1   **The skill of seasonal ensemble low flow forecasts in the Moselle River for three**

2   **different hydrological models**

3

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

5   *Water Engineering and Management, Faculty of Engineering Technology, University of Twente,*

6   *P.O. Box 217, 7500 AE Enschede, the Netherlands.*

7   **Current address: Portland State University, Department of Civil & Environmental Engineering,*

8   *1930 S.W. 4th Avenue, Suite 200, Portland, OR 97201, USA (demirel@pdx.edu)*

9   **Abstract**

10   This paper investigates the skill of 90 day low flow forecasts using two conceptual hydrological

11   models and one data-driven model based on Artificial Neural Networks (ANNs) for the Moselle

12   River. The three models, i.e. HBV, GR4J and ANN-Ensemble (ANN-E), all use forecasted

13   meteorological inputs (Precipitation P and potential evapotranspiration PET), whereby we

14   employ ensemble seasonal meteorological forecasts. We compared low flow forecasts for five

15   different cases of seasonal meteorological forcing: (1) ensemble P and PET forecasts; (2)

16   ensemble P forecasts and observed climate mean PET; (3) observed climate mean P and

17   ensemble PET forecasts; (4) observed climate mean P and PET and (5) zero P and ensemble PET

18   forecasts as input for the models. The ensemble P and PET forecasts, each consisting of 40

19   members, reveal the forecast ranges due to the model inputs. The five cases are compared for a

20   lead time of 90 days based on model output ranges, whereas the models are compared based on

21   their skill of low flow forecasts for varying lead times up to 90 days. Before forecasting, the

22   hydrological models are calibrated and validated for a period of 30 and 20 years respectively. The

23   smallest difference between calibration and validation performance is found for HBV, whereas

24   the largest difference is found for ANN-E. From the results, it appears that all models are prone

25   to over-predict runoff during low flow periods using ensemble seasonal meteorological forcing.

26   The largest range for 90 day low flow forecasts is found for the GR4J model when using

27   ensemble seasonal meteorological forecasts as input. GR4J, HBV and ANN-E under-predicted 90

28   day ahead low flows in the very dry year 2003 without precipitation data. The results of the

29   comparison of forecast skills with varying lead times show that GR4J is less skilful than ANN-E

30   and HBV. Overall, the uncertainty from ensemble P forecasts has a larger effect on seasonal low

31   flow forecasts than the uncertainty from ensemble PET forecasts and initial model conditions.

32

35

36

37

38   **1   INTRODUCTION**

39   Rivers in Western Europe usually experience low flows in late summer and high flows in winter.

40   These two extreme discharge phenomena can lead to serious problems. For example, high flow

41   events are quick and can put human life at risk, whereas streamflow droughts (i.e. low flows)

42   develop slowly and can affect a large area. Consequently, the economic loss during low flow

43   periods can be much bigger than during floods (Pushpalatha et al., 2011;Shukla et al., 2012). In

44   the River Rhine, severe problems for freshwater supply, water quality, power production and

45   river navigation were experienced during the dry summers of 1976, 1985 and 2003. Therefore,

46   forecasting seasonal low flows (Towler et al., 2013;Coley and Waylen, 2006;Li et al., 2008) and

47   understanding low flow indicators (Vidal et al., 2010;Fundel et al., 2013;Demirel et al.,

48   2013b;Wang et al., 2011;Saadat et al., 2013;Nicolle et al., 2013) have both societal and scientific

49   value. The seasonal forecast of water flows is therefore listed as one of the priority topics in EU's

50   Horizon 2020 research program (EU, 2013). Further, there is an increasing interest to incorporate

51   seasonal flow forecasts in decision support systems for river navigation and power plant

52   operation during low flow periods. We are interested in forecasting low flows with a lead time of

53   90 days, and in presenting the effect of ensemble meteorological forecasts for three hydrological

54   models.

55   Generally, two approaches are used in seasonal hydrological forecasting. The first one is a

56   statistical approach, making use of data-driven models based on relationships between river

57   discharge and hydroclimatological indicators (Wang et al., 2011;Van Ogtrop et al., 2011). The

58   second one is a dynamic approach running a hydrological model with forecasted climate input.

59   The first approach is often preferred in regions where significant correlations between river

60   discharge and climatic indicators exist, such as sea surface temperature anomalies (Chowdhury

61 and Sharma, 2009), AMO – Atlantic Multi-decadal Oscillation (Ganguli and Reddy,

62 2013;Giuntoli et al., 2013), PDO – Pacific Decadal Oscillation (Soukup et al., 2009) and warm

63 and cold phases of the ENSO – El Nino Southern Oscillation - index (Chiew et al., 2003;Kalra et

64 al., 2013;Tootle and Piechota, 2004). Kahya and Dracup (1993) identified the lagged response of

65 regional streamflow to the warm phase of ENSO in the south-eastern United States. In the Rhine

66 basin, no teleconnections have been found between climatic indices, e.g. NAO and ENSO, and

67 river discharges (Rutten et al., 2008;Bierkens and van Beek, 2009). However, Demirel et al.

68 (2013b) found significant correlations between hydrological low flow indicators and observed

69 low flows. They also identified appropriate lags and temporal resolutions of low flow indicators

70 (e.g. precipitation, potential evapotranspiration, groundwater storage, lake levels and snow

71 storage) to build data-driven models.

72 The second approach is the dynamic seasonal forecasting approach which has long been explored

73 (Wang et al., 2011;Van Dijk et al., 2013;Gobena and Gan, 2010;Fundel et al., 2013;Shukla et al.,

74 2013;Pokhrel et al., 2013) and has led to the development of the current ensemble streamflow

75 prediction system (ESP) used by different national climate services like the National Weather

76 Service in the United States. The seasonal hydrologic prediction systems are most popular in

77 regions with a high risk of extreme discharge situations like hydrological droughts (Robertson et

78 al., 2013;Madadgar and Moradkhani, 2013). Well-known examples are the NOAA Climate

79 Prediction   Centre's   seasonal   drought   forecasting   system   (available   at

80 http://www.cpc.ncep.noaa.gov), the University of Washington's Surface Water Monitoring

81 system (Wood and Lettenmaier, 2006), Princeton University's drought forecast system (available

82 at   http://hydrology.princeton.edu/forecast)   and   University   of   Utrecht's   global   monthly

83 hydrological forecast system (Yossef et al., 2012). These models provide indications about the

84 hydrologic conditions and their evolution across the modelled domain using available weather

85   ensemble inputs (Gobena and Gan, 2010;Yossef et al., 2012). Moreover, Dutra et al. (2014)

86   showed that global seasonal forecasts of meteorological drought onset are feasible and skilful

87   using the standardized precipitation index (SPI) and two data sets as initial conditions.

