The skill of seasonal ensemble low flow forecasts in the Moselle River for three different hydrological models

MEHMET C. DEMIREL* & MARTIJN J. BOOIJ & ARJEN Y. HOEKSTRA

Water Engineering and Management, Faculty of Engineering Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands.

**Current address: Portland State University, Department of Civil & Environmental Engineering, 1930 S.W. 4th Avenue, Suite 200, Portland, OR 97201, USA (demirel@pdx.edu)*

Abstract

This paper investigates the skill of 90 day low flow forecasts using two conceptual hydrological models and one data-driven model based on Artificial Neural Networks (ANNs) for the Moselle River. The three models, i.e. HBV, GR4J and ANN-Ensemble (ANN-E), all use forecasted meteorological inputs (Precipitation P and potential evapotranspiration PET), whereby we employ ensemble seasonal meteorological forecasts. We compared low flow forecasts for five different cases of seasonal meteorological forcing: (1) ensemble P and PET forecasts; (2) ensemble P forecasts and observed climate mean PET; (3) observed climate mean P and ensemble PET forecasts; (4) observed climate mean P and PET and (5) zero P and ensemble PET forecasts as input for the models. The ensemble P and PET forecasts, each consisting of 40 members, reveal the forecast ranges due to the model inputs. The five cases are compared for a lead time of 90 days based on model output ranges, whereas the models are compared based on their skill of low flow forecasts for varying lead times up to 90 days. Before forecasting, the hydrological models are calibrated and validated for a period of 30 and 20 years respectively. The smallest difference between calibration and validation performance is found for HBV, whereas the largest difference is found for ANN-E. From the results, it appears that all models are prone to over-predict runoff during low flow periods using ensemble seasonal meteorological forcing. The largest range for 90 day low flow forecasts is found for the GR4J model when using ensemble seasonal meteorological forecasts as input. GR4J, HBV and ANN-E under-predicted 90 day ahead low flows in the very dry year 2003 without precipitation data. The results of the comparison of forecast skills with varying lead times show that GR4J is less skilful than ANN-E and HBV. Overall, the uncertainty from ensemble P forecasts has a larger effect on seasonal low flow forecasts than the uncertainty from ensemble PET forecasts and initial model conditions.

Key words: Moselle River, GR4J, HBV, ANN, low flows, ensemble seasonal meteorological forecasts

1 INTRODUCTION

Rivers in Western Europe usually experience low flows in late summer and high flows in winter. These two extreme discharge phenomena can lead to serious problems. For example, high flow events are quick and can put human life at risk, whereas streamflow droughts (i.e. low flows) develop slowly and can affect a large area. Consequently, the economic loss during low flow periods can be much bigger than during floods (Pushpalatha et al., 2011;Shukla et al., 2012). In the River Rhine, severe problems for freshwater supply, water quality, power production and river navigation were experienced during the dry summers of 1976, 1985 and 2003. Therefore, forecasting seasonal low flows (Towler et al., 2013;Coley and Waylen, 2006;Li et al., 2008) and understanding low flow indicators (Vidal et al., 2010;Fundel et al., 2013;Demirel et al., 2013a;Wang et al., 2011;Saadat et al., 2013;Nicolle et al., 2013) have both societal and scientific value. The seasonal forecast of water flows is therefore listed as one of the priority topics in EU's Horizon 2020 research program (EU, 2013). Further, there is an increasing interest to incorporate seasonal flow forecasts in decision support systems for river navigation and power plant operation during low flow periods. We are interested in forecasting low flows with a lead time of 90 days, and in presenting the effect of ensemble meteorological forecasts for three hydrological models.

Generally, two approaches are used in seasonal hydrological forecasting. The first one is a statistical approach, making use of data-driven models based on relationships between river discharge and hydroclimatological indicators (Wang et al., 2011;Van Ogtrop et al., 2011;Förster et al., 2014). The second one is a dynamic approach running a hydrological model with forecasted climate input.

The first approach is often preferred in regions where significant correlations between river discharge and climatic indicators exist, such as sea surface temperature anomalies (Chowdhury and Sharma, 2009), AMO – Atlantic Multi-decadal Oscillation (Ganguli and Reddy, 2013;Giuntoli et al., 2013), PDO – Pacific Decadal Oscillation (Soukup et al., 2009) and warm and cold phases of the ENSO – El Nino Southern Oscillation - index (Chiew et al., 2003;Kalra et al., 2013;Tootle and Piechota, 2004). Kahya and Dracup (1993) identified the lagged response of regional streamflow to the warm phase of ENSO in the south-eastern United States. In the Rhine basin, no teleconnections have been found between climatic indices, e.g. NAO and ENSO, and river discharges (Rutten et al., 2008;Bierkens and van Beek, 2009). However, Demirel et al. (2013a) found significant correlations between hydrological low flow indicators and observed low flows. They also identified appropriate lags and temporal resolutions of low flow indicators (e.g. precipitation, potential evapotranspiration, groundwater storage, lake levels and snow storage) to build data-driven models.

The second approach is the dynamic seasonal forecasting approach which has long been explored (Wang et al., 2011;Van Dijk et al., 2013;Gobena and Gan, 2010;Fundel et al., 2013;Shukla et al., 2013;Pokhrel et al., 2013) and has led to the development of the current ensemble streamflow prediction system (ESP) used by different national climate services like the National Weather Service in the United States. The seasonal hydrologic prediction systems are most popular in regions with a high risk of extreme discharge situations like hydrological droughts (Robertson et al., 2013;Madadgar and Moradkhani, 2013). Well-known examples are the NOAA Climate Prediction Centre's seasonal drought forecasting system (available at http://www.cpc.ncep.noaa.gov), the University of Washington's Surface Water Monitoring system (Wood and Lettenmaier, 2006), Princeton University's drought forecast system (available at http://hydrology.princeton.edu/forecast) and University of Utrecht's global monthly hydrological forecast system (Yossef et al., 2012). These models provide indications about the hydrologic conditions and their evolution across the modelled domain using available weather ensemble inputs (Gobena and Gan, 2010;Yossef et al., 2012). Moreover, Dutra et al. (2014) showed that global seasonal forecasts of meteorological drought onset are feasible and skilful using the standardized precipitation index (SPI) and two data sets as initial conditions.

Many studies have investigated the seasonal predictability of low flows in different rivers such as the Thames and different other rivers in the UK (Bell et al., 2013;Wedgbrow et al., 2002;Wedgbrow et al., 2005), the Shihmen and Tsengwen Rivers in Taiwan (Kuo et al., 2010), the River Jhelum in Pakistan (Archer and Fowler, 2008), more than 200 rivers in France (Sauquet et al., 2008;Giuntoli et al., 2013), five semi-arid areas in South Western Queensland, Australia (Van Ogtrop et al., 2011), five rivers including Limpopo basin and the Blue Nile in Africa (Dutra et al., 2013;Winsemius et al., 2014), the Bogotá River in Colombia (Felipe and Nelson, 2009), the Ohio in the eastern US (Wood et al., 2002;Luo et al., 2007;Li et al., 2009), the North Platte in Colorado, US (Soukup et al., 2009), large rivers in the US (Schubert et al., 2007;Shukla and Lettenmaier, 2011) and the Thur River in the north-eastern part of Switzerland (Fundel et al., 2013). The common result of the above mentioned studies is that the skill of the seasonal forecasts made with global and regional hydrological models is reasonable for lead times of 1-3 months (Shukla and Lettenmaier, 2011;Wood et al., 2002) and these forecasting systems are all prone to large uncertainties as their forecast skills mainly depend on the knowledge of initial hydrologic conditions and weather information during the forecast period (Shukla et al., 2012;Yossef et al., 2013;Li et al., 2009;Doblas-Reyes et al., 2009). In a recent study, Yossef et al. (2013) used a global monthly hydrological model to analyse the relative contributions of initial conditions and meteorological forcing to the skill of seasonal streamflow forecasts. They included 78 stations in large basins in the world including the River Rhine for forecasts with lead times up to 6 months. They found that improvements in seasonal hydrological forecasts in the Rhine depend on better meteorological forecasts, which underlines the importance of meteorological forcing quality particularly for forecasts beyond lead times of 1-2 months.

Most of the previous River Rhine studies use only one hydrological model, e.g. PREVAH (Fundel et al., 2013) or PCR-GLOBWB (Yossef et al., 2013), to assess the value of ensemble meteorological forcing, whereas in this study, we compare three hydrological models with different structures varying from data-driven to conceptual models. The two objectives of this study are to contrast data-driven and conceptual modelling approaches and to assess the effect of ensemble seasonal forecasted precipitation and potential evapotranspiration on low flow forecast quality and skill scores. By comparing three models with different model structures we address the issue of model structure uncertainty, whereas the latter objective reflects the benefit of ensemble seasonal forecasts. Moreover, the effect of initial model conditions is partly addressed using climate mean data in one of the cases.

The analysis complements recent efforts to analyse the effects of ensemble weather forecasts on low flow forecasts with a lead time of 10 days using two conceptual models (Demirel et al., 2013b), by studying the effects of seasonal ensemble weather forecasts on 90 day low flow forecasts using not only conceptual models but also data-driven models.

The outline of the paper is as follows. The study area and data are presented in section 2. Section 3 describes the model structures, their calibration and validation set-ups and the methods employed to estimate the different attributes of the forecast quality. The results are presented in section 4 and discussed in section 5, and the conclusions are summarised in section 6.

2 STUDY AREA AND DATA

2.1 Study area

The study area is the Moselle River basin, the largest sub-basin of the Rhine River basin. The Moselle River has a length of 545 km. The river basin has a surface area of approximately 27,262 km^2. The altitude in the basin varies from 59 to 1326 m, with a mean altitude of 340 m (Demirel et al., 2013a). There are 26 subbasins with surface areas varying from 102 to 3353 km^2. Approximately 410 mm (~130 m^3s^{-1}) discharge is annually generated in the Moselle basin (Demirel et al., 2013b). The outlet discharge at Cochem varies from 14 m^3s^{-1} in dry summers to a maximum of 4000 m^3s^{-1} during winter floods.

The Moselle River has been heavily regulated by dams, power plants, weirs and locks. There are around 12 hydropower plants between Koblenz and Trier producing energy since the 1960s (Bormann, 2010). Moreover, there are 12 locks only on the German part of the river (Bormann et al., 2011).

2.2 Data

2.2.1 Observed data

Observed daily data on precipitation (P), potential evapotranspiration (PET) and the mean altitudes (h) of the 26 sub-basins have been provided by the German Federal Institute of Hydrology (BfG) in Koblenz, Germany (**Table 1**). PET is estimated using the Penman-Wendling equation (ATV-DVWK, 2002) and both variables have been spatially averaged by BfG over 26 Moselle sub-basins using areal weights. Observed data from 12 meteorological stations in the Moselle basin (as part of 49 stations over the Rhine basin), mainly provided by the CHR, the

DWD, Metéo France, are used to estimate the basin averaged input data (Görgen et al., 2010). Observed daily discharge (Q) data at Cochem (station *#6336050*) are provided by the Global Runoff Data Centre (GRDC), Koblenz. The daily observed data (P, PET and Q) are available for the period 1951-2006.

Table 1 Overview of observed data used

Variable	Name	Number of stations/sub-basins	Period	Annual Range (mm)	Time step (days)	Spatial resolution	Source
Q	Discharge	1	1951-2006	163-550	1	Point	GRDC
P	Precipitation	26	1951-2006	570-1174	1	Basin average	BfG
PET	Potential evapotranspiration	26	1951-2006	512-685	1	Basin average	BfG
h	Mean altitude	26	-	-	-	Basin average	BfG

2.2.2 Ensemble seasonal meteorological forecast data

The ensemble seasonal meteorological forecast data, comprising 40 members, are obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) seasonal forecasting archive and retrieval system, i.e. MARS system 3 (ECMWF, 2012). This dataset contains regular 0.25 x 0.25 degree latitude-longitude grids and each ensemble member is computed for a lead time of 184 days using perturbed initial conditions and model physics (**Table 2**). We estimated the PET forecasts using the Penman-Wendling equation requiring forecasted surface solar radiation and temperature at 2 meter above the surface, and the altitude of the sub-basin (ATV-DVWK, 2002). The PET estimation is consistent with the observed PET estimation carried out by BfG (ATV-DVWK, 2002). The grid-based P and PET ensemble forecast data are firstly interpolated over 26 Moselle sub-basins using areal weights. These sub-basin averaged data are then aggregated to the Moselle basin level.

