that the performance of the benchmark (BM) model is slightly worse than the other models, both in terms of NSE and Pbias, while the model calibrated in a more detailed way (M00) is similar to the ensemble models (M01–M10) with slightly better performance. The median NSE at the calibration stations for BM and M00 are 0.39 and 0.59, respectively, while it ranges between 0.38 and 0.5 for the ensemble models. Based on evaluation of the ensemble models against water quality variables, three members were excluded from the ensemble (see Section 3.2.4). The excluded models are marked in Fig. 2. For the remaining model ensemble, the median NSE ranges between 0.44 and 0.5. In terms of Pbias, all models underestimate the mean flow with median values of -17% and -7% for BM and M00, respectively, and between -10 and -7% for the ensemble models. It should, however, be noted that the calibration stations used for BM are different from the stations used to calibrate the other models. Spatially, the model performance shows deterioration from north-west towards south-east across the model domain for all model variants (Fig. 3).

The model performance in terms of NSE at the calibration stations is generally slightly better in the validation period (2001–2013) than in the calibration period for all models, with median NSE of 0.43 and 0.58, respectively, for BM and M00 and ranging between 0.49 and 0.54 for members of the model ensemble. The model bias also shows a slight improvement for many of the models but is not as clear as the performance in terms of NSE.

The performance in terms of NSE at the independent validation stations in the model calibration period is generally similar to the corresponding performance at the calibration stations for the ensemble models, with slightly more spread. The median NSE is slightly

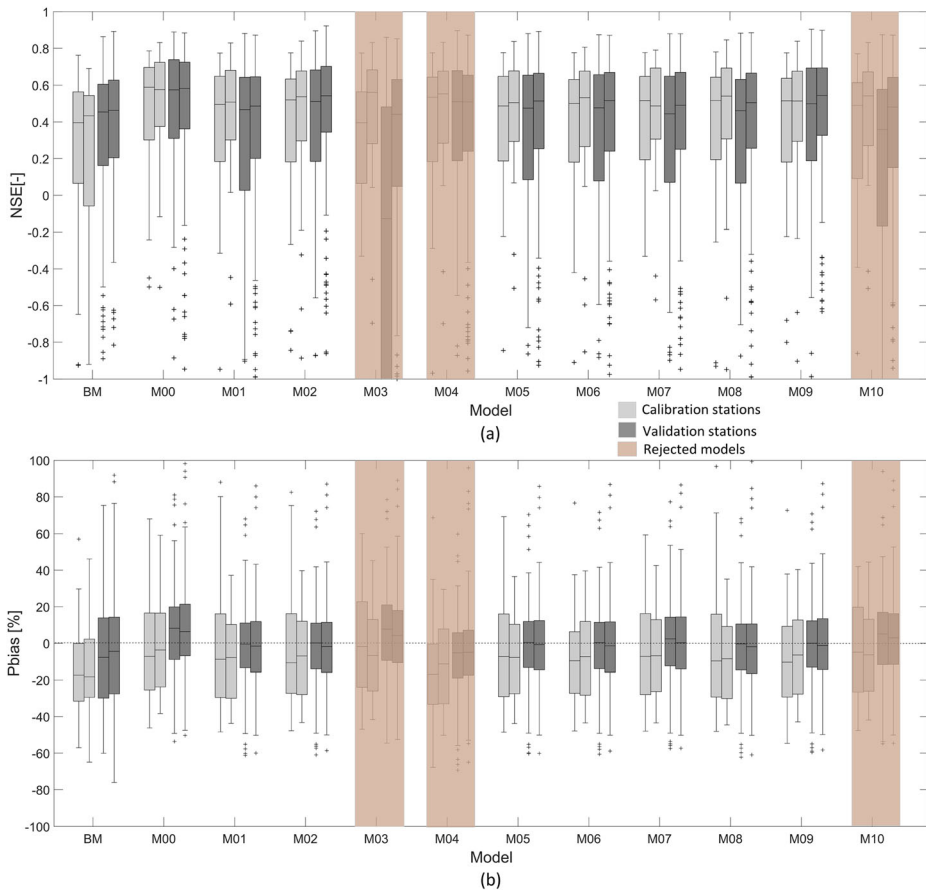


Fig. 2 Distributions of model performance in terms of (a) NSE and (b) Pbias for the calibration and validation stations during the calibration and validation periods (For each group of stations, left box shows calibration period and right box shows validation period)

higher in the validation stations for BM, while it is similar in both the validation and calibration stations for M00. Similar to the calibration stations, the median NSE at the validation stations is slightly better for M00 than for the others. In terms of the model bias, BM generally underestimates the flow at the validation stations with lower median model bias than at the calibration stations but with more spread among stations. M00, on the other hand, slightly overestimates the flow at the validation stations with less spread than in the calibration stations. For the ensemble models, the model bias at the validation stations is very low with less spread than in the calibration stations.