88   Many studies have investigated the seasonal predictability of low flows in different rivers such as

89   the Thames and different other rivers in the UK (Bell et al., 2013;Wedgbrow et al.,

90   2002;Wedgbrow et al., 2005), the Shihmen and Tsengwen Rivers in Taiwan (Kuo et al., 2010),

91   the River Jhelum in Pakistan (Archer and Fowler, 2008), more than 200 rivers in France (Sauquet

92   et al., 2008;Giuntoli et al., 2013), five semi-arid areas in South Western Queensland, Australia

93   (Van Ogtrop et al., 2011), five rivers including Limpopo basin and the Blue Nile in Africa (Dutra

94   et al., 2013;Winsemius et al., 2014), the Bogotá River in Colombia (Felipe and Nelson, 2009),

95   the Ohio in the eastern US (Wood et al., 2002;Luo et al., 2007;Li et al., 2009), the North Platte in

96   Colorado, US (Soukup et al., 2009), large rivers in the US (Schubert et al., 2007;Shukla and

97   Lettenmaier, 2011) and the Thur River in the north-eastern part of Switzerland (Fundel et al.,

98   2013). The common result of the above mentioned studies is that the skill of the seasonal

99   forecasts made with global and regional hydrological models is reasonable for lead times of 1-3

100   months (Shukla and Lettenmaier, 2011;Wood et al., 2002) and these forecasting systems are all

101   prone to large uncertainties as their forecast skills mainly depend on the knowledge of initial

102   hydrologic conditions and weather information during the forecast period (Shukla et al.,

103   2012;Yossef et al., 2013;Li et al., 2009;Doblas-Reyes et al., 2009). In a recent study, Yossef et al.

104   (2013) used a global monthly hydrological model to analyse the relative contributions of initial

105   conditions and meteorological forcing to the skill of seasonal streamflow forecasts. They

106   included 78 stations in large basins in the world including the River Rhine for forecasts with lead

107   times up to 6 months. They found that improvements in seasonal hydrological forecasts in the

108   Rhine depend on better meteorological forecasts, which underlines the importance of

109   meteorological forcing quality particularly for forecasts beyond lead times of 1-2 months.

110   Most of the previous River Rhine studies use only one hydrological model, e.g. PREVAH

111   (Fundel et al., 2013) or PCR-GLOBWB (Yossef et al., 2013), to assess the value of ensemble

112   meteorological forcing, whereas in this study, we compare three hydrological models with

113   different structures varying from data-driven to conceptual models. The two objectives of this

114   study are to contrast data-driven and conceptual modelling approaches and to assess the effect of

115   ensemble seasonal forecasted precipitation and potential evapotranspiration on low flow forecast

116   quality and skill scores. By comparing three models with different model structures we address

117   the issue of model structure uncertainty, whereas the latter objective reflects the benefit of

118   ensemble seasonal forecasts. Moreover, the effect of initial model conditions is partly addressed

119   using climate mean data in one of the cases.

120   The analysis complements recent efforts to analyse the effects of ensemble weather forecasts on

121   low flow forecasts with a lead time of 10 days using two conceptual models (Demirel et al.,

122   2013c), by studying the effects of seasonal ensemble weather forecasts on 90 day low flow

123   forecasts using not only conceptual models but also data-driven models.

124   The outline of the paper is as follows. The study area and data are presented in section 2. Section

125   3 describes the model structures, their calibration and validation set-ups and the methods

126   employed to estimate the different attributes of the forecast quality. The results are presented in

127   section 4 and discussed in section 5, and the conclusions are summarised in section 6.

128

129 **2   STUDY AREA AND DATA**

130 *2.1   Study area*

131 The study area is the Moselle River basin, the largest sub-basin of the Rhine River basin. The

132 Moselle River has a length of 545 km. The river basin has a surface area of approximately 27,262

133 km^2. The altitude in the basin varies from 59 to 1326 m, with a mean altitude of 340 m (Demirel

134 et al., 2013b). There are 26 subbasins with surface areas varying from 102 to 3353 km^2.

135 Approximately 410 mm (~130 m^{3}s^{-1}) discharge is annually generated in the Moselle basin

136 (Demirel et al., 2013c). The outlet discharge at Cochem varies from 14 m^{3}s^{-1} in dry summers to

137 a maximum of 4000 m^{3}s^{-1} during winter floods. The River Rhine, in general, has been heavily

138 canalised for river navigation and flood prevention. There are many dams in the upstream part of

139 the River Rhine in Switzerland. However, the human influence on the Moselle River is assumed

140 to be negligible in this study.

141

142 *2.2   Data*

143 **2.2.1   Observed data**

144 Observed daily data on precipitation (P), potential evapotranspiration (PET) and the mean

145 altitudes (h) of the 26 sub-basins have been provided by the German Federal Institute of

146 Hydrology (BfG) in Koblenz, Germany (**Table 1**). PET is estimated using the Penman-Wendling

147 equation (ATV-DVWK, 2002) and both variables have been spatially averaged by BfG over 26

148 Moselle sub-basins using areal weights. Observed data from 12 meteorological stations in the

149 Moselle basin (as part of 49 stations over the Rhine basin), mainly provided by the CHR, the

150 DWD, Metéo France, are used to estimate the basin averaged input data (Görgen et al., 2010).

151   Observed daily discharge (Q) data at Cochem (station *#6336050*) are provided by the Global

152   Runoff Data Centre (GRDC), Koblenz. The daily observed data (P, PET and Q) are available for

153   the period 1951-2006.

154

155   **Table 1** Overview of observed data used

156

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

157

158

159

2.2.2 Ensemble seasonal meteorological forecast data

160

161 The ensemble seasonal meteorological forecast data, comprising 40 members, are obtained from

162 the European Centre for Medium-Range Weather Forecasts (ECMWF) seasonal forecasting

163 archive and retrieval system, i.e. MARS system 3 (ECMWF, 2012). This dataset contains regular

164 0.25 x 0.25 degree latitude-longitude grids and each ensemble member is computed for a lead

165 time of 184 days using perturbed initial conditions and model physics (**Table 2**). We estimated

166 the PET forecasts using the Penman-Wendling equation requiring forecasted surface solar

167 radiation and temperature at 2 meter above the surface, and the altitude of the sub-basin (ATV-

168 DVWK, 2002). The PET estimation is consistent with the observed PET estimation carried out

169 by BfG (ATV-DVWK, 2002). The grid-based P and PET ensemble forecast data are firstly

170 interpolated over 26 Moselle sub-basins using areal weights. These sub-basin averaged data are

171 then aggregated to the Moselle basin level.

