Climatic Change (2020) 163:1329–1351
https://doi.org/10.1007/s10584-020-02892-2

Performance evaluation of global hydrological models in six large Pan-Arctic watersheds

Anne Gädeke[1] · Valentina Krysanova[1] · Aashutosh Aryal[1] · Jinfeng Chang[2,3,4] · Manolis Grillakis[5,6] · Naota Hanasaki[7] · Aristeidis Koutroulis[5] · Yadu Pokhrel[8] · Yusuke Satoh[3,7] · Sibyll Schaphoff[1] · Hannes Müller Schmied[9,10] · Tobias Stacke[11] · Qiuhong Tang[12] · Yoshihide Wada[3] · Kirsten Thonicke[1]

Received: 15 January 2020 / Accepted: 12 October 2020 / Published online: 24 November 2020

Abstract

Global Water Models (GWMs), which include Global Hydrological, Land Surface, and Dynamic Global Vegetation Models, present valuable tools for quantifying climate change impacts on hydrological processes in the data scarce high latitudes. Here we performed a systematic model performance evaluation in six major Pan-Arctic watersheds for different hydrological indicators (monthly and seasonal discharge, extremes, trends (or lack of), and snow water equivalent (SWE)) via a novel Aggregated Performance Index (API) that is based on commonly used statistical evaluation metrics. The machine learning Boruta feature selection algorithm was used to evaluate the explanatory power of the API attributes. Our results show that the majority of the nine GWMs included in the study exhibit considerable difficulties in realistically representing Pan-Arctic hydrological processes. Average $API_{discharge}$ (monthly and seasonal discharge) over nine GWMs is > 50% only in the Kolyma basin (55%), as low as 30% in the Yukon basin and averaged over all watersheds $API_{discharge}$ is 43%. WATERGAP2 and MATSIRO present the highest ($API_{discharge}$ > 55%) while ORCHIDEE and JULES-W1 the lowest ($API_{discharge}$ ≤ 25%) performing GWMs over all watersheds. For the high and low flows, average $API_{extreme}$ is 35% and 26%, respectively, and over six GWMs API_{SWE} is 57%. The Boruta algorithm suggests that using different observation-based climate data sets does not influence the total score of the APIs in all watersheds. Ultimately, only satisfactory to good performing GWMs that effectively represent cold-region hydrological processes (including snow-related processes, permafrost) should be included in multi-model climate change impact assessments in Pan-Arctic watersheds.

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

✉ Anne Gädeke
a.gaedeke@gmail.com

Extended author information available on the last page of the article

Keywords Global Water Models · Model performance · Model evaluation · Arctic watersheds · Boruta feature selection

1 Introduction

The rapid environmental changes occurring in the Pan-Arctic have triggered increased attention from the scientific community. Such changes include observed decreasing extent and duration of snow cover (Pulliainen et al. 2020), permafrost thaw (Biskaborn et al. 2019), and related changes in soil active layer depth (Walvoord and Kurylyk 2016), increased melting rates of glaciers (Zemp et al. 2019), and changing partitioning of surface and groundwater (Walvoord and Striegl 2007), all of which affect the hydrological processes in Pan-Arctic watersheds. In addition, increasing discharge and subsequent freshwater transport to the Arctic Ocean have been documented (Ahmed et al. 2020), which impact bio-geophysical processes such as sea ice growth (Morison et al. 2012) and ocean circulation (Holliday et al. 2020). The observed changes, and more importantly their rate of change, have the potential for strong feedbacks to terrestrial ecosystems, the global climate system (McGuire et al. 2018; Post et al. 2019), and global freshwater circulation (Bring et al. 2016). Despite the increased scientific attention, our current understanding of the hydrologic cycle in the high latitudes and its linkages to other parts of the earth system still remains limited.

Pan-Arctic hydrological processes are largely controlled by the presence of permafrost, the strong climate seasonality, and the wide fluctuations in surface energy balance (Ge 2013). Annual peak discharge generally occurs following snowmelt, which presents the major hydrological event in Pan-Arctic watersheds, and is often associated with large-scale flooding (Bowling et al. 2003). Most of the snowmelt becomes overland flow as the ground is still frozen constraining infiltration. Hydrological processes in the Pan-Arctic are highly susceptible to climate change, particularly due to the freezing point threshold. To increase our understanding of Pan-Arctic hydrological processes, Global Water Models (GWMs), here including Global Hydrological Models (GHMs), Land Surface Models (LSMs), and Dynamic Global Vegetation model (DGVMs), could provide valuable tools for obtaining estimates of hydrological variables where data availability is poor both spatially and temporally. GWMs simulate the entire water cycle and make use of globally available datasets. Thereby, GWMs can complement the sparse observation records and support climate change impact assessments. A thorough performance evaluation is essential prior to applying models for climate change impact assessments in this region.

Previous model evaluation studies focusing on the Pan-Arctic differ from ours in terms of (i) the number and type of GWMs included, (ii) the spatial area/watersheds covered, (iii) the hydrological indicator(s) analyzed, and (iv) evaluation methods. Slater et al. (2007), for example, evaluated the performance of five LSMs for the period 1980–2001 across the Pan-Arctic drainage system including the Lena, Yenisei, Ob, and Mackenzie watersheds. Their results show that large differences in model performance exist across LSMs in terms of snow hydrological processes, water balance partitioning, discharge seasonality, and baseflow. Similarly, Andresen et al. (2019) found that LSMs tend to agree on decadal discharge trends but underestimate discharge volume when compared to gauge data across the major Arctic watersheds. Zaherpour et al. (2018) highlight the difficulty of GWMs in capturing the timing of the seasonal discharge cycle in northern regions effectively. In a multi-model evaluation

study of daily runoff estimates, Beck et al. (2017) found that uncalibrated GWMs outperform, on average, uncalibrated LSMs in snow-dominated regions.

Global to continental scale multi-model climate change impact assessments are generally performed with GWMs disregarding model performance under historical conditions (e.g., Gosling et al. (2017)). A central tendency of the multi-model ensemble (mean or median) is often assumed as a good predictor due to large variations in performance of individual models and in their projections. Zaherpour et al. (2018) used a novel integrated evaluation method to show, however, that the ensemble mean fails to outperform best individual models for different hydrological indicators that represent mean and extreme discharge conditions. Therefore, using the ensemble mean and not carrying out a thorough model performance evaluation is not recommended.

Krysanova et al. (2018) proposed guidelines consisting of 5 steps for effective evaluation of GHMs to be used prior to climate change impact assessments. Such a thorough model evaluation may suggest applying weighting coefficients to individual models in order to constrain the ensemble to the best performing members instead of using the ensemble mean approach (see Krysanova et al. (2020)). Thereby, confidence in projected impacts under climate change may potentially be increased.

The objective of our study is to contribute to the understanding of how GWMs, LSMs, and a DGVM perform in Pan-Arctic watersheds for different hydrological indicators, including monthly and seasonal discharge, extremes, trends (or lack of), and snow water equivalent (SWE), evaluated via a novel "Aggregated model Performance Index" (API). To reach this objective, we, firstly, systematically evaluated the performance of five global GHMs, three LSMs, and one DGVM using commonly used statistical evaluation metrics for six large watersheds in the Pan-Arctic based on the guidelines for GHM evaluation by Krysanova et al. (2018). After that, we assigned rating scores to each hydrological indicator based on thresholds defined for the statistical evaluation metrics. We calculated three APIs in total: $API_{discharge}$, $API_{extreme}$, and API_{SWE}. The API combines the rating scores for every hydrological indicator in one index. We also applied the machine learning feature selection algorithm Boruta to evaluate the explanatory power of the API attributes (climate forcing, GWM, hydrological indicators, etc.). Our approach is easily interpretable and transferable to other model evaluations and inter-comparisons, and has a potential to deliver more robust multi-model climate change impact assessments.

2 Methods

2.1 Overview of study basins

The six largest watersheds located in the Pan-Arctic serve as a study area for the multi-model GWM performance evaluation: Kolyma, Lena, Yenisei, Ob, Mackenzie, und Yukon (Fig. 1, Table 1). Watershed sizes range between 526,000 and 2,950,000 km^2. The combined discharge from these watersheds is the single largest freshwater source to the Arctic Ocean (Yukon via the Bering Strait). Permafrost covers large parts of the studied watersheds (Fig. 1). Total permafrost coverage, which includes proportions of continuous, discontinuous, sporadic, and/or isolated permafrost, ranges between 34 (Ob) and 100% (Kolyma and Lena). Continuous permafrost covers only 3% in the Ob but the entire (100%) Kolyma watershed (Table 1,

based on Brown et al. (1997)). In the northern, continuous permafrost zone, tundra vegetation dominates, while boreal forests are characteristic for the southern, mostly discontinuous permafrost zone. The climate ranges from polar in the high latitudes to subpolar and continental towards the lower latitudes. Arctic rivers are generally ice-covered for longer than 6 months of the year. Snow covers the Arctic landscapes for most of the year (e.g., 8 months in Arctic Alaska (end of September to May)) and contains a considerable amount of the total annual precipitation at the end-of-winter (Kane et al. 1991). Consequently, snow hydrological processes play an important role in the Arctic hydrological cycle. Population density is low in the study area (Kummu and Varis 2011).