While the model performance in terms of NSE at the validation stations is slightly better in the validation period than in the calibration period for all models, the distribution of the model bias is similar in both periods for all models. Figure S2 also shows comparison of the model performance between BM and M00 at the validation stations by grouping them into the eight catchment groups. Generally, M00 performs better with less spread both in terms of NSE and bias except in one group (group G). This group is characterized by shallow soil in high elevation areas with high mean annual precipitation.

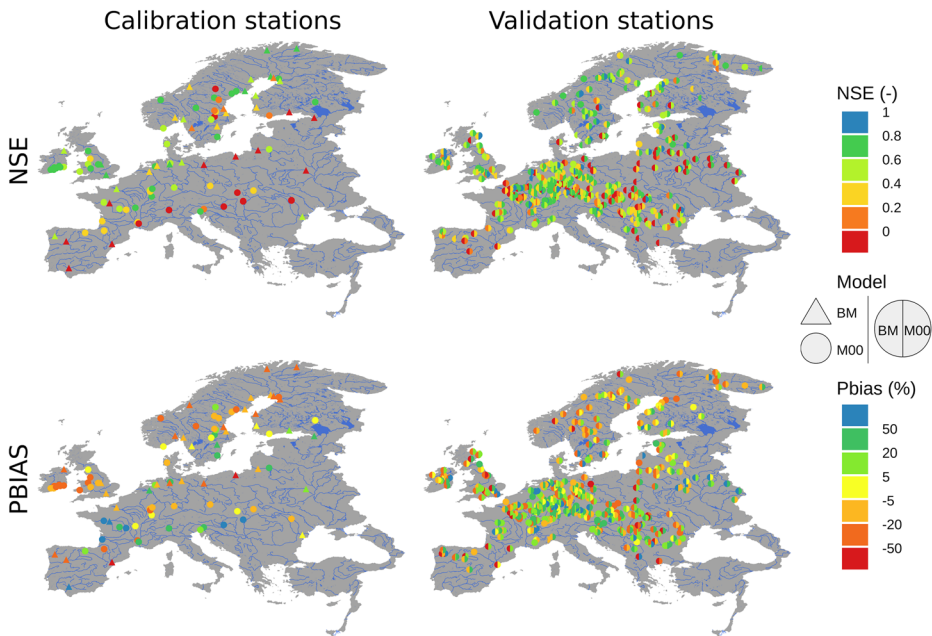


Fig. 3 Spatial pattern of performance of the BM and M00 models in terms of NSE and Pbias

3.2.2 Model performance in terms of flow signatures

M00 and the ensemble models generally reproduce the mean flow at the stations well, as shown in Fig. 4. There is a tendency for all models to underestimate the flow at a few stations with very high specific flow. Variability of the daily flow is reproduced fairly well by M00 with the points scattering well around the 1:1 line (Fig. 4). The ensemble models, however, overestimate the variability. M00 estimates the lower magnitudes of the high flow (Q95) well, while it underestimates the higher magnitudes. The ensemble models also underestimate the higher magnitudes but slightly overestimate the lower magnitudes. M00 generally underestimates the higher magnitudes of the low flow (Q05), while the ensemble models underestimate the low flow over the whole flow spectrum. This reflects the higher variability of the daily flow estimated by the ensemble models.

3.2.3 Trends in annual flow signatures

Trends in the annual mean and extreme daily flow signatures were assessed at 180 stations with at least 20 years of continuous flow data. Significant trends were detected at a small fraction of the stations for all the investigated signatures (Table 1). Significant trend at 5% level was detected at 8.3% of the stations for the annual mean flow for the observed discharge time series, with more positive than negative trends. For the modeled data, significant trend was detected at slightly more stations for both M00 and the ensemble models. However, the trends match between the observed and modeled mean annual flows at only around 1.5% of the stations. The trends are generally similar for Q05 and Q95 as well, with significant trends in Q05 at more stations. The number of stations where the trends in the observed and modeled flow signatures match is also the highest for Q05. No significant trend was detected at the

