Can Continental Models Convey Useful Seasonal Hydrologic Information at the Catchment Scale?

Louise Crochemore, Maria-Helena Ramos, Ilias G Pechlivanidis

AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR025700

Key Points:

- Continental models are often outperformed by catchment-specific models, but models can match when forecasting seasonal streamflow anomalies
- Consistency in the meteorological data sets used in calibration and bias adjustment enables modeling setups to benefit from climate forecasts
- Caution is needed when extracting intermediate hydrologic states such as soil water content from continental models

Correspondence to:
L. Crochemore,
louise.crochemore@smhi.se

Citation:
Crochemore, L., Ramos, M.-H., & Pechlivanidis, I. G. (2020). Can continental models convey useful seasonal hydrologic information at the catchment scale? *Water Resources Research*, *56*, e2019WR025700. https://doi.org/10.1029/2019WR025700

Received 2 JUN 2019
Accepted 25 DEC 2019
Accepted article online 3 JAN 2020

Can Continental Models Convey Useful Seasonal Hydrologic Information at the Catchment Scale?

L. Crochemore[1], M.-H. Ramos[2], and I. G. Pechlivanidis[1]

[1]Swedish Meteorological and Hydrological Institute, Hydrology Research Unit, Norrköping, Sweden, [2]Université Paris-Saclay, INRAE, UR HYCAR, Antony, France

Abstract The development and availability of climate forecasting systems have allowed the implementation of seasonal hydroclimatic services at the continental scale. User guidance and quality of the forecast information are key components to ensure user engagement and service uptake, yet forecast quality depends on the hydrologic model setup. Here, we address how seasonal forecasts from continental services can be used to address user needs at the catchment scale. We compare a continentally calibrated process-based model (E-HYPE) and a catchment-specific parsimonious model (GR6J) to forecast streamflow in a set of French catchments. Results show that despite expected high performance from the catchment setup against observed streamflow, the continental setup can, in some catchments, match or even outperform the catchment-specific setup for 3-month aggregations and threshold exceedance. Forecast systems can become comparable when looking at statistics relative to model climatology, such as anomalies, and adequate initial conditions are the main source of skill in both systems. We highlight the need for consistency in data used in modeling chains and in tailoring service outputs for use at the catchment scale. Finally, we show that the spread in internal model states varies largely between the two systems, reflecting the differences in their setups and calibration strategies, and highlighting that caution is needed before extracting hydrologic variables other than streamflow. We overall argue that continental hydroclimatic services show potential on addressing needs at the catchment scale, yet guidance is needed to extract, tailor and use the information provided.

Plain Language Summary Climatic variations can have a significant impact on a number of water-related sectors. Managing such variations through accurate predictions is thus crucial. Continental hydroclimate services have recently received attention to address various user needs. However, predictions for months ahead can be limited at catchment scale, highlighting the need for data tailoring. Here, we compare the predictions from two hydrologic setups at catchment scale. One setup (E-HYPE) is used in a European hydroclimate service, whereas the other (GR6J) is used for local water-related risk assessment. Our results show that predictions from the continental setup can be as accurate as the predictions from the local model when predicting streamflow averaged over several months and when looking at changes in streamflow rather than absolute values. A good estimation of the hydrologic states, such as soil moisture or lake levels, prior to the prediction is the most important factor in obtaining accurate streamflow predictions in both setups. However, the differences in the setups can result in different uncertainties for variables other than streamflow, like in the case of soil water content. We argue that useful information is provided by continental services, yet guidance for information extraction can result into tailored information for regional needs.

1. Introduction

User needs beyond the local and regional scales have led to the development of hydroclimate services at the continental and global scales (e.g., Arnal et al., 2018; Emerton et al., 2018; Thiemig et al., 2015; Wanders et al., 2018). The aim in developing these services is twofold: to bridge the gap between user needs and the current state of climate knowledge (van den Hurk et al., 2016) and to tailor climate information to be readily available to local users (e.g., in hydropower production; Foster et al., 2018). Opportunities have been identified in sectors, such as water resources management (e.g., flood risk assessment), energy, agriculture, tourism, or transportation.

Despite the availability of continental or global hydroclimate services, barriers related to their uptake have been identified in the literature and highlighted along with recommendations for future developments, for example, the relevance for user needs and the need for user engagement and user guidance (Buontempo et al., 2018; Cavelier et al., 2017; Swart et al., 2017). European users from a wide range of sectors currently obtain their weather, climate, and streamflow information from national meteorological and hydrologic services, rather than other sources (Bruno Soares et al., 2018). Water managers and decision makers often require information from locally set up hydrologic models, which are typically used to assess risks, define local governance, and guide decision making. In such scales, model calibration and verification procedures are usually straightforward and provide adequate performance for modeling purposes. Moreover, the information provided to users needs to (at least) be reliable and precise (Swart et al., 2017), which is a challenge for continental and global services, since their setup is strongly dependent on uncertain open global data sets (Crochemore et al., 2019; Kauffeldt et al., 2013) and hydrological models that can only represent limited but dominant processes (Archfield et al., 2015; Bierkens, 2015; Sood & Smakhtin, 2015).

In this context, the evaluation of hydroclimate services across spatial scales becomes necessary to highlight the types of services that can be of use in the different sectors as well as to guide their future development (Vaughan and Dessai, 2014). In hydrologic forecasting, the quality of the information relies on a forecasting chain that includes (at least) meteorological forcing, initial hydrologic conditions and a hydrologic model (Zappa et al., 2011). To these, data assimilation, postprocessing of meteorological forcing (e.g., downscaling and bias-adjustment) or postprocessing of hydrologic forecasts can be added, which influence the quality and uncertainty of the final forecasts, with an expected added value for decision making (Thiboult et al., 2017; Zalachori et al., 2012). Uncertainties in the forecasting chain generally originate from the meteorological data sets used as forcing in model initialization and forecasting, from the physiographic data sets used in model setup (e.g., topography, land cover, and water bodies), and also from the choices made during model identification (e.g., model structure, parameter calibration, and objective functions) (Mazrooei et al., 2015; Sinha et al., 2014). Nevertheless, the uncertainty in the results of continental and global seasonal hydroclimate services is commonly driven only by uncertainties in meteorological forcing and hydrological model structure.

There is still a lack of guidance on how to extract and use hydrologic predictions available from hydroclimate services for management and planning of local facilities and water resources. Even though extensive work has been done on evaluating large-scale models for local applications (e.g., Gudmundsson, Tallaksen, et al., 2012; Gudmundsson, Wagener, et al., 2012; Stahl et al., 2011), a limited number of studies have focused on comparing hydrologic models across spatial scales (Siqueira et al., 2018; Zhang et al., 2016). The choices made during the setup and parameter identification of large-scale models generally lead to lower performance in comparison to models that were set up at the catchment scale (Zhang et al., 2016). Moreover, some components of continental to global model structures are generally too simplistic and far from local water diversions and structural solutions that truly impact populations and their use of water resources (Nazemi & Wheater, 2015; Pechlivanidis & Arheimer, 2015). Previous studies have focused on comparing responses from models at different scales fed with different climate change scenarios (Gosling et al., 2017, 2011; Krysanova et al., 2017; Pechlivanidis et al., 2017). To the knowledge of the authors, there is a lack of studies in the literature comparing models across scales for seasonal forecasting services. Seasonal scales (several weeks to months ahead) have nevertheless important implications for the strategic management of water resources, including sustainable water allocation and anticipation of future conflicts over water use. Despite the recent development of seasonal forecasting systems at continental or global scales, these systems are often evaluated within their model set up, without considering how users with local needs and local models already available can benefit from the additional information brought by large-scale forecasting models.

The objective of this paper is to highlight the potential added value of a continental hydroclimate forecasting service in terms of seasonal forecast information when a local hydrologic model is already set up for local uses. Moreover, the differences between hydrologic models raise the issue of how useful continental models can be for local decision making or, in other terms, how one can efficiently extract reliable forecast information from these models. The European setup of the HYPE model, used operationally at the Swedish Meteorological and Hydrological Institute, and a locally calibrated parsimonious model, GR6J, developed in France for low-flow forecasting (Crochemore et al., 2016; Nicolle et al., 2014), are compared in 10 French catchments. We first identify the variables and time aggregations to set a framework for the

comparison of the performance between the models. We evaluate the forecasts provided by the two models against both observations (reality) and simulations (model reality). Then, we track the skill within each forecasting system and its components, in order to explain the performance in light of the model setups. Lastly, we investigate how input uncertainties propagate through the models by examining intermediate model variables. This results into better understanding whether intermediate hydrologic variables are comparable between models. The paper is organized as follows. In section 2, the hydrologic model setups are presented, along with the comparison and evaluation frameworks. Section 3 presents the results of the different analyses carried out. A discussion is presented in section 4, and finally, section 5 states the conclusions.

2. Data, Methodology, and Comparison Framework

In this section, we present the data used, the modeling conceptualizations, and the common framework developed to compare and evaluate two hydrologic forecasting chains. The chains are based on two distinct hydrologic models: the GR6J model, calibrated and run locally, and the HYPE model, calibrated and run continentally. The models differ from one another in terms of data requirements, model structure, conceptualization and setup.

2.1. Data

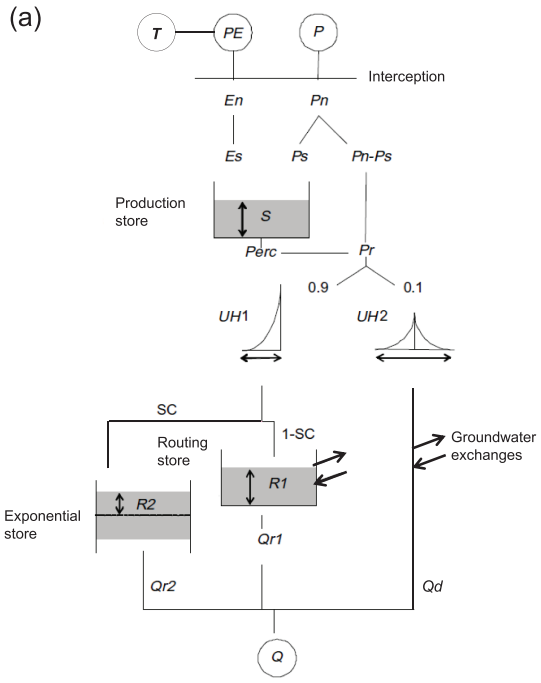
For the setup of GR6J, daily precipitation and temperature were obtained from the SAFRAN reanalysis from Météo-France (Quintana-Seguí et al., 2008; Vidal et al., 2010). This reanalysis covers France at the 8 × 8-km grid resolution. It combines climate model outputs and local (gauge) data. These precipitation and temperature data were used in the calibration of GR6J, as well as in the initialization of the forecast (i.e., a model run forced with observations up to the forecast issue date). Local streamflow observations were also used. They are available from the French national streamflow database (Hydro database: http://www.hydro.eaufrance.fr/). They were used in the calibration of GR6J and as observed reference in the evaluation carried out in this study.

For the setup of E-HYPE, gridded precipitation and temperature were obtained from two global adjusted reanalyzes: Hydro-GFD (Berg et al., 2018) and the WATCH Forcing Data (WFDEI; Weedon et al., 2014). Both are based on the ERA-Interim reanalysis data and are available at a 0.5 × 0.5° resolution (~55 × 55 km). The WFDEI data set was used in the first calibration phase of E-HYPE carried out at the Swedish Meteorological and Hydrological Institute (SMHI) prior to this study. Hydro-GFD was used in a second calibration phase to adjust the water balance. In this study, Hydro-GFD was also used to force E-HYPE up to the forecast issue dates and thus create initial hydrologic states. Streamflow observations used in the calibration of E-HYPE were obtained from the Global Runoff Data Centre (GRDC) Reference Dataset (http://www.bafg.de/GRDC/), which has a global coverage and whose river stations across France come from the Hydro database.