Table 2 Overview of ensemble seasonal meteorological forecast data

Data	**Spatial resolution**	**Ensemble size**	**Period**	**Time step (days)**	**Lead time (days)**
Forecasted P	0.25 x 0.25 degree	39 + 1 control	2002-2005	1	1-90
Forecasted PET	0.25 x 0.25 degree	39 + 1 control	2002-2005	1	1-90

3 METHODOLOGY

3.1 Overview of model structures and forecast scheme

The three hydrological models (GR4J, HBV and ANN-E) are briefly described in sections 3.1.1-3.1.3. Figure 1 shows the simplified model structures. The calibration and validation of the models is described in section 3.1.4. Five cases with different combinations of ensemble meteorological forecast input and climate mean input are introduced in section 3.1.5. We provide a detailed description for each parameter of the three models in section 4.1.

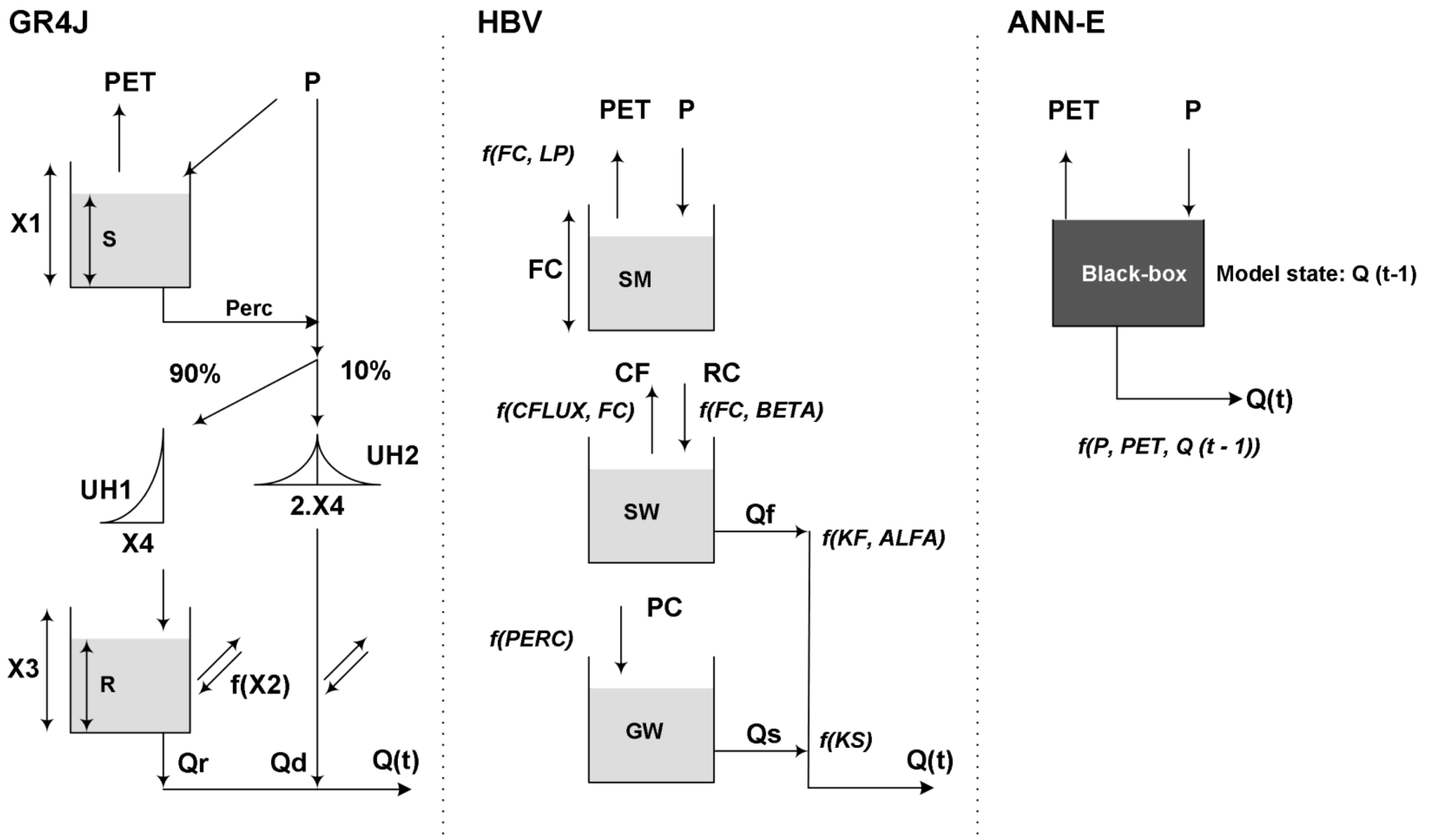


Figure 1 Schematisation of the three models. PET is potential evapotranspiration, P is precipitation and Q is discharge and *t* is the time (day).

3.1.1 GR4J

The GR4J model (Génie Rural à 4 paramètres Journalier) is used as it has a parsimonious structure with only four parameters. The model has been tested over hundreds of basins worldwide, with a broad range of climatic conditions from tropical to temperate and semi-arid basins (Perrin et al., 2003). GR4J is a conceptual model and the required model inputs are daily time series of P and PET (**Table 3**). All four parameters (Figure 1a) are used to calibrate the model. The upper and lower limits of the parameters are selected based on previous works (Perrin et al., 2003;Pushpalatha et al., 2011;Tian et al., 2014).

3.1.2 HBV

The HBV conceptual model (Hydrologiska Byråns Vattenbalansavdelning) was developed by the Swedish Meteorological and Hydrological Institute (SMHI) in the early 1970´s (Lindström et al., 1997). The HBV model consists of four subroutines: a precipitation and snow accumulation and melt routine, a soil moisture accounting routine and two runoff generation routines. The required input data are daily P and PET. The snow routine and daily temperature data are not used in this study as the Moselle basin is a rain-fed basin. Eight parameters (see Figure 1b) in the HBV model are calibrated (Engeland et al., 2010;Van den Tillaart et al., 2013;Tian et al., 2014). The eight parameters are selected for calibration and the parameter ranges are selected based on previous works (Booij, 2005;Eberle, 2005;Tian et al., 2014).

3.1.3 ANN-E

An Artificial Neural Network (ANN) is a data-driven model inspired by functional units (neurons) of the human brain (Elshorbagy et al., 2010). A neural network is a universal approximator capable of learning the patterns and relation between outputs and inputs from historical data and applying it for extrapolation (Govindaraju and Rao, 2000). A three-layer feed-forward neural network (FNNs) is the most widely preferred model architecture for prediction and forecasting of hydrological variables (Adamowski et al., 2012;Shamseldin, 1997;Kalra et al., 2013). Each of these three layers has an important role in processing the information. The first layer receives the inputs and multiplies them with a weight (adds a bias if necessary) before delivering them to each of the hidden neurons in the next layer (Gaume and Gosset, 1999). The weights determine the strength of the connections. The number of nodes in this layer corresponds to the number of inputs. The second layer, the hidden layer, consists of an activation function (also known as transfer function) which non-linearly maps the input data to output target values. In other words, this layer is the learning element of the network which simulates the relationship between inputs and outputs of the model. The third layer, the output layer, gathers the processed data from the hidden layer and delivers the final output of the network.

A hidden neuron is the processing element with n inputs (x_1, x_2, x_3, …, x_n), and one output y using Eq (1).

$$y = f(x_1, x_2, x_3, \dots, x_n) = logsig\left[\left(\sum_{i=1}^{n} x_i w_i\right) + b\right] \tag{1}$$

where w_i are the weights, b is the bias, and *logsig* is the logarithmic sigmoid activation function. We tested the *tansig* and *logsig* activation functions and the latter was selected for this study as it gave better results for low flows. ANN model structures are determined based on the forecast objective. In this study, we used a conceptual type ANN model structure: ANN-Ensemble (ANN-E) which requires daily P, PET and historical Q as input. Observed discharge on the forecast issue day is used to update the model states (**Table 3**). In other words, the ANN-E model receives *Qobs(t)* as input on the time step *t* when the forecast is issued, and then receives the streamflow forecast of the previous time step as input for lead times larger than 1 day. Further, forecasted Q for time step *t+j* is used as input to forecast Q at *t+j+1*.

This is a one day memory which also exists in the conceptual models, i.e. GR4J and HBV (Figure 1). The ANN-E is assumed to be comparable with the conceptual models with similar model structures. The determination of the optimal number of hidden neurons in the second layer is an important issue in the development of ANN models. Three common approaches are ad hoc (also known as trial and error), global and stepwise (Kasiviswanathan et al., 2013). We used a global approach (i.e. Genetic Algorithm) to avoid local minima (De Vos and Rientjes, 2008) and tested the performance of the networks with one, two and three hidden neurons corresponding to a number of parameters (i.e. number of weights and biases) of 6, 11 and 16 respectively. Based on the parsimonious principle, testing ANNs only up to three hidden neurons is assumed to be enough as the number of parameters increases exponentially for every additional hidden neuron.

Table 3 Model descriptions. PET is potential evapotranspiration, P is precipitation and Q is discharge.

Model Type		**Input**	**Temporal resolution of input**	**Lag between forecast issue day and final day of temporal averaging (days)**	**Model time step**	**Model lead time (days)**
Conceptual	**Data-driven**					
GR4J		P: Ensemble PET: Ensemble Q: State update	Daily P Daily PET	P: 0 PET: 0 Q: 1	Daily	1 to 90

HBV	P: Ensemble PET: Ensemble Q: State update	Daily P Daily PET	P: 0 PET: 0 Q: 1	Daily	1 to 90
ANN-E	P: Ensemble PET: Ensemble Q: State update	Daily P Daily PET Daily Q	P: 0 PET: 0 Q: 1	Daily	1 to 90

3.1.4 Calibration and validation of models

A global optimisation method, i.e. Genetic Algorithm (GA) (De Vos and Rientjes, 2008), and historical Moselle low flows for the period from 1971-2001 are used to calibrate the models used in this study. The 30-year calibration period is carefully selected as the first low flow forecast is issued on 01/01/2002. The first three years are used as warm-up period for the hydrological model. For all GA simulations, we use 100 as population size, 5 as reproduction elite count size, 0.7 as cross over fraction, 2000 as maximum number of iterations and 5000 as the maximum number of function evaluations based on the studies by De Vos and Rientjes (2008) and Kasiviswanathan et al. (2013). The evolution starts from the population of 100 randomly generated individuals. The population in each iteration is called a generation and the fitness of every individual in the population is evaluated using the objective function. The best 70 percent of the population (indicated as cross over fraction) survives in the process of 2000 iterations.

The validation period spans from 1951-1970. The definition of low flows, i.e. discharges below the Q75 threshold of ~113 $\mathrm{m^3 s^{-1}}$, is based on previous work by Demirel et al. (2013a). Prior parameter ranges and deterministic equations used for dynamic model state updates of the conceptual models based on observed discharges on the forecast issue day are based on the study by Demirel et al. (2013b). In this study, we use a hybrid Mean Absolute Error (MAE) based on only low flows (MAE_{low}) and inverse discharge values ($MAE_{inverse}$) as objective function (see Eq.(4)).

$$\text{Mean Absolute Error}_{\,low}: \frac{1}{m}\sum_{j=1}^{m}|Q_{sim}(j) - Q_{obs}(j)| \tag{2}$$

where Q_{obs} and Q_{sim} are the observed and simulated values for the *j*-th observed low flow day (i.e. $Q_{obs} < Q_{75}$) and *m* is the total number of low flow days.

$$\text{Mean Absolute Error}_{\,inverse}: \frac{1}{n}\sum_{i=1}^{n}\left|\frac{1}{Q_{sim}(i)+\epsilon} - \frac{1}{Q_{obs}(i)+\epsilon}\right| \tag{3}$$

where *n* is the total number of days (i.e. $m < n$), and ϵ is 1% of the mean observed discharge to avoid infinity during zero discharge days (see Pushpalatha et al.,(2012)). The hybrid Mean Absolute Error is defined as

$$MAE_{hybrid} = MAE_{low} + MAE_{inverse} \tag{4}$$

The MAE_{low} and $MAE_{inverse}$ were not normalised to calculate MAE_{hybrid} metric. It should be noted that we didn't fully neglect the high and intermediate flows using $MAE_{inverse}$, whereas only low flow periods are considered in MAE_{low}. This is one of the advantages of using the MAE_{hybrid} metric and also avoids redundancy.