172
173 **Table 2** Overview of ensemble seasonal meteorological forecast data
174

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

175
176
177

178   **3   METHODOLOGY**

179   *3.1   Overview of model structures and forecast scheme*

180   The three hydrological models (GR4J, HBV and ANN-E) are briefly described in sections 3.1.1-

181   3.1.3. Figure 1 shows the simplified model structures. The calibration and validation of the

182   models is described in section 3.1.4. Five cases with different combinations of ensemble

183   meteorological forecast input and climate mean input are introduced in section 3.1.5.

184
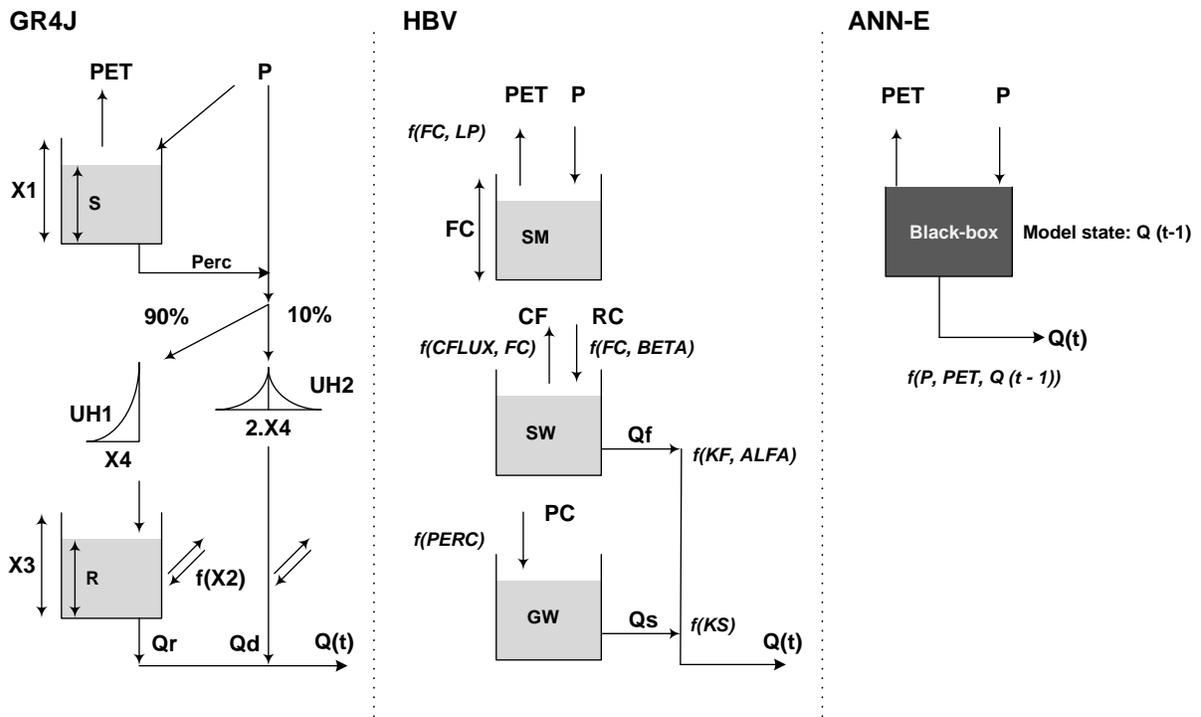
185

186

187   Figure 1 Schematisation of the three models. PET is potential evapotranspiration, P is

188   precipitation and Q is discharge and *t* is the time (day).

189

190

191 **3.1.1   GR4J**

192 The GR4J model (Génie Rural à 4 paramètres Journalier) is used as it has a parsimonious

193 structure with only four parameters as detailed in **Table 7**. The model has been tested over

194 hundreds of basins worldwide, with a broad range of climatic conditions from tropical to

195 temperate and semi-arid basins (Perrin et al., 2003). GR4J is a conceptual model and the required

196 model inputs are daily time series of P and PET (**Table 3**). All four parameters (Figure 1a) are

197 used to calibrate the model. The upper and lower limits of the parameters are selected based on

198 previous works (Perrin et al., 2003;Pushpalatha et al., 2011;Tian et al., 2014).

199 **3.1.2   HBV**

200 The HBV conceptual model (Hydrologiska Byråns Vattenbalansavdelning) was developed by the

201 Swedish Meteorological and Hydrological Institute (SMHI) in the early 1970´s (Lindström et al.,

202 1997). The HBV model consists of four subroutines: a precipitation and snow accumulation and

203 melt routine, a soil moisture accounting routine and two runoff generation routines. The required

204 input data are daily P and PET. The snow routine and daily temperature data are not used in this

205 study as the Moselle basin is a rain-fed basin. Eight parameters (see Figure 1b) in the HBV model

206 are calibrated (Engeland et al., 2010;Van den Tillaart et al., 2013;Tian et al., 2014). The eight

207 parameters for calibration are given in **Table 7** and the parameter ranges are selected based on

208 previous works (Booij, 2005;Eberle, 2005;Tian et al., 2014).

209 **3.1.3   ANN-E**

210 An Artificial Neural Network (ANN) is a data-driven model inspired by functional units

211 (neurons) of the human brain (Elshorbagy et al., 2010). A neural network is a universal

212   approximator capable of learning the patterns and relation between outputs and inputs from

213   historical data and applying it for extrapolation (Govindaraju and Rao, 2000). A three-layer feed-

214   forward neural network (FNNs) is the most widely preferred model architecture for prediction

215   and forecasting of hydrological variables (Adamowski et al., 2012;Shamseldin, 1997;Kalra et al.,

216   2013). Each of these three layers has an important role in processing the information. The first

217   layer receives the inputs and multiplies them with a weight (adds a bias if necessary) before

218   delivering them to each of the hidden neurons in the next layer (Gaume and Gosset, 1999). The

219   weights determine the strength of the connections. The number of nodes in this layer corresponds

220   to the number of inputs. The second layer, the hidden layer, consists of an activation function

221   (also known as transfer function) which non-linearly maps the input data to output target values.