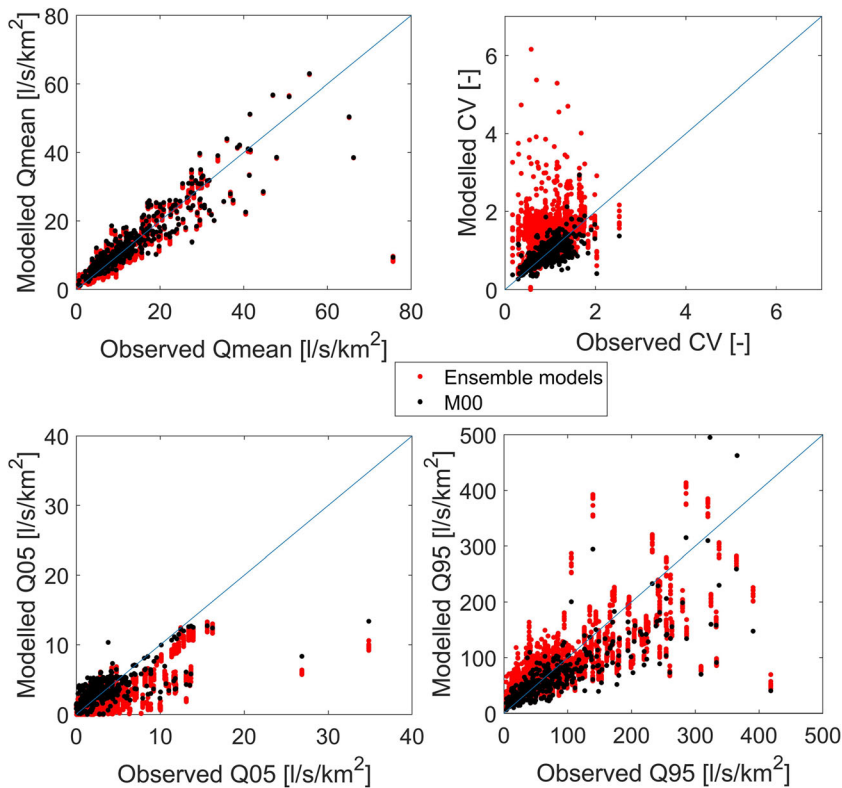


Fig. 4 Scatter plots of observed vs simulated flow signatures for the M00 and ensemble models

majority of the stations for all modeled and observed flow signatures, and the percentage of matching non-trend stations is also high, ranging between 67 and 82% for the different models and flow signatures (Table 1). Similarly, the percentage of stations with matching significant trend and no trend ranges between 69.3 and 83.8%, suggesting that both M00 and the model ensemble perform well in reproducing trends or lack thereof in the flow signatures.

3.2.4 Model evaluation against additional variables

The ensemble models (M01-M10) showed a generally positive bias in the in-stream nutrient concentrations with a wider range of variability across the model domain, while M00 underestimated the concentrations with less variability (Fig. 5a). This systematic difference is a reflection of the discharge bias differences shown in Fig. 2, as larger discharge volumes have a diluting effect on modeled nutrient concentrations. Nutrient loads, i.e., the amount of nutrient transported with stream flow, provide complementary information for evaluating nutrient model performance (Fig. 5b). Load differences compared with the reference model M00 are comparatively large for nitrogen in three members (M03, M04, M10) and phosphorus in one member (M04). Based on this evaluation, these members were considered inadequate for further implementation and were removed from the model ensemble. For M03, poor nutrient performance coincides with a notable performance drop in discharge NSE at the validation stations, whereas for M04 and M10, nutrients added new, albeit subjective, constraints on ensemble members.

Table 1 Percentage of the total number of stations with at least 20 years of continuous data where simulated and observed annual flow signatures show significant (5% level) positive and negative trends and where trends and non-trends of the observed and simulated flow signatures match

Flow signature	Positive observed	Positive M00	Positive ensemble	Negative observed	Negative M00	Negative ensemble	Matching trend M00	Matching trend ensemble	Matching non-trend M00	Matching non-trend ensemble	Matching trend and non-trend M00	Matching trend and non-trend ensemble
Qmean	6.1	10.0	7.8–10.0	2.2	2.8	3.3–5.6	1.7	1.1–1.7	81.0	79.3–82.1	82.7	80.4–83.8
Q05	12.8	13.9	13.3–18.3	3.3	5.0	3.9–5.6	6.7	6.7–7.2	71.3	68.0–73.0	78.0	75.3–80.3
Q95	3.9	10.6	7.2–11.1	3.9	8.9	7.8–19.4	2.2	1.1–2.8	74.9	67.0–74.9	77.1	69.3–77.1

