Climatic Change (2020) 163:1165–1185
https://doi.org/10.1007/s10584-020-02930-z

Does a successful comprehensive evaluation increase confidence in a hydrological model intended for climate impact assessment?

Alexander Gelfan, et al. ***[full author details at the end of the article]***

Received: 2 January 2020 / Accepted: 3 November 2020 / Published online: 27 November 2020

Abstract

The objective of the study is to verify a hypothesis that a hydrological model, which successfully passed a comprehensive evaluation test (CE-test), is more suitable for climate impact study than that which failed the test. In our study, the CE-test is a specially designed model evaluation procedure, including a set of enhanced tests of model performance and robustness. The hypothesis verification is carried out with two models, ECOMAG and SWAP, which are applied for the Lena and Mackenzie River basins. The following three versions of every model are compared: (1) version A with a priori assigned parameters (without any calibration); (2) version B calibrated against streamflow observations at the basin outlets only, and (3) version C calibrated against streamflow observations at several gauges within the basins. We found that the B and C versions were successful in passing the CE-test, while the A versions failed the test. The C versions performed better than the B versions, especially at the monthly time scale. Then, all model versions were forced by global climate model (GCM) ensemble data to simulate flow projections for the twenty-first century and assess the projection uncertainty. Summarizing the results, we found that the differences in projections (in terms of mean annual changes in discharge and their uncertainties) between A version and two other versions were nearly three times larger than the differences between the B and C versions. Thus, the CE-test results together with the estimated differences in projections give us reason to conclude that the successful comprehensive evaluation of a model increases its confidence and suitability for impact assessment.

Keywords Hydrological model · Land surface model · Comprehensive evaluation · Robustness · Climate impact assessment

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

1 Introduction

There is a large mismatch between achievements in improving hydrological models, particularly the ones aimed at climate change impact studies, deepening their physical content, extending their ability to utilize climate model data, etc., on the one hand, and outdated methods of the hydrological model evaluation, on the other hand (see e.g., Krysanova et al., 2018). In the vast majority of impact studies, hydrologists prefer easy-to-pass tests (such as the split-sample test proposed by Klemeš (1986) for a stable climate) to evaluate the model performance. As a result, there are many successfully tested models that pretend to be suitable for impact studies. In order not to get lost in the "jungle of models" (Kundzewicz, 1986), one needs to be able to distinguish between models appropriate for impact studies and unsuitable ones. The question is how should model evaluation be done to draw such a distinction? Klemeš (1986) first outlined two specific tasks that have to be solved when testing the model developed for impact study which are as follows: (1) to evaluate the model transferability in time or the model robustness, i.e., its ability to perform under climatically contrasting conditions and to retain, therein, a stable model structure and parameters, and (2) to evaluate limits of the model applicability, i.e., to specify the range of hydrological variables for which the model performs well (in terms of closeness of simulations to observations). Opportunities for solving these two tasks were proposed in the studies below following the pioneering work of Klemeš (1986).

The first and most known specifically designed robustness test (differential split-sample test; DSS-test) was proposed by Klemeš (1986) and applied by other authors, including Refsgaard and Knudsen (1996), Xu (1999), Seibert (2003), Vaze et al. (2010), Coron et al. (2012), Brigode et al. (2013), and Refsgaard et al. (2013). Most of the authors highlighted a lack of robustness, i.e., a decrease in the model performance when transferring the model between climatically contrasting periods. Updated versions of the DSS-test have been proposed recently, e.g., generalized split-sample test of Coron et al. (2012), sliding window test of Coron et al. (2014), and calibration-evaluation protocol of Thirel et al. (2015). Examples of the application of these tests can be found in the special issue of the *Hydrological Science Journal* (vol. 60; issues 7–8), as well as in Birhanu et al.'s (2018) and Vormoor et al.'s (2018) works, to mention just a few. To estimate robustness, specific measures were developed, e.g., the model robustness criteria (Coron et al. 2012) and the dimensionless bias (Coron et al. 2014), all based on calculating the difference in the model performance measures between climatically contrasting periods. A criterion for judging whether this performance difference is statistically significant or not (i.e., whether the model is robust or not) was first proposed by Gelfan and Millionshchikova (2018). Recently, the DSS-based tests are recognized (Refsgaard et al., 2013 among others) as the best possible evaluation method in order to test the ability of a model to perform well beyond the range of available observations. We emphasize that we do not consider situations where climate changes can lead to such noticeable changes in the land surface characteristics that the time-invariant parameters assumption becomes unacceptable (e.g., Wagener et al., 2003; Merz et al., 2011).

Apart from testing the robustness (a necessary but not sufficient condition for extrapolating the model capabilities to the future (Krysanova et al., 2018)), the evaluation procedure should "safeguard against the use of models for tasks beyond their capabilities" (Klemeš, 1986). Most of the impact studies focus on the evaluation of the ability of a model to reproduce streamflow at a basin outlet only. Such evaluation does not provide any information on the applicability of a model with regard to other streamflow gauges or hydrological variables, and retain uncertain limits of the model applicability for its potential users. Andréassian et al. (2009) justified a methodology named a "crash-test," which is based on the idea that the models intended for impact studies have to be tested most comprehensively and incisively, particularly, on the data of a large number of river basins. The analogous tests that could be called "crash-tests" were developed earlier and involved in strict evaluation protocols within some community initiatives, such as northern hemisphere climate processes land surface experiment (NOPEX) (e.g., Motovilov et al., 1999), DMIP (Smith et al., 2004). These tests were based on the model intercomparison with regard to their ability to reproduce a number of hydrological variables, including streamflow at different gauges, snow and soil moisture characteristics, and groundwater level, to increase confidence in internal consistency of the simulated processes. Recently, studies aimed at the comprehensive model evaluation have been carried out within the ISIMIP framework (see the special issue of *Climatic Change*, vol. 141, 2017) and included, particularly, testing the model performance in terms of several hydrological signatures, e.g., high/low flow indicators and annual flow trends. Krysanova et al. (2018) first proposed a comprehensive evaluation (CE) test (see the "Comprehensive evaluation test" section) that allows one to evaluate both the robustness and limits of the model applicability, i.e., complies with both abovementioned Klemeš's requirements for evaluation of a hydrological model intended for impact study.

A knowledge gap still exists and drives the main motivation for this study. It follows from the fact that, on the one hand, the success of the models in passing the CE-test is evidence that these models perform better and are more robust than the other models that failed the test; but, on the other hand, this is not sufficient for judging if the first model set is more suitable for impact study than the second one. To judge so, one needs to compare these models in terms of their influence on projected impacts. Furthermore, we argue, and this is the key-point in our study logic, that the results of such a comparison can be interpreted as follows. If the models that failed the CE-test give similar projections (in terms of runoff anomalies) to the ones simulated by the models that passed the test, then it is reasonable to conclude that the successful testing does not increase our confidence in the model-based projections. However, if the models that failed the test give different projections than the models that passed the test, then one can conclude that the successfully tested models and the projections based on them deserve more confidence.

The objective of the study is to verify a hypothesis that the model that successfully passed a comprehensive evaluation test is more suitable for impact study than the other model that failed the test. The study was carried out with non-calibrated and calibrated versions of the physically based hydrological model ECOMAG and the land surface model SWAP. The model versions were set up for the two great Arctic basins of the Lena and the Mackenzie Rivers (Fig. 1).

The remaining part of this paper is organized as follows. The case study basins are described in the next section. The "Methods" section involves a description of the models and their versions, the CE-test and measures to compare the evaluation results, numerical experiments for estimating hydrological response to the ensemble of GCM-based climate projections, and method for deriving the estimation uncertainty. The results and discussion are presented in the "Results and discussion" section. The overall conclusions are given in the "Conclusions" section.