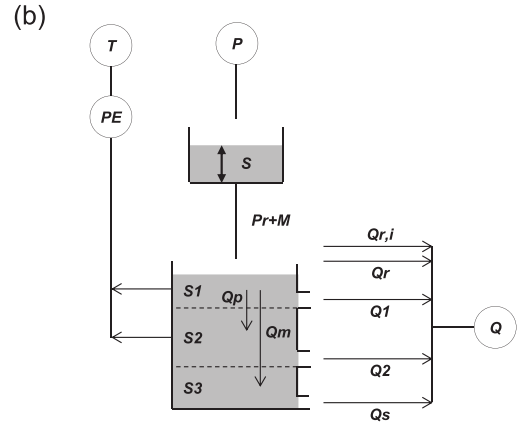
For both models, precipitation and temperature forecasts are obtained from the System 4 European Centre for Medium-Range Weather Forecasts (ECMWF) Global Circulation Model (GCM). System 4 reforecasts for the period 1981–2009 are available at a spatial resolution of about 0.7° and are initialized at the beginning of each month (Molteni et al., 2011). Each forecast ensemble consists of 15 members covering the 7 months ahead. Forecasts were previously bias adjusted based on the Hydro-GFD data set and using the Distribution-based Scaling method (Yang et al., 2010). Both models run with daily weather input at the catchment scale and generate daily hydrological outputs. Gridded forecast weather variables are associated to catchments based on the shortest distance between grid cells and catchment centroids. Our study period ranges from 1981 to 2009.

2.2. Model Specificities: Conceptualization and Calibration Strategy

Differences between the two forecasting systems are partly due to the different model structures (Figure 1). The rainfall-runoff model GR6J (Pushpalatha et al., 2011) is a lumped, conceptual, and parsimonious model. It requires the identification of six parameters. The European setup of the HYPE model (Lindström et al., 2010), known as E-HYPE (Hundecha et al., 2016), is a distributed process-based model with about 35,000 subcatchments and a median resolution of 215 km^2. The contribution of each subcatchment unit is first determined, accounting, for example, for snow, ice, evapotranspiration, soil moisture, groundwater fluctuations, aquifers, and human alterations, before being routed through rivers and lakes. The model requires the



P : Precipitation
T : Temperature
PE : Potential evapotranspiration
Pn : Net rainfall
En : Net evapotranspiration
Ps : Portion of the net rainfall going to the production store
Es : Evaporation applied to the production store
S : Production store level
Perc : Water percolating from production store
Pr : Surface runoff and water percolating
UH1 and UH2: Unit hydrographs 1 and 2
SC : Splitting coefficient of effective rainfall
R1 and R2 : Levels of routing stores 1 and 2
Qr1 : Flow from routing store
Qr2 : Flow from exponential store
Qd : delayed flow contributing to the catchment outlet
Q : Resulting streamflow



P : Precipitation
T : Temperature
PE : Potential evapotranspiration
S : Snow
Pr : Rainfall
M : Snowmelt
S1, S2, S3 : Soil layers 1, 2 and 3
Qr,ii : Surface runoff (low infiltration)
Qr : Surface runoff
Qp : Percolation
Qm : Macropore flow
Q1, Q2 : Runoff layers 1 and 2
Qs : Groundwater flow
Q : Resulting streamflow

Figure 1. Schemas presenting the structures of (a) the GR6J model (modified from Pushpalatha et al., 2011) and (b) the HYPE model (modified from Nijzink et al., 2016, and structure presented in Lindström et al., 2010). The main input data, internal states and flows are indicated.

identification of about 170 parameters, which mainly depend on soil and land use characteristics and which were identified for the European domain prior to this study (Hundecha et al., 2016).

In addition, the models have fundamentally different rationales, which lead to different setups. The GR6J model was developed to represent processes that are integrated at the catchment scale, independently from what is done in neighboring catchments. Its development also relies on parameters identified by optimizing the simulated flow against the observed streamflow at the outlet of the catchment. This model setup has advantages for operational use, when a specific user is interested in having a hydrologic model to simulate streamflow at a specific gauged outlet for a well-defined decision-making problem (e.g., for flood forecasting or reservoir inflow forecasting). In this study, we follow the rationale of this catchment-specific setup. For each catchment, the leave-one-year-out method is used to derive a parameter set for each year of the study period. Calibration for a given year is thus based on data from all other available years. A unique parameter set is obtained in each catchment and for each year. The parameters established for a given year are then used to validate the model when running the forecasts for that year.

The rationale behind the development of the E-HYPE model was that the model should adequately represent a wide range of hydroclimatic variables (e.g., snow, evapotranspiration, streamflow, reservoir regulation, and irrigation management) and, at the same time, provide optimal streamflow and water balance simulations over the whole European continent. Hence, calibration was not performed for each individual catchment and streamflow gauging station. Instead, E-HYPE was calibrated on a set of 115 European catchments, which were chosen as representative of the diversity of climate, soil, land use, and human main impacts over the modeling domain (Hundecha et al., 2016). Parameters are calibrated for combinations of soil and land use types and then propagated to ungauged catchments based on soil and land use information. Model validation is carried out over the set of European catchments for which observed streamflow data are available. In this study, we used the parameter sets identified by Hundecha et al. (2016) over the calibration period (1980–1999) to run the forecasts over the study period (1981–2010). Despite the time overlap between the calibration of E-HYPE and the study period, none of the catchments in this study were directly used to calibrate E-HYPE.

The model structures of GR6J and HYPE also differ despite fundamental similarities. In GR6J (cf. Figure 1a), daily precipitation for the upstream area contributing to the outlet is first intercepted before either contributing directly to surface runoff or reaching a production store and then percolating. The largest part of the resulting flow (90%) is then delayed before being split between two additional stores whose outflows contribute to the river outflow. The remainder (10%) is also delayed before contributing to the river outflow. Groundwater exchanges are taken into account before the final river outflow is computed from all three components. In HYPE (cf. Figure 1b), daily precipitations for the subcatchment area first contribute to a snow store. Rainfall and, if applicable, snow melt then reach the soil reservoir, which is divided in three layers. This water flow can either contribute to runoff directly or transit through the soil layers. In each layer, part of the water contributes to the river flow, part reaches deeper soil layers by percolating or through macropores, and part evaporates (upper two soil layers). The final contribution to the river outflow results from the surface runoff and from the contributions of each soil layer, as well as from the inflow from upstream subcatchments. Note that in both hydrological models, potential evapotranspiration is estimated from temperature using the modified Jensen-Haise model (Oudin et al., 2005).

2.3. A Framework for Continental to Local Analysis

The basic step to establish a common framework to compare the two forecasting chains is to select a joint study area and period for the evaluation of both models. The individual model setups are then used to generate ensemble seasonal forecasts over the chosen area and period.

The study area comprises 10 catchments in France, which are not strongly influenced by snow (less than 10% of solid precipitations) (Table 1 and Figure 2). These catchments were selected based on observed river flow availability, geographical spread, and variability of hydrological behaviors in catchments not affected by snow.

Two additional steps need to be carefully considered to establish an objective framework for comparison. First, it is necessary to check that the chosen catchments are similarly defined in both models. For streamflow simulation, this means ensuring that the upstream areas do not differ much between models. Differences may be observed, for instance, when different digital elevation models or geographic information systems are used to delineate a catchment. Automated techniques for the topographic delimitation of catchment areas, typically used in continental modeling frameworks in order to cover a wide domain, may introduce significant location errors that propagate to the estimation of upstream areas and, consequently, of streamflow. In this study, we verified that the difference in catchment area between the GR6J and E-HYPE setups did not exceed 10% of the total area (an acceptable error for elevation global data sets according to Donnelly et al., 2013; Kauffeldt et al., 2013; cf. Table 1). Second, since the models are going to be evaluated under forecasting mode, it is important to ensure that each model represents reasonably well the variability of the catchment's hydrologic response, when forced by observations. To ensure this aspect, we evaluated the correlation of each model for each catchment, computed over a long time series of observed data. We verified that the correlation of each model was at least greater than 0.7 in all catchments selected in this study.

Table 1
Catchment Name, Rank Number in Terms of Area (as Used in the GR6J Model), Areas From Both GR6J and E-HYPE Setups, and Performance in KGE and Pearson Correlation Coefficient in the 10 Studied Catchments

		Area (km^2)		KGE		Correlation coefficient	
Catchment	#	GR6J	E-HYPE	GR6J	E-HYPE	GR6J	E-HYPE
L'Orne Saosnoise à Montbizot	1	501	525.1	0.89	0.76	0.90	0.81
La Briance à Condat-sur-Vienne	2	605	612.8	0.85	0.63	0.89	0.72
La Seiche à Bruz	3	809	819.1	0.93	0.62	0.93	0.87
La Petite Creuse à Fresselines	4	853	861.7	0.91	0.67	0.94	0.80
La Sèvre Nantaise à Tiffauges	5	872	858.5	0.93	0.69	0.96	0.79
La Vire à Saint-Lô	6	882	895.4	0.96	0.69	0.96	0.89
L'Orge à Morsang-sur-Orge	7	934	943.6	0.87	0.68	0.87	0.75
L'Eyre à Salle	8	1678	1682.2	0.94	0.77	0.97	0.90
La Meuse à Saint-Mihiel	9	2543	2553.6	0.95	0.87	0.96	0.87
L'Oise à Sempigny	10	4320	4262.7	0.94	0.59	0.95	0.87

This study focuses on evaluating model performance for streamflow predictions, based on observed input data (hereafter called "simulations") and on forecast input data (hereafter called "forecasts"). In the 10 studied catchments, models were evaluated based on simulations obtained for the period 1981–2009 and based on the Kling-Gupta efficiency (KGE; Gupta et al., 2009; cf. Appendix A). In calibration and validation in

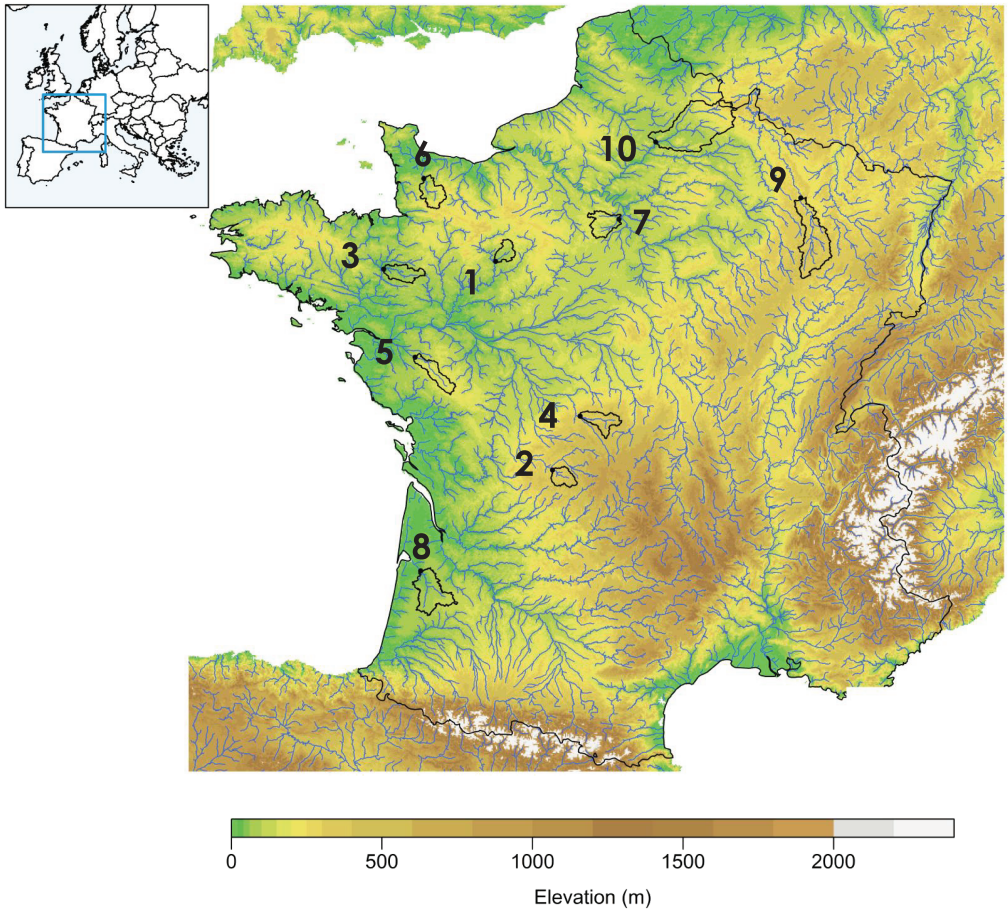


Figure 2. Location of the 10 studied catchments in France. Catchments are numbered according to their upstream areas, from the smallest (#1) to the largest (#10) (see Table 1).

these catchments, the GR6J model reaches an average KGE of 0.92, all values being above 0.85. In simulation, the E-HYPE model reaches an average KGE of 0.72, with all values in the range of 0.59 to 0.87.