222   In other words, this layer is the learning element of the network which simulates the relationship

223   between inputs and outputs of the model. The third layer, the output layer, gathers the processed

224   data from the hidden layer and delivers the final output of the network.

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

226   Eq (1).

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

227   where w_i are the weights, b is the bias, and *logsig* is the logarithmic sigmoid activation function.

228   We tested the *tansig* and *logsig* activation functions and the latter was selected for this study as it

229   gave better results for low flows. ANN model structures are determined based on the forecast

230   objective. In this study, we used a conceptual type ANN model structure: ANN-Ensemble (ANN-

231   E) which requires daily P, PET and historical Q as input. Observed discharge on the forecast

232   issue day is used to update the model states (**Table 3**). In other words, the ANN-E model receives

233   *Qobs(t)* as input on the time step *t* when the forecast is issued, and then receives the streamflow

234 forecast of the previous time step as input for lead times larger than 1 day. Further, forecasted Q

235 for time step *t+j* is used as input to forecast Q at *t+j+1*.

236 This is a one day memory which also exists in the conceptual models, i.e. GR4J and HBV (Figure

237 1). The ANN-E is assumed to be comparable with the conceptual models with similar model

238 structures. The determination of the optimal number of hidden neurons in the second layer is an

239 important issue in the development of ANN models. Three common approaches are ad hoc (also

240 known as trial and error), global and stepwise (Kasiviswanathan et al., 2013). We used a global

241 approach (i.e. Genetic Algorithm) to avoid local minima (De Vos and Rientjes, 2008) and tested

242 the performance of the networks with one, two and three hidden neurons corresponding to a

243 number of parameters (i.e. number of weights and biases) of 6, 11 and 16 respectively. Based on

244 the parsimonious principle, testing ANNs only up to three hidden neurons is assumed to be

245 enough as the number of parameters increases exponentially for every additional hidden neuron.

246

247 **Table 3** Model descriptions. PET is potential evapotranspiration, P is precipitation and Q is
248 discharge.
249

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

250

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 $m^3 s^{-1}$, is based on previous work by Demirel et al. (2013b). 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. (2013c). 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.

272

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

273   where n is the total number of days (i.e. $m < n$), and ϵ is 1% of the mean observed discharge to

274   avoid infinity during zero discharge days (see Pushpalatha et al.,(2012)). The hybrid Mean

275   Absolute Error is defined as

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

276

277   The MAE_{low} and $MAE_{inverse}$ were not normalised to calculate MAE_{hybrid} as the different units

278   had no effect on the calibration results. It should be noted that we didn't fully neglect the high

279   and intermediate flows using $MAE_{inverse}$ metric, whereas only low flow periods are considered

280   in MAE_{low}. This is one of the advantages of using the MAE_{hybrid} metric and also avoids

281   redundancy.

282

283   **3.1.5   Model storage update procedure for HBV and GR4J models**

284   The storages in the two conceptual models are updated based on the observed discharge on the

285   forecast issue day. In our previous study (Demirel et al., 2013a), we derived empirical relations

286   between the simulated discharge and the fast runoff for each model to divide the observed

287   discharge between the fast and slow runoff components (Eq. (5) and (6)).

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

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

288   The Qf and Qs in the HBV model, and Qr and Qd in the GR4J model are estimated using the

289   fractions above and the observed discharge value on the forecast issue day. The routing storage

290   (R) in the GR4J model is updated for a given value of the *X3* parameter using Eq.(7). Moreover,

291   the surface water (*SW*) and groundwater (*GW*) storages in the HBV model are updated for given

292   values of *KF*, *ALFA* and *KS* parameters using Eq. (8) and (9).

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

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

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

293   The remaining two storages *S* (in GR4J) and *SM* (in HBV) are updated using the calibrated model

294   run until the forecast issue day (i.e. top-down approach).

295

296

297   **3.1.6   Case description**

298   In this study, three hydrological models are used for the seasonal forecasts. Five ensemble

299   meteorological forecast input cases for ANN-E, GR4J and HBV models are compared: (1)

300   ensemble P and PET forecasts (2) ensemble P forecasts and observed climate mean PET (3)

301   observed climate mean P and ensemble PET forecasts (4) observed climate mean P and PET (5)

302   zero P and ensemble PET forecasts (**Table 4**). P and PET forecasts are joint forecasts in our

303   modelling practice. For example, if the first ensemble member is called from P then the first

304   member from PET is also called to force the hydrological model.

305   Cases 1-4 are the different possible combinations of ensemble and climate mean meteorological

306   forcing. Case 5 is analysed to determine to which extent the precipitation forecast in a very dry

307   year (2003) is important for seasonal low flow forecasts. It should be noted that all available

308   historical data (1951-2006) were used to estimate the climate mean. For example the climate

309   mean for January 1st is estimated by the average of 55 January 1st values in the available period

310   (1951-2006).

311   **Table 4** Details of the five input cases
312

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

313

314

315   *3.2   Forecast Skill Scores*

316   Three probabilistic forecast skill scores (Brier Skill Score, reliability diagram, hit and false alarm

317   rates) and one deterministic forecast skill score (Mean Forecast Score) are used to analyse the

318   results of low flow forecasts with lead times of 1-90 days. Forecasts for each day in the test

319   period (2002-2005) are used to estimate these scores. The Mean Forecast Score focusing on low

320   flows is introduced in this study, whereas the other three scores have been often used in

321   meteorology (WMO, 2012) and flood hydrology (Velázquez et al., 2010;Renner et al.,

322   2009;Thirel et al., 2008). For the three models, i.e. GR4J, HBV and ANN-E, the forecast

323   probability for each forecast day is estimated as the ratio of the number of ensemble members

324   non-exceeding the preselected thresholds (here Q75) and the total number of ensemble members

325   (i.e. 40 members) for that forecast day.

326   **3.2.1   Brier Skill Score (BSS)**

327   The Brier Skill Score (BSS) (Wilks, 1995) is often used in hydrology to evaluate the quality of

328   probabilistic forecasts (Devineni et al., 2008;Hartmann et al., 2002;Jaun and Ahrens,

329   2009;Roulin, 2007;Towler et al., 2013).

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

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

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

331   where F_t refers to the forecast probability, O_t refers to the observed probability (O_t=1 if the

332   observed flow is below the low flow threshold, 0 otherwise), and *n* is the sample size.