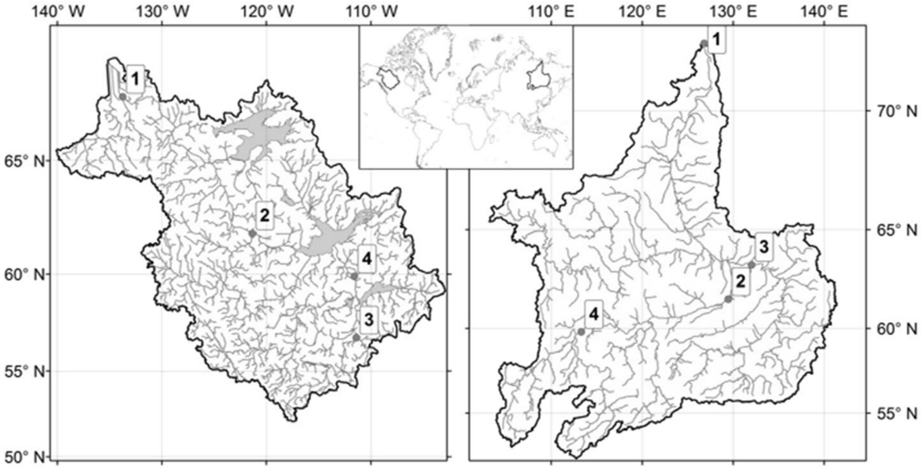

Fig. 1 The study basins: Mackenzie (left) and Lena (right) with calibration and validation gauges. The Mackenzie River basin: 1 – Arctic Red, 2 – Fort Simpson, 3 – McMurray, 4 – Fitzgerald; the Lena River basin: 1 – Stolb, 2 – Tabaga, 3 – Verkhoyanski Perevoz, 4 – Krestovski

2 Basins under study

The Lena River basin (the watershed area is 2,460,000 km^2) is located in the Eastern Siberia between 103 and 142° E and 52–74° N. The climate is extremely continental: cold winter with air temperatures often dropping below – 50 °C and warm summer. The whole territory of the basin is located in the permafrost zone. Four landscape zones are identified in the basin which are as follows: arctic deserts, tundra, forest tundra, and taiga forests covering nearly 70% of the basin. The Lena River runoff is characterized by spring-summer snowmelt flood, summer and autumn rain floods, and extremely low water levels in winter. The mean annual discharge at the outlet gauge Stolb is about 15,000 m^3/s.

The Mackenzie River with a basin area of 1,800,000 km^2 is the largest river in North America flowing into the Arctic Ocean and one of the 10 longest rivers in the world. Its basin is located between 102 and 142° W and 52–69° N and is in the same natural zones as the Lena River basin. The hydrological regime of the Mackenzie River is under the influence of several big lakes. The typical hydrograph of the river is characterized by the extended period of snowmelt flood and rain-induced spring and summer floods. The mean annual discharge of the Mackenzie River at the outlet gauge Arctic Red is about 10,300 m^3/s.

3 Methods

3.1 Models and their versions

The regional physically based hydrological model ECOMAG (ECOlogical Model for Applied Geophysics) (Motovilov et al., 1999; Motovilov, 2016) and the land surface model SWAP (Soil Water - Atmosphere - Plants) (Gusev and Nasonova, 1998, 2003) were used in this study.

Both the ECOMAG and the SWAP models were applied earlier for river basins of various sizes and located in different natural conditions: from small research basins included into the model intercomparison projects, such as NOPEX, PILPS, and MOPEX (e.g., Motovilov et al., 1999; Gusev and Nasonova, 1998, 2003; Schlosser et al., 2000; Gusev et al., 2006; Duan et al., 2006), to the world's largest basins of millions of square kilometer (e.g., Motovilov, 2016; Kalugin and Motovilov, 2018; Gusev et al., 2016), including the Lena and Mackenzie basins within the ISIMIP initiative (Gelfan et al., 2017; Gusev et al., 2018).

The models describe spatially distributed processes of snow accumulation and melt, water infiltration into unfrozen and frozen soil, evapotranspiration, soil freezing and thawing, including hydrothermal processes within the seasonally thawed soils above permafrost, as well as overland, subsurface, groundwater and channel flows. Most of the parameters are physically meaningful and derived from global datasets of the land surface characteristics (topography, soil and vegetation properties). Some key parameters of the models (up to 9) are adjusted through calibration against streamflow data and available measurements of the internal basin variables (snow, soil moisture, etc.). The calibration procedures are described for the ECOMAG model in Motovilov (2016), for the SWAP model—in Gusev et al. (2018).

The ECOMAG model inputs include daily air temperature, air humidity, and precipitation. The SWAP model is forced by precipitation, air temperature, air humidity, incoming shortwave and longwave radiation, atmospheric pressure, and wind speed of different temporal resolutions (from half an hour to a day, depending on available data).

Within the ECOMAG, the discretization of the basins was carried out based on the global (1-km resolution) DEM data from the HYDRO1K database of the US Geological Survey (https://www.usgs.gov/media/files/hydro-1k-readme). To assign soil parameters, the global HSWD database of 1-km resolution (Fischer et al., 2008) was used. Land-use parameters were obtained from the global land cover maps (Bartholomé and Belward, 2005). For the SWAP model, the basins were presented by a set of regular grid cells (Gusev et al., 2016, 2018). For the Mackenzie, most vegetation parameters were derived from the Global ECOCLIMAP dataset (Champeaux et al., 2005), while soil parameters were calculated by making use of equations from Cosby et al. (1984) and data on soil texture taken from ECOCLIMAP. For the Lena basin, the land surface parameters were derived from ISLSCP-II/GSWP-2 global datasets as described in Gusev et al. (2016).

As can be seen from the above description, as well as from the description of the models (Gusev and Nasonova, 2003; Motovilov, 2016), they differ in the information requirements, approaches to the schematization of the catchment, parameterization of hydrological processes, methods of setting parameters, and calibration.

Three versions of each model were constructed and evaluated in this study: (1) the non-calibrated model with parameters all defined a priori from the abovementioned global datasets or from publications (mainly, not related to the study basins) ("Version A" hereafter); (2) the model calibrated against daily streamflow observations at the basin outlet ("Version B" hereafter); (3) the model calibrated against daily streamflow observations at multiple gauges ("Version C" hereafter).

The location of the gauges, whose data were used for calibration, is shown in Fig. 1. The ECOMAG-based simulations for historical period were done using EWEMBI (Lange, 2018) reanalysis data (1979–2012); the calibration period was 1980–1989. We used the WATCH reanalysis dataset (1971–2001) when applied the SWAP-based versions (period of calibration was 1971–1977). The Nash and Sutcliffe efficiency (NSE) and the Percent Bias (PBIAS) efficiency criteria were used for calibration.

Besides the differences in the models listed above, we intentionally used various forcing data and different periods for calibration of the ECOMAG and SWAP models to ensure that if comparison of evaluation results between three versions of each model would lead to similar findings then they are not influenced by the specific features of particular datasets related to the concrete observation period. Consequently, and this is important to emphasize in order to avoid misinterpretation of further results, *the direct comparison of SWAP and ECOMAG performances was not intended: we made comparisons between three versions of both models*. Herewith, using such different models allows us to generalize similarities (if any) in their behavior while testing. Each of the versions A, B, and C was evaluated by the CE-test described in the next subsection.

3.2 Comprehensive evaluation test

The CE-test that is a slight modification of the one proposed by Krysanova et al. (2018) was applied for the constructed model versions A, B, and C. The test begins with checking the quality of the data used (as step 0, done in the previous study and reported here). The used meteorological reanalysis data as well as the global datasets of the basin characteristics were well-verified for the study basins within the ISIMIP project (see special issue in *Climatic Change*, vol. 141, 2017). However, the sparse density of the meteorological stations, especially in mountain regions of the study basins, leads to lower quality of precipitation data in these regions (Yang et al., 2005). Also, the earlier studies (Shiklomanov et al. 2006; Woo and Thorne, 2014) reported that the low flow measurements over the prolonged ice-cover period at the study basins contain large errors (e.g., almost 30% for the Lena River (Shiklomanov et al. 2006)). Winter flow is also influenced by the release of water from the dams (Bennet Dam and Vilyui Dam) located within the study basins (Woo and Thorne, 2014). The dam effects in addition to the large errors in streamflow data may lead to weaker results of winter flow simulation.

The four steps of a model evaluation are the following: (1) evaluation of the model versions for multiple basin gauges and for multiple variables; (2) application of the modified DSS-test (robustness test); (3) evaluation of the model versions for a number of streamflow indicators of interest; and (4) evaluation of the model versions for reproducing the observed trends (or lack of trends) in the annual runoff time series.

At the 1st step, streamflow data observed at the same gauges that were used for calibration (see above) were utilized to evaluate the models for the period of 1990–2012 (for the ECOMAG-based versions) and 1978–2001 (for the SWAP-based versions). The same performance criteria, NSE and PBIAS, as were applied at the calibration stage, were used for the evaluation. We followed the recommendations of Moriasi et al. (2015) (see also Krysanova et al., 2018) to evaluate the A, B, and C version performance as “poor” if NSE < 0.50 or $|\text{PBIAS}| \geq 15\%$ for daily and monthly streamflow data.