For the evaluation of forecasts, four ensemble systems, each with 15 members, are considered (Figure 3). These are based on model runs for the period 1981–2009,which are initialized on the first of each month and cover the 7 months ahead at the daily time step (i.e., 348 forecast start dates):

- "Observed climatology": An ensemble based on observed streamflow data from the local Hydro database is used. For each forecast date, traces are randomly selected within the available streamflow time series for the catchment. Fifteen streamflow traces are selected to match the number of members available from System 4. Candidate time series start on the same day as the forecast date in all available historical years, excluding the forecast year.
- "Simulated climatology": Historical simulated streamflow time series for each hydrologic model are considered. At each forecast date, 15 traces are randomly selected from historical time series of streamflow simulations of each model in order to build an ensemble of possible outcomes that matches the number of System 4 outcomes. Here also, traces are selected from candidate time series that start on the same day as the forecast date in all available historical years, excluding the forecast year.
- Ensemble streamflow prediction (Day, 1985; Wood & Lettenmaier, 2008; abbreviated "ESP" hereafter): The hydrologic model is initialized for the forecast date and then fed with precipitation and temperature traces starting on the forecast date and selected from historical precipitation and temperature records excluding the forecast year. Fifteen precipitation and temperature traces were randomly selected from the historical period to match the number of members available from System 4. In contrast to the "simulated climatology" ensemble, the ESP ensemble takes into account the initialization of the forecast model at the forecast date and hence includes information on initial hydrologic conditions (ICs).
- GCM-based streamflow forecast (abbreviated "SYS4" hereafter): Bias-adjusted ECMWF System 4 precipitation and temperature forecasts are used as input to the hydrologic models.

These different forecasting ensembles cover the possible range of combinations of the main components in a forecasting chain, that is, hydrologic model (HM), initialization of the hydrologic model states to reproduce the hydrologic conditions on the day the forecast is issued (IC), and GCM seasonal forcing (GCM) (see Table 2).

2.4. Evaluation Framework

A range of streamflow-related predictands extracted from model simulation runs (i.e., based on observed input data) are evaluated based on the correlation coefficient to assess timing errors. The predictands computed for each model are the streamflow values, the variation in streamflow from one time step to the next, the sign of this variation, the anomaly in streamflow from the model mean value, the sign of this anomaly, and the detection of the median streamflow and streamflow terciles and quartiles (cf. Appendix A for streamflow characteristics formulation). Hereafter, percentiles are expressed as nonexceedance percentiles, meaning that Q25 (Q75) is the flow exceeded by 75% (25%) of the values and is therefore a low-flow (high-flow) threshold. In addition to the daily time aggregation, streamflow statistics are computed for weekly, monthly, and 3-month streamflow averages to highlight the impact of the time aggregation on model performance.

The overall performance of the forecasts (i.e., based on forecast input data) is evaluated based on the Continuous Ranked Probability Score (CRPS), which compares the forecast distribution to the step function corresponding to the observation (Hersbach, 2000; cf. Appendix A). We also use the interquantile range to quantify the spread of the forecast ensemble members. It is calculated as the difference between the 95th and the 5th percentiles of the forecast distribution. In order to compare the spread of different variables (i.e., precipitation, streamflow, and internal model states), we normalized the interquantile range by dividing it by the ensemble mean for each forecast date and lead time. The adimensional interquantile (AIQ) range is then used as evaluation metric to compare the spread of different variables, that is, precipitation, streamflow, or internal model states. All scores are computed for a given catchment, forecast date and lead time.

Forecast performance is commonly compared to a benchmark to translate quality into gain or loss in performance, that is, forecast skill (Pappenberger et al., 2015). The skill in CRPS (CRPSS hereafter) thus compares the performance of both forecasting system to benchmarks (see Appendix A for formulation). In this study, we use benchmarks that allow us to highlight and isolate forecast skill according to the different components

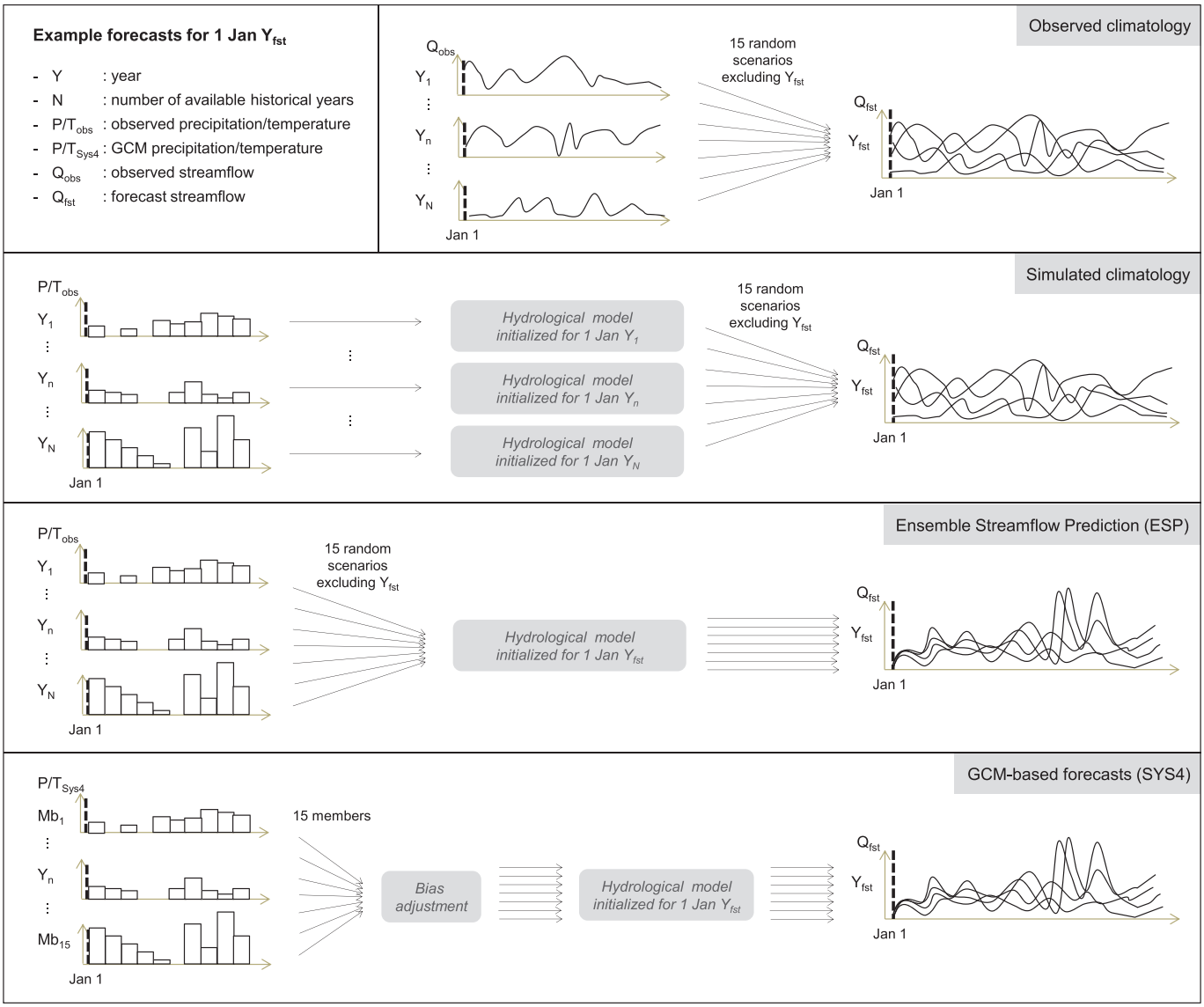


Figure 3. Flowchart describing the construction of the four investigated ensemble forecasts.

in the forecasting chain, namely, the hydrologic model (HM), the initialization of the forecasts (IC), and the use of GCM forcing (GCM). Skill will thus reflect the skill of the components that are present in the evaluated systems but absent in the benchmark system (cf. Table 2).

Figure 4 illustrates the different possible combinations between the forecast ensembles. It also shows the evaluation strategy adopted here: For instance, when we want to evaluate the skill brought by the hydrologic model (HM) alone, we evaluate the "simulated climatology" ensemble system using the "observed climatology" system as benchmark. If we want to evaluate the additional skill brought by constraining the forecasts

Table 2
Forecast Ensembles Investigated and Their Forecasting Chain Components

	Hydrologic model (HM)	Initial hydrologic conditions (IC)	GCM seasonal forcing (GCM)
"Observed climatology"	no	no	no
"Simulated climatology"	yes	no	no
"ESP"	yes	yes	no
"SYS4"	yes	yes	yes

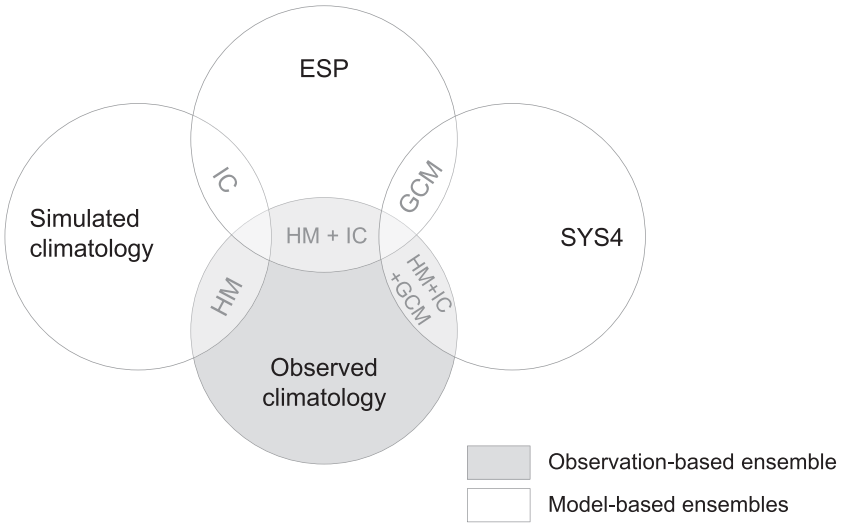


Figure 4. Combinations of ensemble systems and benchmarks for forecast skill evaluation. Intersections between circles indicate the evaluated component of the forecasting chain: the hydrologic model (HM), the initialization of the forecasts (IC) and the use of GCM forcing (GCM).

to depart at the initial conditions (ICs) obtained at the time of forecast, then we evaluate the "ESP" ensemble using the "simulated climatology" ensemble as benchmark, and, finally, in order to evaluate the gain or loss in skill from using GCM seasonal forecasts, we evaluate the "SYS4" system against the "ESP" ensemble. Note that when the "observed climatology" ensemble is used as benchmark in the computation of skill, the reference used is observed streamflow from the Hydro database. When the "simulated climatology" ensemble or "ESP" is used as benchmark, the streamflow simulated with each model is used as reference. In this case, since the simulations depend on the model, different references are used to evaluate the systems depending on the model used. This could lead to pitfalls in our analysis if one of the models failed to represent dominant hydrologic processes and their dynamics. To avoid this and ensure a fair comparison, we selected catchments where both models perform well in terms of the correlation coefficient, since it is a measure that penalizes errors in timing (Table 1).