3.1.5 Model storage update procedure for HBV and GR4J models

The storages in the two conceptual models are updated based on the observed discharge on the forecast issue day. In our previous study (Demirel et al., 2013c), we derived empirical relations between the simulated discharge and the fast runoff for each model to divide the observed discharge between the fast and slow runoff components (Eq. (5) and (6)).

$$k\_GR4J = \frac{Qd}{Qr + Qd} \tag{5}$$

$$k\_HBV = \frac{Qf}{Qf + Qs} \tag{6}$$

The Qf and Qs in the HBV model, and Qr and Qd in the GR4J model are estimated using the fractions above and the observed discharge value on the forecast issue day. The routing storage (*R*) in the GR4J model is updated for a given value of the *X3* parameter using Eq.(7). Moreover, the surface water (*SW*) and groundwater (*GW*) storages in the HBV model are updated for given values of *KF*, *ALFA* and *KS* parameters using Eq. (8) and (9).

$$Qr = R\left\{1 - \left[1 + \left(\frac{R}{X3}\right)^4\right]^{-1/4}\right\} \tag{7}$$

$$SW = \left(\frac{Qf}{KF}\right)^{\left(\frac{1}{(1+ALFA)}\right)} \tag{8}$$

$$GW = \frac{Qs}{KS} \tag{9}$$

The remaining two storages *S* (in GR4J) and *SM* (in HBV) are updated using the calibrated model run until the forecast issue day (i.e. top-down approach).

3.1.6 Case description

In this study, three hydrological models are used for the seasonal forecasts. Five ensemble meteorological forecast input cases for ANN-E, GR4J and HBV models are compared: (1) ensemble P and PET forecasts (2) ensemble P forecasts and observed climate mean PET (3) observed climate mean P and ensemble PET forecasts (4) observed climate mean P and PET (5) zero P and ensemble PET forecasts (**Table 4**). P and PET forecasts are joint forecasts in our modelling practice. For example, if the first ensemble member is called from P then the first member from PET is also called to force the hydrological model.

Cases 1-4 are the different possible combinations of ensemble and climate mean meteorological forcing. Case 5 is analysed to determine to which extent the precipitation forecast in a very dry year (2003) is important for seasonal low flow forecasts. It should be noted that all available historical data (1951-2006) were used to estimate the climate mean. For example the climate mean for January 1st is estimated by the average of 55 January 1st values in the available period (1951-2006).

Table 4 Details of the five input cases

Case	Precipitation (P)	The number of ensemble members (P)	Potential evapotranspiration (PET)	The number of ensemble members (PET)
1	Ensemble forecast	40	Ensemble forecast	40
2	Ensemble forecast	40	Climate mean	1
3	Climate mean	1	Ensemble forecast	40
4	Climate mean	1	Climate mean	1
5	Zero	0	Ensemble forecast	40

3.2 Forecast Skill Scores

Three probabilistic forecast skill scores (Brier Skill Score, reliability diagram, hit and false alarm rates) and one deterministic forecast skill score (Mean Forecast Score) are used to analyse the results of low flow forecasts with lead times of 1-90 days. Forecasts for each day in the test period (2002-2005) are used to estimate these scores. The Mean Forecast Score focusing on low flows is introduced in this study, whereas the other three scores have been often used in meteorology (WMO, 2012) and flood hydrology (Velázquez et al., 2010;Renner et al., 2009;Thirel et al., 2008). For the three models, i.e. GR4J, HBV and ANN-E, the forecast probability for each forecast day is estimated as the ratio of the number of ensemble members non-exceeding the preselected thresholds (here Q75) and the total number of ensemble members (i.e. 40 members) for that forecast day.

3.2.1 Brier Skill Score (BSS)

The Brier Skill Score (BSS) (Wilks, 1995) is often used in hydrology to evaluate the quality of probabilistic forecasts (Devineni et al., 2008;Hartmann et al., 2002;Jaun and Ahrens, 2009;Roulin, 2007;Towler et al., 2013).

$$\text{Brier Skill Score: } 1 - \frac{BS_{forecast}}{BS_{climatology}} \tag{10}$$

where the $BS_{forecast}$ is the Brier Score (BS) for the forecast, defined as:

$$\text{Brier Score: } \frac{1}{n}\sum_{t=1}^{n}(F_t - O_t)^2 \tag{11}$$

where F_t refers to the forecast probability, O_t refers to the observed probability (O_t=1 if the observed flow is below the low flow threshold, 0 otherwise), and n is the sample size. $BS_{climatology}$ is the BS for the climatology, which is also calculated from Eq. (11) for every year using climatological probabilities. BSS values range from minus infinity to 1 (perfect forecast). Negative values indicate that the forecast is less accurate than the climatology and positive values indicate more skill compared to the climatology.

3.2.2 Reliability Diagram

The reliability diagram is used to evaluate the performance of probabilistic forecasts of selected events, i.e. low flows. A reliability diagram represents the observed relative frequency as a function of forecasted probability and the 1:1 diagonal shows the perfect reliability line (Velázquez et al., 2010;Olsson and Lindström, 2008). This comparison is important as reliability is one of the three properties of a hydrological forecast (WMO, 2012). A reliability diagram shows the portion of observed data inside preselected forecast intervals.

In this study, exceedence probabilities of 50%, 75%, 85%, 95%, and 99% are chosen as thresholds to categorize the discharges from mean flows to extreme low flows. The forecasted probabilities are then divided into bins of probability categories; here, five bins (categories) are chosen 0-20%, 20%-40%, 40%-60%, 60%-80% and 80%-100%. The observed frequency for each day is chosen to be 1 if the observed discharge is below the low flow threshold, or 0, if not.

3.2.3 Hit and False Alarm Rates

We used hit and false alarm rates to assess the effect of ensembles on low flow forecasts for varying lead times. The hit and false alarm rates indicate respectively the proportion of events for which a correct warning was issued, and the proportion of non events for which a false warning was issued by the forecast model. These two simple rates can be easily calculated from contingency tables (**Table 5**) using Eq. (12) and (13). These scores are often used for evaluating flood forecasts (Martina et al., 2006), however, they can also be used to estimate the utility of low flow forecasts as they indicate the models' ability to correctly forecast the occurrence or non-occurrence of preselected events (i.e. Q_{75} low flows). There are four cases in a contingency table as shown in **Table 5**.

Table 5 Contingency table for the assessment of low-flow events based on the Q75

	Observed	**Not observed**
Forecasted	*hit*: the event forecasted to occur and did occur	*false alarm*: event forecasted to occur, but did not occur
Not forecasted	*miss*: the event forecasted not to occur, but did occur	*correct negative*: event forecasted not to occur and did not occur

$$hit\ rate = \frac{hits}{(hits + misses)} \tag{12}$$

$$false\ alarm\ rate = \frac{false\ alarms}{(correct\ negatives + false\ alarms)} \tag{13}$$

3.2.4 Mean Forecast Score (MFS)

The Mean Forecast Score (MFS) is a new skill score which can be derived from either probabilistic or deterministic forecasts. The probabilities are calculated for the days when low flow occurred. In this study we used a deterministic approach for calculating the observed frequency for all three models. For all three models, ensembles are used for estimating forecast probabilities. The score is calculated as below only for deterministic observed low flows.

$$\text{Mean Forecast Score}:\frac{1}{m}\sum_{j=1}^{m}F_j \tag{14}$$

where F_j is the forecast probability for the *j*-th observed low flow day (i.e. $O_j \leq Q_{75}$) and *m* is the total number of low flow days. The probability of a deterministic forecast can be 0 or 1, whereas it varies from 0 to 1 for ensemble members. For instance, if 23 of the 40 ensemble forecast members indicate low flows for the *j*-th low flow day then F_j = 23/40. It should be noted that this score is not limited to low flows as it has a flexible forecast probability definition which can be adapted to any type of discharges. MFS values range from zero to 1 (perfect forecast).

4 RESULTS

4.1 Calibration and validation

Table 6 shows the parameter ranges and the best performing parameter sets of the three models. The GR4J and HBV models have both well-defined model structures; therefore, their calibration was more straightforward than the calibration of the ANN models. Calibration of the ANN-E model was done in two steps. First, the number of hidden neurons was determined by testing the performance of the ANN-E model with one, two and three hidden neurons.

Table 6 Parameter ranges and calibrated values of the pre-selected three models

Parameter	Unit	Range	Calibrated value	Description
			GR4J model	
X1	[mm]	10-2000	461.4	Capacity of the production store
X2	[mm]	-8 to +6	-0.3	Groundwater exchange coefficient
X3	[mm]	10-500	80.8	One day ahead capacity of the routing store
X4	[d]	0-4	2.2	Time base of the unit hydrograph
			HBV model	
FC	[mm]	200-800	285.1	Maximum soil moisture capacity
LP	[-]	0.1-1	0.7	Soil moisture threshold for reduction of evapotranspiration
BETA	[-]	1-6	2.2	Shape coefficient
CFLUX	[mm/d]	0.1-1	1.0	Maximum capillary flow from upper response box to soil moisture zone
ALFA	[-]	0.1-3	0.4	Measure for non-linearity of low flow in quick runoff reservoir
KF	$[d^{-1}]$	0.005-0.5	0.01	Recession coefficient for quick flow reservoir
KS	$[d^{-1}]$	0.0005-0.5	0.01	Recession coefficient for base flow reservoir
PERC	[mm/d]	0.3-7	0.6	Maximum flow from upper to lower response box
			ANN-E model	
W1	[-]	-10 to +10	-2.3	Weight of connection between 1st input node (P) and hidden neuron
W2	[-]	-10 to +10	0.03	Weight of connection between 2nd input node (PET) and hidden neuron
W3	[-]	-10 to +10	-0.02	Weight of connection between 3rd input node (*Q(t-1)*) and hidden neuron
W4	[-]	-10 to +10	3.7	Weight of connection between hidden neuron and output node
B1	[-]	-10 to +10	0.02	Bias value in hidden layer

B2	[-]	-10 to +10	1.1	Bias value in output layer

Second, daily P, PET and Q are used as three inputs for the tested ANN-E model with one, two and three hidden neurons due to the fact that these inputs are comparable with the inputs of the GR4J and HBV models. Figure 2a shows that the performance of the ANN-E model does not improve with additional hidden neurons. Based on the performance in the validation period, one hidden neuron is selected. GR4J and HBV are also calibrated. The results of the three models used in this study are presented in Figure 2b.

The performances of GR4J and HBV are similar in the calibration period, whereas HBV performs better in the validation period (Figure 2b). This is not surprising, since HBV has a more sophisticated model structure than GR4J.

It should be noted that the effect of anthropogenic activities (e.g. flood preventive regulations and urbanisation) on the alteration of flow magnitude and dynamics is not obvious as we found weak positive trends in all P, PET and Q series ($p<0.025$ for the three variables using Man Kendall method) which might be caused by climatic changes. Other studies reported that the trends in flood stages in Moselle River were not significant (Bormann et al., 2011).