At the 2nd step, the modified DSS-test proposed by Gelfan and Millionshchikova (2018) was applied for evaluation of the model versions’ robustness in terms of both daily and monthly streamflow hydrographs. The essence of the test is in the following. The whole simulation period (calibration period plus evaluation period, i.e., 1980–2012 for the ECOMAG-based versions and 1971–2001 for the SWAP-based versions), is divided into four contrasting climate periods with regard to the ratios of the annual air temperature (T^j_{annual}, where j is the number of year) and annual precipitation (P^j_{annual}) to the corresponding mean annual values averaged over the whole simulation period (T_{cl} and P_{cl}, respectively). The 1st climate period includes non-continuous sequence of years with $T^j_{\text{annual}} > T_{\text{cl}}$ and $P^j_{\text{annual}} > P_{\text{cl}}$ and is indicated as a “warm-wet” (WW) period; the 2nd one is a “cold-wet” (CW) period consisting of years with $T^j_{\text{annual}} < T_{\text{cl}}$ and $P^j_{\text{annual}} > P_{\text{cl}}$; the 3rd one is a “warm-dry” (WD) period with $T^{j}_{\mathrm{annual}} > T_{\mathrm{cl}}$ and $P^{j}_{\mathrm{annual}} < P_{\mathrm{cl}}$; and the 4th one is a "cold-dry" (CD) period with $T^{j}_{\mathrm{annual}} < T_{\mathrm{cl}}$ and $P^{j}_{\mathrm{annual}} < P_{\mathrm{cl}}$. Herewith, the annual values are obtained from the used meteorological data averaged over the basin area. For each selected period, the NSE criterion is estimated from the hydrographs observed in the corresponding years with respective hydrographs simulated for these years at the previous 1st step of the CE-test. As a result, four NSE values are estimated: $\mathrm{NSE}_{\mathrm{WW}}$, $\mathrm{NSE}_{\mathrm{CW}}$, $\mathrm{NSE}_{\mathrm{WD}}$, and $\mathrm{NSE}_{\mathrm{CD}}$. Gelfan and Millionshchikova (2018) developed a modified DSS-test assessing statistical significance of the NSE changes from one climatic period to another. Within the test, NSE is considered a random variable, whose sampling variance depends on the statistics of the observed and simulated discharge data (see Eq. S5 in the Supplementary information). The authors (Gelfan and Millionshchikova 2018) used the NSE-decomposition formula (see e.g., Gupta et al. 2009) and showed that the desired statistical significance of the NSE changes depends on mean values of the simulated and observed discharge data, variances of these data, and the covariance between the simulated and the observed discharges. Finally, the modified DSS-test is considered successfully passed (i.e., a model is robust) if and only if the NSE values between periods differ not significantly in the statistical sense (as measured by criterion in Eq. S3) for each of six possible differences: ($\mathrm{NSE}_{\mathrm{WW}}$ - $\mathrm{NSE}_{\mathrm{CW}}$), ($\mathrm{NSE}_{\mathrm{WD}}$ - $\mathrm{NSE}_{\mathrm{CW}}$), ($\mathrm{NSE}_{\mathrm{CD}}$ - $\mathrm{NSE}_{\mathrm{CW}}$), etc.

At the 3rd step, the A, B, and C versions' performance was evaluated in terms of the two flow indicators proposed by Yilmaz et al. (2008), namely, FHV_{10} that is the percent bias in the flow-duration curve (FDC) high-segment volume for flows with exceedance probability lower than 10%, and FLV_{70} that is the percent bias in the FDC low-segment volume (70–100% exceedance probability). We evaluated the performance as "poor" in terms of high or low flow indicators if $|\mathrm{FHV}_{10}| \geq 15\%$ or $|\mathrm{FLV}_{70}| \geq 15\%$. Formulas (S6)–(S7) for calculation of FHV_{10} and FLV_{70} are presented in the Supplementary information.

At the 4th step, we applied the test described by Santer et al. (2000) to evaluate the model versions' ability in reproducing the observed trends in annual flow. First, we fitted linear trends to the time series of the observed and simulated annual flows and assessed statistical significance of these trends through the modified Student's t criteria adjusted for autocorrelation of residuals. The corresponding formulas (S8)–(S13) are presented in the Supplementary information. As a result of the first test, we evaluated the model performance as "poor" in terms of trend reproduction, if one of the two compared trends was significant while the other one was not. If both trends were assessed as statistically insignificant, we considered that the model version successfully passed the 4th step of the evaluation test. If both trends were assessed as statistically significant, then one should apply the second, difference series test (Santer et al. 2000), because in this case, the *difference* between the observed and simulated trends may be significant. If this is the case, the model version performance is evaluated as "poor" in terms of trend reproduction, in spite of the fact that the corresponding version had passed the first trend test.

Figure 2 presents the flowchart of the used CE-test and summarizes conditions that should be met by the model version in order to pass this test.

3.3 Modeling hydrological projections and uncertainty assessment

First, the compared A, B, and C versions were driven by the bias-corrected (to WATCH reanalysis data) climate data from four global climate models (GCMs): GFDL-ESM2M,

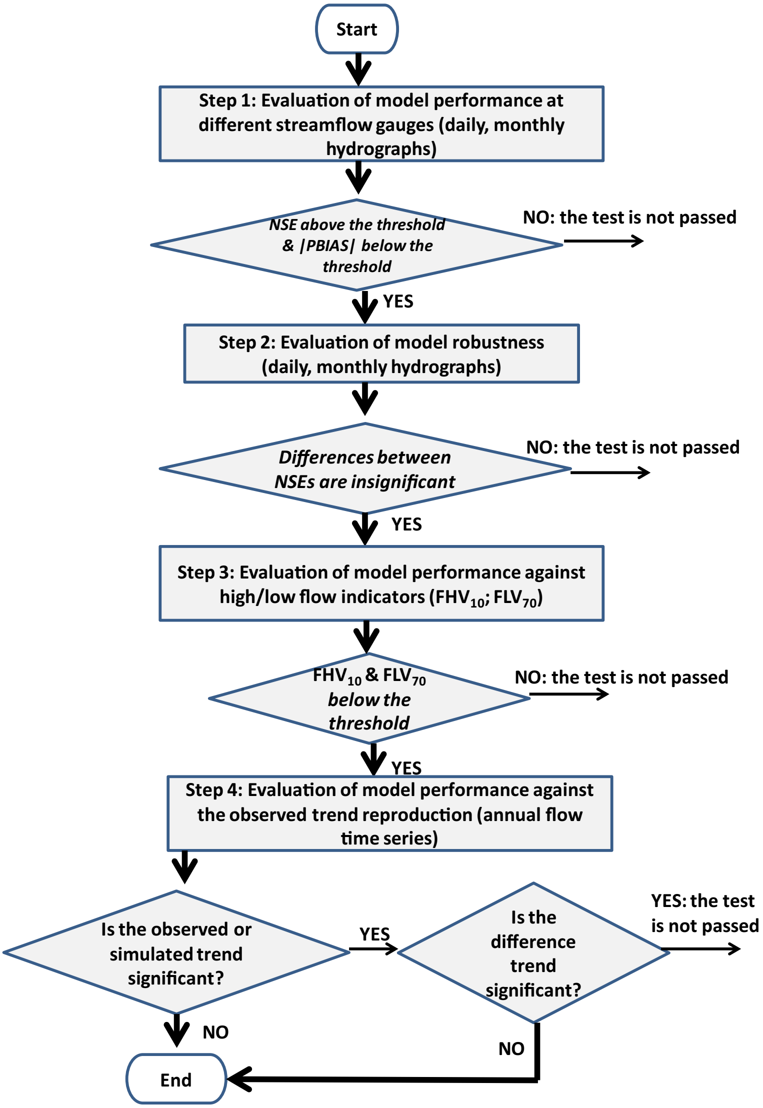
Start
Step 1: Evaluation of model performance at different streamflow gauges (daily, monthly hydrographs)
NSE above the threshold & |PBIAS| below the threshold
NO: the test is not passed
YES
Step 2: Evaluation of model robustness (daily, monthly hydrographs)
Differences between NSEs are insignificant
NO: the test is not passed
YES
Step 3: Evaluation of model performance against high/low flow indicators (FHV10; FLV70)
FHV10 & FLV70 below the threshold
NO: the test is not passed
YES
Step 4: Evaluation of model performance against the observed trend reproduction (annual flow time series)
Is the observed or simulated trend significant?
YES
Is the difference trend significant?
YES: the test is not passed
NO
NO
End

Fig. 2 Flowchart of the comprehensive evaluation test and the conditions that should be met by the model versions in order to pass the test

HadGEM2-ES, IPSL-CM5A-LR, MIROC-ESM-CHEM) for 1976–2005, allowing to obtain mean annual simulated flows for the historical period. Then, we simulated hydrological response to climate projections derived from the four GCMs and four Representative Concentration Pathway (RCP) scenarios. The annual flow time series simulated under each of 16 GCM-RCP-based climate scenarios were compared with the mean annual flows simulated under the corresponding GCM outputs for the historical period. As a result, the 16 time series (trajectories) of annual flow anomalies were constructed as percentages of the future flow to the simulated historical flow. We smoothed the constructed trajectories by the moving average technique with a 30-year sliding window. The technique was applied for 16 94-year series (2006–2099) of annual flow anomalies simulated by each model version. The smoothed trajectories were averaged and the mean of the ensemble of trajectories was assumed as a representation of the future flow changes. The spread of the modeled trajectories was considered a proxy measure of uncertainty (Kundzewicz et al., 2018).