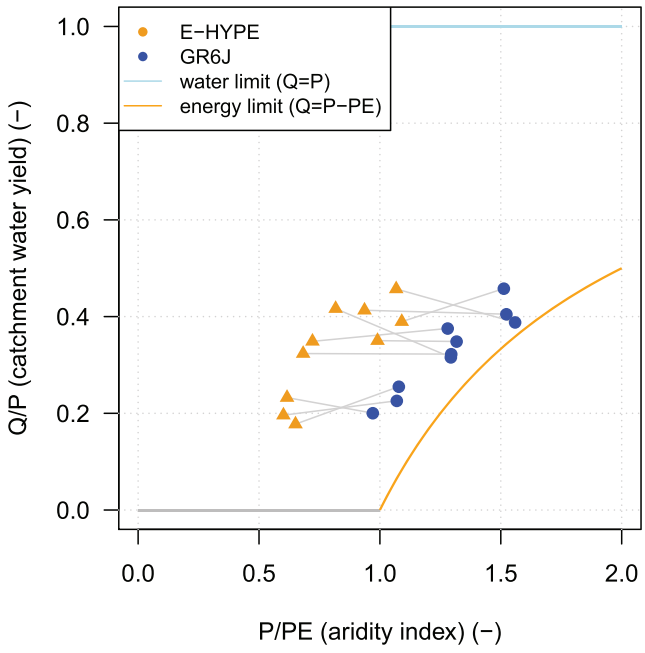


Figure 5. Turk-Budyko representation of the 10 catchments as seen by E-HYPE (orange triangles) and GR6J (blue circles). Dots corresponding to a same catchment in E-HYPE and GR6J are linked by a straight line.

3. Results

We first analyze the model data and observations in the historical period to highlight errors in model setups, structure, and parameterization, assuming that forcing input is perfect (section 3.1). We then add the forecasting components (i.e., the GCM and ESP forcing) and focus on the model performance in forecasting streamflow (section 3.2). Lastly, we analyze the states conceptualizing soil water content in the two models in order to identify whether results can extend to internal states (section 3.3).

3.1. How Do the Models Compare in Simulation?

3.1.1. Water Balance Analysis

Each and every model, depending on the resolution and origin of the data used in its setup, will have a different representation of the fluxes. One way to highlight this is to look at the main elements of the water balance, that is, precipitation, evapotranspiration, and streamflow. The model differences in water balance components in the 10 studied catchments are depicted using the Turc-Budyko diagram (Figure 5). Water yield is generally similar between the two model setups (with the exception of two catchments in which runoff coefficient between the models differs

maximum by 0.1); however, the aridity index differs in a consistent way, with higher values in the GR6J setup. Higher potential evapotranspiration values are estimated in the E-HYPE than in the GR6J setup (on average 380-mm difference). These differences in the aridity indices indicate that the model states (i.e., soil moisture) might vary between the two models, further affecting the forecasts.

In both E-HYPE and GR6J, potential evapotranspiration is estimated using the modified Jensen-Haise model but with the existing E-HYPE parameterization leading to higher potential evapotranspiration than in GR6J. Moreover, in E-HYPE, parameters are identified for each land use type against satellite-based potential evapotranspiration data (Moderate Resolution Imaging Spectroradiometer global data set; Mu et al., 2011), and annual potential evapotranspiration from E-HYPE and Moderate Resolution Imaging Spectroradiometer should thus match over the entire model domain. In model identification, parameters can compensate for meteorological biases and reach similar streamflow results nonetheless. For instance, the 380-mm difference between E-HYPE and GR6J in terms of potential evapotranspiration reduces to 110 mm in actual evapotranspiration. E-HYPE's multivariable calibration strategy (tune parameters to simultaneously respect streamflow, evapotranspiration, and snow) differs from the GR6J calibration approach, which tunes the parameters to fit only streamflow, hence resulting into these internal flux differences.

3.1.2. Streamflow Evaluation—Statistics and Time Aggregations

We next evaluate the performance of the two models in reproducing key streamflow statistics often encountered in forecasting services. The objective is to compare how the catchment and continental models perform with perfect input data and detect whether the continental model can yield local information with a quality equivalent to the one displayed by the catchment model.

As expected, GR6J outperforms E-HYPE in terms of correlation (r; Figure 6), given that GR6J is identified in each catchment based on local climatological observations and against observed streamflow time series. Time aggregations have a clear impact on the model performance, with larger aggregations ensuring better performance for both models and reducing the gap between the two models. Nevertheless, this varies with the statistics. In terms of streamflow value, aggregations over longer time windows increase performance overall. Except for the sign of the variation in streamflow, which has a continuous improvement with larger time aggregations, and threshold detection, statistics tend to reach a plateau without significant improvement beyond a certain point. For the variation and anomaly in streamflow, streamflow values themselves, and for the sign of the anomaly, the monthly aggregation already provides high performance.

An analysis of differences in correlation, for all statistics and all time aggregations showed that the number of cases (i.e., combinations of statistics, time aggregation, and catchment) in which E-HYPE performs as well as or better than GR6J increases with larger time aggregations. At the 3-month time step, E-HYPE performs better than GR6J in 16% of the cases. These cases occur for when detecting thresholds or variations in flows. Longer time aggregations also reduce the difference in performance between GR6J and E-HYPE in terms of streamflow values and anomalies in flows. Reasons for this increase in performance in E-HYPE simulations at larger time aggregations may be explained by a parameter identification that prioritizes annual streamflow regimes rather than daily dynamics, as well as the fact that Hydro-GFD is optimized against local observations at the monthly time step.

3.2. How Do the Models Compare in Forecasting?

Factors apart from meteorological data, model structure, and model parameters affect the skill of hydrologic forecasting systems. Here, we first present an overall evaluation of the two forecasting systems and then look at the sensitivity of their skill to the two components added when forecasting, that is, the initialization of the hydrologic model states and the input of GCM forecasts.

3.2.1. Overall Performance in Forecasting Mode

Two different benchmarks are used in the computation of the skill based on (1) “observed climatology” and observed streamflow and (2) “simulated climatology” and simulated streamflow (cf. Figure 4). When comparing “SYS4” to “observed climatology” (Figure 7, top row), the forecasts produced by GR6J are skillful sometimes up to 4 months ahead, whereas E-HYPE forecasts have little to no skill (up to 1 month). This result also observed in the previous section was expected since an evaluation of skill using observed climatology takes into account the performance of the hydrologic model with respect to local observations. However, when the benchmark is “simulated climatology” and hence model performance is omitted from

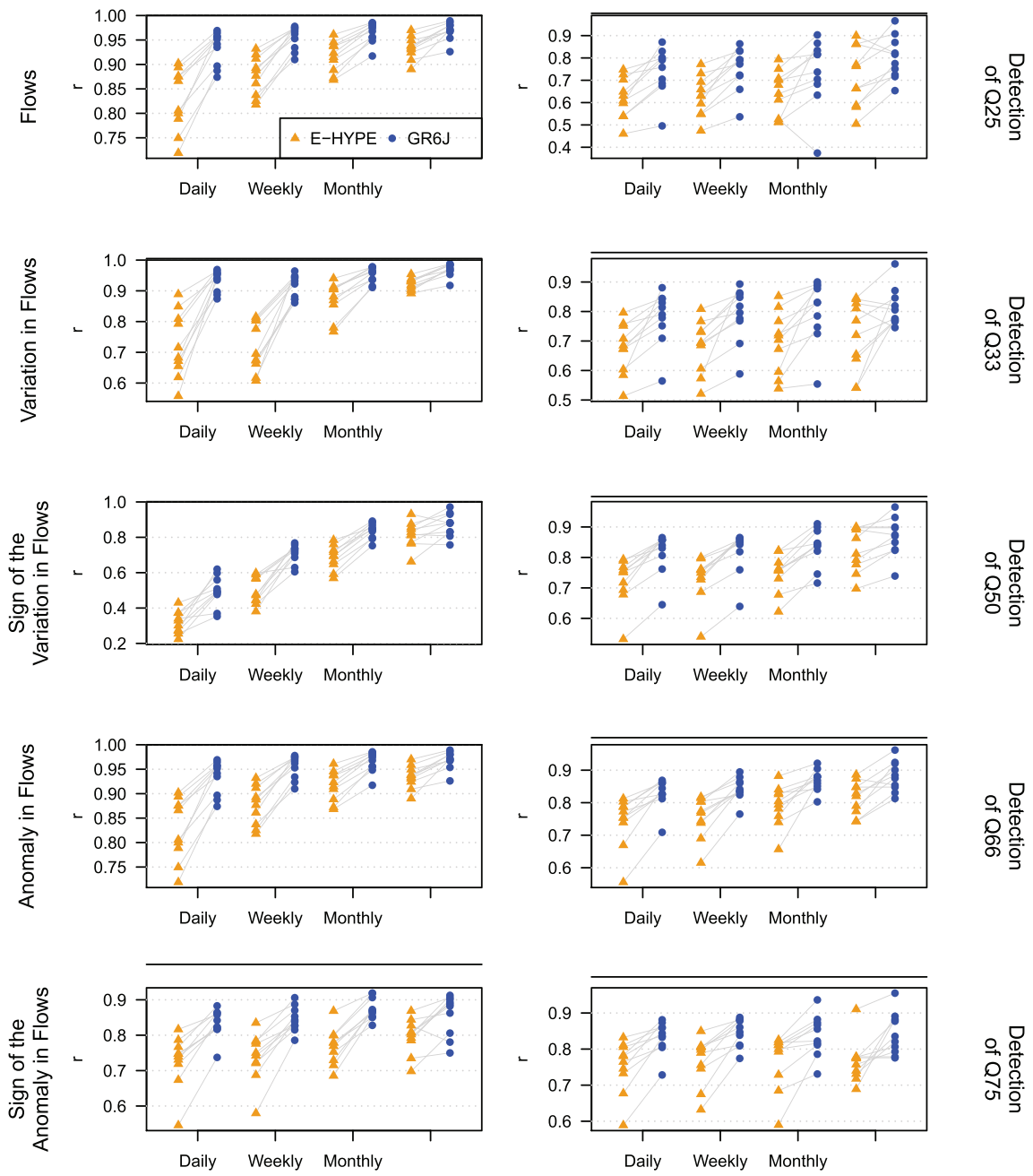


Figure 6. E-HYPE (orange triangles) and GR6J (blue circles) performance (in terms of correlation) for 10 streamflow statistics. Four temporal aggregations are applied: daily, weekly, monthly, and 3-month. Within each aggregation, for the same catchment, GR6J and E-HYPE performance are linked with a straight line.

the skill analysis (Figure 7, bottom), skill is accounted for by the meteorological forecast inputs and the process representation influencing the initial hydrologic conditions. Here, the two models have similar skill and their values are generally higher than when "observed climatology" is used as benchmark. Both systems generally have skill up to 5 to 10 weeks, and in some catchments up to 25 weeks for a forecast issued in May. Results highlight that when a hydrologic model has a good performance, information is extracted from both benchmarks.