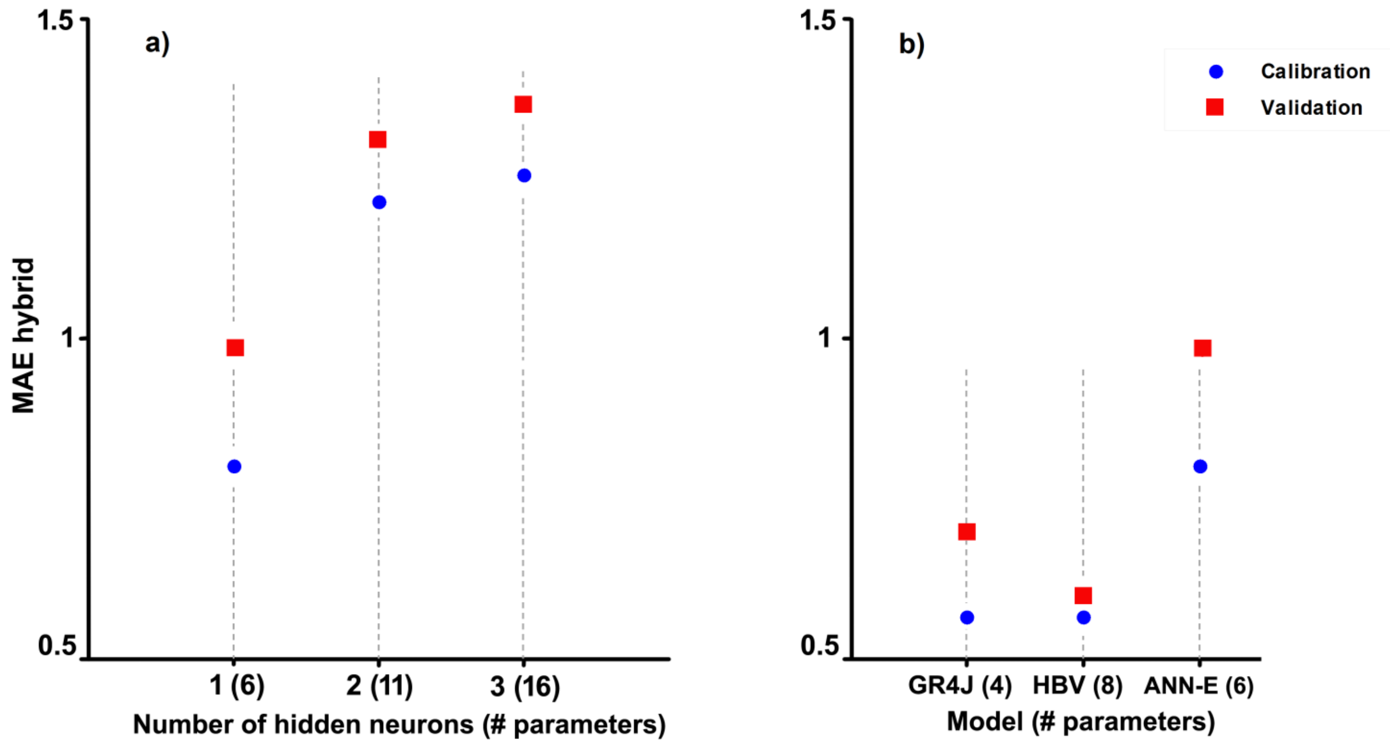


Figure 2 Calibration and validation results of **a)** the ANN-E model with one, two and three hidden neurons and **b)** the three models used in this study. The same calibration (1971-2001) and validation (1951-1970) periods are used for both plots.

4.2 Effect of ensembles on low flow forecasts for 90 day lead time

The effect of ensemble P and PET on GR4J, HBV and ANN-E is presented as a range bounded by the lowest and highest forecast values in Figure 3a and b. The two years, i.e. 2002 and 2003, are carefully selected as they represent a relatively wet year and a very dry year respectively. Figure 3a shows that there are significant differences between the three model results. The 90 day ahead low flows in 2002 are mostly over-predicted by the ANN-E model, whereas GR4J and HBV over-predict low flows observed after August. The over-prediction of low flows is more pronounced for GR4J than for the other three models. The over-prediction of low flows by ANN-E is mostly at the same level. This less sensitive behaviour of ANN-E to the forecasted ensemble inputs shows the effect of the logarithmic sigmoid transfer function on the results. Due to the nature of this algorithm, input is rescaled to a small interval [0, 1] and the gradient of the sigmoid function at large values approximates zero (Wang et al., 2006). Further, ANN-E is also not sensitive to the initial model conditions updated on every forecast issue day. The less pronounced over-prediction of low flows by HBV compared to GR4J may indicate that the slow responding groundwater storage in HBV is less sensitive to different forecasted ensemble P and PET inputs (Demirel et al., 2013b).

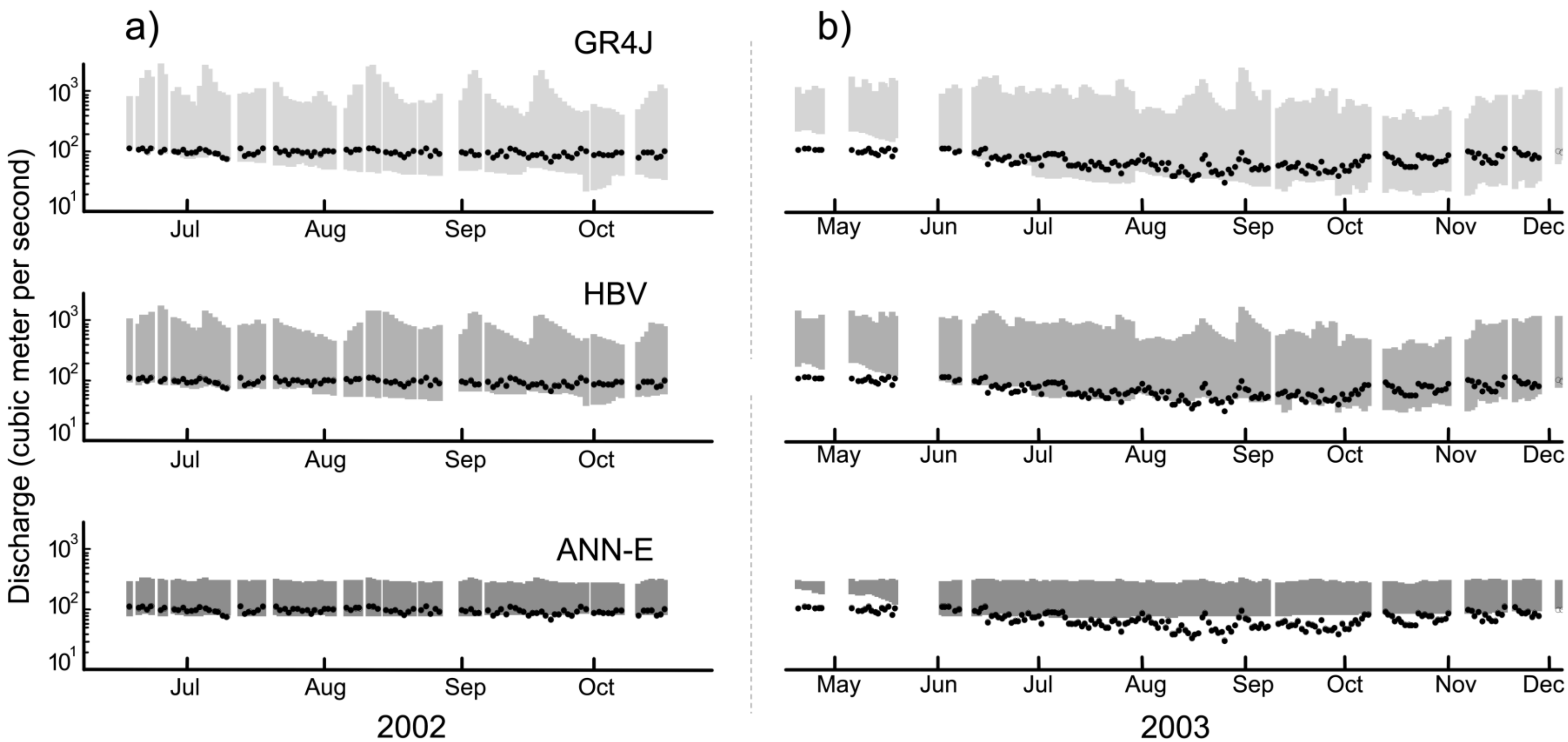


Figure 3 Range (shown as grey shade) of low flow forecasts in **a)** 2002 (the wettest year of the test period with 101 low flow days) **b)** 2003 (the driest year of the test period with 192 low flow days) for a lead time of 90 days using ensemble P and PET as input for GR4J, HBV and ANN-E models (case 1 – 2002 and 2003). The gaps in the figures indicate non-low flow days (i.e. censored).

The results for 2003 are slightly different than those for 2002. As can be seen from Figure 3b the number of low flow days has increased in the dry year, i.e. 2003, and the low flows between August and November are not captured by any of the 40-ensemble forecasts using ANN-E. The most striking result in Figure 3b is that the low flows observed in the period between April and May are not captured by any of the three models, i.e. GR4J, HBV and ANN-E. The poor performance of the models during the spring period can be explained by the high precipitation amount in this period. The poor simulation of high flows in the preceding winter months can have an effect on the forecasts too. The 90 day low flows between October and November are better forecasted by GR4J and HBV than the ANN-E model. The two hydrological models used in this study have well defined surface and ground water components. Therefore, they react to the weather inputs in a physically meaningful way. However, in black box models, the step functions (transfer functions or activation functions) may affect the model behaviour. The ANN model will then react to a certain range of inputs based on the objective function. This feature of ANN is the main reason for the erratic behaviour in Figure 4b and the small (and uniform) uncertainty range in the figures (e.g. Figure 3).

For the purpose of determining to which extent ensemble P and PET inputs and different initial conditions affect 90 day low flow forecasts, we run the models with different input combinations such as ensemble P or PET and climate mean P or PET and zero precipitation. Figure 4a shows the forecasts using ensemble P and climate mean PET as input for three models. The picture is very similar to Figure 3b as most of the observed low flows fall within the constructed forecast range by GR4J and HBV. The forecasts issued by GR4J are better than those issued by the other two models. However, the range of forecasts using GR4J is larger than for the other models showing the sensitivity of the model for different precipitation inputs. It is obvious that most of the range in all forecasts is caused by uncertainties originating from ensemble precipitation input.