4 Results and discussion

4.1 Comprehensive evaluation test results

The results of the CE-test applied for the A, B, and C versions are presented below in accordance with the step chain described in the "Comprehensive evaluation test" section (see also Fig. 2).

4.1.1 First step: Evaluation of model performance at different streamflow gauges

Table 1 presents the performance of the A, B, and C versions evaluated in terms of the NSE and PBIAS measures for the whole simulation periods (1980–2012 for the ECOMAG-based versions and 1971–2001 for the SWAP-based versions). Let us repeat that the model performance is considered "poor" if one or both measures do not satisfy the pre-determined conditions: NSE < 0.50, |PBIAS| ≥ 15%. One can see from Table 1 that the performance of the A versions is not satisfactory for both basins. For almost all streamflow gauges of the Lena basin (except Krestovski), the C versions perform similarly well as the corresponding B versions. This is quite unexpected because the B versions were calibrated against streamflow observations at the basin outlet only, not for other gauges. The obtained result demonstrates that the runoff generation process is more or less homogeneous over the almost whole Lena basin except for the Krestovski sub-basin, most part of which is located in the mountainous area. We believe that the weaker model performance for this sub-basin can be explained by the low density of the meteorological stations in the mountains leading to lower quality of precipitation data in such regions (Yang et al., 2005) and, as the consequence, to increasing errors in the respective reanalysis data. Unlike the C version, the calibration of the B model version did not consider streamflow data at the Krestovski gauge, and therefore, the evaluation results are weaker there.

This is not the case for the Mackenzie basin, where large lakes influence the upper parts of the basin more than the lower parts. As one can see from Table 1, the SWAP-based versions appeared to be sensitive to this difference. The reason is in the fact that in the SWAP model, the lake effect is not taken into account but corrected in the calibration process. As the B version was calibrated against the outlet streamflow hydrograph, which is less influenced by the lakes than the Fitzgerald and Fort Simpson hydrographs, the B version performed better at the outlet gauge and poorly at the latter ones. The SWAP-based C version calibration improved the Fitzgerald and Fort Simpson hydrograph simulations, and as a result, C version performed better than the B version.

Thus, the B and C versions have successfully passed the first step of the evaluation test for the Lena basin (except the Krestovski sub-basin). For the Mackenzie basin, the C versions as well as the ECOMAG-based B version perform well. The A versions have not passed the test for any gauge.

Table 1 Performance of the A, B, and C versions evaluated in terms of the NSE and PBIAS measures: 1st step of the evaluation procedure (shaded cells mean that the performance is evaluated as "poor" according to the accepted criteria NSE < 0.50, |PBIAS| ≥ 15%; see also Fig. 2)

MACKENZIE RIVER BASIN									
ECOMAG, 1979 - 2012									
Gauges	Version A			Version B			Version C		
	NSE day	NSE month	PBIAS, %	NSE day	NSE month	PBIAS, %	NSE day	NSE month	PBIAS, %
Arctic Red	0.52	0.53	25.3	0.88	0.93	2.3	0.86	0.92	-0.6
Fort Simpson	0.57	0.60	26.9	0.83	0.89	7.1	0.83	0.91	3.5
Fitzgerald	0.26	0.27	13.8	0.54	0.59	2.1	0.56	0.65	-0.5
McMurray	0.42	0.45	15.4	0.70	0.78	5.2	0.68	0.81	2.2
SWAP, 1971-2001									
Gauges	Version A			Version B			Version C		
	NSE day	NSE month	PBIAS, %	NSE day	NSE month	PBIAS, %	NSE day	NSE month	PBIAS, %
Arctic Red	-1.05	-1.11	-14.6	0.38	0.49	0.3	0.69	0.76	2.8
Fort Simpson	0.03	0.09	-28.4	0.06	0.33	-2.5	0.81	0.89	0.4
Fitzgerald	-4.1	-3.85	-8.0	-1.12	-0.85	7.8	0.32	0.40	-1.3
McMurray	0.03	0.32	-4.6	0.41	0.55	22.1	0.47	0.56	1.2
LENA RIVER BASIN									
ECOMAG, 1979-2012									
Gauges	Version A			Version B			Version C		
	NSE day	NSE month	PBIAS, %	NSE day	NSE month	PBIAS, %	NSE day	NSE month	PBIAS, %
Stolb	0.44	0.54	-24.0	0.89	0.96	6.0	0.85	0.93	11.1
Tabaga	0.42	0.49	-34.1	0.89	0.94	-12.0	0.90	0.96	-7.3
Krestovski	0.19	0.35	-40.0	0.79	0.86	-29.0	0.81	0.88	-14.9
Verkhoyanski Perevoz	0.40	0.53	-30.0	0.84	0.9	2.0	0.83	0.89	6.8
SWAP, 1971 - 2001									
Gauges	Version A			Version B			Version C		
	NSE day	NSE month	PBIAS, %	NSE day	NSE month	PBIAS, %	NSE day	NSE month	PBIAS, %
Stolb	0.13	0.16	-66.8	0.83	0.93	-3.2	0.82	0.94	-0.5
Tabaga	0.02	0.01	-37.9	0.61	0.75	-14.5	0.74	0.87	-0.3
Krestovski	0.30	0.32	-44.9	0.05	0.35	-30.0	0.50	0.64	-13.5
Verkhoyanski Perevoz	0.29	0.45	-36.9	0.66	0.75	-14.3	0.83	0.91	-3.0

4.1.2 Second step: Evaluation of model robustness

The full second-step results related to the Mackenzie and Lena basins are in Tables 2 and 3, respectively (see evaluation approach in the Supplementary information, part 1). The A

Table 2 NSE values estimated from observed vs. simulated hydrographs for the contrasting climate periods* in the Mackenzie River basin. Shaded cells mean that the necessary condition (Eq. S3) of the model robustness was not met at the 95% confidence level

MACKENZIE RIVER BASIN									
ECOMAG (Version B)									
		Arctic Red		Fort Simpson		Fitzgerald		McMurray	
		warm	cold	warm	cold	warm	cold	warm	cold
NSE day	wet	0.90	0.87	0.86	0.81	0.66	0.60	0.68	0.72
	dry	0.88	0.86	0.78	0.85	0.72	0.64	0.64	0.72
	robustness	YES		NO		NO		NO	
NSE month	wet	0.95	0.92	0.91	0.88	0.28	0.41	0.76	0.78
	dry	0.94	0.92	0.89	0.90	0.30	0.46	0.75	0.80
	robustness	YES		YES		NO		YES	
ECOMAG (Version C)									
		warm	cold	warm	cold	warm	cold	warm	cold
NSE day	wet	0.87	0.85	0.84	0.82	0.65	0.65	0.68	0.73
	dry	0.85	0.85	0.82	0.84	0.31	0.42	0.59	0.68
	robustness	YES		YES		NO		NO	
NSE month	wet	0.94	0.91	0.92	0.90	0.76	0.72	0.79	0.83
	dry	0.93	0.91	0.93	0.90	0.37	0.52	0.78	0.82
	robustness	YES		YES		NO		YES	
SWAP (Version B)									
		Arctic Red		Fort Simpson		Fitzgerald		McMurray	
		warm	cold	warm	cold	warm	cold	warm	cold
NSE day	wet	0.32	0.44	0.05	-0.08	-2.56	-0.23	0.26	0.58
	dry	0.35	0.45	0.16	0.02	-1.98	-0.39	0.30	0.46
	robustness	NO		NO		NO		NO	
NSE month	wet	0.42	0.55	0.29	0.18	-2.16	-0.02	0.36	0.71
	dry	0.46	0.58	0.42	0.37	-1.65	-0.20	0.48	0.60
	robustness	YES		YES		NO		NO	
SWAP (Version C)									
		warm	cold	warm	cold	warm	cold	warm	cold
NSE day	wet	0.65	0.69	0.80	0.83	-0.07	0.59	0.45	0.48
	dry	0.66	0.66	0.80	0.82	-0.12	0.62	0.44	0.45
	robustness	YES		YES		NO		NO	
NSE month	wet	0.72	0.79	0.88	0.90	0.04	0.64	0.53	0.58
	dry	0.74	0.83	0.88	0.90	-0.01	0.70	0.54	0.54
	robustness	YES		YES		NO		YES	

* The mean annual temperature (T) and precipitation (P) during the contrasting climate periods are $T_{WW} = -2.8$ °C; $T_{WD} = -2.6$ °C; $T_{CW} = -4.4$ °C; $T_{CD} = -4.0$ °C; $P_{WW} = 488$ mm; $P_{WD} = 425$ mm; $P_{CW} = 482$ mm; $P_{CD} = 422$ mm

versions have not passed the 2nd evaluation step (appeared non-robust for all gauges) and are not shown in these Tables.