Figure 8 presents the difference in skill between "SYS4" produced with E-HYPE and GR6J, showing the better system in forecasting the 1-week, 1-month, and 3-month ahead streamflow. The model identified at catchment scale outperforms the continental model when "observed climatology" is the benchmark,

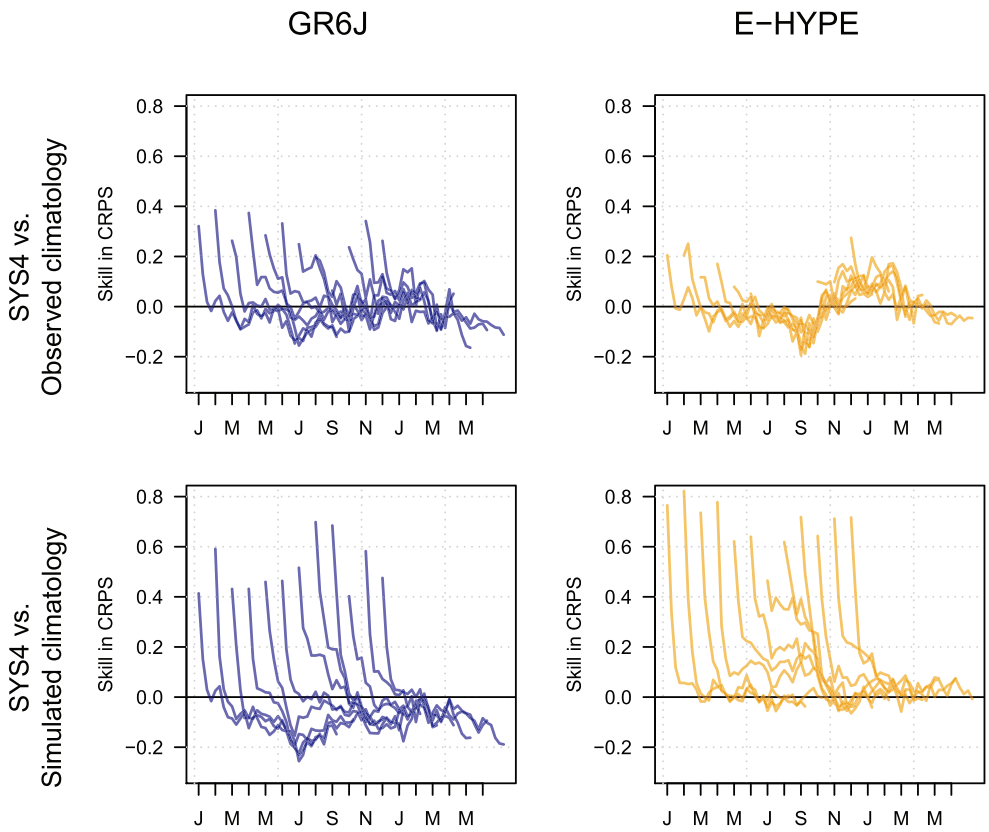


Figure 7. CRPSS of "SYS4" generated with GR6J (blue, left) and E-HYPE (orange, right) against lead time. Skill scores are computed against "Observed climatology" (upper row) and "Simulated climatology" (lower row).

especially in the dry summer season. In the winter season, observed differences in skill are very small and both systems are close with regard to observations, especially when forecasting the 3 months ahead. When "simulated climatology" is used as benchmark, the continental model clearly achieves higher skill than the catchment model, in all months, except in summer when differences are smaller, especially when forecasting the week ahead.

3.2.2. Skill Mainly Comes From Initializing Hydrologic Conditions

An in-depth diagnostic can be carried out by looking at the different components of the skill. Here we decompose the forecasting skill of "SYS4" and assess its sensitivity to meteorological forecasts and ICs (Figure 9).

Results show that in all seasons and for both models, the highest gain in skill comes from the initial hydrologic conditions. The gain is higher in spring and summer than in autumn and winter and can be observed up to 25 weeks. GCM forecasts yield some additional gain for the first month in the continental setup, with almost equivalent gain to ICs in the autumn and winter seasons. However, for GR6J, the GCM forecasts do not yield gain, except for the very first weeks of winter. Here, it is assumed that the gain in E-HYPE forecasting skill using GCM forecasts is due to the meteorological data set used in the calibration and warmup. Indeed, ECMWF System 4 forecasts are bias adjusted to Hydro-GFD, which was also used in the E-HYPE calibration and warmup. In the case of GR6J, which was calibrated and warmed up with local observations, the large-scale data set is not close enough to local observations for bias adjustment to bring the forecasts closer to the local model climatology. Consequently, in the case of the catchment model, running the hydrologic model with past precipitation and temperature climatology is the best option overall.

3.3. Can Internal States Explain Differences Between Models?

Model structure has a major role in predicting catchment hydrologic response and thus in the skill of the forecasting systems. Here, we analyze the precipitation, soil water content, and streamflow to understand how uncertainty propagates throughout model components and further assess the impact of model

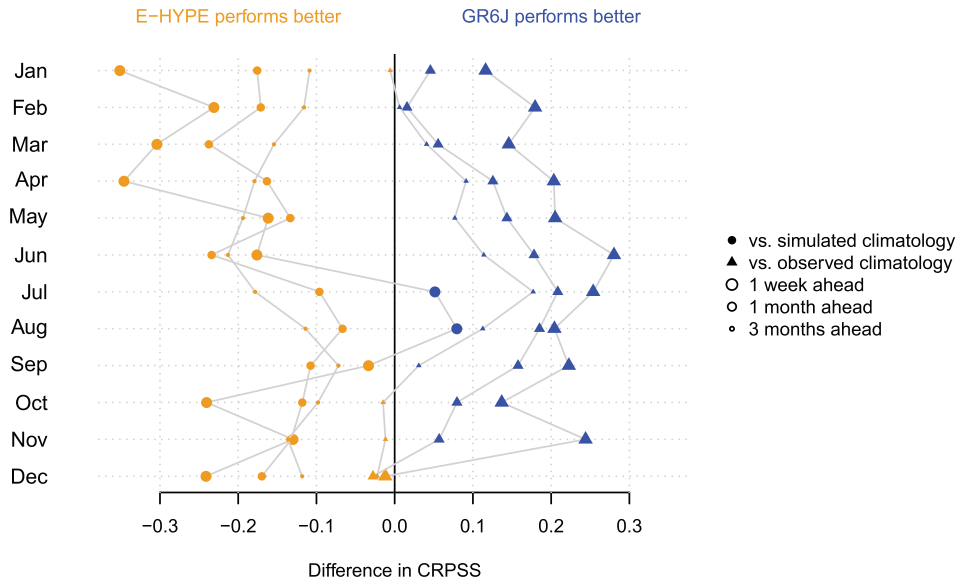


Figure 8. Difference in CRPSS of "SYS4" averaged over the first week, month, and 3 months, respectively, for each month of the year. Each line represents the skill based on a different benchmark and time horizon: "Simulated climatology" (dots) and "observed climatology" (triangles). When GR6J (E-HYPE) performs better, dots are blue (orange) on the right (left) hand side.

structure on the SYS4 forecast spread. In GR6J, the soil water content is defined as the content of the production store, while in E-HYPE, it is the soil moisture root zone (upper two soil layers) as a fraction of the soil available for evapotranspiration but not for runoff (cf. Figure 1b). To allow for intercomparability of these variables, we analyze the evolution in time of the standardized variables rather than absolute values in both models.

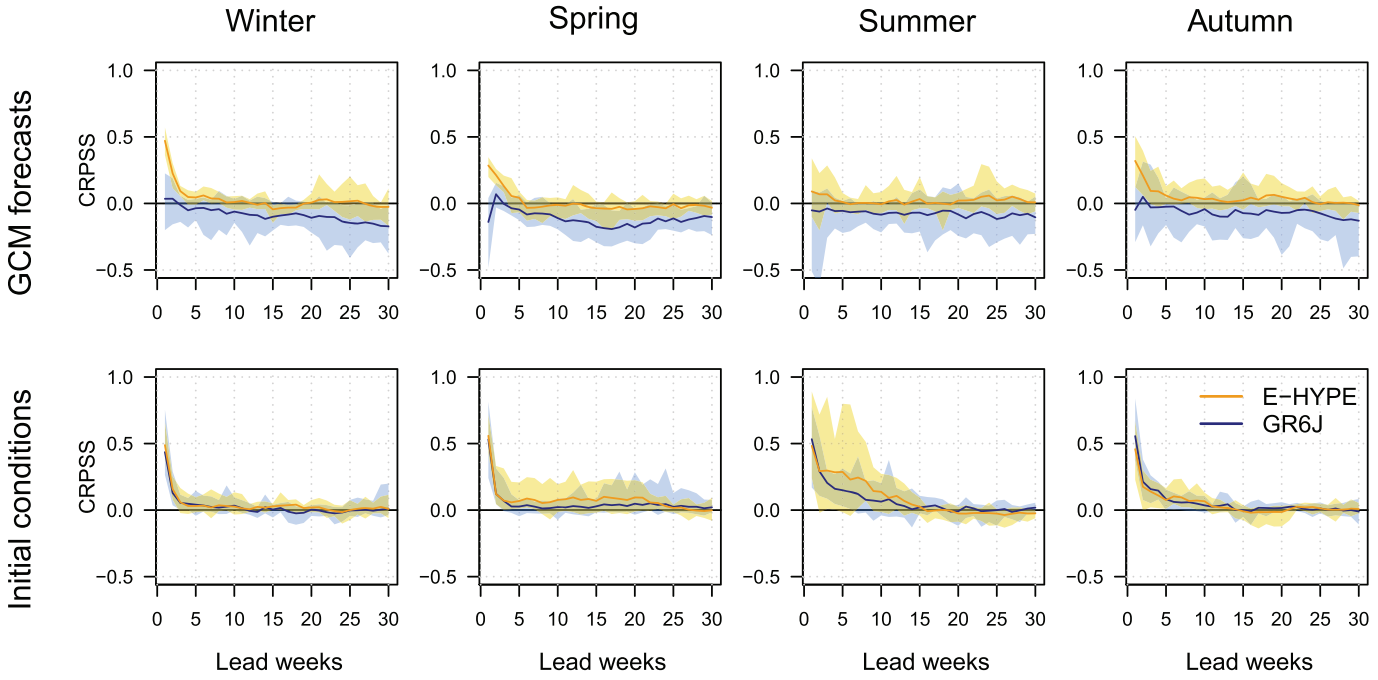


Figure 9. CRPSS for different lead times and elements of the forecasting chains for E-HYPE (yellow) and GR6J (blue). The upper row presents the gain in skill from using GCM forecasts as forcing ("SYS4" evaluated against "ESP"), while the bottom row presents the gain in skill due to initializing the model states ("ESP" evaluated against "simulated climatology"). The shades represent the range in skill from the 10 catchments and the solid line represents the mean skill.

Figure 10 shows the correspondence between the two model variables in each catchment and season. Results show that the two variables are generally strongly correlated (Spearman correlation varying between 0.82 and 0.98) with the trends during spring and summer being nonlinear. This nonlinearity is likely linked to the emptying of the soil reservoir and may be attributed to the lower aridity index in the E-HYPE setup (cf. Figure 4) allowing for a faster drying of the soil than in the GR6J setup. Three catchments located in the northwestern part of France, namely, Catchments 1, 3, and 6 (cf. Figure 2), stand out with close-to-linear relationships between the model variables (Pearson correlation of 0.98). An analysis of these three catchments in the E-HYPE setup showed that they have simple land use and soil structures: They are almost solely occupied by rainfed agricultural lands and pastures, and their soil is predominantly medium fine to very fine. All other catchments generally exhibit a large heterogeneity with multiple fragmented land uses and coarse soils. The E-HYPE soil parameters allow for a slower recession and a more limited percolation in medium fine to fine soils than in coarse soils, leading to a slower emptying of the soil reservoirs. This difference in soil type is likely to be the main reason for the close behavior of the two models in these three catchments and the fast emptying of soil reservoirs in all other catchments. Indeed, these three catchments do not otherwise stand out from the others in terms of climatology, seasonality, upstream area, nor model performance.

Figure 11 presents the forecasts spread (AIQ) in different components (precipitation, soil water content, and streamflow) of the hydrologic models. Overall, the forecast uncertainty is greater in precipitation in comparison to the other variables, and results show that the catchment is behaving as a filter reducing significantly this spread. It is interesting to note that the uncertainty in streamflow is always greater than that in soil water content for the GR6J model; however, the analysis showed a different pattern in E-HYPE and particularly in summer. The uncertainty in forecast precipitation and water content is greater in summer when soils are not saturated, whereas the uncertainty in forecast streamflow is greater in the other seasons when flows are typically higher. Therefore, in both models, uncertainties in model states do not translate directly through the model and higher uncertainties in inputs or model states do not necessarily lead to higher uncertainty in outputs. Finally, we highlight that the spread of the results between catchments is higher for E-HYPE than for GR6J. Catchments 1, 3, and 6 (having similar patterns in soil water content in E-HYPE and GR6J) follow similar paths in E-HYPE and GR6J, although this does not necessarily imply similar uncertainties in forecast streamflow. This could be related to the model structure and how explicitly catchment heterogeneity is represented in the model.