2.2 Models and data

Model evaluation was based on measured discharge at 18 gaging stations (two to four gauging stations in each watershed: Fig. 1, Table 1) and estimates of SWE. Discharge measurements were retrieved from "The Global Streamflow Indices and Metadata Archive" (GSIM) (Do et al. 2018; Gudmundsson et al. 2018). Additionally, daily discharge data, used for the extreme discharge analysis, was provided by GRDC (Global Runoff Data Centre, 56068 Koblenz, Germany) at the outlet stations (highlighted in italics in Table 1). Estimates of total monthly

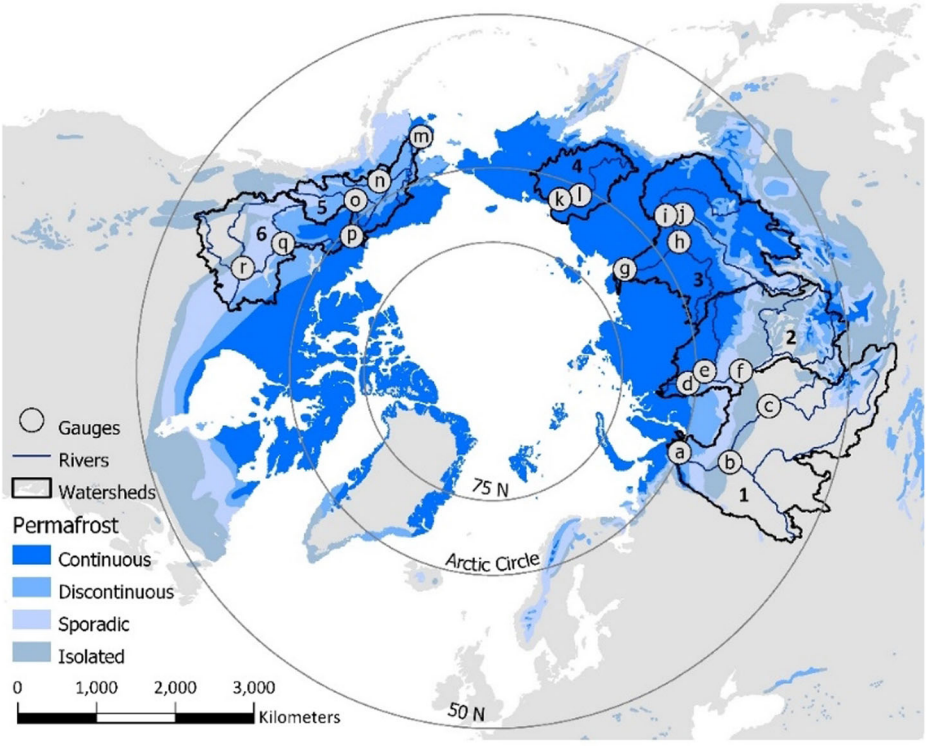

Fig. 1 Overview of study area including watershed outlines, gauges, and permafrost extent and type (Brown et al. 1997). Watersheds (number) and gauging stations (letters) are detailed in Table 1

Table 1 Study area details and gauging stations used (italicized ones represent the outlet/most downstream stations). Permafrost coverage in each watershed was calculated based on permafrost extent in Brown et al. (1997). Total permafrost coverage includes continuous, discontinuous, isolated, and sporadic permafrost. The locations (longitude and latitude of the gauging stations as represented in the models is displayed in Table S1 in the supplementary material). Watershed number and gauging station letter are in accordance with Fig. 1

Watersheds (numbered as in Fig. 1)	Countries	Basin area (M km^2)	Gauging stations (river)	Total permafrost coverage (continuous permafrost coverage) (%)
1 Ob	Russia, Kazakhstan, China, Mongolia	2.95	*a) Salekhard (Ob)* b) Hanti-Mansisk (Irtysh) c) Kolpashevo (Ob)	34 (3)
2 Yenisei	Russia Mongolia	2.4	*d) Igarka (Yenisei)* e) Bol. Porog (Nizhnyaya Tunguska) f) Pod. Tunguska (Yenisei)	90 (33)
3 Lena	Russia	2.43	*g) Kusur (Lena)* h) Hatyrik-Homo (Vilyuy) i) Verkhoyanski Perevoz (Aldan) j) Tabaga (Lena)	100 (80)
4 Kolyma	Russia	0.53	*k) Kolymskaya (Kolyma)* l) Sredne-Kolymsk (Kolyma)	100 (100)
5 Yukon	Canada USA	0.83	*m) Pilot Point AK (Yukon)* n) Nenana AK (Tanana) o) Eagle AK (Yukon)	99 (23)
6 Mackenzie	Canada	1.66	*p) Arctic Red River (Mackenzie)* q) Fort Simpson (Mackenzie) r) Peace Point Alberta (Peace)	83 (15)

SWE were obtained from the remote sensing product GlobSnow-2 (Metsämäki et al. 2015) for the period 1980–2000. The SWE estimates were produced using a combination of passive microwave radiometer and ground-based weather station data.

Model performance is evaluated for nine GWMs (4 GHMs, 4 LSMs, 1 DGVM) that participated in the global water sector of ISIMIP2a (Gosling et al. 2019): the GHMs WaterGAP2, H08, MPI-HM, PCR-GLOBWB, the DGVM LPJmL, and the LSMs DBH, JULES-W1, MATSIRO, and ORCHIDEE (here all referred to as GWMs). The participating GWMs and their main characteristics are detailed in Table 2. The simulations are based on a common modeling protocol (ISIMIP2a 2018) which guarantees, as far as possible, consistent spatial (0.5°) and temporal model resolution as well as input and output datasets. All GWMs simulate the major global terrestrial hydrological processes, though using different algorithms and mathematical formulae (Table 2). Simulated daily discharge was available from all nine GWMs, and total monthly snow water equivalent (SWE) from six GWMs. MATSIRO, JULES-W1, and LPJmL represent permafrost temperatures and soil freeze and thaw processes that affect hydrological processes such as infiltration and water flow through permafrost. Three other GWMs (WaterGAP2, PCR-GLOBWB, MPI-HM) present permafrost coverage statically (fixed in space/time, by, e.g., reducing the maximum water holding capacity of the soil) without dynamic feedbacks/linkages to hydrology. We evaluated the simulations that do not consider the human influences on the water cycle, such as irrigation and dams. Apart from WaterGAP2, the GWMs were not calibrated. The calibration of WaterGAP2 solely focused on matching average long-term annual observed discharge by varying up to three parameters. Additional information can be found in the respective model description papers (Table 2) and

Table 2 Main characteristics of the participating Global Water Models (GWMs), including model type (Global Hydrological Model (GHM), Land Surface Model (LSM), Dynamic Global Vegetation Model (DGVM)). Input climate variables include precipitation (P), mean air temperature (T), maximum air temperature (Tmax), minimum air temperature (Tmin), longwave downward radiation (LW), longwave net radiation (LWnet), shortwave downward radiation (SW), relative humidity (Q), surface pressure (SP), near-surface wind speed (W), snowfall rate (S), potential evapotranspiration (PET). For MPI-HM, potential evapotranspiration was computed during pre-processing based on LW, SW, T, W, SP, and Q. The river network is based on the 30′ global drainage direction map DDM30 (Döll and Lehner 2002)

Model name (type)	Input climate variables	Discharge routing (river network)	PET method	Snow scheme	Variables used	Dynamic Permafrost
WaterGAP2 (GHM) (Müller Schmied et al. 2016)	P, T, LW, SW	Linear reservoir (DDM30)	Priestley Taylor	Degree-day	Q, SWE	No*
DBH (GHM) (Tang et al. 2007)	Tmax, T, Tmin, LW, Q, SW, SP, P	Linear reservoir (DDM30)	Energy balance	Energy balance	Q, SWE	No
H08 (GHM) (Hanasaki et al. 2008)	T, LW, W, SW, S, SP, P	Linear reservoir (DDM30)	Bulk Formula	Energy balance	Q	No
MPI-HM (GHM) (Stacke and Hagemann 2012)	T, P	Linear reservoir (DDM30)	Penman-Monteith	Degree-day	Q, SWE	No*
PCR-GLOBWB (GHM) (Wada et al. 2014)	T, P, PET	Travel time routing	Hamon	Degree-day	Q, SWE	No*
LPJmL (DGVM) (Schaphoff et al. 2013; Sitch et al. 2003)	T, LWnet, SW, P	Linear reservoir (DDM30)	Energy balance	Energy balance	Q, SWE	Yes
MATSIRO (LSM) (Pokhrel et al. 2015)	T, LW, Q, SW, S, SP, P	TRIP (Oki et al. 1999)(DDM30)	Bulk Formula	Energy balance	Q, SWE	Yes
ORCHIDEE (LSM) (Traore et al. 2014)	Tmax, Tmin, LW, W, Q, SW, SP, P	Linear reservoir (DDM30)	Penman-Monteith	Energy balance	Q	No
JULES-W1 (LSM) (Best et al. 2011)	Tmax, T, Tmin, LW, W, Q, SW, SP, P	CaMaFlood Routing Model	Penman-Monteith	Energy balance	Q	Yes

*Models present permafrost coverage statically (fixed in space/time, by, e.g., reducing the maximum water holding capacity of the soil) without dynamic feedbacks/linkages to hydrology

for six of the GWMs (WaterGAP2, DBH, H08, PCR-GLOBWB, MATSIRO, LPJmL) in the supplementary material of Zaherpour et al. (2018).

Four common observation-based climate datasets forced the GWMs (Table S2): Global Soil Wetness Project Phase 3 (GSWP3), Princeton, WATCH, and WFDEI. Müller Schmied et al. (2016) provided a more detailed description of the climate forcing datasets for hydrological studies. The GWM JULES-W1, however, provided simulation results for only three of the four climate forcing data (WATCH missing). In total, 35 model simulation combinations (4 forcing data sets for 8 GWMs and 3 forcing data sets for JULES-W1) were available for the hydrological model performance assessments.

2.3 Hydrological indicators

The hydrological indicators used in this study are detailed in Table 3. Monthly discharge, long-term mean monthly discharge (seasonal dynamics), and mean annual discharge were computed based on measured monthly and simulated daily discharge records. For 10 gauging stations, which include the outlet stations, the analysis period covers 30 years (1971–2000). For the remaining eight stations, the measured discharge record is shorter (between 20 and 29 years, Table S3). The calculation of the flow percentiles for high and low flows (based on daily measured and simulated discharge) is based on the daily 30-year record and was limited to the outlet gauging stations due to the data availability. Estimated (GlobSnow-2) and simulated total monthly SWE data were used to calculate long-term total monthly SWE (1980–2000).

2.4 Evaluating model performance

Our GWM performance evaluation approach for Pan-Arctic watersheds, based on guidelines provided in Krysanova et al. (2018), is summarized in Fig. 2. The model performance was evaluated for 14 different hydrological indicators (Table 3) at different locations within the watersheds in order to check internal consistency of the simulated hydrological processes.