As an example, Fig. 3 presents the NSE values and their differences estimated from observed vs. simulated daily hydrographs for the contrasting climate periods in the Mackenzie River basin (only the ECOMAG-based B and C versions are shown). One can see from Fig. 3 that the NSE values are quite similar, and their differences are close to zero for the Arctic Red and Fort Simpson gauges, no matter for what climate period they are estimated (note that the climate conditions are visibly different for these periods: $T_{WW} = -2.8$ °C; $T_{WD} = -2.6$ °C; $T_{CW} = -4.4$ °C; $T_{CD} = -4.0$ °C; $P_{WW} = 488$ mm; $P_{WD} = 425$ mm; $P_{CW} = 482$ mm; $P_{CD} =$ 422 mm). This is not the case for the other two gauges: one or more NSE differences are

Table 3 NSE values estimated from observed vs. simulated hydrographs for the contrasting climate periods* in the Lena River basin. Shaded cells mean that the necessary condition (Eq. S3) of the model robustness was not met at the 95% confidence level

LENA RIVER BASIN									
ECOMAG (Version B)									
		Stolb		Tabaga		Krestovski		VPerevoz	
		warm	cold	warm	cold	warm	cold	warm	cold
NSE day	wet	0.88	0.91	0.90	0.87	0.79	0.77	0.84	0.85
	dry	0.89	0.88	0.90	0.87	0.83	0.77	0.89	0.80
	robustness	YES		YES		NO		NO	
NSE month	wet	0.95	0.96	0.95	0.93	0.85	0.86	0.87	0.91
	dry	0.96	0.96	0.96	0.93	0.89	0.84	0.95	0.90
	robustness	YES		YES		YES		YES	
ECOMAG (Version C)									
		warm	cold	warm	cold	warm	cold	warm	cold
NSE day	wet	0.84	0.88	0.9	0.88	0.81	0.79	0.83	0.85
	dry	0.87	0.84	0.91	0.89	0.86	0.79	0.89	0.80
	robustness	YES		YES		NO		NO	
NSE month	wet	0.92	0.93	0.96	0.95	0.87	0.88	0.86	0.90
	dry	0.94	0.94	0.96	0.95	0.92	0.87	0.94	0.86
	robustness	YES		YES		YES		YES	
SWAP (Version B)									
		Stolb		Tabaga		Krestovski		VPerevoz	
		warm	cold	warm	cold	warm	cold	warm	cold
NSE day	wet	0.83	0.85	0.56	0.64	0.09	0.01	0.63	0.67
	dry	0.86	0.78	0.67	0.60	0.09	-0.08	0.68	0.66
	robustness	NO		NO		NO		YES	
NSE month	wet	0.93	0.94	0.71	0.76	0.42	0.27	0.70	0.73
	dry	0.94	0.91	0.80	0.74	0.40	0.24	0.77	0.81
	robustness	YES		YES		YES		YES	
SWAP (Version C)									
		warm	cold	warm	cold	warm	cold	warm	cold
NSE day	wet	0.83	0.81	0.73	0.77	0.48	0.51	0.83	0.83
	dry	0.86	0.76	0.73	0.74	0.50	0.41	0.88	0.80
	robustness	NO		YES		NO		NO	
NSE month	wet	0.94	0.94	0.88	0.84	0.69	0.60	0.89	0.89
	dry	0.94	0.92	0.90	0.89	0.66	0.54	0.95	0.93
	robustness	YES		YES		YES		YES	

* The mean annual temperature (T) and precipitation (P) during the contrasting climate periods are $T_{WW} = -8.4$ °C; $T_{WD} = -8.6$ °C; $T_{CW} = -9.7$ °C; $T_{CD} = -10.1$ °C; $P_{WW} = 434$ mm; $P_{WD} = 360$ mm; $P_{CW} = 421$ mm; $P_{CD} = 375$ mm

noticeably far from zero. A formal statistical test confirmed this visual analysis-based finding (see Table 2).

Let us analyze the robustness results presented in Tables 2 and 3 from the two viewpoints: (1) comparison of the B and C versions of the same origin (e.g., SWAP-based); and (2) comparison between the basins.

Generally, the ECOMAG-based B version appeared almost as robust as the C version of the same origin. In terms of reproducing daily and monthly streamflow hydrographs at the Mackenzie basin outlet, the B and C ECOMAG-based versions performed well and statistically similar for contrasting climate periods (Table 2). The same can be said regarding the robustness of the ECOMAG-based versions evaluated for the Lena basin (Table 3): the B and C version results were not sensitive to change in the climate period when simulating daily

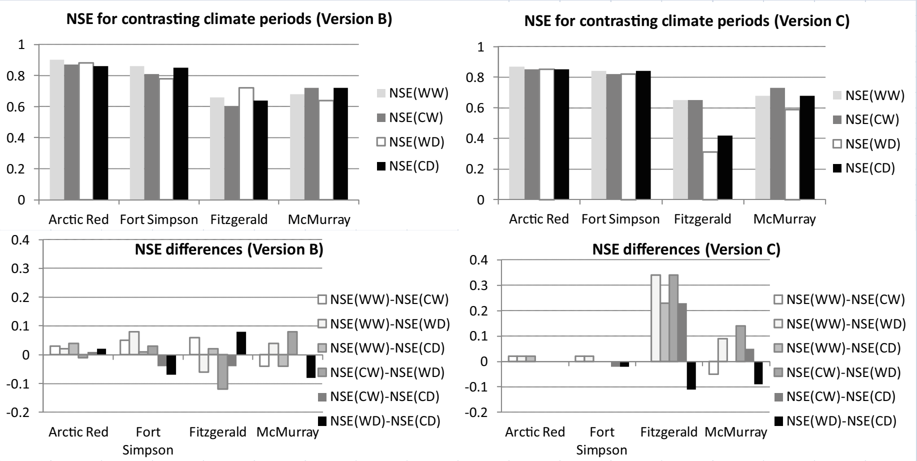

Fig. 3 NSE values and their differences calculated by ECOMAG for the contrasting climate periods at the Mackenzie River basin (daily streamflow)

streamflow hydrographs in the Stolb (outlet) and Tabaga gauges. In terms of the monthly streamflow simulation, the ECOMAG-based versions were robust for all gauges of the Lena basin. Both daily and monthly streamflow simulations by the SWAP-based C version appeared to be more robust than the B version simulations for Mackenzie (Table 2), but their robustness was comparable for Lena, and much higher for the monthly simulations (Table 3). Interestingly, that the SWAP-based B version was evaluated as robust but poor for the Krestovski gauge (Lena basin) and the Fort Simpson gauge (Mackenzie basin), in other words, NSE values estimated for climate contrasting periods were close to each other, but small (Tables 2, and 3).

Thus, for monthly simulations, the B and C version robustness was quite similar, irrespective of their origin. In terms of daily simulations, the C versions were slightly more robust than the B versions for the Mackenzie basin, but robustness was similar for the Lena basin. Totally, at the monthly scale, the model versions demonstrated higher robustness for the Lena (all gauges) than for the Mackenzie (mostly 3 gauges of four).

4.1.3 Third step: Evaluation of the model's skill to simulate indicators of extremes

At this step, the performance of the A, B, and C versions was evaluated as their skill for simulating indicators of hydrological extremes: the high-flow (FHV_{10}) and low-flow (FLV_{70}) indicators (see Supplementary information, part 2). The version was considered "poor" in relation to the corresponding indicator if the values of this indicator exceeded 15%. The versions A demonstrated poor performance at this step, and their results are not discussed hereafter. The evaluation results for the B and C versions s are shown in Fig. 4.