4. Discussion

The issue of the added value of continental models where catchment-specific models are available is increasingly relevant. For users to optimally extract the seasonal information, it is necessary to provide a framework on the setup of large-scale systems and on assessing expected information content from different forecasting systems.

4.1. Information Extraction to Address User Needs

Despite the high performance of the catchment-specific GR6J model for all statistics and time aggregations, outputs from the continental model could yield more accurate hydrologic information depending on the time aggregation and statistics. Here, E-HYPE performs equally to GR6J in almost half of the cases at the 3-monthly time step, highlighting that the information yielded by continental models is highly dependent on time aggregation and model statistics.

The continental E-HYPE model generally showed better performance in the "model reality" (i.e., when compared against model simulations) and, therefore, in predicting anomalies or any statistics relative to model climatology. This finding is consistent with the results from Zhang et al. (2016), who identified potential in predicting trends in streamflow based on historical simulations from large-scale models. Anomalies are not penalized by constant biases (amplitude) in model simulations; however, they are penalized by errors in timing, for example, due to a wrong representation of some processes, such as aquifer recharge. Therefore, a continental model may not reproduce the water balance correctly (in specific regions) and still provide better performance in terms of anomalies. It can also be argued that a model that forecasts anomalies is not necessarily expected to forecast the dynamics precisely. A correct signal could be enough to forecast anomalies

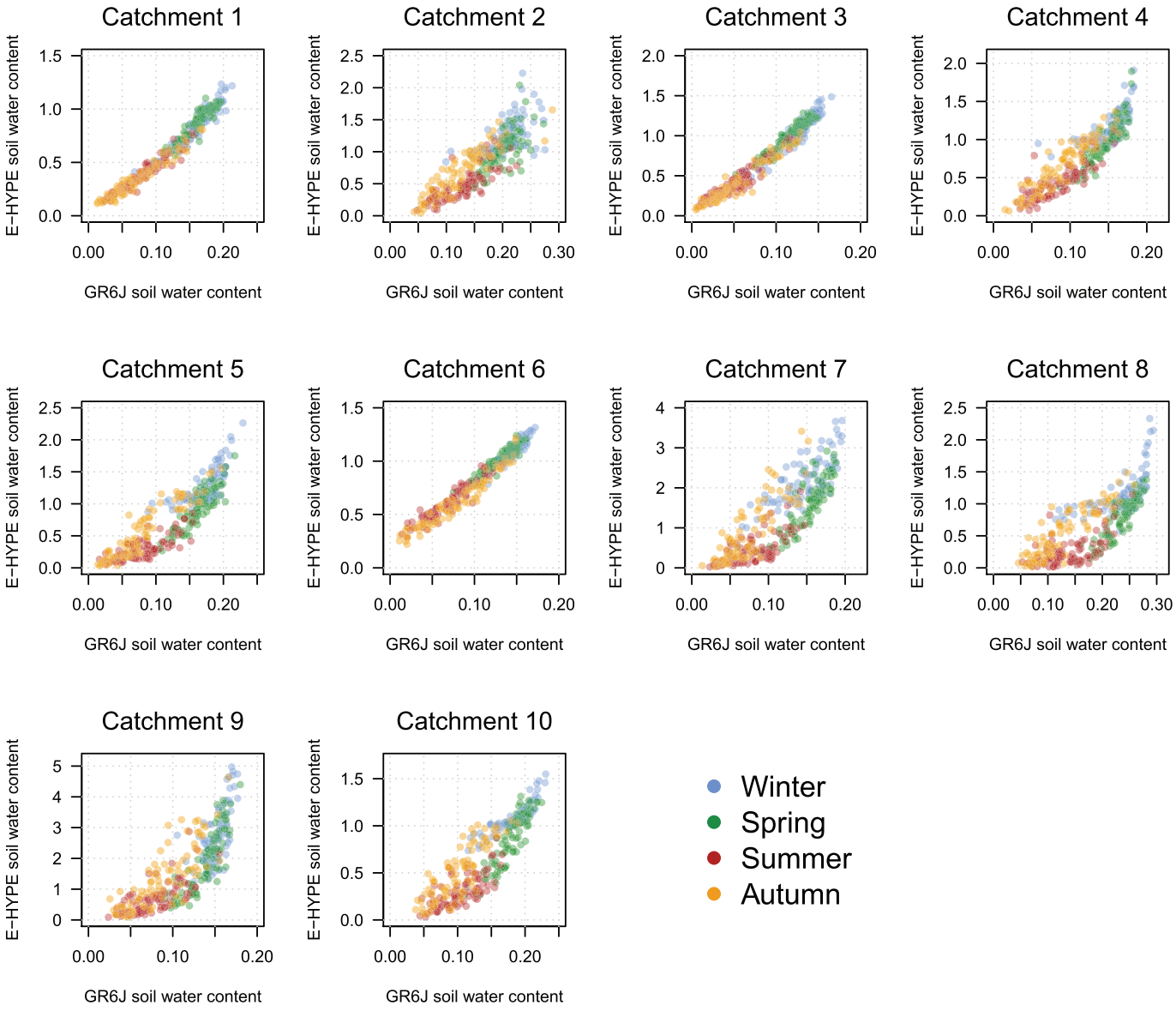


Figure 10. Scatterplots of the soil water content (internal model variable) for E-HYPE (*y* axis) and GR6J (*x* axis) in each catchment. Colors indicate seasons. Each dot corresponds to a forecast date (here the average of the first three lead days was used).

and meet operational needs if needs are restricted to such statistics (i.e., relative information and probabilities of being above/near/below normal conditions).

4.2. Link to Physiographic Representation

GR6J does not explicitly distinguish between soil and land use classes in the catchment but instead conceptualizes these physiographic characteristics in its catchment-specific parameters. In E-HYPE, however, a more explicit and complex setup is available to cover catchment physiographic variability over the continent. This leads to differences in the way catchments are treated in terms of internal states, but not necessarily in terms of output streamflow. Nevertheless, it is a common practice in continental-scale models to account more explicitly for differences in land use and soil types and react internally accordingly. This practice is driven by the need to set a model capable of being regionalized to a large variety of catchments.

In our set of catchments, the hydrologic response of the catchment-specific model was similar to the response of the continental model in catchments where medium fine to very fine soils dominate. In

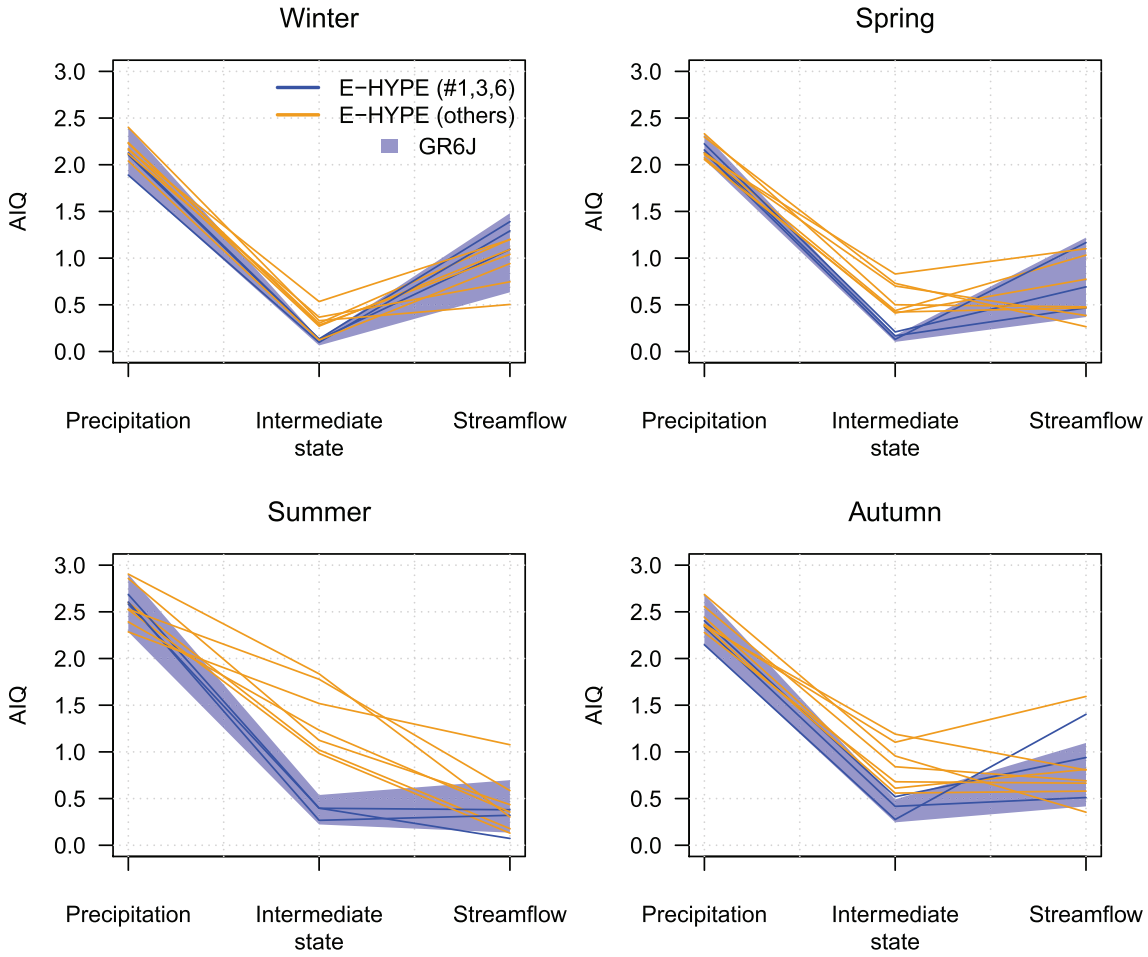


Figure 11. Forecasts spread (AIQ) for different model components of the GR6J (blue shade) and E-HYPE (blue and orange lines) models, and each season, and averaged over the first month lead time. The blue shade represents the range of AIQ for the 10 catchments. The blue lines represent the evolution of the AIQ in Catchments 1, 3, and 6, and the orange lines represent the evolution of the AIQ in the other catchments.

catchments dominated by coarse soils, the response of the two models differed with a faster emptying of the soil reservoir in the continental model setup. This difference in behavior was due to the parametrization of the continental model where a faster soil reservoir recession and larger amounts of water could percolate than in catchments dominated by fine soils, where the response of the two models was similar, even in terms soil water content. This highlights that the compartmentalization of catchments by soil and land use types in the continental model setup may be too drastic in some regions where soil responses do not justify very different model parameters.

4.3. Recommendations in Operational Forecasting Services

Our results raise the issue of usability of large-scale data sets versus local data sets. In the current state of knowledge and data availability (open data), the performance of a model driven by global data is likely to be lower than that based on local (and commonly quality-controlled) data. Here, the water balance analysis highlighted that models describe catchments differently depending on the reference data. Therefore, biases in water balance may already be present due to the meteorological data set used when setting up any hydrologic model (Donnelly et al., 2013). Models can to a certain extent compensate for these errors during parameter identification, which however can lead to unrealistic model fluxes or states (Andersson et al., 2015).

Nevertheless, a forecasting service setup, initialized and run with consistent meteorological data sets can reach similar performances to catchment-tailored services by learning from the system's dynamics and potential biases. E-HYPE takes better advantage of bias-adjusted GCM seasonal forecasts, because the data set used as reference in the bias-adjustment was also the one used in the setup and initialization of the

model. For the catchment-specific model, climatology could yield more or equivalent information than GCM forecasts bias-adjusted based on large-scale reanalysis. Indeed, ESP, which is the model fed with climatology, has been shown to be already a good setup, which can be hard to outperform (Arnal et al., 2018). For hydroclimate service users, this means that bias adjustment based on local observations is crucial in order to gain from meteorological forecasts. However, if continental/global hydrologic forecast products are produced from consistent production chains involving consistent reference data for model setup, initialization, and meteorological forecast adjustment, they can be also useful to users in their operations.