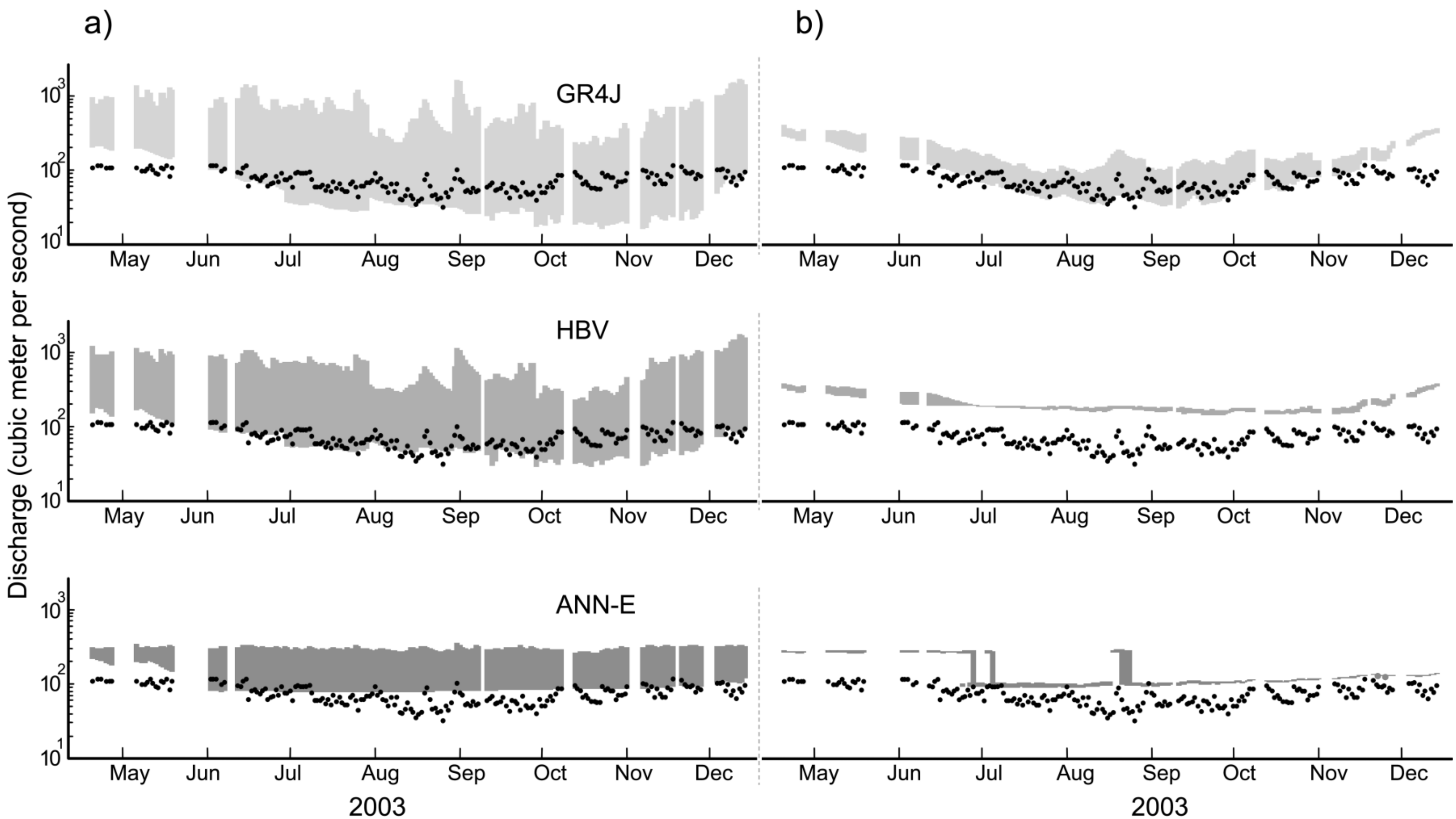


Figure 4 Range (shown as grey shade) of low flow forecasts in 2003 for a lead time of 90 days using **a)** ensemble P and climate mean PET (case 2) **b)** climate mean P and ensemble PET as input for GR4J, HBV and ANN-E models (case 3). The gaps in the figures indicate non-low flow days (i.e. censored).

Figure 4b shows the forecasts using climate mean P and ensemble PET as input for three models, i.e. GR4J, HBV and ANN-E. Interestingly, only GR4J could capture the 90 day low flows between July and November using climate mean P and ensemble PET showing the ability of the model to handle the excessive rainfall. None of the low flows were captured by HBV, whereas very few low flow events were captured by ANN-E (Figure 4b). The precipitation information is crucial for the conceptual models to forecast low flows for a lead time of 90 days. The narrow uncertainty band indicates that the effect of the PET ensemble on the forecasts is less pronounced as compared to the effect of the P ensemble.

Figure 5a shows the forecasts using climate mean P and PET as input for three models. The results are presented by point values without a range since only one deterministic forecast is issued. There are significant differences in the results of the three models. For instance, all 90 day ahead low flows in 2003 are over-predicted by HBV, whereas the over-prediction of low flows is less pronounced for ANN-E. It is remarkable that GR4J can forecast a very dry year accurately using the climate mean. The low values of the calibrated maximum soil moisture capacity and percolation parameters of HBV (*FC* and *PERC*) can be the main reason for over-prediction of all low flows as the interactions of parameters with climate mean P input can result in higher model outputs.

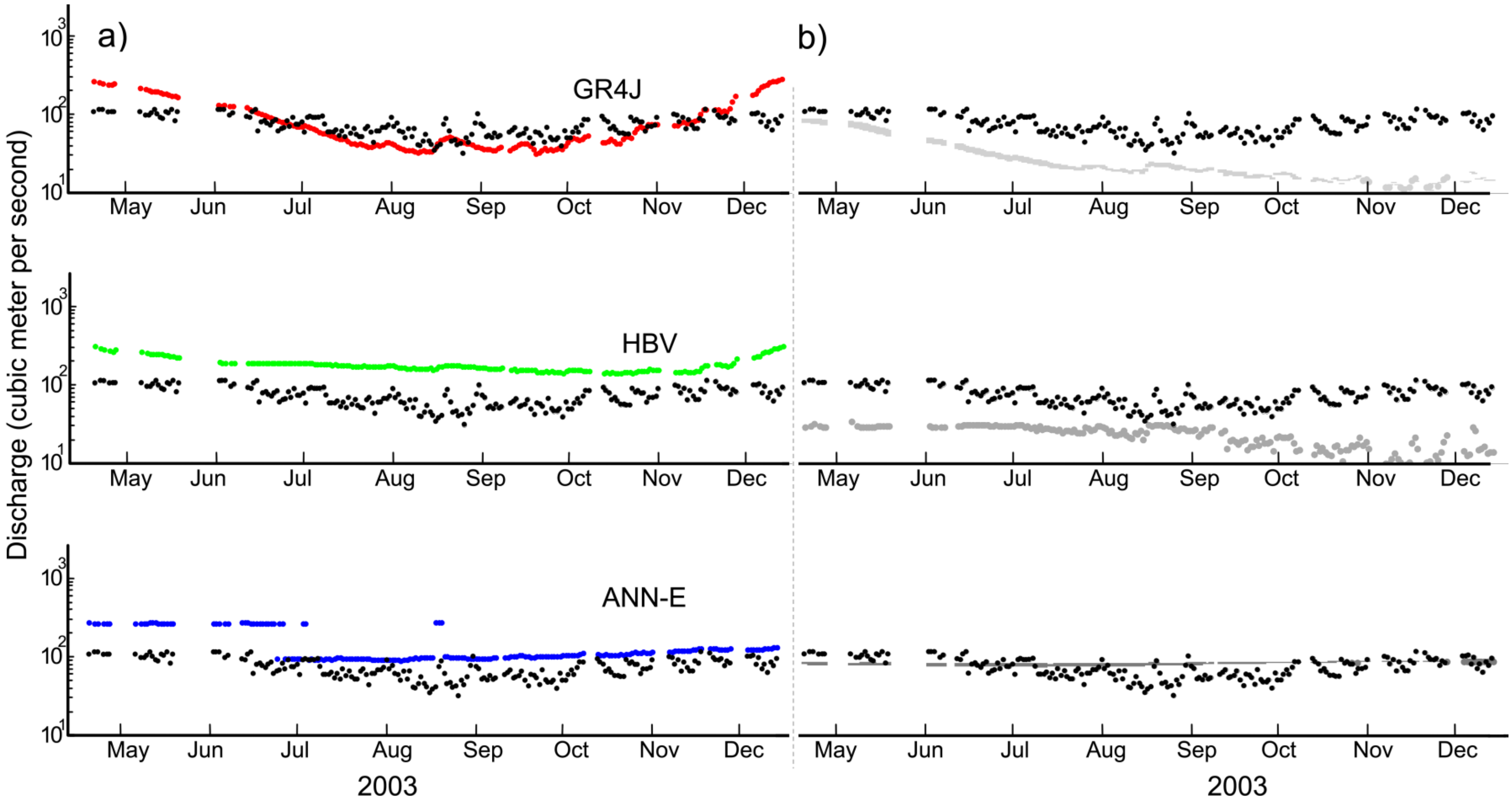
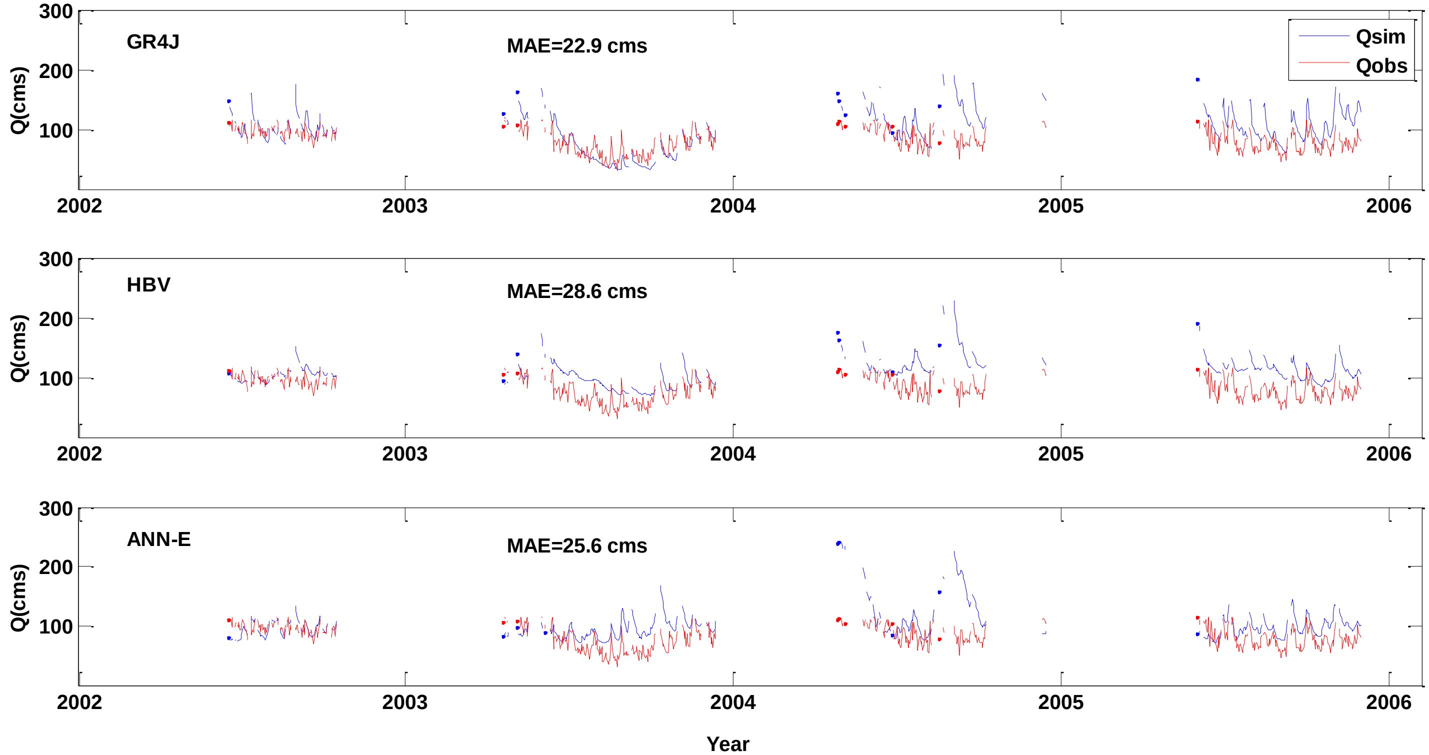


Figure 5 Low flow forecasts in 2003 for a lead time of 90 days using **a)** both climate mean P and PET (case 4) and **b)** zero P and ensemble PET (case 5) as input for GR4J, HBV and ANN-E models. The gaps in the figures indicate non-low flow days (i.e. censored).

We also assessed the seasonal forecasts using zero P and ensemble PET as inputs for three models (Figure 5b). Not surprisingly, both GR4J and HBV under-predicted most of the low flows when they are run without precipitation input. The results of the case 5 confirm that the P input is very crucial for improving low flow forecasts although obviously less precipitation is usually observed in a low flow period compared to other periods.

Figure 6 shows the performance of the three models in the test period using perfect P and PET forecasts as input. This is an idealistic case showing that GR4J model performs better than the other two models. It is interesting to note that ANN-E model does not produce constant predictions as in the previous figures showing the ability of this black box model to perform comparable to the conceptual models when configured and trained properly.

505



506

507 Figure 6 Benchmark reference forecasts using the three models (GR4J, HBV and ANN-E) using observed P and PET (i.e. perfect
508 forecasts)
509

We also show the minimum and maximum prediction errors for each case in **Table 7**. There are large differences in case 1 and 2 as compared to the other cases. It is also obvious that the uncertainty range is larger in case 1 than in case 2 for the conceptual models. This is also what we see in Figure 3 and Figure 4 above.