One can see that both versions performed much better in reproducing high flow than low flow. Both ECOMAG-based versions successfully passed the high-flow test for all gauges of the basins. The SWAP-based B versions were worse for both basins for high flow than the C versions, which also passed the test (except for one gauge in the Mackenzie). With regard to the low flow indicator, almost all versions performed poorly. The same result (worse performance in reproducing low flow) was obtained for the regional hydrological models in Huang et al.'s study (2017). In our case, the result can be explained by neglecting the ice phenomena

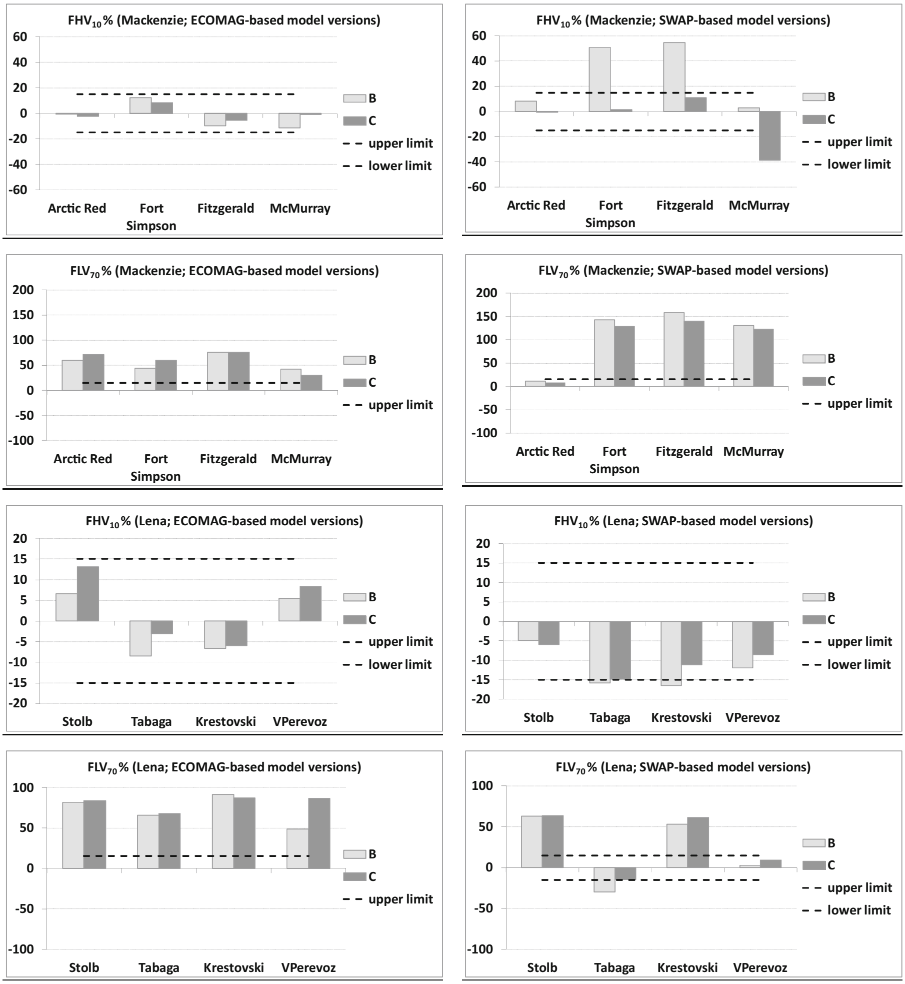

Fig. 4 Performance of the B and C versions evaluated in terms of reproducing the high-flow (FHV_{10}) and low-flow (FLV_{70}) indicators calculated from the daily streamflow hydrographs. A rebound out of the dashed lines indicates a model as "poor" in reproducing the corresponding indicator (see the "Comprehensive evaluation test" section)

in the used models, and, to a larger extent, by the low quality of streamflow data during the prolonged ice-covered period at the rivers under study (Shiklomanov et al. 2006; Woo and Thorne, 2014).

4.1.4 Fourth step: Evaluation of the skill of the model to simulate trends in the annual runoff

The correspondence of the linear trends fitted to the time series of the observed and simulated annual flows was tested by computing a signal-to-noise ratio (the ratio between the estimated trend slope and its standard deviation) under the assumption that this ratio is distributed as Student's *t* random variable. The assessed trend slope values and their standard deviations are presented in Table S1 (supplementary information). The time series of the annual flows and the corresponding trends are shown in Figs. S1 and S2 (Supplementary information).

The main finding is that all the trends fitted to the observed annual flow series are statistically insignificant, and all the trends fitted to the corresponding simulated series, irrespective of the version used (even the non-calibrated A versions), are insignificant too. The insignificance is due to the large residual variance, which was corrected to account for temporal autocorrelation in the residual series and to avoid thereby underestimation of the variance (Santer et al., 2000; see also Supplementary information, part 3). Adjusting the actual sample size (around 30 years for the studied annual discharge series) to an effective sample size (Eq. S13 in the Supplementary information) resulted in a decrease of some series down to 30% of their actual sizes.

Thus, the trends were assessed as statistically insignificant, i.e., all the model versions successfully passed the trend evaluation test without the necessity of applying the difference series method (see the "Comprehensive evaluation test" section). Some previous studies detected significance of the annual runoff trend for the Lena basin (Shi et al., 2019; Tananaev et al., 2016). One reason for such discrepancy is that the statistical test we applied (Santer et al., 2000) complicates rejection of the null hypothesis, i.e., the test is more conservative than those (Mann-Kendall and the Spearman tests) used in the mentioned studies. Another reason could be that the discharge time series analyzed in our study are shorter, which also makes rejecting of the null hypothesis less likely.

4.1.5 Summary of the evaluation results

Tables S2 and 4 summarize the results of the 4-step CE-test at the daily and monthly time scales, respectively. First, the model versions performed better in reproducing monthly streamflow series than the daily series. Second, the non-calibrated A versions performed worse than the B versions, which, generally, performed worse than the C versions, especially at the monthly time scale (Tables 1, 2, and 4). Third, from the evaluation results, we identified the preferable (in terms of the assigned criteria) versions and established the limits of their applicability. For instance, both C versions passed all tests for all gauges for the Lena and for two gauges (Arctic Red and Fort Simpson) for the Mackenzie and are considered preferable for simulations at the monthly time scale (Table 4).

4.2 Hydrological projections and their uncertainty

The evaluation results summarized above indicate which model versions performed better in the historical period, but this is not sufficient for judging if these versions are more suitable for impact study (see the "Introduction" section, par. 4). To judge so, let us repeat, one needs to compare these model versions in terms of their influence on projected impacts. The projection comparison is needed also due to the fact that sometimes models are used for impact assessment without checking their performance in advance or ignoring their weak or poor performance in the historical period, because modelers assume that changes between the future and reference periods would be still meaningful, because their models include all relevant processes and default parameters. This is especially true regarding application of the global hydrological models for impact assessment: almost all such models are not calibrated and are often applied for impact studies without evaluation of their performance in the historical period.

Figure 5 presents trajectories of the annual flow anomalies projected for the basin outlets by the ECOMAG-based and SWAP-based versions. The uncertainty bounds, i.e., the spread of 16 trajectories projected under the input data corresponding to 16 combinations of the GCMs and RCP-scenarios, are also shown in Fig. 5. The results for the other gauges can be found in the Supplementary information (Figs. S3–S6).

Generally, Fig. 5 demonstrates that the basin outlet flow projections simulated by the non-calibrated A versions significantly differ from the projections of the calibrated B and C versions. First, one can see that the projection uncertainty is larger for the A versions. For Lena, for instance, at the end of the century, the uncertainty band (spread of the modeled trajectories) for the SWAP-based A versions equals to 223% (for the B and C versions 52% and 51%, respectively); for the ECOMAG-based versions, the corresponding uncertainty bands are 28%, 20%, and 19%, respectively. This finding is connected with the large difference between the non-calibrated vs. calibrated SWAP-based versions (see 1–3 steps of

Table 4 Summary of the comprehensive evaluation test results (monthly streamflow simulation). Shaded cells identify that the corresponding test was not passed by the corresponding model version for the corresponding streamflow gauge

		ECOMAG			SWAP		
		Version A	Version B	Version C	Version A	Version B	Version C
MACKENZIE RIVER BASIN							
Arctic Red	Multi-site evaluation	▨			▨	▨	
	Robustness test	▨			▨		
	Trend test (annual flow)						
Fort Simpson	Multi-site evaluation	▨			▨	▨	
	Robustness test	▨			▨		
	Trend test (annual flow)						
Fitzgerald	Multi-site evaluation	▨			▨	▨	▨
	Robustness test	▨	▨	▨	▨	▨	▨
	Trend test (annual flow)						
McMurray	Multi-site evaluation	▨			▨	▨	▨
	Robustness test	▨			▨	▨	
	Trend test (annual flow)						
LENA RIVER BASIN							
Stolb	Multi-site evaluation	▨			▨		
	Robustness test	▨			▨		
	Trend test (annual flow)						
Tabaga	Multi-site evaluation	▨			▨		
	Robustness test	▨			▨		
	Trend test (annual flow)						
Krestovski	Multi-site evaluation	▨	▨		▨	▨	
	Robustness test	▨			▨		
	Trend test (annual flow)						
V. Perevoz	Multi-site evaluation	▨			▨		
	Robustness test	▨			▨		
	Trend test (annual flow)						