Evaluation of M00 and the ensemble models against the MODIS actual evapotranspiration shows that the models are capable of representing the long-term actual evapotranspiration and its temporal dynamics (Fig. S3). All models tend to underestimate actual evapotranspiration, with the median values varying between -10 and 0% . However, the models can adequately represent the actual evapotranspiration temporal dynamics, with the median correlation coefficient above 0.9 for all models. In particular, the models tend to underestimate actual evapotranspiration in northern Europe and overestimate in Eastern Europe and the Mediterranean. The dynamics are well represented over the entire domain, with deterioration of performance (i.e., correlation coefficient) in the Mediterranean region (see Fig. S4).

3.3 Comparison of projection of climate impact

Differences in the climate change impacts were investigated across the differently calibrated models. Table 2 gives an overview of the results for all hydrological indicators on an annual basis, supplemented by Table S3 with results for each season. M00 and the median result of the model ensemble ($\langle\text{MXX}\rangle$) are similar for all indicators, except for the slightly higher change in annual maximum discharge for $\langle\text{MXX}\rangle$. However, results for BM differ distinctly from that of M00 and $\langle\text{MXX}\rangle$ for different indicators. The median changes projected for mean annual discharge and runoff by BM are close to zero, but the distributions are biased towards negative changes with the 5th and 95th percentile changes around -50% and 25% , respectively. M00, on the other hand, projected a moderate increase (median values of 8% and

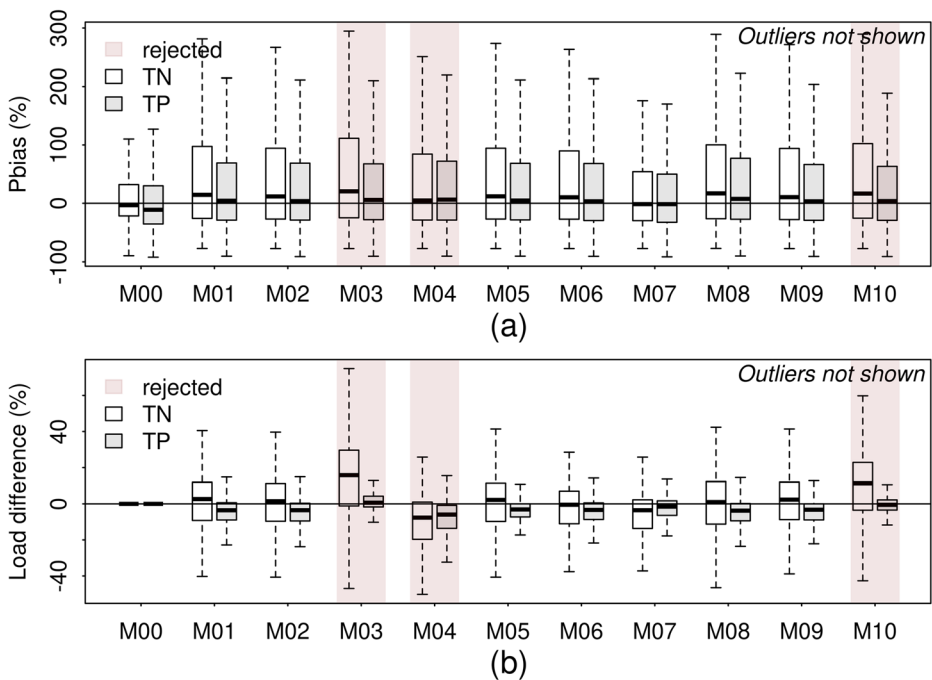


Fig. 5 Performance of model M00 and the ensemble models (M01-M10) when evaluated using total nitrogen (TN) and total phosphorus (TP). Distributions of performances at all available sites as (a) percentage bias in concentrations and (b) as difference in load estimates (using modeled discharge) compared with M00 load estimates

6% for annual runoff and discharge, respectively), with the 95th percentile changes comparable with that of BM. However, the 5th percentile change is around -40% , suggesting that BM projects a drier pattern. This is further confirmed by the projected changes in soil moisture. A major portion of the distribution of the change projected for soil moisture by BM is negative, with the 95th percentile change around 0% . The median and 5th percentile changes projected by M00 are comparable with that of BM, but the 95th percentile change is considerably higher (10% , see also Table 2). Both BM and M00 projected an increase in aridity with similar ranges of change. The median change projected by BM is, however, slightly higher than that of M00 (9% versus 5%). Both BM and M00 projected comparable increase in the annual maximum discharge with a median change of 11% (See Table 2).