Table 7 Minimum and maximum prediction errors for low flow forecasts for a lead time of 90 days during the test period 2002-2005

Model	Minimum , Median and Maximum MAE (m^3/s)				
	Case 1	**Case 2**	**Case 3**	**Case 4**	**Case 5**
HBV	[23 101 **785**]	[23 72 **600**]	[108 119 135]	[105 105 105]	[57 57 57]
GR4J	[33 122 **906**]	[36 75 **646**]	[46 61 111]	[44 44 44]	[55 58 59]
ANN-E	[17 94 227]	[18 72 221]	[65 73 80]	[65 65 65]	[16 16 17]

4.3 Effect of ensembles on low flow forecast skill scores

Figure 7 compares the three models and the effect of ensemble P and PET on the skill of probabilistic low flow forecasts with varying lead times. In this figure, four different skill scores are used to present the results of probabilistic low flow forecasts issued by GR4J, HBV and ANN-E. From an operational point of view, the main purpose of investigating the effect of ensembles and model initial conditions on ensemble low flow forecasts with varying lead times is to improve the forecast skills (e.g. hit rate, reliability, BSS and MFS) and to reduce false alarms and misses. From Figure 7 we can clearly see that the results of GR4J show the lowest BSS, MFS and hit rate. The false alarm rate of forecasts using GR4J is also the lowest compared to those using other models. The decrease in false alarm rates after a lead time of 20 days shows the importance of initial condition uncertainty for short lead time forecasts. The limit is around 20 days for ANN-E and shorter for the other two models. When the forecast is issued on day (t), the model states are updated using the observed discharge on that day (t). For GR4J and HBV we used the deterministic state update procedure described in section 3.1.5. However, the models probably spin-up after some days and improve the results for false alarm rate are improved. For longer lead times the error is better handled by the models. We further analysed the forecasted meteorological forcing data (P and PET) to see if there is any difference between the short lead time (~20 days) and long lead time (e.g. 90 days). This is done for three different lead times for each model when the false alarm rate was highest (i.e. 12, 15 and 21 days based on the false alarm rates of GR4J, HBV and ANN-E respectively.). We compared the boxplots from these problematic lead times with the 90 day lead time (not shown here but available in the review reports). It is interesting to note that the ranges for P and PET are larger at 90 day lead time as compared to shorter lead times. However, the observed P and PET values (i.e. perfect forecasts) are covered by the large ranges resulting in higher hit rates (i.e. lower false alarm rates). In other words, for short lead times, 12, 15 and 21 days in particular, the ranges for P and PET are smaller than those for the 90 day lead time but the observed P and PET values are usually missed causing higher false alarm rates in the results.

It appears from the results that ANN-E and HBV show a comparable skill in forecasting low flows up to a lead time of 90 days.

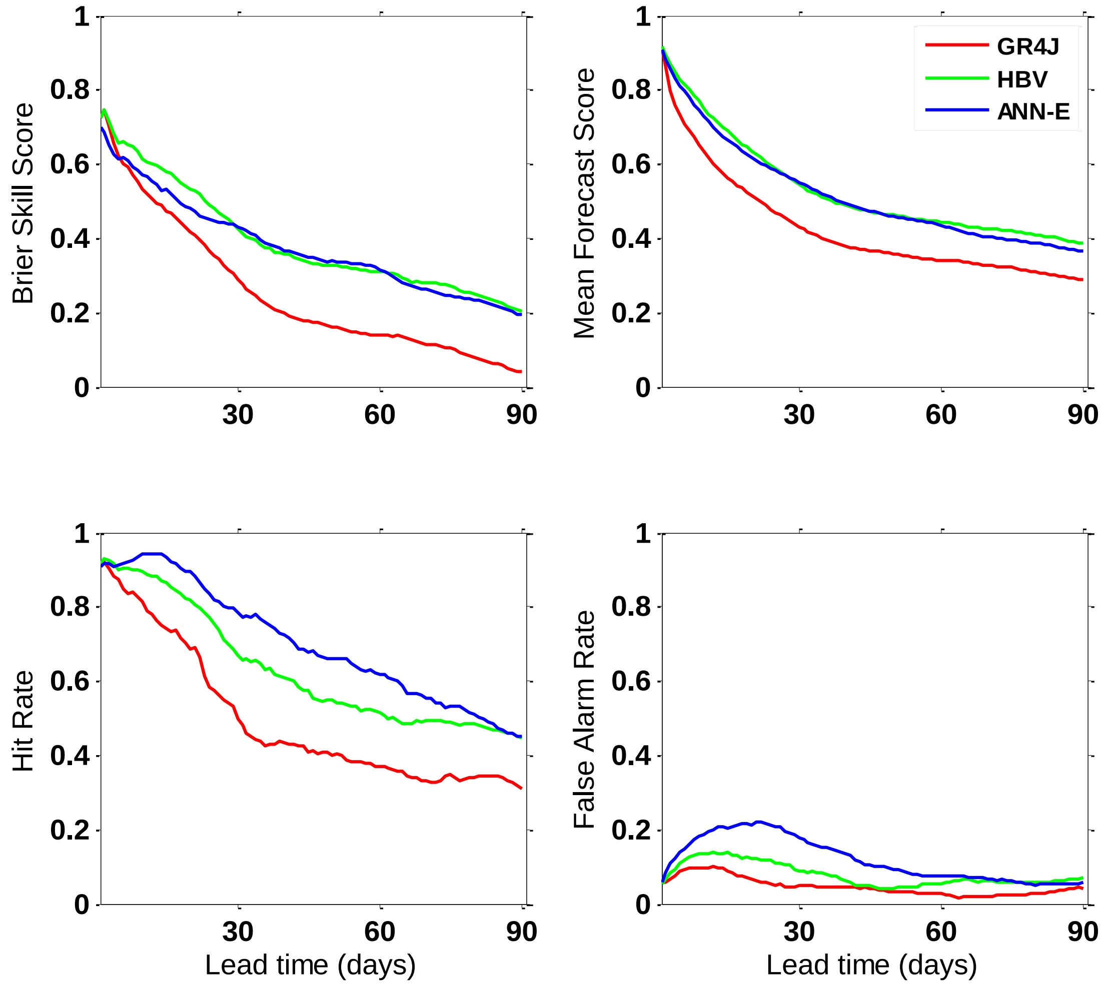


Figure 7 Skill scores for forecasting low flows at different lead times for three different hydrological models for the test period 2002-2005. Note that all forecasts (including high and low flow time steps) are used to estimate these skill scores.

Figure 8 compares the reliability of probabilistic 90 day low flows forecasts below different thresholds (i.e. Q75, Q90 and Q95) using ensemble P and PET as input for three models. The figure shows that the Q75 and Q90 low flow forecasts issued by the HBV model are more reliable compared to the other models. Moreover, all three models under-predict most of the forecast intervals. It appears from Figure 8c that very critical low flows (i.e. Q99) are under-predicted by the GR4J model.

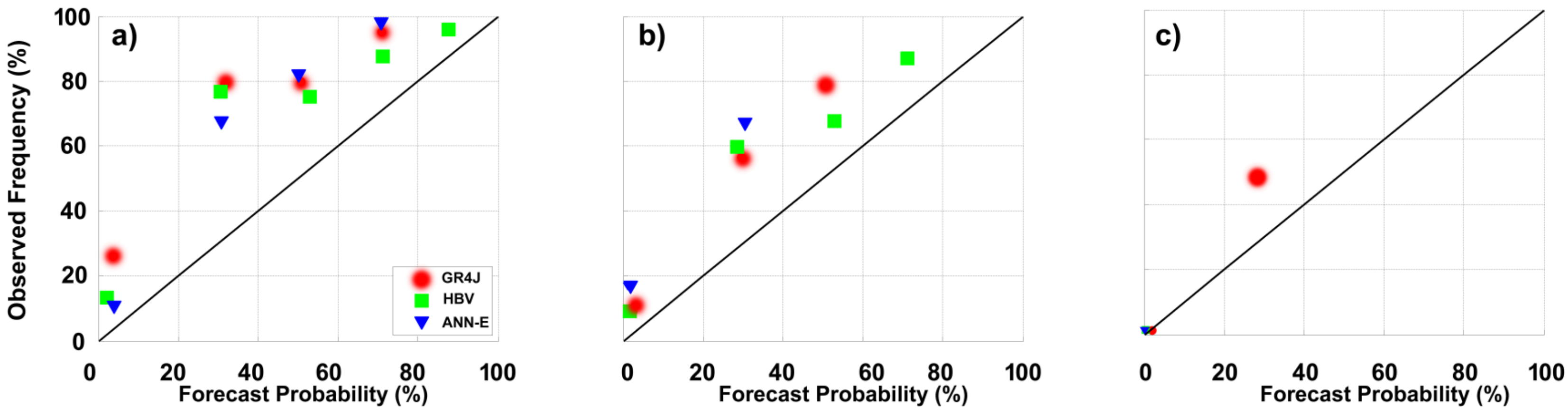


Figure 8 Reliability diagram for different low flow forecasts **a)** Low flows below Q75 threshold (584 observed events in the test period 2002-2005) **b)** Low flows below Q90 threshold (250 observed events) **c)** Low flows below Q99 threshold (20 observed events). The forecasts are issued for a lead time of 90 days for the test period 2002-2005 using ensemble P and PET as input for GR4J, HBV and ANN-E models.

5 DISCUSSION

To compare data-driven and conceptual modelling approaches and to evaluate the effects of seasonal meteorological forecasts on low flow forecasts, 40-member ensembles of ECMWF seasonal meteorological forecasts were used as input for three low flow forecast models.

These models were calibrated using a hybrid low flow objective function. Although combining two metrics offered a selective evaluation of low flows, we have noted an important caveat using the second component of the hybrid metric as it is less sensitive as compared to the first part of the hybrid metric resulting in higher (optimistic) values for most cases. The different units had no effect on our calibration results as the ultimate calibration target value is zero (i.e. unit independent). Other studies also combined different metrics with different units (Nash Sutcliffe, RMSE, R^2 and NumSC, i.e. the number of sign changes in the errors) into one objective function (Hamlet et al., 2013). However, the modellers should carefully use the hybrid function introduced in this study, in particular when comparing different model results. Plotting the two parts of this hybrid function as a Pareto front can lead to a more clear picture than simply summing the two metrics.

In this study, different input combinations were compared to distinguish between the effects of ensemble P and PET and model initial conditions on 90 day low flow forecasts. The models could reasonably forecast low flows when ensemble P was introduced into the models. This result is in line with that of Shukla and Lettenmaier (2011) who found that seasonal meteorological forecasts have a greater influence than initial model conditions on the seasonal hydrological forecast skills. Moreover, our analyses show that the better forecast performance for longer lead times is an obvious artefact since the higher hit rates are the result of a more uncertain (larger range) forecasts. The probabilistic skill scores focuses on the forecasts, the uncertainty in the meteorological forcing data should be carefully scrutinized using different quantitative screening methods e.g. box plots.

Two other related studies also showed that the effect of a large spread in ensemble seasonal meteorological forecasts is larger than the effect of initial conditions on hydrological forecasts with lead times longer than 1-2 months (Li et al., 2009;Yossef et al., 2013). The encouraging results of low flow forecasts using ensemble seasonal precipitation forecasts for the hydrological models confirm the utility of seasonal meteorological forcing for low flow forecasts. Shukla et al. (2012) also found useful forecast skills for both runoff and soil moisture forecasting at seasonal lead times using the medium range weather forecasts.

In this study, we also assessed the effects of ensemble P and PET on the skill scores of low flow forecasts with varying lead times up to 90 days. In general, the four skill scores show similar results. Not surprisingly, all models under-predicted low flows without precipitation information (zero P). The most evident two patterns in these scores are that first, the forecast skill drops sharply until a lead time of 30 days and second, the skill of probabilistic low flow forecasts issued by GR4J is the lowest, whereas the skill of forecasts issued by ANN-E is the highest compared to the other two models. Further, our study showed that data-driven models can be good alternatives to conceptual models for issuing seasonal low flow forecasts (e.g. Figure 6).

The two hydrological models used in this study have well defined surface and ground water components. Therefore, they react to the weather inputs in a physically meaningful way. However, in black box models, the step functions (transfer functions or activation functions) may limit model sensitivity after the training. The ANN model will then react to a certain range of inputs based on the objective function. This feature of an ANN is the main reason for the small (and uniform) uncertainty range in the figures. The over prediction of the models is closely related to the over prediction of the P by the ensembles. Low flows are usually over predicted by the models for the entire period. However, there are under-predictions of low flows for some days in November-December as well. Before June, none of the low flows are captured by the ensemble members. The best performing period is the fall and the worst performing period is the spring period for the models. The poor performance of the models during the spring period can be explained by the high precipitation amount in this period. Since the first part of the objective function used in this study solely focuses on low flows, the high flow period is less important in the calibration. The low flows occurring in the spring period are, therefore, missed in the forecasts. The simulation of snow cover during winter and snow melt during the spring can both have effects on the forecasts too.