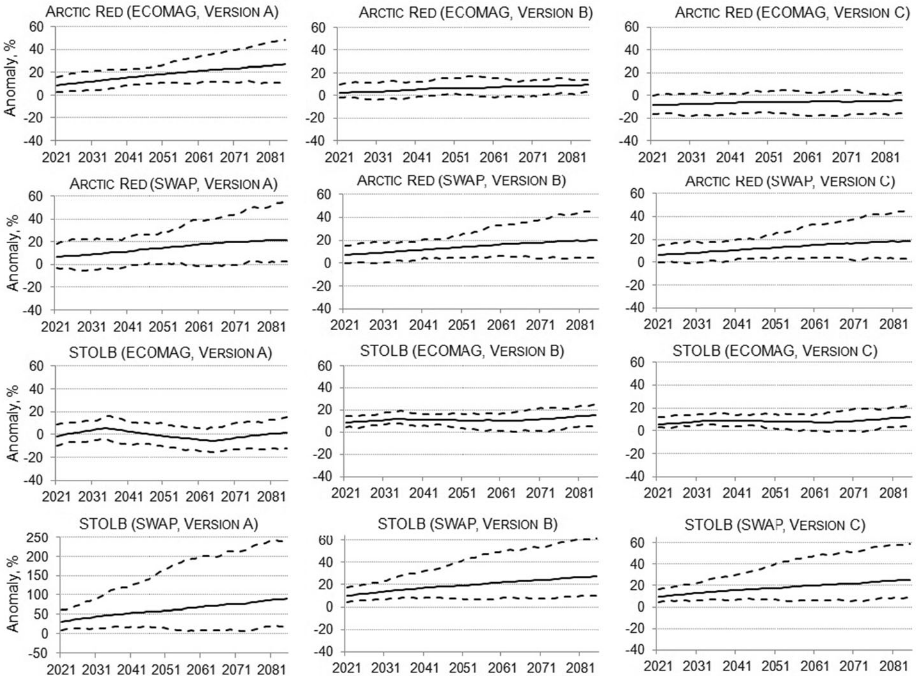

Fig. 5 Annual flow anomaly projections smoothed by the moving average technique with a 30-year sliding window: solid line - ensemble mean; dashed line - ensemble spread

CE-test). The wide uncertainty band obtained for the SWAP-based A version projections of the Lena River flow is due to the fact that this model underestimated the historical mean flow very much. Second, the ensemble means of flow projections simulated by the A versions differ from the corresponding projections of the B and C versions, though this difference is not so marked. For example, at the end of the century, the ECOMAG-based A version simulates 2% changes in the mean annual discharge at the Lena River basin outlet, while the B and C versions simulate 15% and 12%, respectively. The corresponding SWAP-based projections are 91%, 27%, and 28%. For the Mackenzie, the ECOMAG-based A version projects 27% changes, while the B and C versions project 9% and − 4%, respectively. The corresponding SWAP-based projections are 21%, 20%, and 18%.

The main features of the flow anomaly projections for the other gauges (see Figs. S3–S6) do not contradict, as a whole, those detected for the basin outlets. Again, the difference between the B and C versions is smaller than between them and the A versions. Indeed, at the end of the century, the ECOMAG- and SWAP-based A versions simulate, on the average, 43% change in the mean annual discharge for all four gauges of the Lena basin, while the B and C versions simulate 24% and 19%, respectively. The corresponding uncertainty bands are 125%, 63%, and 37%, respectively. For all streamflow gauges within the Mackenzie basin, mean annual runoff anomaly simulated by the A versions is 15% on the average, while the B and C versions give 8% and 2%, respectively. The corresponding uncertainty bands are 42%, 30%, and 36%, respectively. Thus, the mean difference between the end-century annual runoff anomalies projected by the B and C versions equals 5% for both basins, while the mean difference between these anomalies and ones projected by the A versions is 11–22%.

However, there are projection peculiarities whose explanation requires further and more thorough analysis.

Summarizing the obtained results, we can say that the similarity of the projections simulated by the calibrated and successfully evaluated B and C model versions, together with the difference of these projections from those simulated by the A versions, which failed the evaluation tests, gives us reason to believe that the B and C projections of future mean annual flows are more credible than the A projections. Herewith, the credibility of the projections simulated by the C versions with their multi-site calibration and better evaluation results is assumed to be higher than that of the B version projections.

5 Conclusions

The two models—ECOMAG and SWAP—were used to reveal whether a model which passed a comprehensive evaluation test is more suitable for hydrological projections as compared with a model that failed the test. For this purpose, three versions of each model were considered: the non-calibrated model version (version A), the model version calibrated against observed streamflow at the basin outlet (version B) and the one calibrated against observed streamflow at several gauges (version C). All the versions have undergone the comprehensive evaluation test, which included the following four steps: (1) multi-site evaluation of the model performance in terms of daily and monthly hydrograph simulation, (2) the model robustness test evaluating the model ability to perform similarly under the contrasting climate conditions, (3) evaluation of the model skill in simulating the high/low flow indices, and (4) evaluation of the model skill in reproducing the observed trends (or absence of the trends) in the annual streamflow series.

The main findings can be summarized as follows:

1. The non-calibrated ECOMAG- and SWAP-based A versions have not passed the proposed comprehensive evaluation test for both the basin outlet streamflow gauges and the interior basin gauges. The A versions have failed all the specific tests, based on both daily and monthly data, included into the CE-test. The only test, which the A versions have successfully passed together with the B and C versions, was done for detecting lack of trends in the annual runoff.
2. The model calibration has led to a significant improvement of the model ability to reproduce the historical streamflow and almost all hydrological indicators referenced in this paper (except the low flow index). Regardless of the origin (either the ECOMAG or the SWAP), the B and C versions were more successful in passing the CE-test than the A versions. Herewith, the multi-site-calibrated C versions were slightly or notably better than the single-site-calibrated B versions, especially at the monthly time scale.
3. On the basis of the test results, the B and C versions were considered the more preferable candidates for the impact study than the A versions, which failed the tests, and the C versions were assumed as the most preferable. The limits of the model applicability were established: the C versions performed better and more robustly than the B versions in the simulation of monthly hydrographs except a few gauges depending on the basin and the version origin.
4. Comparison of the future flow anomalies projected by the different model versions has demonstrated that the projections simulated by the B and C versions turned out to be closer to each other than to the corresponding projections simulated by the A versions. This

similarity is manifested in either the mean ensemble changes or the projection uncertainty or both indicators. The averaged over all gauge difference between the end-century annual runoff anomalies projected by the B and C versions equals to 5%, while the average difference between these anomalies and ones projected by the A versions is 16%. For the projection uncertainties, the corresponding estimates are 16% and 51%. Thus, difference (in terms of the mean ensemble projections and their uncertainties) between the A versions and other versions is nearly 3 times larger than the difference between the B and C versions.

5. The CE-test results together with the similarity of the B and C projections and difference of these projections from those simulated by the A versions give us reason to believe that the former projections are more credible than the latter. Making this conclusion, we proceed from the logical statement: if one model set performs better against the historical data than the other model set and if the projections simulated by the first models are notably different from the second set projections, then the first projections are more credible. We assume that the credibility of the projections simulated by the C versions with better evaluation results is higher than that of the B version projections, as calibration at intermediate gauges increases the projection credibility in comparison with the single-site calibration.
6. Thus, under the study conditions (used models, studied basins), we answer “yes” to the question posed in the title of the paper. We emphasize that we formulated this conclusion with the help of the models, which significantly differ from each other in the model structure, the parameters, forcing datasets used, and in terms of the calibration and evaluation periods. The similarity of the results of such different models allows us to generalize the conclusion obtained and to recommend the described procedure for testing models aimed at impact study.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10584-020-02930-z.

Acknowledgments The authors are grateful to three anonymous reviewers and guest editor Dr. V. Krysanova for their critical comments. Also, we would like to thank Dr. V. Krysanova for constructive discussions that motivated us to think about the study issue far before the earliest draft of the paper, as well as for her invitation to contribute in the Climatic Change Special Issue.

The numerical experiments were designed within the framework of the State Assignment theme № 0147-2019-0001.

The present work was carried out within the framework of the Panta Rhei Research Initiative of the International Association of Hydrological Sciences (IAHS).

Funding Simulations by the SWAP model were financially supported by the Russian Science Foundation (Grant 16-17-10039). Simulations by the ECOMAG model were financially supported by the Russian Science Foundation (Grant 19-17-00215). Analysis of hydrological projections and their uncertainties was financially supported by the Ministry of Science and Higher Education of the Russian Federation (Grant MK-1753.2020.5).