333   $BS_{climatology}$ is the BS for the climatology, which is also calculated from Eq. (11) for every year

334   using climatological probabilities. BSS values range from minus infinity to 1 (perfect forecast).

335   Negative values indicate that the forecast is less accurate than the climatology and positive values

336   indicate more skill compared to the climatology.

337   3.2.2   Reliability Diagram

338   The reliability diagram is used to evaluate the performance of probabilistic forecasts of selected

339   events, i.e. low flows. A reliability diagram represents the observed relative frequency as a

340   function of forecasted probability and the 1:1 diagonal shows the perfect reliability line

341   (Velázquez et al., 2010;Olsson and Lindström, 2008). This comparison is important as reliability

342   is one of the three properties of a hydrological forecast (WMO, 2012). A reliability diagram

343   shows the portion of observed data inside preselected forecast intervals.

344   In this study, exceedence probabilities of 50%, 75%, 85%, 95%, and 99% are chosen as

345   thresholds to categorize the discharges from mean flows to extreme low flows. The forecasted

346   probabilities are then divided into bins of probability categories; here, five bins (categories) are

347   chosen 0-20%, 20%-40%, 40%-60%, 60%-80% and 80%-100%. The observed frequency for

348   each day is chosen to be 1 if the observed discharge is below the low flow threshold, or 0, if not.

349   3.2.3   Hit and False Alarm Rates

350   We used hit and false alarm rates to assess the effect of ensembles on low flow forecasts for

351   varying lead times. The hit and false alarm rates indicate respectively the proportion of events for

352   which a correct warning was issued, and the proportion of non events for which a false warning

353   was issued by the forecast model. These two simple rates can be easily calculated from

354   contingency tables (**Table 5**) using Eq. (12) and (13). These scores are often used for evaluating

355   flood forecasts (Martina et al., 2006), however, they can also be used to estimate the utility of low

356   flow forecasts as they indicate the models' ability to correctly forecast the occurrence or non-

357   occurrence of preselected events (i.e. Q_{75} low flows). There are four cases in a contingency table

358   as shown in **Table 5**.

359
360   **Table 5** Contingency table for the assessment of low-flow events based on the Q75
361

	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

362
363

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

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

364   **3.2.4   Mean Forecast Score (MFS)**

365   The Mean Forecast Score (MFS) is a new skill score which can be derived from either

366   probabilistic or deterministic forecasts. The probabilities are calculated for the days when low

367   flow occurred. **Table 6** shows the low flow contingency table for calculating MFS. In this study

368   we used a deterministic approach for calculating the observed frequency for all three models. For

369   all three models, ensembles are used for estimating forecast probabilities.

370

371

372   **Table 6** Low flow contingency table for the assessment of forecasts

373

		Observed low flows	
		Deterministic	**Probabilistic**
Forecasted low flows	**Deterministic**	$O_j=1$ (Low flow observed) $F_j=1$ (Low flow forecasted if more than half of the ensemble members indicate low flows) otherwise 0	$O_j=$ Observed frequency based on long term climate (e.g. 34/50 years indicates low flow for day j) $F_j=1$ or 0
	Probabilistic	$O_j = 1$ $F_j=$Forecast frequency based on 40 ensemble members (e.g. 23/40 members indicate low flows for day j)	$O_j=$ Observed frequency based on long term climate $F_j=$Forecast frequency based on 40 ensemble members

374

375   The score is calculated as below only for deterministic observed low flows (left column in **Table**

376   **6**).

377

378

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

379   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

380   total number of low flow days. The probability of a deterministic forecast can be 0 or 1, whereas

381   it varies from 0 to 1 for ensemble members. For instance, if 23 of the 40 ensemble forecast

382   members indicate low flows for the *j*-th low flow day then $F_j = 23/40$. It should be noted that this

383   score is not limited to low flows as it has a flexible forecast probability definition which can be

384   adapted to any type of discharges. MFS values range from zero to 1 (perfect forecast).

385

386 **4   RESULTS**

387 *4.1   Calibration and validation*

388 **Table 7** shows the parameter ranges and the best performing parameter sets of the three models.

389 The GR4J and HBV models have both well-defined model structures; therefore, their calibration

390 was more straightforward than the calibration of the ANN models. Calibration of the ANN-E

391 model was done in two steps. First, the number of hidden neurons was determined by testing the

392 performance of the ANN-E model with one, two and three hidden neurons.

393

394 **Table 7** 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 |

395

396   Second, daily P, PET and Q are used as three inputs for the tested ANN-E model with one, two

397   and three hidden neurons due to the fact that these inputs are comparable with the inputs of the

398   GR4J and HBV models. Figure 2a shows that the performance of the ANN-E model does not

399   improve with additional hidden neurons. Based on the performance in the validation period, one

400   hidden neuron is selected. GR4J and HBV are also calibrated. The results of the three models

401   used in this study are presented in Figure 2b.

402   The performances of GR4J and HBV are similar in the calibration period, whereas HBV

403   performs better in the validation period (Figure 2b). This is not surprising, since HBV has a more

404   sophisticated model structure than GR4J.

405
406
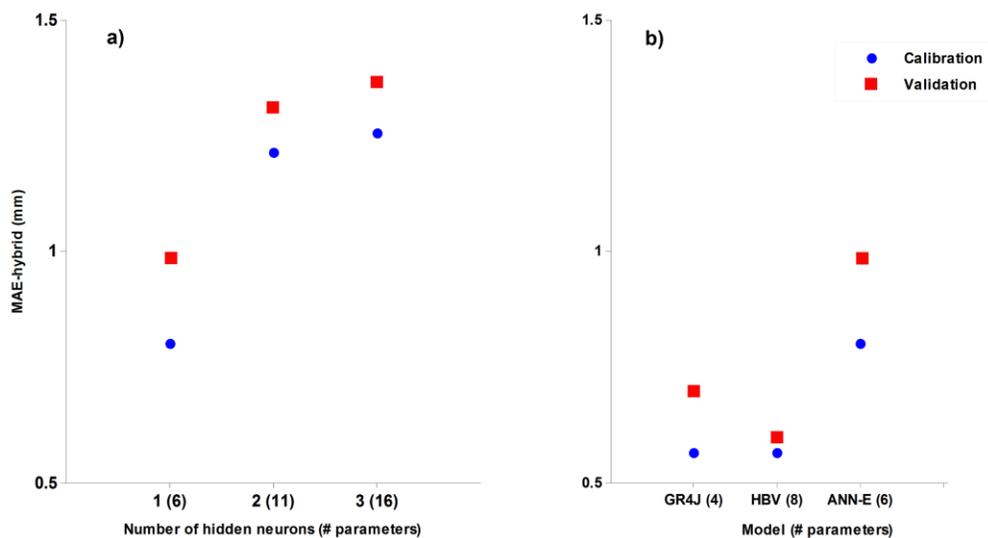
407
408
409   Figure 2 Calibration and validation results of **a)** the ANN-E model with one, two and three

410   hidden neurons and **b)** the three models used in this study. The same calibration (1971-2001) and

411   validation (1951-1970) periods are used for both plots.