Table 3 Overview of the hydrological indicators used in this study and the statistical evaluation metrics applied (*NSE* Nash Sutcliffe Efficiency, *PBIAS* percent bias, *SD* standard deviation). Discharge related indicators were calculated for the time period 1971–2000 (or shorter, depending on data availability, see Table S3) and SWE for 1980–2000. The indicators monthly, seasonal, and annual discharge were evaluated at 18 gaging stations and the extremes at 6 gauging stations (outlets, Table 1). Seasonal SWE was evaluated at 4 points in each watershed (24 in total, locations defined in Table S4)

Indicator abbreviation	Description of indicator	Statistical evaluation metrics
Monthly	Monthly discharge	NSE, PBIAS
Seasonal	Long-term mean monthly discharge (seasonal dynamics of discharge)	NSE, BIAS in SD
Annual	Mean annual discharge	Linear trend analysis
Q_{10}, Q_5, Q_1, $Q_{0.1}$, $Q_{0.01}$	The magnitude of daily discharge that is exceeded 10%, 5%, 1%, 0.1%, and 0.01% of the time in the daily time series of 30 years (indicator of high flow)	PBIAS
Q_{90}, Q_{95}, Q_{99}, $Q_{99.9}$, $Q_{99.99}$	The magnitude of daily discharge that is exceeded 90%, 95%, 99%, 99.9%, and 99.99% of the time in the daily time series of 30 years (indicator of low flow)	PBIAS
Seasonal SWE	Long-term total monthly snow water equivalent (seasonal dynamics of SWE)	NSE, BIAS in SD

Three different APIs ($API_{discharge}$, $API_{extreme}$, API_{SWE}) were developed based on assigning individual rating scores considering threshold values of the statistical evaluation metrics (Table 4). A rating score of 1 is associated with good model performance, 0.5 with weak/satisfactory, and 0 with poor model performance. The statistical evaluation metrics used include percent bias (PBIAS), bias in standard deviation (bias in SD), and Nash and Sutcliffe Efficiency (NSE) (Nash and Sutcliffe 1970). The NSE (Eq. (1) in the supplementary material), a dimensionless model efficiency criterion, assesses overall model fit and is not very sensitive towards over- and underestimation (details in Krause et al. (2005)). Therefore, the monthly discharge performance evaluation was complemented by PBIAS (Eq. (2) in the supplementary material). The bias in SD (Eq. (3) in the supplementary material) assesses the standard deviation of the mean annual cycle between measured and simulated time series (MMD) and is therefore a suitable metric to evaluate model performance in terms of reproducing the seasonality (amplitude). The thresholds for the statistical evaluation metrics were initially oriented on widely used recommendations by Moriasi et al. (2007) and Moriasi et al. (2015) and by considering suggestions of Krysanova et al. (2018). In this study, the thresholds for the statistical performance were adjusted that means we made them less strict for GWMs. For example, the NSE and PBIAS thresholds in Moriasi et al. (2015) for good performance of monthly runoff in hydrological models are $NSE \geq 0.70$ and $PBIAS < \pm 10$, and for satisfactory performance, $NSE > 0.55$ and $\pm 10 \leq PBIAS \leq \pm 15$. In this study, we defined a good model performance of monthly runoff simulated by GWMs when $NSE \geq 0.5$ and PBIAS is within ± 25%. Table 4 details the thresholds defined for this study.

The $API_{discharge}$ consists of four different statistical evaluation metrics: $NSE_{monthly}$ and PBIAS were calculated for monthly discharge, and $NSE_{seasonal}$ and BIAS in SD for long-

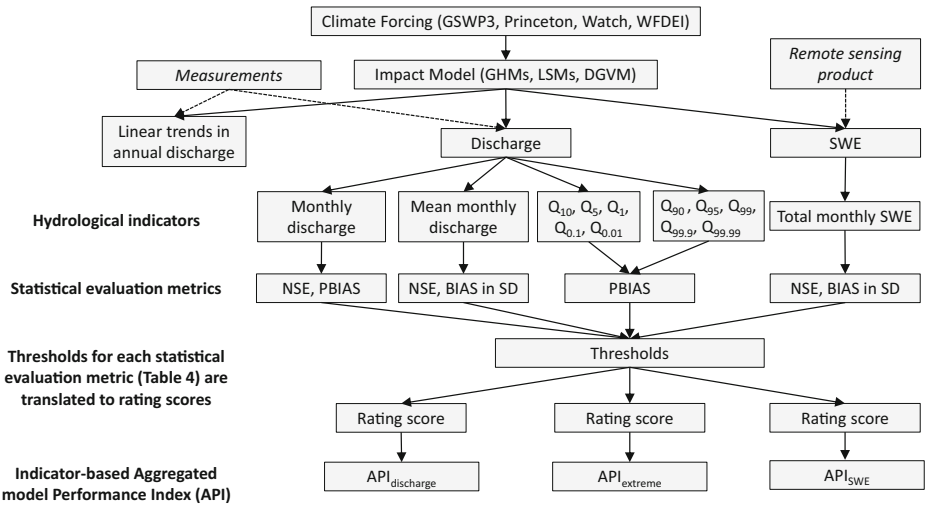

Fig. 2 Overview of study approach: a set of hydrological indicators were calculated based on observed (measured discharge, SWE from remote sensing product) and simulated discharge/SWE. Statistical evaluation metrics (NSE: Nash-Sutcliffe Efficiency, PBIAS: percent bias, and bias in SD (standard deviation)) are used to evaluate model performance for each hydrological indicator. Based on threshold values for each statistical evaluation metrics, rating scores are assigned for each climate forcing/model/gauging station/indicator for good, weak, and poor performance. The individual scores are aggregated to obtain an overall aggregated performance index (API). Aggregation is carried out separately for hydrological indicators related to monthly discharge, long-term mean monthly discharge, extremes (high and low flows), and SWE

Table 4 Rating scores and thresholds used for the statistical performance criteria. Discharge was analyzed in terms monthly ($NSE_{monthly}$, PBIAS) and long-term mean monthly (seasonal dynamics, $NSE_{seasonal}$, and BIAS in SD) temporal resolution. Snow water equivalent (SWE) was only evaluated for long-term mean monthly (seasonal dynamics, $NSE_{seasonal}$, and bias in SD) temporal resolution. A rating score of 1 corresponds to a good, 0.5 a weak/satisfactory, and 0 to a poor performance. The values presented in the brackets and italics show the thresholds suggested by Moriasi et al. (2007) and (Moriasi et al. 2015)

Rating scores	$NSE_{monthly}$	$NSE_{seasonal}$	PBIAS and bias in SD
1	≥ 0.5	≥ 0.7	≤ − 25%, ≥ + 5%
0.5	(0.3, 0.5)	(0.5, 0.7)	(− 50%, − 25%) or (+ 25%, + 50%)
0	≤ 0.3	≤ 0.5	≤ − 50%, ≥ + 50%

term mean monthly (seasonal) discharge. The rating scores were computed for each model simulation (nine GWMs forced by four climate datasets each) at 18 gaging stations and for four different metrics. An example of how the statistical evaluation metric $NSE_{monthly}$ is translated into a rating score is presented in Table S5 for the gauging station Kusur, Lena basin. In total, 2592 discharge rating scores were computed. For JULES-W1, we averaged over the statistical evaluation metrics ($NSE_{monthly}$, PBIAS, $NSE_{seasonal}$, BIAS in SD) of the three available climate forcing data sets to represent the missing WATCH-JULES-W1 for consistency. We then summed up the rating scores for each climate forcing (maximum score 4 for each model), for each gauging station within a watershed, and for all statistical evaluation criteria. For each watershed, between 288 and 576 rating scores, depending on the number of gauging stations (2–4), form the basis of the watershed specific $API_{discharge}$. The rating scores were aggregated to 54 rating scores (9 rating scores for each watershed) which were then divided by the maximum possible score and transferred in % to get the $API_{discharge}$ for each model and watershed. An $API_{discharge}$ of 100% for one model means that for monthly discharge, $NSE_{monthly}$ is > 0.5 and PBIAS is within ± 25% and that for long-term mean monthly discharge $NSE_{seasonal}$ is > 0.7 and BIAS in standard deviation is within ± 25% at all gauging stations within a watershed.

The $API_{extreme}$ was computed based on the statistical evaluation metric PBIAS for 10 percentile values (5 for low and 5 for high flow conditions) from the flow duration curve (Table 3), similarly as presented in Liersch et al. (2018). The percentiles were calculated based on daily measured and simulated discharge for a 30-year period (1971–2000) at the outlet stations. The magnitude of daily discharge that is exceeded 10%, 5%, 1%, 0.1%, and 0.01% and 90%, 95%, 99%, 99.9%, and 99.99% of the time corresponds to high flows and low flows, respectively. The assignment of rating scores was done by computing the PBIAS for each flow percentile individually. As a result, a total of 1080 scores for high and low flow were calculated (6 gauging stations (only outlets), 4 climate forcing datasets, 9 GWMs, 5 flow percentiles for high and low flow each, 1 statistical evaluation metric). For each watershed, 180 scores were aggregated to 54 model performance indices (9 for each watershed) for high and flow flows each.

The API_{SWE} consists of the BIAS in SD and $NSE_{seasonal}$ between total monthly estimated (GlobSnow-2) and simulated SWE at four to five representative grid cells covering all cardinal directions in each watershed (Table S4). The location of the points is shown in Fig. S1. For SWE, rating scores were computed for 4 climate forcing datasets, 6 (out of 9) GWMs (Tables 2, 4 GHMs, 2 LSM), at 24 locations (4–5 locations in 6 watersheds), and 2 scores ($NSE_{seasonal}$, BIAS in SD), totaling to 1152 scores. For each watershed, 192 to 240 scores (depending on the number of points) were aggregated to 36 model performance indices. Model

analysis was restricted to the period 1980–2000, due to the data availability of the GlobSnow-2 product.

The Boruta feature selection algorithm (Kursa et al. 2010) was used to estimate the relevance of each attribute to the total score of the APIs ($API_{discharge}$, $API_{extreme}$, API_{SWE}).

The attributes consisted of:

- climate forcing data (4)
- GWMs (6 for API_{SWE}, 9 for $API_{discharge}$ and $API_{extreme}$)
- statistical performance criteria (4 for $API_{discharge}$, 1 for each percentile for $API_{extreme}$, 2 for API_{SWE})
- gauging station per watershed (2–4 for $API_{discharge}$, 1 for $API_{extreme}$)/SWE location (4–5))

For this purpose, we used the Boruta package in R. The analysis was carried out for each API and watershed separately.