6 CONCLUSIONS

Three hydrological models have been compared regarding their performance in the calibration, validation and forecast periods, and the effect of seasonal meteorological forecasts on the skill of low flow forecasts has been assessed for varying lead times. The comparison of three different models help us to contrast data-driven and conceptual models in low flow forecasts, whereas running the models with different input combinations, e.g. climate mean precipitation and ensemble potential evapotranspiration, help us to identify which input source led to the largest range in the forecasts. A new hybrid low flow objective function, comprising the mean absolute error of low flows and the mean absolute error of inverse discharges, is used for comparing low flow simulations, whereas the skill of the probabilistic seasonal low flow forecasts has been evaluated based on the ensemble forecast range, Brier Skill Score, reliability, hit/false alarm rates and Mean Forecast Score. The latter skill score (MFS) focusing on low flows is firstly introduced in this study. In general our results showed that;

- Based on the results of the calibration and validation, one hidden neuron in ANN was found to be enough for seasonal forecasts as additional hidden neurons did not increase the simulation performance. The difference between calibration and validation performances was smallest for the HBV model, i.e. the most sophisticated model used in this study.
- Based on the results of the comparison of different model inputs for two years (i.e. 2002 and 2003), the largest range for 90 day low flow forecasts is found for the GR4J model when using ensemble seasonal meteorological forecasts as input. Moreover, the uncertainty arising from ensemble precipitation has a larger effect on seasonal low flow forecasts than the effects of ensemble potential evapotranspiration. All models are prone

to over-predict low flows using ensemble seasonal meteorological forecasts. However, the precipitation forecasts in the forecast period are crucial for improving the low flow forecasts. As expected, all three models, i.e. GR4J, HBV and ANN-E under-predicted 90 day ahead low flows in 2003 without rainfall data.

- Based on the results of the comparison of forecast skills with varying lead times, the false alarm rate of GR4J is the lowest indicating the ability of the model of forecasting non-occurrence of low flow days. The low flow forecasts issued by HBV are more reliable compared to the other models. The hit rate of ANN-E is higher than that of the two conceptual models used in this study. Overall, the ANN-E and HBV models are the best performing two of the three models using ensemble P and PET.

Further work should examine the effect of model parameters and initial conditions on the seasonal low flow forecasts as the values of the maximum soil moisture and percolation related parameters of conceptual models can result in over- or under-prediction of low flows. The uncertainty increases in seasonal meteorological forecasts can lead to better skill scores as an artefact of large ranges in input. Therefore, the quality of the model inputs should be assessed in addition to the model outputs. It is noteworthy to mention that the data-driven model developed in this study, i.e. ANN-E, can be applied to other large river basins elsewhere in the world. Surprisingly, ANN-E and HBV showed a similar skill for seasonal forecasts, where a priori we expected that the two conceptual models, GR4J and HBV, would show similar results up to a lead time of 90 days.

ACKNOWLEDGEMENTS

We acknowledge the financial support of the Dr. Ir. Cornelis Lely Stichting (CLS), Project No. 20957310. The research is part of the programme of the Department of Water Engineering and Management at the University of Twente and it supports the work of the UNESCO-IHP VII FRIEND-Water programme. Discharge data for the River Rhine were provided by the Global Runoff Data Centre (GRDC) in Koblenz (Germany). Areal precipitation and evapotranspiration data were supplied by the Federal Institute of Hydrology (BfG), Koblenz (Germany). REGNIE grid data were extracted from the archive of the Deutscher Wetterdienst (DWD: German Weather Service), Offenbach (Germany). ECMWF ENS data used in this study have been obtained from the ECMWF seasonal forecasting system, i.e. Mars System 3. We thank Dominique Lucas from ECMWF who kindly guided us through the data retrieval process. The GIS base maps with delineated 134 basins of the Rhine basin were provided by Eric Sprokkereef, the secretary general of the Rhine Commission (CHR). The GR4J and HBV model codes were provided by Ye Tian. We are grateful to the members of the Referat M2 – Mitarbeiter/innen group at BfG, Koblenz, in particular Peter Krahe, Dennis Meißner, Bastian Klein, Robert Pinzinger, Silke Rademacher and Imke Lingemann, for discussions on the value of seasonal low flow forecasts.
The constructive review comments of Kerstin Stahl (Associate Editor), Renata Romanowicz, Stefanie Jörg-Hess and one anonymous reviewer significantly improved this paper.

REFERENCES

Adamowski, J., Chan, H. F., Prasher, S. O., Ozga-Zielinski, B., and Sliusarieva, A.: Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada, Water Resour. Res., 48, W01528, 10.1029/2010wr009945, 2012.

Archer, D. R., and Fowler, H. J.: Using meteorological data to forecast seasonal runoff on the River Jhelum, Pakistan, J. Hydrol., 361, 10-23, 10.1016/j.jhydrol.2008.07.017, 2008.

ATV-DVWK: Verdunstung in Bezug zu Landnutzung, Bewuchs und Boden, Merkblatt ATV-DVWK-M 504, Hennef, 2002.

Bell, V. A., Davies, H. N., Kay, A. L., Marsh, T. J., Brookshaw, A., and Jenkins, A.: Developing a large-scale water-balance approach to seasonal forecasting: application to the 2012 drought in Britain, Hydrol. Processes, 10.1002/hyp.9863, 2013.

Bierkens, M. F. P., and van Beek, L. P. H.: Seasonal Predictability of European Discharge: NAO and Hydrological Response Time, J. Hydrometeorol., 10, 953-968, Doi 10.1175/2009jhm1034.1, 2009.

Booij, M. J.: Impact of climate change on river flooding assessed with different spatial model resolutions, J. Hydrol., 303, 176-198, DOI 10.1016/j.jhydrol.2004.07.013, 2005.

Bormann, H.: Runoff regime changes in German rivers due to climate change, Erdkunde, 257-279, 2010.

Bormann, H., Pinter, N., and Elfert, S.: Hydrological signatures of flood trends on German rivers: Flood frequencies, flood heights and specific stages, J. Hydrol., 404, 50-66, http://dx.doi.org/10.1016/j.jhydrol.2011.04.019, 2011.

Chiew, F. H. S., Zhou, S. L., and McMahon, T. A.: Use of seasonal streamflow forecasts in water resources management, J. Hydrol., 270, 135-144, 2003.

Chowdhury, S., and Sharma, A.: Multisite seasonal forecast of arid river flows using a dynamic model combination approach, Water Resour. Res., 45, W10428, 10.1029/2008wr007510, 2009.

Coley, D. M., and Waylen, P. R.: Forecasting dry season streamflow on the Peace River at Arcadia, Florida, USA, J. Am. Water Resour. Assoc., 42, 851-862, 2006.

De Vos, N. J., and Rientjes, T. H. M.: Multiobjective training of artificial neural networks for rainfall-runoff modeling, Water Resour. Res., 44, W08434, 10.1029/2007wr006734, 2008.

Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: Identification of appropriate lags and temporal resolutions for low flow indicators in the River Rhine to forecast low flows with different lead times, Hydrol. Processes, 27, 2742-2758, 10.1002/hyp.9402, 2013a.

Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models, Water Resour. Res., 49, 10.1002/wrcr.20294, 2013b.

Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: Effect of different uncertainty sources on the skill of 10 day ensemble low flow forecasts for two hydrological models, Water Resour. Res., 49, 4035-4053, 10.1002/wrcr.20294, 2013c.

Devineni, N., Sankarasubramanian, A., and Ghosh, S.: Multimodel ensembles of streamflow forecasts: Role of predictor state in developing optimal combinations, Water Resour. Res., 44, W09404, 10.1029/2006wr005855, 2008.

Doblas-Reyes, F. J., Weisheimer, A., Déqué, M., Keenlyside, N., McVean, M., Murphy, J. M., Rogel, P., Smith, D., and Palmer, T. N.: Addressing model uncertainty in seasonal and annual dynamical ensemble forecasts, Q. J. R. Meteorol. Soc., 135, 1538-1559, 10.1002/qj.464, 2009.

Dutra, E., Di Giuseppe, F., Wetterhall, F., and Pappenberger, F.: Seasonal forecasts of droughts in African basins using the Standardized Precipitation Index, Hydrol. Earth Syst. Sci., 17, 2359-2373, 10.5194/hess-17-2359-2013, 2013.

Dutra, E., Pozzi, W., Wetterhall, F., Di Giuseppe, F., Magnusson, L., Naumann, G., Barbosa, P., Vogt, J., and Pappenberger, F.: Global meteorological drought – Part 2: Seasonal forecasts, Hydrol. Earth Syst. Sci., 18, 2669-2678, 10.5194/hess-18-2669-2014, 2014.

Eberle, M.: Hydrological Modelling in the River Rhine Basin Part III - Daily HBV Model for the Rhine Basin BfG-1451, Institute for Inland Water Management and Waste Water Treatment (RIZA) and Federal Institute of Hydrology (BfG) Koblenz, Germany 2005.

ECMWF: Describing ECMWF's forecasts and forecasting system, ECMWF newsletter 133, Available from: http://old.ecmwf.int/publications/manuals/mars/ (last access: 26/07/2014), 2012.

Elshorbagy, A., Corzo, G., Srinivasulu, S., and Solomatine, D. P.: Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - Part 1: Concepts and methodology, Hydrol. Earth Syst. Sci., 14, 1931-1941, 10.5194/hess-14-1931-2010, 2010.

Engeland, K., Renard, B., Steinsland, I., and Kolberg, S.: Evaluation of statistical models for forecast errors from the HBV model, J. Hydrol., 384, 142-155, 2010.

EU: Horizon 2020 – Work Programme 2014-2015: Water 7_2015: Increasing confidence in seasonal-to-decadal predictions of the water cycle. http://www.aber.ac.uk/en/media/departmental/researchoffice/funding/UKRO-Horizon-2020_climate_change_draft_wp.pdf, (Accessed September 4, 2013).
, 2013.

Felipe, P.-S., and Nelson, O.-N.: Forecasting of Monthly Streamflows Based on Artificial Neural Networks, J. Hydrol. Eng., 14, 1390-1395, 2009.

Förster, K., Meon, G., Marke, T., and Strasser, U.: Effect of meteorological forcing and snow model complexity on hydrological simulations in the Sieber catchment (Harz Mountains, Germany), Hydrol. Earth Syst. Sci., 18, 4703-4720, 10.5194/hess-18-4703-2014, 2014.

Fundel, F., Jörg-Hess, S., and Zappa, M.: Monthly hydrometeorological ensemble prediction of streamflow droughts and corresponding drought indices, Hydrol. Earth Syst. Sci., 17, 395-407, 10.5194/hess-17-395-2013, 2013.

Ganguli, P., and Reddy, M. J.: Ensemble prediction of regional droughts using climate inputs and SVM-copula approach, Hydrol. Processes, (accepted), 10.1002/hyp.9966, 2013.

Gaume, E., and Gosset, R.: Over-parameterisation, a major obstacle to the use of artificial neural networks in hydrology?, Hydrol. Earth Syst. Sci., 7, 693-706, 10.5194/hess-7-693-2003, 1999.

Giuntoli, I., Renard, B., Vidal, J. P., and Bard, A.: Low flows in France and their relationship to large-scale climate indices, J. Hydrol., 482, 105-118, 10.1016/j.jhydrol.2012.12.038, 2013.

Gobena, A. K., and Gan, T. Y.: Incorporation of seasonal climate forecasts in the ensemble streamflow prediction system, J. Hydrol., 385, 336-352, 10.1016/j.jhydrol.2010.03.002, 2010.

Görgen, K., Beersma, J., Brahmer, G., Buiteveld, H., Carambia, M., de Keizer, O., Krahe, P., Nilson, E., Lammersen, R., Perrin, C., and Volken, D.: Assessment of Climate Change Impacts on Discharge in the Rhine River Basin: Results of the RheinBlick 2050 Project, Lelystad, CHR, ISBN 978-90-70980-35-1, 211p. Available from: http://www.news.admin.ch/NSBSubscriber/message/attachments/20770.pdf (last access: 30/10/2014), 2010.