Seasonally, all models projected a strong increase in runoff and discharge in winter (median increase of between 30 and 43%), with M00 projecting slightly more increase than BM (Table S3). They projected a moderate decrease in summer with comparable median values between the models and stronger decrease in discharge than runoff (around 20% versus -11%). They all projected a strong and comparable decrease in soil moisture in summer, with median values between -23 and -29% . BM projected a slight decrease, while M00 and <MXX> projected a slight increase in winter (See Table S3). Aridity increases in all seasons, but the change is the strongest in winter, with median increases of around 250% by all models. In summer, the median increase projected by BM is twice as much as that of M00 (23% versus 12%).

As shown in Fig. 6, the main patterns of the change signals in annual mean discharge projected by M00 and BM have the same overall picture of a wetter north and drier south. This is mainly due to a strong winter increase in the north (Figs. S9, S21), which is to a less extent offset by a decrease in summer, except for central Europe (Figs. S11, S23). The differences between the models are generally second order to this but can have large impacts on particular catchments and also seasonal dependency (see Figs. S9–S24). The summer changes in

Table 2 Overview of ensemble mean relative changes between the end of century (2071–2100) and the historical period (1971–2000) ($(\text{future} - \text{historical})/\text{historical} \times 100$), for different indicators for the three main model setups

[%]		BM	M00	<MXX>
Runoff	P5	-46.7	-32.8	-36.3
	P50	1.1	8.0	7.7
	P95	28.1	29.4	30.8
Discharge	P5	-50.9	-37.9	-40.4
	P50	-0.6	6.0	5.6
	P95	26.0	28.2	28.5
Aridity-actual	P5	0.2	-0.5	-0.4
	P50	9.1	5.2	5.7
	P95	36.7	32.8	34.3
Aridity-potential	P5	11.8	11.7	11.7
	P50	27.4	27.5	27.5
	P95	68.0	67.6	67.6
Soil moisture	P5	-30.3	-30.4	-30.5
	P50	-9.4	-8.3	-8.6
	P95	0.4	10.3	12.5
Ann. max discharge	P5	-26.7	-26.7	-24.6
	P50	10.5	11.1	15.9
	P95	52.4	56.9	63.0

The multi-model ensemble is shown as the median of the seven members. The 5th, 50th and 95th percentiles of the spatial patterns are presented. All results are for RCP8.5

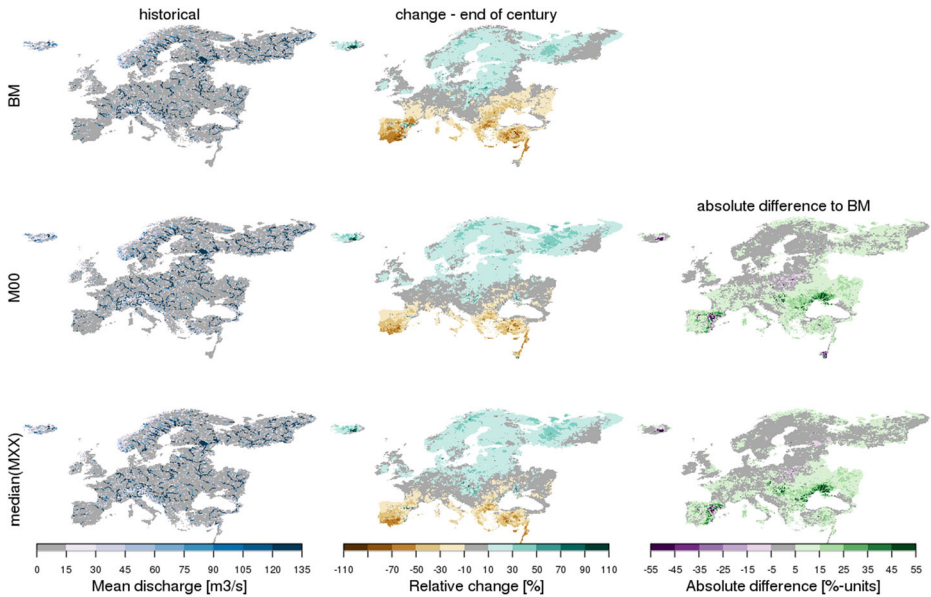


Fig. 6 Annual mean discharge for the BM (top row), M00 (middle row), and median of the multi-model ensemble (bottom row) for the historical period 1971–2000 (left column), relative change by the end of the century 2071–2100 under RCP8.5 (middle column), and differences between the relative changes of M00 and MXX to BM (right column), e.g., the bottom right panel is the absolute difference between the bottom center panel and the top center panel, in percentage units