In addition, the observed and simulated mean annual discharge time series were analyzed for possible trends (or lack of trend) using a simple linear regression analysis with a significance level of 0.05. Simulations for time periods without available measurements were excluded for consistency. The linear trend analysis is not part of the APIs, but a separate analysis step in accordance with the approach suggested by Krysanova et al. (2018).

3 Results

3.1 Mean monthly discharge and seasonal dynamics

The performance of the GWMs regarding the statistical evaluation metrics $NSE_{monthly}$ (Fig. 3), PBIAS (Fig. 4), and $NSE_{seasonal}$ (Fig. S2) and BIAS in SD (Fig. S3) shows large differences across GWMs and climate forcing data set. When averaged over all climate forcing data and GWMs at all gauging stations, $NSE_{monthly}$ varies between 0.94 (WFDEI-WaterGAP2 at Igarka (Yenisei)) and − 28 (WATCH-LPJmL at Yukon (Eagle)), averaging to − 0.22. $NSE_{seasonal}$ averages to − 0.29 with a maximum of 0.98 (WATCH-MPI-HM at Hatyrik-Homo (Lena)) and a minimum of − 28 (WATCH-LPJmL at Eagle AK (Yukon)). Systematic under-/overestimation (PBIAS monthly discharge) varies between + 150% (WATCH-DBH at Hanti-Mansisk (Ob)) and − 87% (Princeton-ORCHIDEE at Nenana AK (Yukon)), averaging to 31%. The bias in SD averages to 50%, ranging from + 420% (WATCH-LPJmL at Hanti-Mansisk (Ob)) to − 99% (Princeton-ORCHIDEE at Pilot Point AK (Yukon)). Performance is, on average (over all statistical evaluation metrics, GWMs, and in all watersheds), not higher at outlet compared to upstream stations. Variability in discharge across GWMs is larger compared to the climate forcing data. No climate forcing data set consistently outperforms the other for all statistical metrics in all basins, though our analysis suggests that GWMs forced by GSWP3 show better results for bias in SD and PBIAS compared to when forced by the other climate data sets. GWMs forced by Princeton are more likely to perform poorer regarding PBIAS and $NSE_{monthly}$.

Based on the assigned rating scores (Table 4) for each statistical evaluation metric, model performance regarding discharge (monthly and seasonal) was summarized for each watershed (Fig. 5(a)) and each GWM (Fig. 5(d)) via the $API_{discharge}$. WaterGAP2 outperformed the other GWMs in all basins except in Kolyma. The $API_{discharge}$ of WaterGAP2 ranged between 38

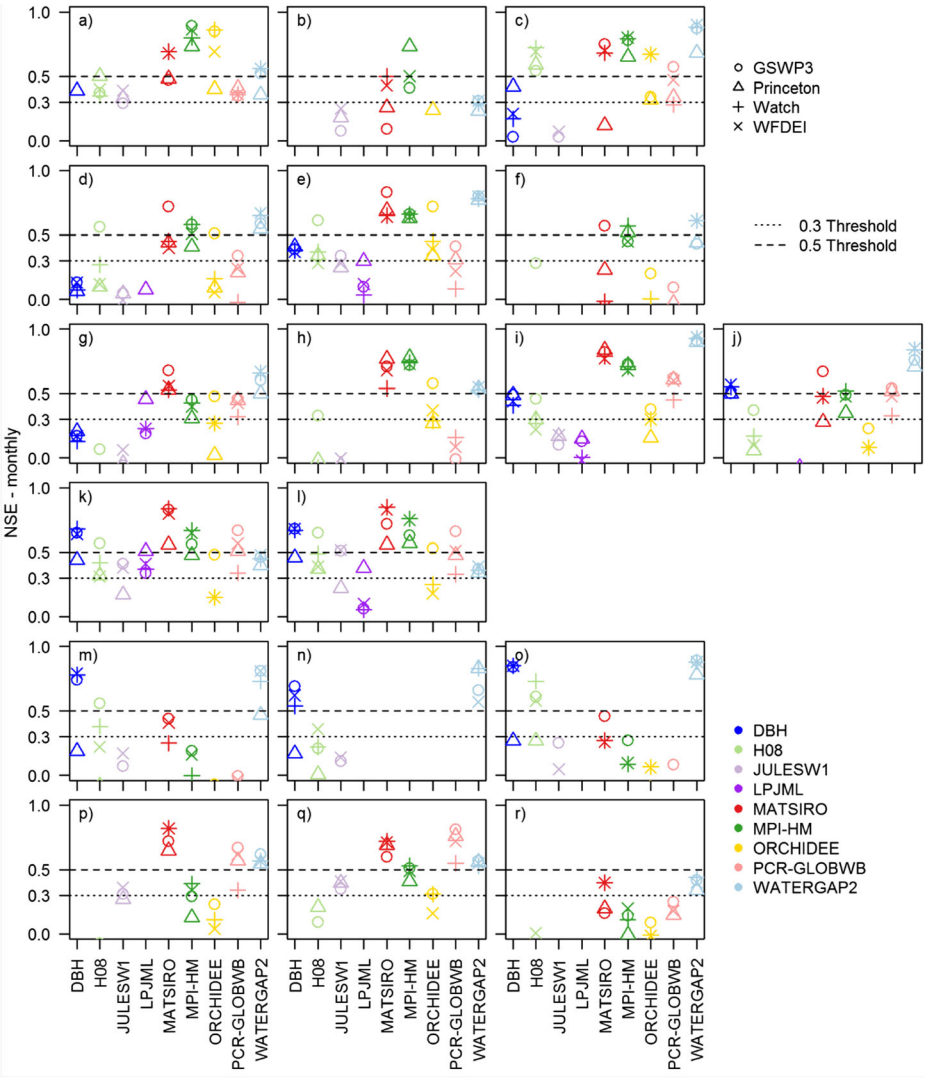

Fig. 3 Model performance evaluated using the statistical evaluation metric "Nash-Suitcliffe Efficiency (NSE)" based on simulated and measured monthly discharge for each GWM forced by four observation-based climate datasets (GSWP3, Princeton, WATCH, WFDEI). Each row presents the results for one watershed (row 1: Ob; row 2: Yenisei; row 3: Lena; row 4: Kolyma; row 5: Yukon; row 6: Mackenzie) and each letter (a–r) refers to one gauging station from the outlet (left column) to the upstream basins (Table 1). The dotted lines at 0.3 and 0.5 present the thresholds for assigning rating scores. The *y*-axis was adjusted to only represent the range 0–1

(Kolyma) and 93% (Yukon) and averaging to 72% (Table S6). MATSIRO and MPI-HM also had an average $API_{discharge}$ above 50%, exceeding 60% in four basins. ORCHIDEE, JULES-W1, and the DGVM LPJmL have rather low average $API_{discharge}$ of 25%, 16%, and 32% respectively. For JULES-W1, $API_{discharge}$ was below 32% in all basins, averaging to 16% (Fig. 5(a)).

Considering that reaching a $API_{discharge}$ of 50% can be treated as an "acceptable model," 6 GWMs in Kolyma basin, 4 GWMs in Lena basin, 5 GWMs in Ob basin, 3 GWMs in Yenisei basin, 3 GWMs in Mackenzie basin, and 2 GWMs in Yukon basin meet the criterion. The

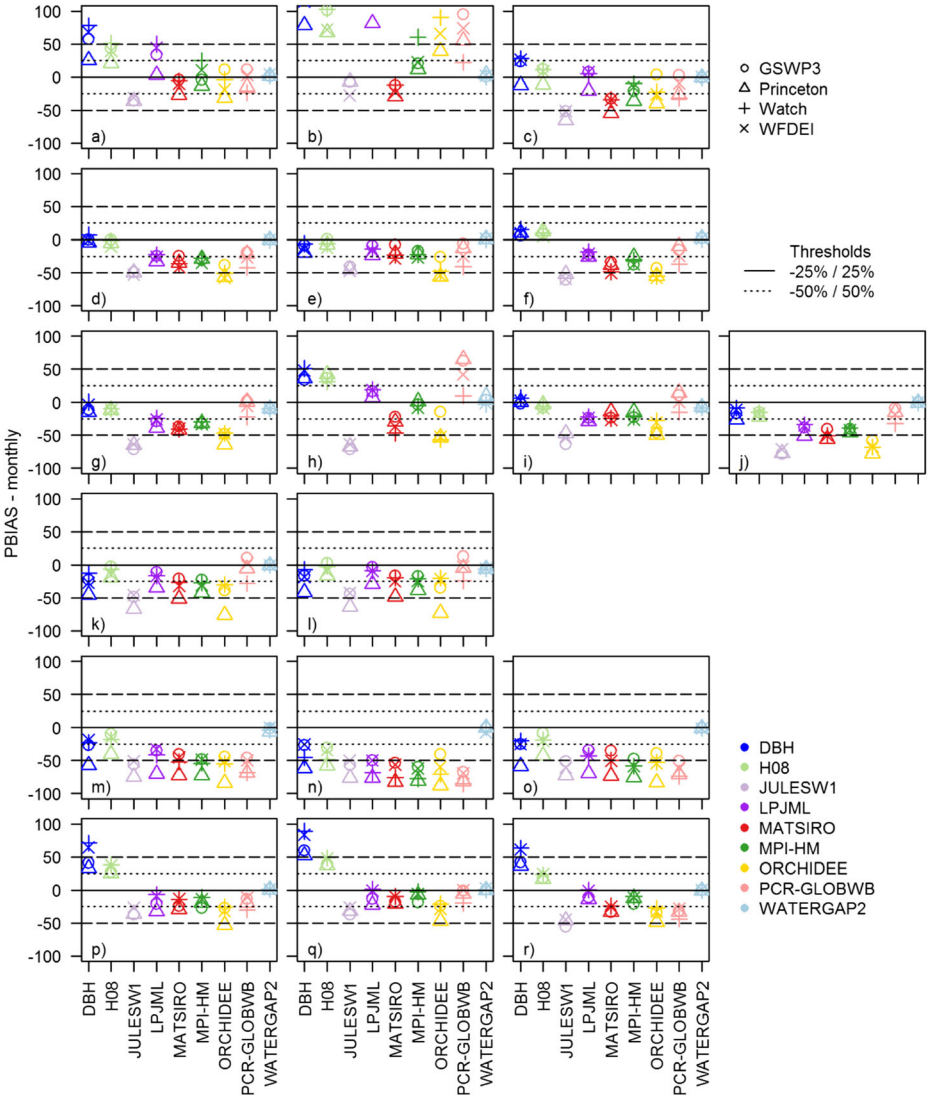

Fig. 4 Model performance evaluated using the statistical evaluation metric "Percent Bias" (Eq. (2) based on simulated and measured monthly discharge for each GWM forced by four observation-based climate datasets (GSWP3, Princeton, WATCH, WFDEI). Each row presents the results for one watershed (row 1: Ob; row 2: Yenisei; row 3: Lena; row 4: Kolyma; row 5: Yukon; row 6: Mackenzie) and each letter (a–r) refers to one gauging station from the outlet (left column) to the upstream basins (Table 1). The dotted lines at ± 25 and ± 50 present the thresholds for assigning rating scores. The *y*-axis was adjusted to only represent the range − 100 to + 100%

average GWM performance is best for Kolyma basin ($API_{discharge} = 55\%$), followed by Lena, Ob, Yenisei, and Mackenzie ($API_{discharge} = 40\%$). In the Yukon watershed, $API_{discharge}$ is 30%. WaterGAP2 and MATSIRO demonstrated good or acceptable performance in five, MPI-HM in four, DBH in three, and H08 and PCR-GLOBWB in two basins. ORCHIDEE and LPJmL each performed well in only one basin, and all JULES-W1 results were below the acceptable level of 50% in all six basins (Fig. 5(b)).