In terms of skill sensitivity, the initial hydrological conditions showed the largest impact on the overall skill, which is consistent with results from pan-European and global studies (Wanders et al., 2018; Yossef et al., 2013). This result highlights the added value of forecasting methods relying on ICs, such as ESP, and of techniques to increase the accuracy of ICs, such as data assimilation. The extent of the influence of ICs in the forecast performance was longer in summer dry months for both models; similar conclusions were drawn by Harrigan et al. (2018) and Staudinger and Seibert (2014). The impact of meteorological forcing on the forecast skill was lesser than that of ICs. Nevertheless, the use of GCM forecasts did improve the forecast skill, especially during wet months.

5. Conclusions

In this study, we addressed the question of assessing the quality and usefulness of hydrological seasonal forecasts based on both continental model outputs and local model outputs. We proposed a step-by-step methodology and illustrated it by comparing two forecasting chains based on two hydrologic model setups in 10 catchments in France. The models were the catchment-specific GR6J model and the continental E-HYPE model. Comparisons of input data, water balance representation, and forecasting chain elements allowed the identification of (dis)similarities between the continental and local model setups. We further proposed a methodology to identify the source of seasonal skill in the forecasting chains. It can also be applied to determine the skill of a current system and how much skill comes from the use of meteorological forecasts or from the assessment of hydrologic ICs.

The main conclusions from this study are as follows:

1. The catchment-specific seasonal forecasting system outperformed the continental forecasting system when forecasts were evaluated against observations. This was mainly due to the fact that the calibration of the former model targeted local observations. When evaluating the forecasting chains against model simulations, the continental system performed and the catchment-specific system or outperformed it at least in the first forecast month. This suggests that the continental model can provide statistics relative to model climatology, such as anomalies, with good quality. The comparative performance was also sensitive to the time aggregation, and, at large time aggregations, it was seen that the continental model could at times outperform the catchment-specific model.
2. The results highlight the importance of having consistency among all the meteorological data used throughout a hydrologic forecasting system (e.g., input data used for calibration/simulation, data used for bias adjustment, and climatological data used for reference in forecast evaluation). In case of significant inconsistency in data sets (i.e., data from different sources), users can benefit from using the hydrologic outputs from continental/global models from hydroclimate services, even if they lack local accuracy, since these usually ensure consistency in meteorological inputs. For some applications, this can be more efficient than having to apply a postprocessing routine (e.g., bias adjustment or downscaling) to meteorological outputs before using them in a catchment-specific hydrologic model. This also highlights the need for further research on methods for tailoring continental hydroclimate service outputs to different user needs at the local scale.
3. In this set of French catchments, the initialization of the hydrologic models was the main source of skill in both systems and in all seasons. The forecasting systems also made better use of GCM forecasts during wet months.
4. The comparison between outputs from continental/global models and local/catchment-specific models is not straightforward. Despite comparable inputs and outputs of the same variable, the uncertainty and variability in internal model states can be distinct, due to the different rationales of the models. The explicit representation of catchment characteristics used for regionalization in the continental model setup,

Table A1
Formulation of the Streamflow Characteristics Investigated

Name	Formulation
Flows	$(X_i)_{i \in [1;\, N]} = (Q_i)_{i \in [1;\, N]}$
Variation in flows	$(X_i)_{i \in [1;\, N-1]} = (Q_i - Q_{i-1})_{i \in [2;\, N]}$
Sign of the variation in flows	$(X_i)_{i \in [1;\, N-1]} = sign(Q_i - Q_{i-1})_{i \in [2;\, N]}$
Anomaly in flows	$(X_i)_{i\in[1;N]} = (Q_i - \overline{Q})_{i\in[1;N]}$
Sign of the anomaly in flows	$(X_i)_{i\in[1;N]} = sign(Q_i - \overline{Q})_{i\in[1;N]}$
Detection of Q25	$(X_i)_{i\in[1;N]} = \left(f(Q_i) = \begin{cases} 0, & Q_i < Q_{25} \\ 1, & Q_i \geq Q_{25} \end{cases}\right)_{i\in[1;N]}$
Detection of Q33	$(X_i)_{i\in[1;N]} = \left(f(Q_i) = \begin{cases} 0, & Q_i < Q_{33} \\ 1, & Q_i \geq Q_{33} \end{cases}\right)_{i\in[1;N]}$
Detection of Q50	$(X_i)_{i\in[1;N]} = \left(f(Q_i) = \begin{cases} 0, & Q_i < Q_{50} \\ 1, & Q_i \geq Q_{50} \end{cases}\right)_{i\in[1;N]}$
Detection of Q66	$(X_i)_{i\in[1;N]} = \left(f(Q_i) = \begin{cases} 0, & Q_i < Q_{66} \\ 1, & Q_i \geq Q_{66} \end{cases}\right)_{i\in[1;N]}$
Detection of Q75	$(X_i)_{i\in[1;N]} = \left(f(Q_i) = \begin{cases} 0, & Q_i < Q_{75} \\ 1, & Q_i \geq Q_{75} \end{cases}\right)_{i\in[1;N]}$

Note. Q_i is the streamflow at time step i, N is the total number of time steps, $\overline{Q}$ is the long-term averaged streamflow, and Q_{25}, Q_{33}, Q_{50}, Q_{66}, and Q_{75} are the 25th, 33rd, 50th, 66th, and 75th nonexceedance percentiles, respectively.

such as soil types, leads to a wider variability in internal processes and state values, such as soil drainage and soil water content, than in the parsimonious catchment-specific model. Therefore, even if streamflow is comparable between continental and catchment-specific models, conclusions on similarity among models cannot always be extended to other, intermediate model variables. Caution is therefore needed when extracting and comparing variables such as soil water content from different modeling systems.

Finally, it must be noted that our results are related to the hydrological models used and their respective setups as well as the hydroclimatic conditions of the set of studied catchments. They are representative of a large number of situations encountered in practice (e.g., operational forecasting using conceptual models, high sensitivity of hydrological seasonal forecasts to ICs, discrepancies encountered between global and local model flow magnitudes, and acknowledged importance of using consistent meteorological data sets throughout the forecasting chain). The methodology developed could be applied to other hydroclimatic contexts and other global/continental or local models. For instance, further research could be done in river systems in central Europe or northern Russia, where E-HYPE performance is, in general, higher than its performance in France. It would also be interesting to study the added value of global or continental model simulations to local water management in heavy regulated catchments, where hydrological models have usually lower performance, independently of their scales.

Appendix A: Evaluation Metrics and Target Variables

A.1. Validation of Model Simulations

The modified version of the KGE criterion (Gupta et al., 2009; Kling et al., 2012) is defined as follows:

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2}$$

where r is the Pearson correlation coefficient representing the error in timing and β and γ represent the error in volume and in variability, respectively, and are defined as follows:

$$\left\{ \beta = \overline{X}_{sim} / \overline{X}_{obs} \gamma = CV_{sim} / CV_{obs} \right.$$

where $\overline{X}_{sim}$ and $\overline{X}_{obs}$ refer to the mean simulated and observed streamflow characteristics and CV_{sim} and CV_{obs} refer to the coefficients of variation of the simulated and observed streamflow characteristics.

Table A presents the list of the 10 evaluated formulations of X.

A.2. Forecast Evaluation

The Continuous Rank Probability Score (CRPS; Hersbach, 2000) assesses overall forecast performance and is defined as follows:

$$CRPS = \frac{1}{N}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\left(F_i^{fst}(x) - F_i^{ref}(x)\right)^2 \mathrm{d}x$$

where F_i^{fst} is the cumulative distribution of the forecast at time step i, F_i^{ref} is the step function corresponding to the reference at time step i (i.e., 0 if x is smaller than the reference value at time step i, 1 otherwise), and N is the number of time steps used in the evaluation.

The normalized version of the skill in CRPS (CRPS) is subsequently computed to compare the forecast system to a benchmark:

$$\mathrm{CRPSS} = \frac{\mathrm{CRPS}_{\mathrm{bench}} - \mathrm{CRPS}_{\mathrm{syst}}}{\mathrm{CRPS}_{\mathrm{bench}} + \mathrm{CRPS}_{\mathrm{syst}}}$$

where $CRPS_{syst}$ and $CRPS_{bench}$ are the CRPS for the forecast system and the benchmark, respectively. Both CRPS and CRPSS vary with the forecast horizon.

Acknowledgments

We thank Dr. Shaun Harrigan, Dr. Niko Wanders and an anonymous referee for their constructive feedback to help improve this paper. This study was funded by the IMPREX, CLARA, and S2S4E projects. The IMPREX project has received funding from the European Union's Horizon 2020 research and innovation program under the Grant Agreement 641811 (www.imprex.eu). The CLARA project has received funding from the European Union's Horizon 2020 research and innovation program under the Grant Agreement 730482 (www.clara-project.eu). The S2S4E project has received funding from the European Union's Horizon 2020 research and innovation program under the Grant Agreement 776787 (www.s2s4e.eu). Readers can access streamflow observations used in this study at the website (http://www.hydro.eaufrance.fr/). A more recent version of the seasonal meteorological forecasts SEAS5 from the European Centre for Medium-Range Weather Forecasts is freely accessible from the Copernicus Climate Data Store (https://cds.climate.copernicus.eu/). The GR models, including GR6J, are available from the airGR R package (https://cran.r-project.org/web/packages/airGR/index.html). The HYPE model code is available from the HYPEweb portal (http://hypeweb.smhi.se/model-water/). Real-time seasonal forecasts obtained through E-HYPE are openly available also on the HYPEweb portal (http://hypeweb.smhi.se/explore-water/forecasts/seasonal-forecasts-europe/).

References

Andersson, J. C. M., Pechlivanidis, I. G., Gustafsson, D., Donnelly, C., & Arheimer, B. (2015). Key factors for improving large-scale hydrological model performance. *European Water*, *49*, 77–88.

Archfield, S. A., Clark, M., Arheimer, B., Hay, L. E., McMillan, H., Kiang, J. E., et al. (2015). Accelerating advances in continental domain hydrologic modeling. *Water Resources Research*, *51*, 10,078–10,091. https://doi.org/10.1002/2015WR017498

Arnal, L., Cloke, H. L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., et al. (2018). Skilful seasonal forecasts of streamflow over Europe? *Hydrology and Earth System Sciences*, *22*, 2057–2072. https://doi.org/10.5194/hess-22-2057-2018

Berg, P., Donnelly, C., & Gustafsson, D. (2018). Near-real-time adjusted reanalysis forcing data for hydrology. *Hydrology and Earth System Sciences*, *22*, 989–1000. https://doi.org/10.5194/hess-22-989-2018

Bierkens, M. F. P. (2015). Global hydrology 2015: State, trends, and directions. *Water Resources Research*, *51*, 4923–4947. https://doi.org/10.1002/2015WR017173

Bruno Soares, M., Alexander, M., & Dessai, S. (2018). Sectoral use of climate information in Europe: A synoptic overview. *Climate Services*, *9*, 5–20. https://doi.org/10.1016/j.cliser.2017.06.001

Buontempo, C., Hanlon, H. M., Soares, M. B., Christel, I., Soubeyroux, J.-M., Viel, C., et al. (2018). What have we learnt from EUPORIAS climate service prototypes? *Climate Services*, *9*, 21–32. https://doi.org/10.1016/j.cliser.2017.06.003

Cavelier, R., Borel, C., Charreyron, V., Chaussade, M., Le Cozannet, G., Morin, D., & Ritti, D. (2017). Conditions for a market uptake of climate services for adaptation in France. *Climate Services*, *6*, 34–40. https://doi.org/10.1016/j.cliser.2017.06.010

Crochemore, L., Isberg, K., Pimentel, R., Pineda, L., Hasan, A., & Arheimer, B. (2019). Lessons learnt from checking the quality of openly accessible river flow data worldwide. *Hydrological Sciences Journal*, *1–13*, 1–13. https://doi.org/10.1080/02626667.2019.1659509

Crochemore, L., Ramos, M.-H., & Pappenberger, F. (2016). Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrology and Earth System Sciences*, *20*, 3601–3618. https://doi.org/10.5194/hess-20-3601-2016