discharge in central Europe, for instance, are less pronounced in M00 compared with BM. A complete presentation of the absolute values for the historical period, the relative changes by the end of the century, and the difference to BM is shown in the supplementary material for annual mean (Figs. S5–S8) and by season (Figs. S9–24).

The differences in the change signals (Fig. 7) in the mean annual runoff and discharge between M00 and BM show similar patterns with a wetter M00 in most of the domain and especially in southern Europe. Soil moisture shows the opposite pattern but with stronger differences in Scandinavia (Fig. S7). Aridity shows clear boundaries between the sub-areas of BM used for the calibration (Fig. S8). The mean annual maximum discharge shows a distinctly

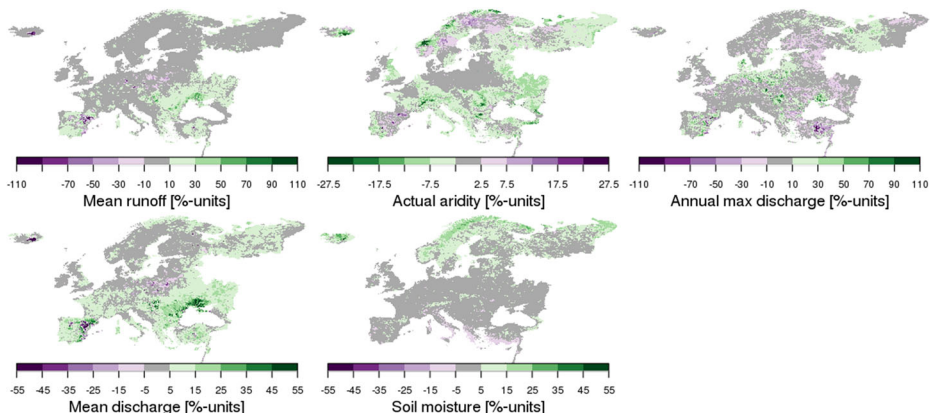


Fig. 7 Difference between the relative changes M00 and BM in percentage units, as in the right column of Fig. 7

different pattern compared with mean discharge but with spatial homogeneities that suggest a model calibration issue, rather than random noise (Fig. S6). Seasonally, the catchment scale variability in the differences between the projections of BM and M00 becomes more apparent for all indicators, as the variability becomes visible even over smaller regions (Figs. S9–S24).

3.4 Discussion

The changes projected in the investigated hydrological indicators under a climate change scenario by the hydrological models calibrated in different ways show similar large-scale spatial patterns. These patterns are generally in agreement with previous European studies on the impact of climate change, indicating a generally wetter north and drier south (e.g., Donnelly et al. 2017; Lobanova et al. 2017; Schneider et al. 2013). For many of the indicators, however, differences appear in the magnitudes of the projected changes between the BM and M00 models that have distinct spatial pattern. BM projects a stronger decrease in runoff in the southern part of Europe in comparison with M00. In some areas of central Eastern Europe and eastern Spain, where the models project increase in runoff, BM projects less increase than M00, as shown in Fig. 6 (right column). This highlights that BM generally projects a drier condition than M00. This general pattern mirrors the BM model's tendency to underestimate the flow more than M00 does (see Section 3.2.1).

The impact of the scale at which model calibration was performed on the climate projection is more apparent when one examines the projections made for the aridity index. Unlike runoff and discharge, the difference in the projections between the different models does not follow a pattern along the direction of the change signal. Whether one of the models projects a stronger change signal than the other appears to be unique to each of the major basins used for the calibration of BM (see Fig. 7). The difference in the model simulations of the index is controlled by the simulated evapotranspiration. The region-dependent evapotranspiration parameter has the same value in all subcatchments of a major basin for BM, while it can be different in each subcatchment for M00 and the ensemble models. Differences in the change signals between subcatchments within a major basin originate from this difference in the evapotranspiration parameter.