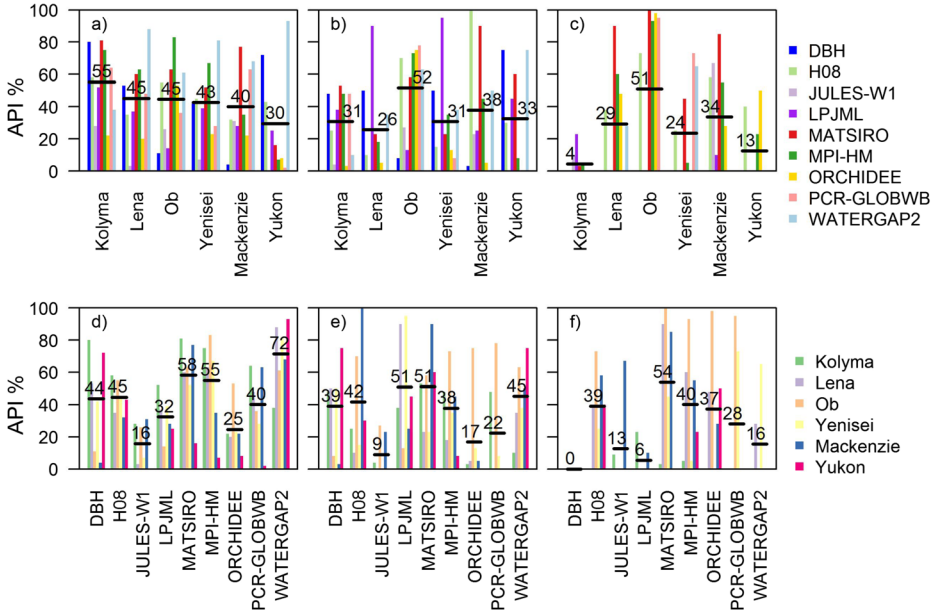

Fig. 5 Aggregated Performance Indices for monthly and seasonal discharge ($API_{discharge}$) and the extremes ($API_{extreme}$) organized by watershed (top row) and by Global Water Model (GWM) (bottom row). $API_{discharge}$ is displayed by watershed (a) and Global Water Model (GWM) (d). $API_{extreme}$ is displayed in for high flows (including the percentiles Q_{10}, Q_5, Q_1, $Q_{0.1}$, $Q_{0.01}$) (b, e) and low flows (Q_{90}, Q_{95}, Q_{99}, $Q_{99.9}$, $Q_{99.99}$) (c, f) by watershed (b,c) and by GWM (e, f). The black horizontal line (and number displayed) presents the average for a watershed (a–c) and a GWM (d–f). Table S6 summarizes the underlying values for $API_{discharge}$ and Table S7 for $API_{extreme}$

Figure 6 displays the observed and simulated mean seasonal discharge of the two best performing and two worst performing GWMs, based on the API, for each watershed. In the six watersheds, WaterGAP2 is four times among the best performing models, MATSIRO and MPI-HM twice, and DBH once. ORCHIDEE belongs to the poorest performing models in all watersheds, expect in Ob, followed by JULES-W1 (four times), DBH (twice), and LPJmL (once). The best performing models reproduce the seasonal dynamics satisfactorily, although the snow melt peak is, in the majority of the cases, underestimated and late summer discharge overestimated. The poorly performing GWMs do not reproduce the snowmelt peak neither in terms of timing (lag (DBH Mackenzie), lead (LPJmL in Ob)), nor magnitude (e.g., ORCHIDEE, JULES-W1, overestimation although timing is correct (DBH in Ob)). Consequently, the seasonal dynamic of the Pan-Arctic watersheds is not represented well by the GWMs as reflected in high absolute values of the BIAS in standard deviation (Fig. S3). Figure 6 also shows that the uncertainty caused by the choice of climate forcing datasets (shaded area around the mean) is highly variable across watersheds and GWMs.

3.2 Extremes

The $API_{extreme}$ aggregated for high and low flows, each including 5 percentiles, is summarized in Fig. 5(b, c, e, f). The $API_{extreme}$ is displayed separately for each percentile in Table 5 for high

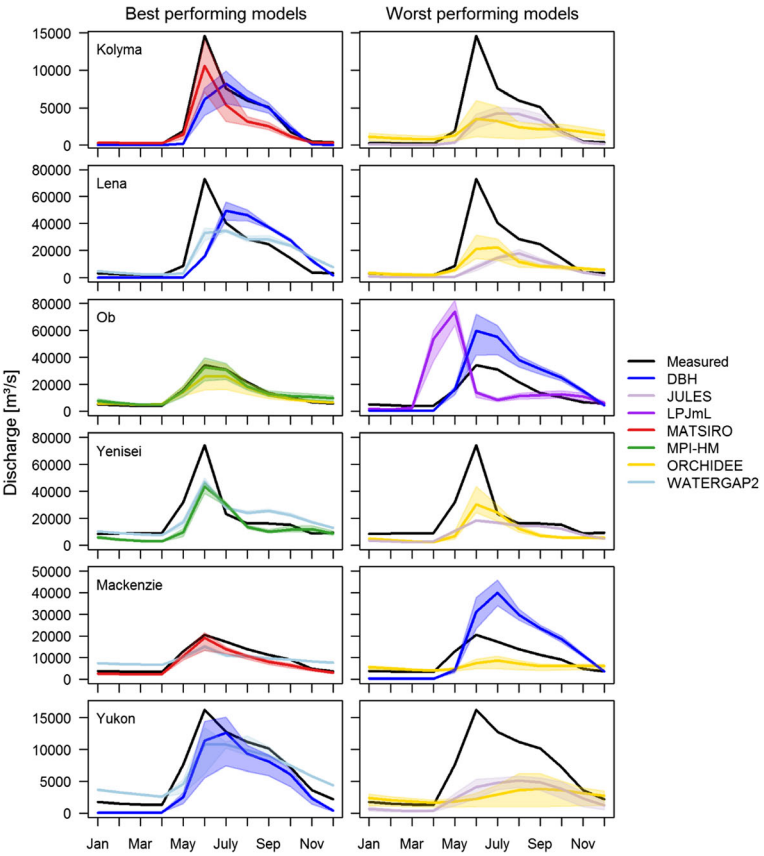

Fig. 6 Average mean monthly discharge of the two best (left column) and worst (right column) performing Global Water Models (GWMs) each in six Pan-Arctic watersheds. The categories best and worst are based on the model performance analysis detailed in the method section. The shaded area presents the variability range caused by the four different observation-based climate forcing datasets, the thick line presents the mean

and in Table S7 for low flow, each containing 270 values. Model performance is lower (average 35% for high and 26% for low flows) compared to mean discharge (43%, Fig. 5). The $API_{extreme}$, on average, decreases from the less (Q_{10}, Q_{90}) to the most ($Q_{0.01}$, $Q_{99.99}$) extreme flow percentiles (Table 5, Table S7). For Q_{10}, for example, $API_{extreme}$ is > 50% for 36 out of 54 GWM and watershed combinations, while for $Q_{0.01}$, it is only in 15 out of 54 cases. Similarly, for low flows, the number of cases that $API_{extreme}$ is > 50% reduces from 19 out of 54 (Q_{90}) to 12 out of 54 ($Q_{99.99}$). Among all GWMs, only MATSIRO has, on average, over all flow percentiles, an $API_{extreme}$ > 50% for both high and low flows. LPJmL reaches, consistently across all high flow percentiles, an $API_{extreme}$ of 100% and of > 50% in the Yenisei and Lena basin, respectively (Table 5). High and low flows are, on average over all high and low flow percentiles, best represented in the Ob watershed, with an $API_{extreme}$ > 50%. In all other basins, average $API_{extreme}$ ranges between 26 and 38% for high flows (Fig. 5(b)). Average $API_{extreme}$ for low flows ranges between 4 (Kolyma) and 51% (Ob) when categorized by watershed and between 0 (DBH) and 54% (MATSIRO) when categorized by GWM (Fig. 5(c, f)).