412   ***4.2   Effect of ensembles on low flow forecasts for 90 day lead time***

413   The effect of ensemble P and PET on GR4J, HBV and ANN-E is presented as a range bounded

414   by the lowest and highest forecast values in Figure 3a and b. The two years, i.e. 2002 and 2003,

415   are carefully selected as they represent a relatively wet year and a very dry year respectively.

416   Figure 3a shows that there are significant differences between the three model results. The 90 day

417   ahead low flows in 2002 are mostly over-predicted by the ANN-E model, whereas GR4J and

418   HBV over-predict low flows observed after August. The over-prediction of low flows is more

419   pronounced for GR4J than for the other three models. The over-prediction of low flows by ANN-

420   E is mostly at the same level. This less sensitive behaviour of ANN-E to the forecasted ensemble

421   inputs shows the effect of the logarithmic sigmoid transfer function on the results. Due to the

422   nature of this algorithm, input is rescaled to a small interval [0, 1] and the gradient of the sigmoid

423   function at large values approximates zero (Wang et al., 2006). Further, ANN-E is also not

424   sensitive to the initial model conditions updated on every forecast issue day. The less pronounced

425   over-prediction of low flows by HBV compared to GR4J may indicate that the slow responding

426   groundwater storage in HBV is less sensitive to different forecasted ensemble P and PET inputs

427   (Demirel et al., 2013c).

428

a)   b)

GR4J

HBV

ANN-E
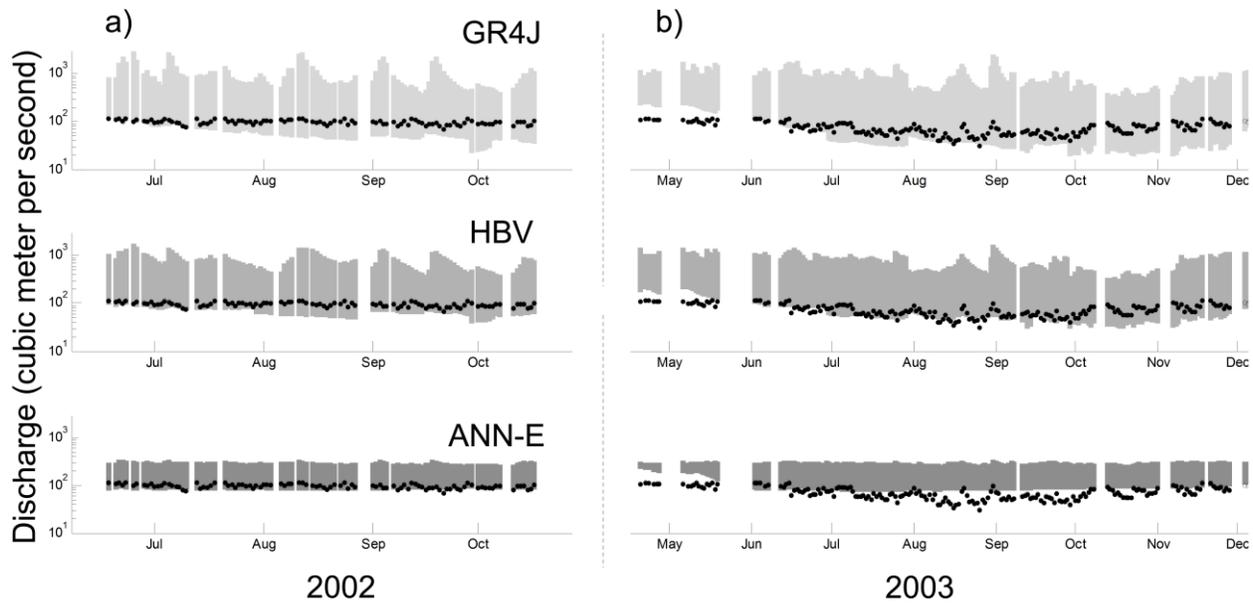
429   2002   2003
430
431   Figure 3 Range (shown as grey shade) of low flow forecasts in **a)** 2002 (the wettest year of the
432   test period with 101 low flow days) **b)** 2003 (the driest year of the test period with 192 low flow
433   days) for a lead time of 90 days using ensemble P and PET as input for GR4J, HBV and ANN-E
434   models (case 1 − 2002 and 2003). The gaps in the figures indicate non-low flow days (i.e.
435   censored).
436
437   The results for 2003 are slightly different than those for 2002. As can be seen from Figure 3b the

438   number of low flow days has increased in the dry year, i.e. 2003, and the low flows between

439   August and November are not captured by any of the 40-ensemble forecasts using ANN-E. The

440   most striking result in Figure 3b is that the low flows observed in the period between April and

441   May are not captured by any of the three models, i.e. GR4J, HBV and ANN-E. The poor

442   performance of the models during the spring period can be explained by the high precipitation

443   amount in this period. The poor simulation of high flows in the preceding winter months can have

444   an effect on the forecasts too. The 90 day low flows between October and November are better

445   forecasted by GR4J and HBV than the ANN-E model. The two hydrological models used in this

446   study have well defined surface and ground water components. Therefore, they react to the

447   weather inputs in a physically meaningful way. However, in black box models, the step functions

448   (transfer functions or activation functions) may affect the model behaviour. The ANN model will

449   then react to a certain range of inputs based on the objective function. This feature of ANN is the

450   main reason for the erratic behaviour in Figure 4b and the small (and uniform) uncertainty range

451   in the figures (e.g. Figure 3).