The similarity in the distributions of the changes projected by the M00 model and the model ensembles (see Table 2) highlights how model parameter uncertainty propagates into the projection of the hydrological indicators relative to the uncertainty of the climate models. The stations used for model calibration for the M00 and the ensemble models are the same, and the only differences are the model calibration procedures. Ensembles of parameter sets were sampled based on the GLUE approach. Transfer of model parameters to other catchments was performed using the same regionalization scheme as in the M00 model. The result suggests that despite the variation in the model performance between M00 and the model ensemble, as well as among the ensemble members introduced by the differences in the parameter sets, the uncertainty introduced to the climate change projection by the model ensemble is not that significant.

Although it is difficult to state which of the models provide a more credible projection in the climate change signals on the hydrological indicators, it is important to highlight here that M00 results in a more realistic representation of the hydrological processes at a more spatial detail than BM. M00 is likely to provide a physically consistent projection across the model domain due to the incorporation of catchment physiographic features in the estimation of parameters in addition to the finer spatial scale of the catchments used for model calibration. The BM model may miss local variations in certain processes due to the lumped larger scale parameterization of some of the processes (Krysanova et al. 2018). Furthermore, the absence

of variation in the climate impact projection signals between M00 and the ensemble models, which were all parametrized at a similar spatial scale despite the difference in the calibration procedures, increases our confidence in the projection made by M00.

4 Conclusions

The large-scale patterns of the change signals for different hydrological indicators projected at a pan-European scale by a regional hydrological model calibrated with different levels of detail were found to be similar and are generally consistent with the results of several previous European studies. However, the change signal strength varies between the models. For the long-term discharge and runoff, the difference in the change signals appears to be related to the model bias in terms of reproducing the long-term mean runoff. The BM model, which more strongly underestimates the long-term runoff, projects a stronger change signal where the model projects a drier signal and weaker signal where the model projects a wetter signal. For some of the indicators, the differences in the projected signals mirror the spatial differences in the model parameters that control processes that affect the indicators, suggesting the need for a robust model parameter estimation scheme that captures the local variability of the parameters. We can conclude that the calibration strategy can have a large impact at the catchment scale but that the overall pan-European message of the projections remains similar.

The distributions of the median change signals projected by an ensemble of models introduced by sampling parameters using the GLUE approach from the regionalized model were found to be similar to the distributions of change signals projected by the regionalized model for all hydrological indicators. This suggests that the projected climate change signals are less sensitive to the absolute magnitudes of the model parameters than their spatial variability as long as the model performance does not strongly vary between the models. This, however, needs further investigation by setting up a simulation experiment over a wider range of model parameters. Finally, although our results suggest that the hydrological model introduces less uncertainty to the climate impact projection than the climate models, a more comprehensive uncertainty analysis would enable to quantify the contribution of the different components of the model chain.

Acknowledgments We acknowledge the use of the EFAS Meteo dataset, which has been produced as part of the Copernicus Emergency Management Service. We also acknowledge the E-OBS dataset from the EU-FP6 project UERRA (<http://www.uerra.eu>) and the Copernicus Climate Change Service, and the data providers in the ECA&D project (<https://www.ecad.eu>)”.