Table 5 The $API_{extreme}$ for the high flow discharge percentiles (Q_{10}, Q_5, Q_1, $Q_{0.1}$, $Q_{0.01}$) for each GWM and watershed as visualized in Fig. 5. The darker blue color highlights an API > 50%, the lighter blue color an API = 50%, API < 50% are presented in gray. The yellow color highlights a performance in >5 0% averaged over all GWMs (per watershed, last column) and over all watersheds (per GWM, last row). The orange color highlights the overall average (over all watersheds and GWMs)

Q_{10}	WATERGAP2	DBH	H08	MPI-HM	PCR-GLOBWB	MATSIRO	ORCHIDEE	LPJML	JULES-W1	*Average*
Lena	100	100	50	50	0	50	12.5	87.5	0	***50***
Kolyma	50	75	62.5	62.5	62.5	50	12.5	50	12.5	***49***
Yenisei	100	100	62.5	100	37.5	62.5	50	100	0	***68***
Ob	50	12.5	100	87.5	75	87.5	62.5	62.5	25	***63***
Mackenzie	50	0	100	50	0	87.5	25	100	50	***51***
Yukon	87.5	87.5	37.5	0	0	50	0	75	0	***38***
Average	***73***	***63***	***69***	***58***	***29***	***65***	***27***	***79***	***15***	***53***
Q_5	WATERGAP2	DBH	H08	MPI-HM	PCR-GLOBWB	MATSIRO	ORCHIDEE	LPJML	JULES-W1	
Lena	50	87.5	0	37.5	0	37.5	12.5	100	0	***36***
Kolyma	0	50	25	62.5	37.5	50	0	62.5	0	***32***
Yenisei	50	62.5	12.5	50	0	12.5	12.5	100	0	***33***
Ob	50	12.5	87.5	87.5	75	75	62.5	0	25	***53***
Mackenzie	50	0	100	50	0	87.5	0	25	12.5	***36***
Yukon	87.5	75	50	12.5	0	50	0	62.5	0	***38***
Average	***48***	***48***	***46***	***50***	***19***	***52***	***15***	***58***	***6***	***38***
Q_1	WATERGAP2	DBH	H08	MPI-HM	PCR-GLOBWB	MATSIRO	ORCHIDEE	LPJML	JULES-W1	
Lena	25	37.5	0	0	0	12.5	0	100	0	***19***
Kolyma	0	37.5	12.5	37.5	25	50	0	37.5	0	***22***
Yenisei	12.5	37.5	0	0	0	12.5	0	100	0	***18***
Ob	50	12.5	62.5	87.5	75	62.5	87.5	0	37.5	***53***
Mackenzie	50	0	100	50	0	100	0	0	12.5	***35***
Yukon	75	75	37.5	25	0	75	0	37.5	0	***36***
Average	***35***	***33***	***35***	***33***	***17***	***52***	***15***	***46***	***8***	***31***
$Q_{0.1}$	WATERGAP2	DBH	H08	MPI-HM	PCR-GLOBWB	MATSIRO	ORCHIDEE	LPJML	JULES-W1	
Lena	0	25	0	0	0	12.5	0	100	0	***15***
Kolyma	0	37.5	12.5	37.5	50	50	0	25	0	***24***
Yenisei	0	25	0	12.5	0	12.5	0	100	0	***17***
Ob	75	0	50	50	75	37.5	75	0	37.5	***44***
Mackenzie	50	0	100	37.5	0	87.5	0	0	12.5	***32***
Yukon	62.5	75	12.5	0	0	62.5	0	25	0	***26***
Average	***31***	***27***	***29***	***23***	***21***	***44***	***13***	***42***	***8***	***26***
$Q_{0.01}$	WATERGAP2	DBH	H08	MPI-HM	PCR-GLOBWB	MATSIRO	ORCHIDEE	LPJML	JULES-W1	
Lena	0	0	0	0	0	0	0	62.5	0	***7***
Kolyma	0	37.5	12.5	37.5	62.5	62.5	0	12.5	0	***25***
Yenisei	25	25	0	12.5	0	12.5	0	100	0	***19***
Ob	87.5	0	50	50	87.5	25	87.5	0	37.5	***47***
Mackenzie	50	12.5	100	37.5	0	87.5	0	0	12.5	***33***
Yukon	62.5	62.5	12.5	0	0	62.5	0	25	0	***25***
Average	***38***	***23***	***29***	***23***	***25***	***42***	***15***	***33***	***8***	***26***

3.3 Trends mean annual discharge

Trends in measured mean annual discharge are found to be significant ($p < 0.05$) only at two stations: Igarka (Yenisei) and Sredne-Kolymsk (Kolyma) (Fig. S4). At Igarka, all simulations, except WFDEI-PCR-GLOBWB, agree with the measurements in simulating a negative trend in mean annual discharge despite difference in the magnitude (Fig. S4c). All simulated trends are also significant except for WFDEI-PCR-GLOBWB and LPJmL (for all climate data forcing sets) (Fig. S4 a). At Sredne-Kolymsk (Kolyma), all simulations, except WFDEI-

PCR-GLOBWB, agree on a negative trend in mean annual discharge (Fig. S4d) but only 17 (out of 40, 42.5%) are also significant (Fig. S4b). At all other gauging stations, trends in measured mean annual discharge are not significant.

3.4 Snow water equivalent

The performance index regarding seasonal SWE (API_{SWE}) is displayed in Fig. S5 and the corresponding values in Table S8. Average API_{SWE} is 57%. An $API_{SWE} > 50\%$ is reached in 27 out of 36 cases. These numbers are higher compared to the averages for discharge and extremes, but it cannot be directly compared to the $API_{discharge}$, as only six (compared to nine) models provided SWE output, only four grid cells are considered in each basin, and the analysis period differs slightly. GWMs reproduce SWE best in Mackenzie watershed (72%), followed by Lena (62%) and poorest in the Yukon basin (44%). All GWMs reach an $API_{SWE} \geq 50\%$. The simulated seasonal dynamics of SWE is compared to the observations for each watershed in Fig. S6 to Fig. S11.

3.5 Boruta feature selection

For all APIs ($API_{discharge}$, $API_{extreme}$, API_{SWE}), the climatic forcing was consistently detected as not relevant by the Boruta algorithm across all watersheds. This implies that forcing the GWMs with four different (instead of only one) observation-based climate forcing data sets has a low relevance for the overall API score in this study. All other attributes, e.g., the GWMs and the statistical performance criteria, are confirmed relevant for the overall API score. For API_{SWE}, other attributes, such as the statistical performance criteria, were, in some cases, also found unimportant in addition to the climate forcing data. The GWM is identified the most important attribute for API_{SWE} in all watersheds except in Lena and Mackenzie, where the GWMs is, however, still among the three most important attributes (out of 10 in total). For the calculation of API_{SWE} and $API_{discharge}$ in the Kolyma watershed, the data available to train the Boruta algorithm was likely not sufficient (for API_{SWE} only 6 GWMs, for $API_{discharge}$ Kolyma only two gauging stations).

4 Discussion

The GWMs often have a considerable bias (mostly systematic underestimation) and difficulties in reproducing the seasonal discharge cycle when compared against observations in Pan-Arctic watersheds. Overall GWM performance, assessed for different hydrological indicators with several statistical evaluation metrics for up to four gauges in each watershed, ranges from satisfactory to poor. However, in some cases, API is larger than 70% (9 of 54 for the monthly and seasonal discharge, 10 of 54 for high flows and 8 of 54 for low flow, 3 out of 36 for SWE). No GWM consistently outperforms the other models in all watersheds and for all indicators, and model performance, on average, does not increase with basin size. This is in line with other model inter-comparison studies (e.g., Slater et al. (2007)), where also no model was the best or worst performing when compared to a range of observations and in different watersheds across the Pan-Arctic. Our results, satisfactory to poor performance of GWMs, are also consistent with global studies that also include watersheds located in temperate and tropical climates (Krysanova et al. n.d.). In the study by Krysanova et al. (n.d.), the best (WaterGAP2

and MATSIRO) and poorest (e.g., LPJmL) performing GWMs match with our study while two GWMs, H08 and DBH, performed slightly better in the Artic compared to other climate zones.

We also demonstrate that the variability across the observation-based climate forcing data is smaller compared to that across GWMs. This is also confirmed by the feature selection using the Boruta algorithm. The large variability of performance across GWMs is most likely related to model structural differences and/or lack of physical process representation for some processes, difficulties to represent some processes with a relatively coarse resolution of 0.5°, and missing calibration (except WaterGAP2, which was calibrated) as well as no targeted model setup/parameterization focusing on Arctic hydrological processes.

Most GWMs struggle to simulate the snowmelt peak, the most important hydrological event in (sub) Arctic rivers, both in terms of absolute discharge amount and timing. This is directly linked to the GWMs rather simple representation of snow hydrological processes including the onset of snowmelt (isothermal phase change of the snowpack), the fate of snowmelt (infiltration into soils, refreezing over cold periods), snow compaction, and redistribution of snow on the landscape. Additionally, processes related to and affecting river routing, such as ice jams and dams, are highly complex and often are not considered or only included very simplistic in GWMs. Dams are not considered in the runs without human influences that we analyzed here. This likely explains the relatively poor model performance in the Ob (particularly at Hanti-Mansisk (Irtysch River)) and in the Yenisei watershed where the impact of dams on changes in the seasonal discharge has been documented (Adam et al. 2007). Concurrently, general errors in the forcing data, which are consistent across all datasets, such as snowfall underestimation (Beck et al. 2017; Hancock et al. 2014) and uncertainties in wind speed, amplify the rather poor simulation of snowmelt peak flow. Strong winds that are characteristic for Arctic tundra environments enhance sublimation and could therefore add to the general underestimation of the snowmelt peak by GWMs.

Except for many GWMs in the Ob watersheds (particularly at gauges Salekhard and Hanti-Mansisk) and for DBH and H08 in Mackenzie and Lena (gauge: Hatyrik-Homo), the GWMs have a tendency to underestimate measured monthly discharge in this region. This phenomena has already been documented by others, e.g., Andresen et al. (2019) and Lohmann et al. (2004). Lohmann et al. (2004) highlighted that measured discharge is underestimated by LSMs in areas with significant snowfall, and that snowmelt peak timing can be off by up to 4 months. Beck et al. (2017) and Hancock et al. (2014) attributed an early bias in spring snowmelt peak to precipitation underestimation that leads to insufficient snow accumulation and subsequently to too rapid snow melt. In our study, GWMs forced by Princeton underestimate, on average, discharge (and snowmelt peak) more significantly. The GWMs forced by WFDEI, WATCH, and GSWP3 perform better, as precipitation is corrected for snow undercatch and scaled to the monthly precipitation sums of Global Precipitation Climatology Centre (GPCC). GWM improvements, particularly, related to snow hydrological processes are, however, limited by sparse data availability and the challenges in measuring snow-related processes effectively over larger spatial and temporal scales.