452

453   For the purpose of determining to which extent ensemble P and PET inputs and different initial

454   conditions affect 90 day low flow forecasts, we run the models with different input combinations

455   such as ensemble P or PET and climate mean P or PET and zero precipitation. Figure 4a shows

456   the forecasts using ensemble P and climate mean PET as input for three models. The picture is

457   very similar to Figure 3b as most of the observed low flows fall within the constructed forecast

458   range by GR4J and HBV. The forecasts issued by GR4J are better than those issued by the other

459   two models. However, the range of forecasts using GR4J is larger than for the other models

460   showing the sensitivity of the model for different precipitation inputs. It is obvious that most of

461   the range in all forecasts is caused by uncertainties originating from ensemble precipitation input.

462
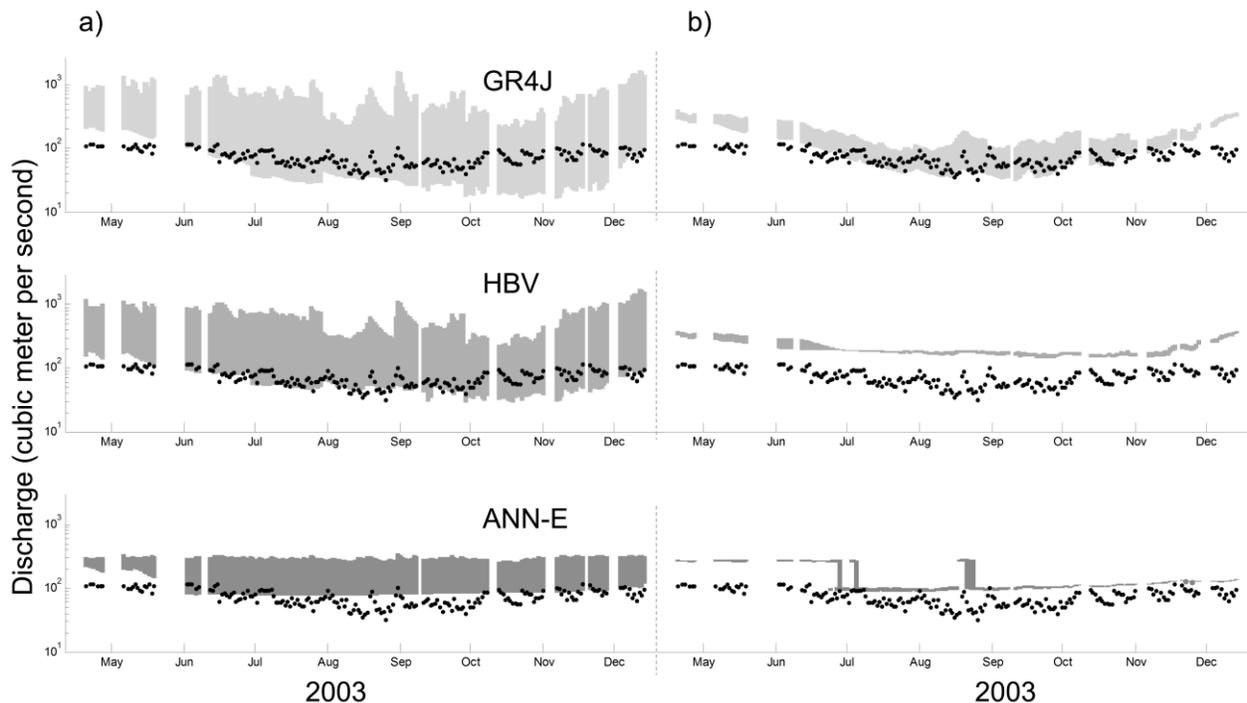
463

464 Figure 4 Range (shown as grey shade) of low flow forecasts in 2003 for a lead time of 90 days
465 using **a)** ensemble P and climate mean PET (case 2) **b)** climate mean P and ensemble PET as
466 input for GR4J, HBV and ANN-E models (case 3). The gaps in the figures indicate non-low flow
467 days (i.e. censored).
468
469 Figure 4b shows the forecasts using climate mean P and ensemble PET as input for three models,

470 i.e. GR4J, HBV and ANN-E. Interestingly, only GR4J could capture the 90 day low flows

471 between July and November using climate mean P and ensemble PET showing the ability of the

472 model to handle the excessive rainfall. None of the low flows were captured by HBV, whereas

473 very few low flow events were captured by ANN-E (Figure 4b). The precipitation information is

474 crucial for the conceptual models to forecast low flows for a lead time of 90 days. The narrow

475 uncertainty band indicates that the effect of the PET ensemble on the forecasts is less pronounced

476 as compared to the effect of the P ensemble.

477 Figure 5a shows the forecasts using climate mean P and PET as input for three models. The

478 results are presented by point values without a range since only one deterministic forecast is

479 issued. There are significant differences in the results of the three models. For instance, all 90 day

480 ahead low flows in 2003 are over-predicted by HBV, whereas the over-prediction of low flows is

481 less pronounced for ANN-E. It is remarkable that GR4J can forecast a very dry year accurately

482 using the climate mean. The low values of the calibrated maximum soil moisture capacity and

483 percolation parameters of HBV (*FC* and *PERC*) can be the main reason for over-prediction of all

484 low flows as the interactions of parameters with climate mean P input can result in higher model

485 outputs.