Govindaraju, R. S., and Rao, A. R.: Artificial Neural Networks in Hydrology, Kluwer Academic Publishers Norwell, MA, USA, 329 pp., 2000.

Hamlet, A. F., Elsner, M. M., Mauger, G. S., Lee, S.-Y., Tohver, I., and Norheim, R. A.: An Overview of the Columbia Basin Climate Change Scenarios Project: Approach, Methods, and Summary of Key Results, Atmosphere-Ocean, 51, 392-415, 10.1080/07055900.2013.819555, 2013.

Hartmann, H. C., Pagano, T. C., Sorooshian, S., and Bales, R.: Confidence builders: Evaluating seasonal climate forecasts from user perspectives, Bulletin of the American Meteorological Society, 83, 683-698, 2002.

Jaun, S., and Ahrens, B.: Evaluation of a probabilistic hydrometeorological forecast system, Hydrol. Earth Syst. Sci., 13, 1031-1043, 2009.

Kahya, E., and Dracup, J. A.: U.S. streamflow patterns in relation to the El Niño/Southern Oscillation, Water Resour. Res., 29, 2491-2503, 10.1029/93wr00744, 1993.

Kalra, A., Ahmad, S., and Nayak, A.: Increasing streamflow forecast lead time for snowmelt-driven catchment based on large-scale climate patterns, Adv. Water Resour., 53, 150-162, 10.1016/j.advwatres.2012.11.003, 2013.

Kasiviswanathan, K. S., Raj, C., Sudheer, K. P., and Chaubey, I.: Constructing prediction interval for artificial neural network rainfall runoff models based on ensemble simulations, J. Hydrol., 499, 275-288, 10.1016/j.jhydrol.2013.06.043, 2013.

Kuo, C.-C., Gan, T. Y., and Yu, P.-S.: Seasonal streamflow prediction by a combined climate-hydrologic system for river basins of Taiwan, J. Hydrol., 387, 292-303, 2010.

Li, H., Luo, L., and Wood, E. F.: Seasonal hydrologic predictions of low-flow conditions over eastern USA during the 2007 drought, Atmospheric Science Letters, 9, 61-66, 2008.

Li, H., Luo, L., Wood, E. F., and Schaake, J.: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting, J. Geophys. Res., 114, D04114, 10.1029/2008jd010969, 2009.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergstrom, S.: Development and test of the distributed HBV-96 hydrological model, J. Hydrol., 201, 272-288, 1997.

Luo, L., Wood, E. F., and Pan, M.: Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions, J. Geophys. Res., 112, D10102, 10.1029/2006jd007655, 2007.

Madadgar, S., and Moradkhani, H.: A Bayesian Framework for Probabilistic Seasonal Drought Forecasting, J. Hydrometeorol., 14, 1685-1705, 10.1175/JHM-D-13-010.1, 2013.

Martina, M. L. V., Todini, E., and Libralon, A.: A Bayesian decision approach to rainfall thresholds based flood warning, Hydrol. Earth Syst. Sci., 10, 413-426, 10.5194/hess-10-413-2006, 2006.

Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., Le Lay, M., Besson, F., Soubeyroux, J. M., Viel, C., Regimbeau, F., Andréassian, V., Maugis, P., Augeard, B., and Morice, E.: Benchmarking hydrological models for low-flow simulation and forecasting on French catchments, Hydrol. Earth Syst. Sci. Discuss., 10, 13979-14040, 10.5194/hessd-10-13979-2013, 2013.

Olsson, J., and Lindström, G.: Evaluation and calibration of operational hydrological ensemble forecasts in Sweden, J. Hydrol., 350, 14-24, 2008.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, J. Hydrol., 279, 275-289, 2003.

Pokhrel, P., Wang, Q. J., and Robertson, D. E.: The value of model averaging and dynamical climate model predictions for improving statistical seasonal streamflow forecasts over Australia, Water Resour. Res., 49, 6671-6687, 10.1002/wrcr.20449, 2013.

Pushpalatha, R., Perrin, C., Moine, N. L., Mathevet, T., and Andréassian, V.: A downward structural sensitivity analysis of hydrological models to improve low-flow simulation, J. Hydrol., 411, 66–76, 2011.
Pushpalatha, R., Perrin, C., Moine, N. L., and Andréassian, V.: A review of efficiency criteria suitable for evaluating low-flow simulations, J. Hydrol., 420–421, 171-182, 10.1016/j.jhydrol.2011.11.055, 2012.
Renner, M., Werner, M. G. F., Rademacher, S., and Sprokkereef, E.: Verification of ensemble flow forecasts for the River Rhine, J. Hydrol., 376, 463-475, 2009.
Robertson, D. E., Pokhrel, P., and Wang, Q. J.: Improving statistical forecasts of seasonal streamflows using hydrological model output, Hydrol. Earth Syst. Sci., 17, 579-593, 10.5194/hess-17-579-2013, 2013.
Roulin, E.: Skill and relative economic value of medium-range hydrological ensemble predictions, Hydrol. Earth Syst. Sci., 11, 725-737, 2007.
Rutten, M., van de Giesen, N., Baptist, M., Icke, J., and Uijttewaal, W.: Seasonal forecast of cooling water problems in the River Rhine, Hydrol. Processes, 22, 1037-1045, 2008.
Saadat, S., Khalili, D., Kamgar-Haghighi, A., and Zand-Parsa, S.: Investigation of spatio-temporal patterns of seasonal streamflow droughts in a semi-arid region, Natural Hazards, 1-24, 10.1007/s11069-013-0783-y, 2013.
Sauquet, E., Lerat, J., and Prudhomme, C.: La prévision hydro-météorologique à 3-6 mois. Etat des connaissances et applications, La Houille Blanche, 77-84, 2008.
Schubert, S., Koster, R., Hoerling, M., Seager, R., Lettenmaier, D., Kumar, A., and Gutzler, D.: Predicting Drought on Seasonal-to-Decadal Time Scales, Bulletin of the American Meteorological Society, 88, 1625-1630, 10.1175/bams-88-10-1625, 2007.
Shamseldin, A. Y.: Application of a neural network technique to rainfall-runoff modelling, J. Hydrol., 199, 272-294, 10.1016/s0022-1694(96)03330-6, 1997.
Shukla, S., and Lettenmaier, D. P.: Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill, Hydrol. Earth Syst. Sci., 15, 3529-3538, DOI 10.5194/hess-15-3529-2011, 2011.
Shukla, S., Voisin, N., and Lettenmaier, D. P.: Value of medium range weather forecasts in the improvement of seasonal hydrologic prediction skill, Hydrol. Earth Syst. Sci., 16, 2825-2838, 10.5194/hess-16-2825-2012, 2012.
Shukla, S., Sheffield, J., Wood, E. F., and Lettenmaier, D. P.: On the sources of global land surface hydrologic predictability, Hydrol. Earth Syst. Sci., 17, 2781-2796, 10.5194/hess-17-2781-2013, 2013.
Soukup, T. L., Aziz, O. A., Tootle, G. A., Piechota, T. C., and Wulff, S. S.: Long lead-time streamflow forecasting of the North Platte River incorporating oceanic-atmospheric climate variability, J. Hydrol., 368, 131-142, 2009.
Thirel, G., Rousset-Regimbeau, F., Martin, E., and Habets, F.: On the Impact of Short-Range Meteorological Forecasts for Ensemble Streamflow Predictions, J. Hydrometeorol., 9, 1301-1317, 10.1175/2008jhm959.1, 2008.
Tian, Y., Booij, M. J., and Xu, Y.-P.: Uncertainty in high and low flows due to model structure and parameter errors, Stoch. Environ. Res. Risk Assess., 28, 319-332, 10.1007/s00477-013-0751-9, 2014.
Tootle, G. A., and Piechota, T. C.: Suwannee River Long Range Streamflow Forecasts Based On Seasonal Climate Predictors, JAWRA Journal of the American Water Resources Association, 40, 523-532, 2004.

Towler, E., Roberts, M., Rajagopalan, B., and Sojda, R. S.: Incorporating probabilistic seasonal climate forecasts into river management using a risk-based framework, Water Resour. Res., 49, 4997–5008, 10.1002/wrcr.20378, 2013.

Van den Tillaart, S. P. M., Booij, M. J., and Krol, M. S.: Impact of uncertainties in discharge determination on the parameter estimation and performance of a hydrological model, Hydrology Research, 44, 454–466 2013.

Van Dijk, A. I. J. M., Peña-Arancibia, J. L., Wood, E. F., Sheffield, J., and Beck, H. E.: Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide, Water Resour. Res., 49, 2729–2746, 10.1002/wrcr.20251, 2013.

Van Ogtrop, F. F., Vervoort, R. W., Heller, G. Z., Stasinopoulos, D. M., and Rigby, R. A.: Long-range forecasting of intermittent streamflow, Hydrol. Earth Syst. Sci., 15, 3343-3354, 10.5194/hess-15-3343-2011, 2011.

Velázquez, J. A., Anctil, F., and Perrin, C.: Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments, Hydrol. Earth Syst. Sci., 14, 2303-2317, 10.5194/hess-14-2303-2010, 2010.

Vidal, J. P., Martin, E., Franchistéguy, L., Habets, F., Soubeyroux, J. M., Blanchard, M., and Baillon, M.: Multilevel and multiscale drought reanalysis over France with the Safran-Isba-Modcou hydrometeorological suite, Hydrol. Earth Syst. Sci., 14, 459-478, 2010.

Wang, E., Zhang, Y., Luo, J., Chiew, F. H. S., and Wang, Q. J.: Monthly and seasonal streamflow forecasts using rainfall-runoff modeling and historical weather data, Water Resour. Res., 47, W05516, 10.1029/2010wr009922, 2011.

Wang, W., Gelder, P. H. A. J. M. V., Vrijling, J. K., and Ma, J.: Forecasting daily streamflow using hybrid ANN models, J. Hydrol., 324, 383-399, 10.1016/j.jhydrol.2005.09.032, 2006.

Wedgbrow, C. S., Wilby, R. L., Fox, H. R., and O'Hare, G.: Prospects for seasonal forecasting of summer drought and low river flow anomalies in England and Wales, Int. J. Climatol., 22, 219-236, 10.1002/joc.735, 2002.

Wedgbrow, C. S., Wilby, R. L., and Fox, H. R.: Experimental seasonal forecasts of low summer flows in the River Thames, UK, using Expert Systems, Clim. Res., 28, 133-141, 2005.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Elsevier, New York., 1995.

Winsemius, H. C., Dutra, E., Engelbrecht, F. A., Archer Van Garderen, E., Wetterhall, F., Pappenberger, F., and Werner, M. G. F.: The potential value of seasonal forecasts in a changing climate in southern Africa, Hydrol. Earth Syst. Sci., 18, 1525-1538, 10.5194/hess-18-1525-2014, 2014.

Wood, A. W., Maurer, E. P., Kumar, A., and Lettenmaier, D. P.: Long-range experimental hydrologic forecasting for the eastern United States, J. Geophys. Res, 107, 4429, 10.1029/2001JD000659, 2002.

Wood, A. W., and Lettenmaier, D. P.: A Test Bed for New Seasonal Hydrologic Forecasting Approaches in the Western United States, Bulletin of the American Meteorological Society, 87, 1699-1712, 10.1175/bams-87-12-1699, 2006.

Yossef, N. C., van Beek, L. P. H., Kwadijk, J. C. J., and Bierkens, M. F. P.: Assessment of the potential forecasting skill of a global hydrological model in reproducing the occurrence of monthly flow extremes, Hydrol. Earth Syst. Sci., 16, 4233-4246, 10.5194/hess-16-4233-2012, 2012.

Yossef, N. C., Winsemius, H., Weerts, A., van Beek, R., and Bierkens, M. F. P.: Skill of a global seasonal streamflow forecasting system, relative roles of initial conditions and meteorological forcing, Water Resour. Res., 49, 4687–4699, 10.1002/wrcr.20350, 2013.