Under historic climate conditions, the GHMs, on average, performed better than the LSMs in the Pan-Arctic watersheds, with the exception (in many cases) of MATSIRO. Beck et al. (2017) suggest that the differences in model performance are caused by the snow routines with the simple conceptual degree-day approach (GHMs) outperforming the physically based energy balance approach (LSMs). In our case, however, DBH performed reasonably well regarding SWE (Fig. S5), despite relying on the more complex energy balance approach. This

suggests that the reasons for better/poorer performance are more complex. For example, the overall annual flow volume differs in many cases considerably between GWMs and observations as well as across GWMs. We also suggest to additionally evaluate the entire water balance as well as the sensitivity of water balance components to changes in the climatic drivers to gain a deeper understanding of the GWM structural differences in Arctic watersheds. Hattermann et al. (2017) have for example shown that the climate sensitivity of uncalibrated GWMs in the historical period is comparable to calibrated regional models. However, the differences in processes implementation (particularly permafrost) could make larger differences under climate change conditions. The consideration of permafrost temperatures and annual soil freeze/thaw processes in JULES-W1, MATSIRO, and LPJmL add additional complexity but is imperative for climate change impact assessments as increased surface and subsurface water interactions will have the potential to change the discharge regimes of high-latitude watersheds with large-scale consequences for land-atmosphere biochemical processes.

Model performances concerning monthly and long-term mean monthly discharge are poorer for the watersheds located in North America (Yukon followed by Mackenzie). These watersheds are characterized by a higher percentage of glacier coverages compared to the watersheds located in Asia. Glaciers, even with a low overall coverage, influence the discharge regime considerably, especially under drought/low rainfall conditions (Huss 2011). None of the GWMs evaluated here explicitly simulate glacier-related hydrological processes, which most likely also contributes to their weak or poor performance in these watersheds.

WaterGAP2 (calibrated only for the long-term average annual discharge) outperformed the other GWMs when analyzing hydrological indicators based on monthly and long-term mean monthly discharge, which underscores the importance of targeted calibration for the hydrological indicator of interest. However, our study showed, similarly to Zaherpour et al. (2018), that WaterGAP2's performance for extremes is only average. For SWE, the performance of WaterGAP2 ($API_{SWE} = 48\%$) is below the average ($API_{SWE} = 57\%$). MATSIRO outperforms the other GWMs for discharge extremes ($API_{extreme} = 51\%$ for high and 54% for low flows), and is the second best performing model for discharge ($API_{discharge} = 58\%$). Zaherpour et al. (2018) highlighted the superior, physically based snow and soil scheme of MATSIRO. In our analysis, MATSIRO did not outperform the other GWMs regarding SWE. However, we have to admit that our SWE evaluation approach was rather simplistic, and may not represent the model performance accurately enough, since we compared the simulated SWE output to the GlobSnow-2 product at only four randomly chosen grid cells within the entire watershed. Additionally, GLobSnow-2, a remote sensing product, is known to inherit uncertainties (Metsämäki et al. 2015).

DGVMs are developed to explain the changes in vegetation dynamics and associated impacts on and linkages to the hydrological and carbon cycles, while LSMs are designed to simulate the exchange of water, carbon, and energy in General Circulation and Earth System Models. Therefore, hydrological processes per se and the interaction with snow and permafrost were not the main objectives for neither their development nor their setup/parameterization. This may explain a weak or poor performance of ORCHIDEE, LPJmL, and JULES-W1 for discharge and snow seasonality in many cases. Updated model versions of ORCHIDEE, such as reported in Wang et al. (2013) and Guimberteau et al. (2018), improved model performance regarding Artic hydrological processes considerably (such as snow depth, SWE, snow albedo, and snowmelt runoff) which highlights the potential of targeted model development to improve model performance independently of explicit parameter calibration.

5 Conclusions and outlook

Our results show that the majority of the nine GWMs included in the study exhibit considerable difficulties in realistically representing Pan-Arctic hydrological processes. The performance evaluation showed that no GWM outperformed other GWMs in all watersheds and for all hydrological indicators. This also implies that we could not identify any links between the model efficiency and model structure/parameterization. We also highlighted that no climate forcing dataset was better suited for all indicators and statistical evaluation metrics, though results based on Princeton showed a larger PBIAS on average. Thus, there is an urgent need to refine the representation of cold-region hydrological processes in GWMs to improve their performance, as changes in these processes will govern the fate of the Arctic. This includes direct process representation (snow, permafrost, glaciers, river routing accounting for ice cover/different flow velocities) and a more appropriate model setup and parameterization. The GHM WaterGAP2 outperformed the more complex LSMs for mean flow, while the LSM MATSIRO showed a better performance than other models for extremes. The GHMs often lack relevant cold-region processes (permafrost, glaciers), have simple snow schemes, and represent vegetation statically. Many LSMs, on the other hand, represent vegetation dynamically which is highly relevant under climate change as vegetation cover strongly influences the partitioning of precipitation into evapotranspiration and discharge.

Considering the relatively weak or poor performance of most GWMs in most watersheds under current climate conditions, a comprehensive model evaluation is recommended before conducting climate impact assessments. Only GWMs that show good or satisfactory evaluation results (i.e., $API > 0.5$) and that represent cold-region hydrological processes (snow hydrological processes, permafrost) effectively should be included in multi-model climate change impact assessments in Pan-Arctic watersheds. Models meeting these criteria include WaterGAP2 (though requiring a dynamic permafrost module), MATSIRO, MPI-HM for monthly/seasonal dynamics, MATSIRO and LPJmL for high flows, and MATSIRO for low flows. All other models require performance improvement and all except JULES-W1 the incorporation of dynamic permafrost. Tools, such as the API presented here, are promising to define weighting coefficients based on model performance, and to identify GWMs that potentially could be applied for impact assessment, and GWMs that should be excluded from the ensemble. Model exclusion should trigger an analysis about the model's shortcomings and result in model refinement. Therefore, a close collaboration between impact modelers and the model development teams is crucial. Ultimately, the model weighting approach could provide more trustworthy results of impact assessment and reduce large uncertainty ranges which are generally characteristic for multi-model climate change impact assessments. Further, the API makes the interpretation of model performance involving a large ensemble of participating GWMs transparent and easily understandable by a wide range of audience.

Our study can be extended by defining model weighting coefficients based on the API and comparing this approach to the traditional ensemble mean approach in climate change impact assessments. Another interesting study could be done by comparing GWM performances with and without including human influences, which could be also done using our suggested API approach. Other statistical evaluation metrics and additional hydrological indicators could be also included in the evaluation. Considering the rapid rate of environmental changes occurring in the Pan-Arctic, we urgently need to increase our understanding of the hydrological cycle and its linkages to other parts of the earth system, and GWMs are suitable tools to do so.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10584-020-02892-2.

Acknowledgments We would like to thank the German Federal Ministry of Education and Research (BMBF) and the European Research Area for Climate Services ERA4CS with project funding reference number 01LS1711C for funding the ISIPedia project (Yoshihide Wada was supported by the same project under grant no. 690462). This study also benefited from funding of JSPS KAKENHI (grant no.: 17K12820), the Key Research Program of the Chinese Academy of Sciences (grant no. ZDRW-ZS-2017-4), and the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA19070302). We would like to thank Maria del Rocio Rivas Lopez for downloading and extracting the GlobSnow-2 dataset. Faruque Abdullah supported Figs. 3 and 4 and Figs. S1 and S2.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Adam JC, Haddeland I, Su F, Lettenmaier DP (2007) Simulation of reservoir influences on annual and seasonal streamflow changes for the Lena, Yenisei, and Ob' rivers. J Geophys Res: Atmos 112. https://doi.org/10.1029/2007jd008525

Ahmed R, Prowse T, Dibike Y, Bonsal B, O'Neil H (2020) Recent trends in freshwater influx to the Arctic Ocean from four major arctic-draining rivers. Water 12:1189

Andresen CG et al (2019) Soil moisture and hydrology projections of the permafrost region: a model intercomparison. Cryosphere 3(2):591–609. discussion 2019:1–20. https://doi.org/10.5194/tc-2019-144

Beck HE, van Dijk AIJM, de Roo A, Dutra E, Fink G, Orth R, Schellekens J (2017) Global evaluation of runoff from 10 state-of-the-art hydrological models. Hydrol Earth Syst Sci 21:2881–2903. https://doi.org/10.5194/hess-21-2881-2017

Best MJ et al (2011) The Joint UK Land Environment Simulator (JULES), model description - Part 1: Energy and water fluxes. Geosci Model Dev 4:677–699. https://doi.org/10.5194/gmd-4-677-2011

Biskaborn BK et al (2019) Permafrost is warming at a global scale. Nat Commun 10:264. https://doi.org/10.1038/s41467-018-08240-4

Bowling LC, Kane DL, Gieck RE, Hinzman LD, Lettenmaier DP (2003) The role of surface storage in a low-gradient Arctic watershed. Water Resour Res 39:1087. https://doi.org/10.1029/2002WR001466

Bring A et al (2016) Arctic terrestrial hydrology: a synthesis of processes, regional effects, and research challenges. J Geophys Res Biogeosci 121:621–649. https://doi.org/10.1002/2015JG003131

Brown J, Ferrians OJ, Heginbottom JJA, Melnikov ES (1997) Circum-Arctic map of permafrost and ground-ice conditions. Washington, DC: U.S. Geological Survey in Cooperation with the Circum-Pacific Council for Energy and Mineral Resources. Circum-Pacific Map Series CP-45. https://doi.org/10.3133/cp45

Do HX, Gudmundsson L, Leonard M, Westra S (2018) The global streamflow indices and metadata archive (GSIM) - part 1: the production of a daily streamflow archive and metadata. Earth Syst Sci Data 10:765–785. https://doi.org/10.5194/essd-10-765-2018

Döll P, Lehner B (2002) Validation of a new global 30-min drainage direction map. J Hydrol 258:214–231. https://doi.org/10.1016/S0022-1694(01)00565-0

Ge S (2013) Permafrost hydrology. By Ming-ko Woo Arct Antarct Alp Res 45:615–616. https://doi.org/10.1657/1938-4246-45.4.615

Gosling SN et al (2017) A comparison of changes in river runoff from multiple global and catchment-scale hydrological models under global warming scenarios of 1 °C, 2 °C and 3 °C. Clim Chang 141:577–595. https://doi.org/10.1007/s10584-016-1773-3