References

Andréassian V, Perrin C, Berthet L et al (2009) Crash tests for a standardized evaluation of hydrological models. Hydrol Earth Syst Sci 13(10):1757–1764. https://doi.org/10.5194/hess-13-1757-2009

Bartholomé E, Belward A (2005) GLC2000: a new approach to global land cover mapping from Earth observation data. Int J Remote Sens 26(9):1959–1977

Birhanu D, Kim H, Jang C, Park S (2018) Does the complexity of evapotranspiration and hydrological models enhance robustness? Sustainability 10:2837. https://doi.org/10.3390/su10082837

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/10.2016/j.jhydrol.2012.11.012

Champeaux JL, Masson V, Chauvin F (2005) ECOCLIMAP: a global database of land surface parameters at 1 km resolution. Meteorol Appl 12(1):29–32. https://doi.org/10.1017/S1350482705001519

Coron L, Andréassian V, Perrin C et al (2012) Crash testing hydrological models in contrasted climate conditions: an experiment on 216 Australian catchments. Water Resour Res 48(W05552). https://doi.org/10.1029/2011WR011721

Coron L, Andréassian V, Perrin C et al (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 18:727–746. https://doi.org/10.5194/hess-18-727-2014

Cosby B, Hornberger GM, Clapp RB, Ginn TR (1984) A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils. Water Resour Res 20(6):682–690. https://doi.org/10.1029/WR020i006p00682

Duan Q, Schaake J, Andréassian V et al (2006) Model Parameter Estimation Experiment (MOPEX): an overview of science strategy and major results from the second and third workshops. J Hydrol 320(1–2):3–17. https://doi.org/10.1016/j.jhydrol.2005.07.031

Fischer G, Velthuizen H, Shah M, Nachtergaele F (2008) Global agro-ecological zones assessment for agriculture (GAEZ 2008) IIASA. Laxenburg and FAO, Austria and Rome

Gelfan A, Gustafsson D, Motovilov Y et al (2017) Climate change impact on the water regime of two great Arctic rivers: modeling and uncertainty issues. Clim Chang 141(3):499–515. https://doi.org/10.1007/s10584-016-1710-5

Gelfan A, Millionshchikova T (2018) Validation of a hydrological model intended for impact study: problem statement and solution example for Selenga River basin. Water Res 45(S1):90–101. https://doi.org/10.1134/S0097807818050354

Gupta HV, Kling H, Yilmaz KK et al (2009) Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J Hydrol 377:80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gusev YM, Nasonova ON (1998) The land surface parameterization scheme SWAP: description and partial validation. Glob Planet Chang 19(1–4):63–86

Gusev YM, Nasonova ON (2003) Modelling heat and water exchange in the boreal spruce forest by the land-surface model SWAP. J Hydrol 280(1–4):162–191

Gusev EM, Nasonova ON, Dzhogan LY (2006) The simulation of runoff from small catchments in the permafrost zone by the SWAP model. Water Res 33(2):115–126. https://doi.org/10.1134/S0097807806020011

Gusev EM, Nasonova ON, Dzhogan LY (2016) Physically based modeling of many-year dynamics of daily streamflow and snow water equivalent in the Lena R. basin. Water Res 43(1):21–32. https://doi.org/10.1134/S0097807816010085

Gusev YM, Nasonova ON, Kovalev EE, Aizel GV (2018) Modelling river runoff and estimating its weather-related uncertainty for 11 large-scale rivers located in different regions of the globe. Hydrol Res 49(4):1072–1087. https://doi.org/10.2166/nh.2017.015

Huang S, Kumar R, Flörke M et al (2017) Evaluation of an ensemble of regional hydrological models in 12 large-scale river basins worldwide. Clim Chang 141:381–397. https://doi.org/10.1007/s10584-016-1841-8

Kalugin AS, Motovilov YG (2018) Runoff formation model for the Amur River basin. Water Res 45(2):149–159. https://doi.org/10.1134/S0097

Klemeš V (1986) Operational testing of hydrological simulation models. Hydrol Sci J 31:13–24

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

Kundzewicz ZW (1986) The hydrology of tomorrow. Hydrol Sci J 31(2):223–235

Kundzewicz ZW, Krysanova V, Benestad RE et al (2018) Uncertainty in climate change impacts on water resources. Environ Sci Pol 79:1–8. https://doi.org/10.1016/j.envsci.2017.10.008

Lange S (2018) Bias correction of surface downwelling longwave and shortwave radiation for the EWEMBI dataset. Earth Sys Dynam 9(2):627–645

Merz R, Parajka J, Blöschl G (2011) Time stability of catchment model parameters: implications for climate impact analyses. Water Resour Res 47(W02531). https://doi.org/10.1029/2010WR009505

Moriasi DN, Zeckoski RW, Arnold JG et al (2015) Models: performance measures and evaluation criteria. Trans ASABE 58(6):1763–1785. https://doi.org/10.13031/trans.58.10715

Motovilov YG (2016) Hydrological simulation of river basins at different spatial scales: 1. Generalization and averaging algorithms. Water Res 43(3):429–437. https://doi.org/10.1134/S0097807816030118

Motovilov YG, Gottschalk L, Engeland K, Rodhe A (1999) Validation of a distributed hydrological model against spatial observations. Agric For Meteorol 98-99:257–277. https://doi.org/10.1016/S0168-1923(99)00102-1

Refsgaard JC, 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 JC, Madsen H, Andréassian V et al (2013) A framework for testing the ability of models to project climate change and its impacts. Clim Chang 122:271–282. https://doi.org/10.1007/s10584-013-0990-2

Santer BD, Wigley TML, Boyle JS et al (2000) Statistical significance of trends and trend differences. J Geophys Res 105(D6):7337–7356. https://doi.org/10.1029/1999JD901105

Schlosser CA, Slater AG, Robock A et al (2000) Simulations of a boreal grassland hydrology at Valdai, Russia: PILPS phase 2(d). Mon Weather Rev 128(2):301–321

Seibert J (2003) Reliability of model predictions outside calibration conditions. Nord Hydrol 34:477–492. https://doi.org/10.2166/nh.2003.0019

Shi X, Qin T, Nie H et al (2019) Changes in major global river discharges directed into the ocean. J Environ Res Publ Health 16:1469. https://doi.org/10.3390/ijerph16081469

Shiklomanov AI et al (2006) Cold region river discharge uncertainty estimates from large Russian rivers. J Hydrol 326:231–256

Smith MB, Seo DJ, Koren VI et al (2004) The distributed model intercomparison project (DMIP): motivation and experiment design. J Hydrol 298(1–4):4–26. https://doi.org/10.1016/j.jhydrol.2004.03.040

Tananaev NI, Makarieva OM, Lebedeva LS (2016) Trends in annual and extreme flows in the Lena River basin, Northern Eurasia. Geophys Res Lett 43(10):764–772. https://doi.org/10.1002/2016GL070796

Thirel G, Andréassian V, Perrin C et al (2015) 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

Vaze J, Post DA, Chiew FHS et al (2010) Climate nonstationarity – validity of calibrated rainfall-runoff models for use in climatic changes studies. J Hydrol 394(3–4):447–457. https://doi.org/10.1016/j.jhydrol.2010.09.018

Vormoor K, Heistermann M, Bronstert A, Lawrence D (2018) Hydrological model parameter (in)stability – “crash testing” the HBV model under contrasting flood seasonality conditions. Hydrol Sci J 63(7):991–1007. https://doi.org/10.1080/02626667.2018.1466056

Wagener T, McIntyre M, Lees MJ et al (2003) Towards reduced uncertainty in conceptual rainfall-runoff modeling: dynamic identifiability analysis. Hydrol Proced 17:455–476

Woo MK, Thorne R (2014) Winter flows in the Mackenzie drainage system. Arctic 67:238–256

Xu C (1999) Operational testing of a water balance model for predicting climate change impacts. Agric For Meteorol 98-99:295–304. https://doi.org/10.1016/S0168-1923(99)00106-9

Yang D et al (2005) Bias-corrections of long-term (1973-2004) daily precipitation data over the northern regions. Geophys Res Lett 32:L19501. https://doi.org/10.1029/2005GL024057

Yilmaz KK, Gupta HV, Wagener T et al (2008) A process-based diagnostic approach to model evaluation: application to the NWS distributed hydrologic model. Water Resour Res 44(W09417). https://doi.org/10.1029/2007WR006716

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

Affiliations

Alexander Gelfan[1,2] • **Andrey Kalugin**[1] • **Inna Krylenko**[1,2] • **Olga Nasonova**[1] • **Yeugeniy Gusev**[1] • **Evgeny Kovalev**[1]

✉ Alexander Gelfan
hydrowpi@mail.ru

[1] Water Problems Institute of Russian Academy of Sciences, Moscow, Russia

[2] Faculty of Geography, Lomonosov Moscow State University, Moscow, Russia