Day, G. (1985). Extended streamflow forecasting using NWSRFS. *Journal of Water Resources Planning and Management*, *111*(2), 157–170. https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157)

Donnelly, C., Rosberg, J., & Isberg, K. (2013). A validation of river routing networks for catchment modelling from small to large scales. *Hydrology Research*, *44*, 917–925. https://doi.org/10.2166/nh.2012.341

Emerton, R., Zsoter, E., Arnal, L., Cloke, H. L., Muraro, D., Prudhomme, C., et al. (2018). Developing a global operational seasonal hydro-meteorological forecasting system: GloFAS-Seasonal v1.0. *Geoscientific Model Development*, *11*, 3327–3346. https://doi.org/10.5194/gmd-11-3327-2018

Foster, K., Bertacchi Uvo, C., & Olsson, J. (2018). The development and evaluation of a hydrological seasonal forecast system prototype for predicting spring flood volumes in Swedish rivers. *Hydrology and Earth System Sciences*, *22*, 2953–2970. https://doi.org/10.5194/hess-22-2953-2018

Gosling, S. N., Taylor, R. G., Arnell, N. W., & Todd, M. C. (2011). A comparative analysis of projected impacts of climate change on river runoff from global and catchment-scale hydrological models. *Hydrology and Earth System Sciences*, *15*, 279–294. https://doi.org/10.5194/hess-15-279-2011

Gosling, S. N., Zaherpour, J., Mount, N. J., Hattermann, F. F., Dankers, R., Arheimer, B., 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. *Climatic Change*, *141*, 577–595. https://doi.org/10.1007/s10584-016-1773-3

Gudmundsson, L., Tallaksen, L. M., Stahl, K., Clark, D. B., Dumont, E., Hagemann, S., et al. (2012). Comparing large-scale hydrological model simulations to observed runoff percentiles in Europe. *Journal of Hydrometeorology*, *13*, 604–620. https://doi.org/10.1175/JHM-D-11-083.1

Gudmundsson, L., Wagener, T., Tallaksen, L. M., & Engeland, K. (2012). Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe. *Water Resources Research*, *48*, W11504. https://doi.org/10.1029/2011WR010911

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1-2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Harrigan, S., Prudhomme, C., Parry, S., Smith, K., & Tanguy, M. (2018). Benchmarking ensemble streamflow prediction skill in the UK. *Hydrology and Earth System Sciences*, *22*, 2023–2039. https://doi.org/10.5194/hess-22-2023-2018

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2

Hundecha, Y., Arheimer, B., Donnelly, C., & Pechlivanidis, I. (2016). A regional parameter estimation scheme for a pan-European multi-basin model. *Journal of Hydrology: Regional Studies*, *6*, 90–111. https://doi.org/10.1016/j.ejrh.2016.04.002

Kauffeldt, A., Halldin, S., Rodhe, A., Xu, C.-Y., & Westerberg, I. K. (2013). Disinformative data in large-scale hydrological modelling. *Hydrology and Earth System Sciences*, *17*, 2845–2857. https://doi.org/10.5194/hess-17-2845-2013

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, *424–425*, 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011

Krysanova, V., Vetter, T., Eisner, S., Huang, S., Pechlivanidis, I., Strauch, M., et al. (2017). Intercomparison of regional-scale hydrological models and climate change impacts projected for 12 large river basins worldwide—a synthesis. *Environmental Research Letters*, *12*, 105002. https://doi.org/10.1088/1748-9326/aa8359

Lindström, G., Pers, C., Rosberg, J., Strömqvist, J., & Arheimer, B. (2010). Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales. *Hydrology Research*, *41*(3-4), 295–319. https://doi.org/10.2166/nh.2010.007

Mazrooei, A., Sinha, T., Sankarasubramanian, A., Kumar, S., & Peters-Lidard, C. D. (2015). Decomposition of sources of errors in seasonal streamflow forecasting over the U.S. Sunbelt. *Journal of Geophysical Research: Atmospheres*, *120*, 11,809–11,825. https://doi.org/10.1002/2015JD023687

Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., et al. (2011). The new ECMWF seasonal forecast system (System 4). ECMWF Tech Memo 656, 49 pp.

Mu, Q., Zhao, M., & Running, S. W. (2011). Improvements to a MODIS global terrestrial evapotranspiration algorithm. *Remote Sensing of Environment*, *115*, 1781–1800. https://doi.org/10.1016/j.rse.2011.02.019

Nazemi, A., & Wheater, H. S. (2015). On inclusion of water resource management in Earth system models—Part 2: Representation of water supply and allocation and opportunities for improved modeling. *Hydrology and Earth System Sciences*, *19*, 63–90. https://doi.org/10.5194/hess-19-63-2015

Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., et al. (2014). Benchmarking hydrological models for low-flow simulation and forecasting on French catchments. *Hydrology and Earth System Sciences*, *18*, 2829–2857. https://doi.org/10.5194/hessd-10-13979-2013

Nijzink, R., Hutton, C., Pechlivanidis, I., Capell, R., Arheimer, B., Freer, J., et al. (2016). The evolution of root-zone moisture capacities after deforestation: A step towards hydrological predictions under change? *Hydrology and Earth System Sciences*, *20*, 4775–4799. https://doi.org/10.5194/hess-20-4775-2016

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., & Loumagne, C. (2005). Which potential evapotranspiration input for a lumped rainfall-runoff model? Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall-runoff modelling. *Journal of Hydrology*, *303*(1-4), 290–306. https://doi.org/10.1016/j.jhydrol.2004.08.026

Pappenberger, F., Ramos, M. H., Cloke, H. L., Wetterhall, F., Alfieri, L., Bogner, K., et al. (2015). How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *Journal of Hydrology*, *522*, 697–713. https://doi.org/10.1016/j.jhydrol.2015.01.024

Pechlivanidis, I. G., & Arheimer, B. (2015). Large-scale hydrological modelling by using modified PUB recommendations: the India-HYPE case. *Hydrology and Earth System Sciences*, *19*, 4559–4579. https://doi.org/10.5194/hess-19-4559-2015

Pechlivanidis, I. G., Arheimer, B., Donnelly, C., Hundecha, Y., Huang, S., Aich, V., et al. (2017). Analysis of hydrological extremes at different hydro-climatic regimes under present and future conditions. *Climatic Change*, *141*(3), 467–481. https://doi.org/10.1007/s10584-016-1723-0

Pushpalatha, R., Perrin, C., Mathevet, T., & Andreassian, V. (2011). A downward structural sensitivity analysis of hydrological models to improve low-flow simulation. *Journal of Hydrology*, *411*, 66–76. https://doi.org/10.1016/j.jhydrol.2011.09.034

Quintana-Seguí, P., Le Moigne, P., Durand, Y., Martin, E., Habets, F., Baillon, M., et al. (2008). Analysis of near-surface atmospheric variables: Validation of the SAFRAN analysis over France. *Journal of Applied Meteorology and Climatology*, *47*, 92–107. https://doi.org/10.1175/2007JAMC1636.1

Sinha, T., Sankarasubramanian, A., & Mazrooei, A. (2014). Decomposition of sources of errors in monthly to seasonal streamflow forecasts in a rainfall–runoff regime. *Journal of Hydrometeorology*, *15*, 2470–2483. https://doi.org/10.1175/JHM-D-13-0155.1

Siqueira, V. A., Paiva, R. C. D., Fleischmann, A. S., Fan, F. M., Ruhoff, A. L., Pontes, P. R. M., et al. (2018). Toward continental hydrologic–hydrodynamic modeling in South America. *Hydrology and Earth System Sciences*, *22*, 4815–4842. https://doi.org/10.5194/hess-22-4815-2018

Sood, A., & Smakhtin, V. (2015). Global hydrological models: A review. *Hydrological Sciences Journal*, *60*, 549–565. https://doi.org/10.1080/02626667.2014.950580

Stahl, K., Tallaksen, L. M., Gudmundsson, L., & Christensen, J. H. (2011). Streamflow data from small basins: A challenging test to high-resolution regional climate modeling. *Journal of Hydrometeorology*, *12*, 900–912. https://doi.org/10.1175/2011JHM1356.1

Staudinger, M., & Seibert, J. (2014). Predictability of low flow—An assessment with simulation experiments. *Journal of Hydrology*, *519*, 1383–1393. https://doi.org/10.1016/j.jhydrol.2014.08.061

Swart, R. J., de Bruin, K., Dhenain, S., Dubois, G., Groot, A., & von der Forst, E. (2017). Developing climate information portals with users: Promises and pitfalls. *Climate Services*, *6*, 12–22. https://doi.org/10.1016/j.cliser.2017.06.008

Thiboult, A., Anctil, F., & Ramos, M. H. (2017). How does the quantification of uncertainties affect the quality and value of flood early warning systems? *Journal of Hydrology*, *551*, 365–373. https://doi.org/10.1016/j.jhydrol.2017.05.014

Thiemig, V., Bisselink, B., Pappenberger, F., & Thielen, J. (2015). A pan-African medium-range ensemble flood forecast system. *Hydrology and Earth System Sciences*, *19*, 3365–3385. https://doi.org/10.5194/hess-19-3365-2015

van den Hurk, B. J. J. M., Bouwer, L. M., Buontempo, C., Döscher, R., Ercin, E., Hananel, C., et al. (2016). Improving predictions and management of hydrological extremes through climate services: www.imprex.eu. *Climate Services*, *1*, 6–11. https://doi.org/10.1016/j.cliser.2016.01.001

Vaughan, C., & Dessai, S. (2014). Climate services for society: Origins, institutional arrangements, and design elements for an evaluation framework. *Wiley Interdisciplinary Reviews: Climate Change*, *5*(5), 587–603. https://doi.org/10.1002/wcc.290

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., & Soubeyroux, J.-M. (2010). A 50-year high-resolution atmospheric reanalysis over France with the Safran system. *International Journal of Climatology*, *30*, 1627–1644. https://doi.org/10.1002/joc.2003

Wanders, N., Thober, S., Kumar, R., Pan, M., Sheffield, J., Samaniego, L., & Wood, E. F. (2018). Development and evaluation of a Pan-European multimodel seasonal hydrological forecasting system. *Journal of Hydrometeorology*, *20*, 99–115. https://doi.org/10.1175/JHM-D-18-0040.1

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., & Viterbo, P. (2014). The WFDEI meteorological forcing data set: WATCH forcing data methodology applied to ERA-Interim reanalysis data. *Water Resources Research*, *50*, 7505–7514. https://doi.org/10.1002/2014WR015638

Wood, A. W., & Lettenmaier, D. P. (2008). An ensemble approach for attribution of hydrologic prediction uncertainty. *Geophysical Research Letters*, *35*, L14401. https://doi.org/10.1029/2008GL034648

Yang, W., Andréasson, J., Phil Graham, L., Olsson, J., Rosberg, J., & Wetterhall, F. (2010). Distribution-based scaling to improve usability of regional climate model projections for hydrological climate change impacts studies. *Hydrology Research*, *41*, 211–229. https://doi.org/10.2166/nh.2010.004

Yossef, N. C., Winsemius, H., Weerts, A., van Beek, R., & Bierkens, M. F. P. (2013). Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing. *Water Resources Research*, *49*, 4687–4699. https://doi.org/10.1002/wrcr.20350

Zalachori, I., Ramos, M.-H., Garçon, R., Mathevet, T., & Gailhard, J. (2012). Statistical processing of forecasts for hydrological ensemble prediction: A comparative study of different bias correction strategies. *Advances in Science and Research*, *8*, 135–141. https://doi.org/10.5194/asr-8-135-2012

Zappa, M., Jaun, S., Germann, U., Walser, A., & Fundel, F. (2011). Superposition of three sources of uncertainties in operational flood forecasting chains. *Atmospheric Research*, *100*, 246–262. https://doi.org/10.1016/j.atmosres.2010.12.005

Zhang, Y., Zheng, H., Chiew, F. H. S., Arancibia, J. P., & Zhou, X. (2016). Evaluating regional and global hydrological models against streamflow and evapotranspiration measurements. *Journal of Hydrometeorology*, *17*, 995–1010. https://doi.org/10.1175/JHM-D-15-0107.1