Gosling S et al (2019) ISIMIP2a simulation data from water (global) sector (V. 1.1). https://doi.org/10.5880/PIK.2019.003
Gudmundsson L, Do HX, Leonard M, Westra S (2018) The global streamflow indices and metadata archive (GSIM) – part 2: quality control, time-series indices and homogeneity assessment. Earth Syst Sci Data 10: 787–804. https://doi.org/10.5194/essd-10-787-2018
Guimberteau M et al (2018) ORCHIDEE-MICT (v8.4.1), a land surface model for the high latitudes: model description and validation. Geosci Model Dev 11:121–163. https://doi.org/10.5194/gmd-11-121-2018
Hanasaki N et al (2008) An integrated model for the assessment of global water resources – part 1: model description and input meteorological forcing. Hydrol Earth Syst Sci 12:1007–1025. https://doi.org/10.5194/hess-12-1007-2008
Hancock S, Huntley B, Ellis R, Baxter R (2014) Biases in reanalysis snowfall found by comparing the JULES land surface model to globsnow. J Clim 27:624–632. https://doi.org/10.1175/jcli-d-13-00382.1
Hattermann FF et al (2017) Cross-scale intercomparison of climate change impacts simulated by regional and global hydrological models in eleven large river basins. Clim Chang 141(3):561–576. https://doi.org/10.1007/s10584-016-1829-4
Holliday NP et al (2020) Ocean circulation causes the largest freshening event for 120 years in eastern subpolar North Atlantic. Nat Commun 11:585. https://doi.org/10.1038/s41467-020-14474-y
Huss M (2011) Present and future contribution of glacier storage change to runoff from macroscale drainage basins in Europe. Water Resour Res 47:W07511. https://doi.org/10.1029/2010WR010299
ISIMIP2a (2018) The Inter-Sectoral Impact Model Intercomparison Project (ISIMIP): ISIMIP2a simulation protocol. Authors: ISIMIP Coordination Team, Sectoral Coordinators & Scientific Advisory Board. https://www.isimip.org/#isimip2a/. [Online Accessed on 5 April 2019]
Kane DL, Hinzman LD, Benson CS, Liston GE (1991) Snow hydrology of a headwater Arctic basin: 1. Physical measurements and process studies. Water Resour Res 27:1099–1109. https://doi.org/10.1029/91WR00262
Krause P, Boyle DP, Bäse F (2005) Comparison of different efficiency criteria for hydrological model assessment. Adv Geosci 5:89–97. https://doi.org/10.5194/adgeo-5-89-2005
Krysanova V, Donnelly C, Gelfan A, Gerten D, Arheimer B, Hattermann F, Kundzewicz ZW (2018) How the performance of hydrological models relates to credibility of projections under climate change. Hydrol Sci J 63:696–720. https://doi.org/10.1080/02626667.2018.1446214
Krysanova V et al (2020) How evaluation of global hydrological models can help to improve credibility of river discharge projections under climate change. Clim Change. https://doi.org/10.1007/s10584-020-02840-0
Kummu M, Varis O (2011) The world by latitudes: a global analysis of human population, development level and environment across the north–south axis over the past half century. Appl Geogr 31:495–507. https://doi.org/10.1016/j.apgeog.2010.10.009
Kursa MB, Jankowski A, Rudnicki WR (2010) Boruta – a system for feature selection. Fundamenta Inform 101: 271–285. https://doi.org/10.3233/FI-2010-288
Liersch S et al (2018) Are we using the right fuel to drive hydrological models?. A climate impact study in the Upper Blue Nile. Hydrol Earth Syst Sci 22:2163–2185. https://doi.org/10.5194/hess-22-2163-2018
Lohmann D et al (2004) Streamflow and water balance intercomparisons of four land surface models in the North American Land Data Assimilation System project. J Geophys Res Atmos 109:D07S91. https://doi.org/10.1029/2003jd003517
McGuire AD et al (2018) Dependence of the evolution of carbon dynamics in the northern permafrost region on the trajectory of climate change. Proc Natl Acad Sci U S A 115:3882–3887. https://doi.org/10.1073/pnas.1719903115
Metsämäki S et al (2015) Introduction to GlobSnow snow extent products with considerations for accuracy assessment. Remote Sens Environ 156:96–108. https://doi.org/10.1016/j.rse.2014.09.018
Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans ASABE 50:885–900. https://doi.org/10.13031/2013.23153
Moriasi DN, Gitau MW, Pai N, Daggupati P (2015) Hydrologic and water quality models: performance measures and evaluation criteria. Trans ASABE 58:1763–1785. https://doi.org/10.13031/trans.58.10715
Morison J, Kwok R, Peralta-Ferriz C, Alkire M, Rigor IG, Andersen R, Steele M (2012) Changing Arctic ocean freshwater pathways. Nature 481:66–70 doi:https://doi.org/10.1038/nature10705
Müller Schmied H et al (2016) Variations of global and continental water balance components as impacted by climate forcing uncertainty and human water use. Hydrol Earth Syst Sci 20:2877–2898. https://doi.org/10.5194/hess-20-2877-2016
Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I — a discussion of principles. J Hydrol 10:282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Oki T, Nishimura T, Dirmeyer P (1999) Assessment of annual runoff from land surface models using total runoff integrating pathways (TRIP). J Meteorol Soc Jpn Ser II 77:235–255. https://doi.org/10.2151/jmsj1965.77.1B_235

Pokhrel YN, Koirala S, Yeh PJ-F, Hanasaki N, Longuevergne L, Kanae S, Oki T (2015) Incorporation of groundwater pumping in a global land surface model with the representation of human impacts. Water Resour Res 51:78–96. https://doi.org/10.1002/2014wr015602

Post E et al (2019) The polar regions in a 2°C warmer world. Sci Adv 5:eaaw9883. https://doi.org/10.1126/sciadv.aaw9883

Pulliainen J et al (2020) Patterns and trends of northern hemisphere snow mass from 1980 to 2018. Nature 581:294–298. https://doi.org/10.1038/s41586-020-2258-0

Schaphoff S, Heyder U, Ostberg S, Gerten D, Heinke J, Lucht W (2013) Contribution of permafrost soils to the global carbon budget. Environ Res Lett 8:014026. https://doi.org/10.1088/1748-9326/8/1/014026

Sitch S et al (2003) Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. Glob Chang Biol 9:161–185. https://doi.org/10.1046/j.1365-2486.2003.00569.x

Slater AG, Bohn TJ, McCreight JL, Serreze MC, Lettenmaier DP (2007) A multimodel simulation of pan-Arctic hydrology. J Geophys Res Biogeosci 112:G04S45. https://doi.org/10.1029/2006JG000303

Stacke T, Hagemann S (2012) Development and evaluation of a global dynamical wetlands extent scheme. Hydrol Earth Syst Sci 16:2915–2933. https://doi.org/10.5194/hess-16-2915-2012

Tang Q, Oki T, Kanae S, Hu H (2007) The influence of precipitation variability and partial irrigation within grid cells on a hydrological simulation. J Hydrometeorol 8:499–512. https://doi.org/10.1175/jhm589.1

Traore AK et al (2014) Evaluation of the ORCHIDEE ecosystem model over Africa against 25 years of satellite-based water and carbon measurements. J Geophys Res Biogeosci 119:1554–1575. https://doi.org/10.1002/2014JG002638

Wada Y, Wisser D, Bierkens MFP (2014) Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources. Earth Syst Dynam 5:15–40. https://doi.org/10.5194/esd-5-15-2014

Walvoord MA, Kurylyk BL (2016) Hydrologic impacts of thawing permafrost—a review. Vadose Zone J 15:1–20. https://doi.org/10.2136/vzj2016.01.0010

Walvoord MA, Striegl RG (2007) Increased groundwater to stream discharge from permafrost thawing in the Yukon River basin: potential impacts on lateral export of carbon and nitrogen. Geophys Res Lett 34:L12402. https://doi.org/10.1029/2007GL030216

Wang T et al (2013) Evaluation of an improved intermediate complexity snow scheme in the ORCHIDEE land surface model. J Geophys Res Atmos 118:6064–6079. https://doi.org/10.1002/jgrd.50395

Zaherpour J et al (2018) Worldwide evaluation of mean and extreme runoff from six global-scale hydrological models that account for human impacts. Environ Res Lett 13:065015. https://doi.org/10.1088/1748-9326/aac547

Zemp M et al (2019) Global glacier mass changes and their contributions to sea-level rise from 1961 to 2016. Nature 568:382–386. https://doi.org/10.1038/s41586-019-1071-0

Affiliations

Anne Gädeke[1] · Valentina Krysanova[1] · Aashutosh Aryal[1] · Jinfeng Chang[2,3,4] · Manolis Grillakis[5,6] · Naota Hanasaki[7] · Aristeidis Koutroulis[5] · Yadu Pokhrel[8] · Yusuke Satoh[3,7] · Sibyll Schaphoff[1] · Hannes Müller Schmied[9,10] · Tobias Stacke[11] · Qiuhong Tang[12] · Yoshihide Wada[3] · Kirsten Thonicke[1]

[1] Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Telegrafenberg, 14412 Potsdam, Germany

[2] Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA–CNRS–UVSQ, Université Paris-Saclay, 91191 Gif-sur-Yvette, France

[3] International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, A-2361 Laxenburg, Austria

[4] College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China

[5] School of Environmental Engineering, Technical University of Crete, 73100 Chania, Greece

[6] Lab of Geophysical-Remote Sensing & Archaeoenvironment, Institute for Mediterranean Studies, Foundation for Research & Technology Hellas, 74100 Rethimnon, Greece

[7] National Institute for Environmental Studies, Japan 16-2 Onogawa, Tsukuba, Ibaraki 305-8506, Japan

[8] Department of Civil and Environmental Engineering, Michigan State University, East Lansing, MI 48824, USA

[9] Institute of Physical Geography, Goethe-University of Frankfurt, 60438 Frankfurt am Main, Germany

[10] Senckenberg Leibniz Biodiversity and Climate Research Centre (SBiK-F) Frankfurt, 60325 Frankfurt am Main, Germany

[11] Helmholtz-Zentrum Geesthacht, Institute of Coastal Research, 21502 Geesthacht, Germany

[12] Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Science and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China