Funding This work was partly 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), and ANR (FR) with co-funding by the European Commission (grant no. 690462).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Arheimer B, Lindström G (2015) Climate impact on floods – changes of high-flows in Sweden for the past and future (1911–2100). *Hydrol Earth Syst Sci* 19:771–784
- Arheimer B, Donnelly C, Lindström G (2017) Regulation of snow-fed rivers affects flow regimes more than climate change. *Nat Commun* 8(62):1–9. <https://doi.org/10.1038/s41467-017-00092-8>
- Beven K, Binley A (1992) The future of distributed models: model calibration and uncertainty prediction. *Hydrol Process* 6(3):279–298
- Cornes R, van der Schrier G, van den Besselaar EJM, Jones PD (2018) An ensemble version of the E-OBS temperature and precipitation datasets. *J Geophys Res Atmos* 123. <https://doi.org/10.1029/2017JD028200>
- Dankers R, Arnell NW, Clark DB et al (2014) First look at changes in flood hazard in the inter-sectoral impact model intercomparison project ensemble. *Proc Natl Acad Sci U S A* 111:3257–3261
- Donnelly C, Andersson JCM, Arheimer B (2016) Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe. *Hydrol Sci J* 61(2):255–273. <https://doi.org/10.1080/02626667.2015.1027710>
- Donnelly C, Greuell W, Andersson J et al (2017) Impacts of climate change on European hydrology at 1.5, 2 and 3 degrees mean global warming above preindustrial level. *Clim Chang* 143:13–26. <https://doi.org/10.1007/s10584-017-1971-7>
- Forzieri G, Feyen L, Rojas R, Flörke M, Wimmer F, Bianchi A (2014) Ensemble projections of future streamflow droughts in Europe. *Hydrol Earth Syst Sci* 18(1):85–108. <https://doi.org/10.5194/hess-18-85-2014>
- Good P (1994) *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer, New York
- Hagemann S, Chen C, Clark DB 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>
- Hundecha Y, Bárdossy A (2004) Modeling the effect of land use changes on runoff generation of a river basin through parameter regionalization of a watershed model. *J Hydrol* 292:281–295
- Hundecha Y, Arheimer B, Donnelly C, Pechlivanidis I (2016) A regional parameter estimation scheme for a pan-European multi-basin model. *J Hydrol: Regional Studies* 6:90–111. <https://doi.org/10.1016/j.ejrh.2016.04.002>
- Kendall MG (1975) *Rank correlation methods*. Griffin, London
- Krysanova V, Vetter T, Eisner S et al (2017) Intercomparison of regional-scale hydrological models in the present and future climate for 12 large river basins worldwide - a synthesis. *Environ Res Lett* 12:105002. <https://doi.org/10.1088/1748-9326/aa8359>
- Krysanova V, Donnelly C, Gelfan A et al (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>
- 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>
- Lobanova A, Liersch S, Nunes JP et al (2017) Hydrological impacts of moderate and high-end climate change across European river basins. *J Hydrol: Reg Stud* 18:15–30. <https://doi.org/10.1016/j.ejrh.2018.05.003>
- Mu Q, Zhao M, Running SW (2011) Improvements to a MODIS global terrestrial evapotranspiration algorithm. *Remote Sens Environ* 115:1781–1800. <https://doi.org/10.1016/j.rse.2011.02.019>
- Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I – A discussion of principles. *Journal of Hydrology* 10(3):282–290
- Ntegeka V, Salamon P, Gomes G et al (2013) EFAS-Meteo: a European daily high-resolution gridded meteorological data set for 1990–2011. *Report Eur* 26408
- Olsson J, Arheimer B, Borris M et al (2016) Hydrological climate change impact assessment at small and large scales: key messages from recent progress in Sweden. *Climate* 4(3):39. <https://doi.org/10.3390/cli4030039>
- Pechlivanidis IG, Arheimer B, Donnelly C et al (2017) Analysis of hydrological extremes at different hydro-climatic regimes under present and future conditions. *Clim Chang* 141(3):467–481. <https://doi.org/10.1007/s10584-016-1723-0>
- Pechlivanidis IG, 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* 54:1–14. <https://doi.org/10.1029/2017WR022035>
- Pianosi F, Fanny S, Wagener T (2015) A Matlab toolbox for global sensitivity analysis. *Environ Model Softw* 70:80–85

- Prudhomme C, Giuntoli I, Robinson EL et al (2014) Hydrological droughts in the twenty-first century, hotspots and uncertainties from a global multimodel ensemble experiment. *Proc Natl Acad Sci U S A* 111:3262–3267. <https://doi.org/10.1073/pnas.1222473110>
- Schneider C, Laizé CLR, Acreman MC, Flörke M (2013) How will climate change modify riverflow regimes in Europe? *Hydrol Earth Syst Sci* 17:325–339. <https://doi.org/10.5194/hess-17-325-2013>
- Ter Braak CJF (2006) A Markov chain Monte Carlo version of the genetic algorithm differential evolution: easy Bayesian computing for real parameter spaces. *Stat Comput* 16(3):239–249. <https://doi.org/10.1007/s11222-006-8769-1>
- Vormoor K, Lawrence D, Heistermann M, Bronstert A (2015) Climate change impacts on the seasonality and generation processes of floods – projections and uncertainties for catchments with mixed snowmelt/rainfall regimes. *Hydrol Earth Syst Sci* 19:913–931. <https://doi.org/10.5194/hess-19-913-2015>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Yeshewatesfa Hundecha¹ · Berit Arheimer¹ · Peter Berg¹ · René Capell¹ · Jude Musuuza¹ · Ilias Pechlivanidis¹ · Christiana Photiadou¹

¹ Swedish Meteorological and Hydrological Institute, Folkborgsvägen 17, 601 76 Norrköping, Sweden