486
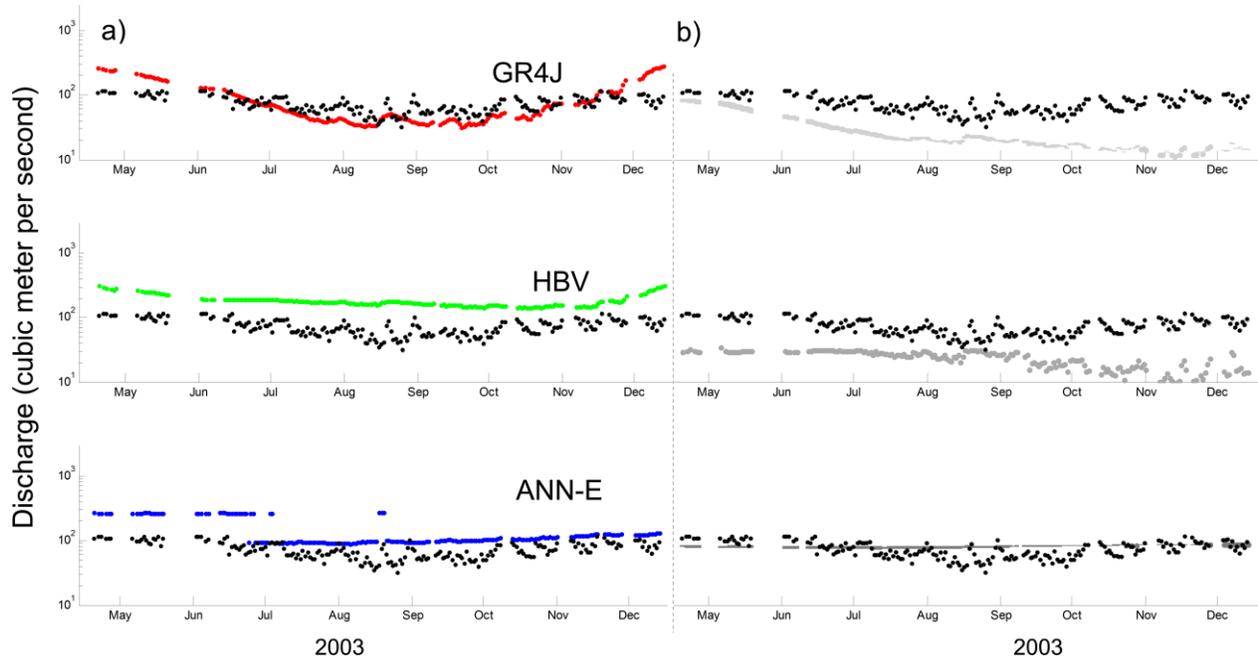
Figure 5 Low flow forecasts in 2003 for a lead time of 90 days using **a)** both climate mean P and PET (case 4) and **b)** zero P and ensemble PET (case 5) as input for GR4J, HBV and ANN-E models. The gaps in the figures indicate non-low flow days (i.e. censored).

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.

504
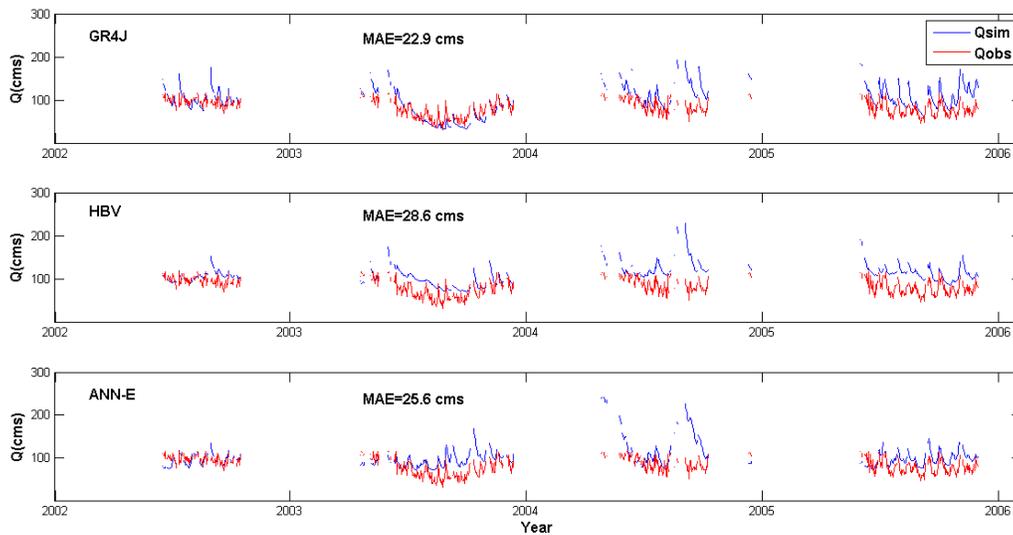
Figure 6 Benchmark reference forecasts using three models (GR4J, HBV and ANN-E) using observed P and PET (i.e. perfect forecasts)

507
508
509 We also show the minimum and maximum prediction errors for each case in **Table 8**. There are

510 large differences in case 1 and 2 as compared to the other cases. It is also obvious that the

511 uncertainty range is larger in case 1 than in case 2 for the conceptual models. This is also what

512 we see in Figure 3 and Figure 4 above.

513

514 **Table 8** Minimum and maximum prediction errors for low flow forecasts for a lead time of 90
515 days during the test period 2002-2005
516

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]

517

518 *4.3   Effect of ensembles on low flow forecast skill scores*

519 Figure 7 compares the three models and the effect of ensemble P and PET on the skill of

520 probabilistic low flow forecasts with varying lead times. In this figure, four different skill scores

521   are used to present the results of probabilistic low flow forecasts issued by GR4J, HBV and

522   ANN-E. From an operational point of view, the main purpose of investigating the effect of

523   ensembles and model initial conditions on ensemble low flow forecasts with varying lead times is

524   to improve the forecast skills (e.g. hit rate, reliability, BSS and MFS) and to reduce false alarms

525   and misses. From Figure 7 we can clearly see that the results of GR4J show the lowest BSS, MFS

526   and hit rate. The false alarm rate of forecasts using GR4J is also the lowest compared to those

527   using other models. The decrease in false alarm rates after a lead time of 20 days shows the

528   importance of initial condition uncertainty for short lead time forecasts. The limit is around 20

529   days for ANN-E and shorter for the other two models. When the forecast is issued on day (t), the

530   model states are updated using the observed discharge on that day (t). For GR4J and HBV we

531   used the deterministic state update procedure described in section 3.1.5. However, the models

532   probably spin-up after some days and improve the results for false alarm rate are improved. For

533   longer lead times the error is better handled by the models. We further analysed the forecasted

534   meteorological forcing data (P and PET) to see if there is any difference between the short lead

535   time (~20 days) and long lead time (e.g. 90 days). This is done for three different lead times for

536   each model when the false alarm rate was highest (i.e. 12, 15 and 21 days based on the false

537   alarm rates of GR4J, HBV and ANN-E respectively.). We compared the boxplots from these

538   problematic lead times with the 90 day lead time (not shown here but available in the review

539   reports). It is interesting to note that the ranges for P and PET are larger at 90 day lead time as

540   compared to shorter lead times. However, the observed P and PET values (i.e. perfect forecasts)

541   are covered by the large ranges resulting in higher hit rates (i.e. lower false alarm rates). In other

542   words, for short lead times, 12, 15 and 21 days in particular, the ranges for P and PET are smaller

543   than those for the 90 day lead time but the observed P and PET values are usually missed causing

544   ==higher false alarm rates in the results.== It appears from the results that ANN-E and HBV show a

545   comparable skill in forecasting low flows up to a lead time of 90 days.

546

547
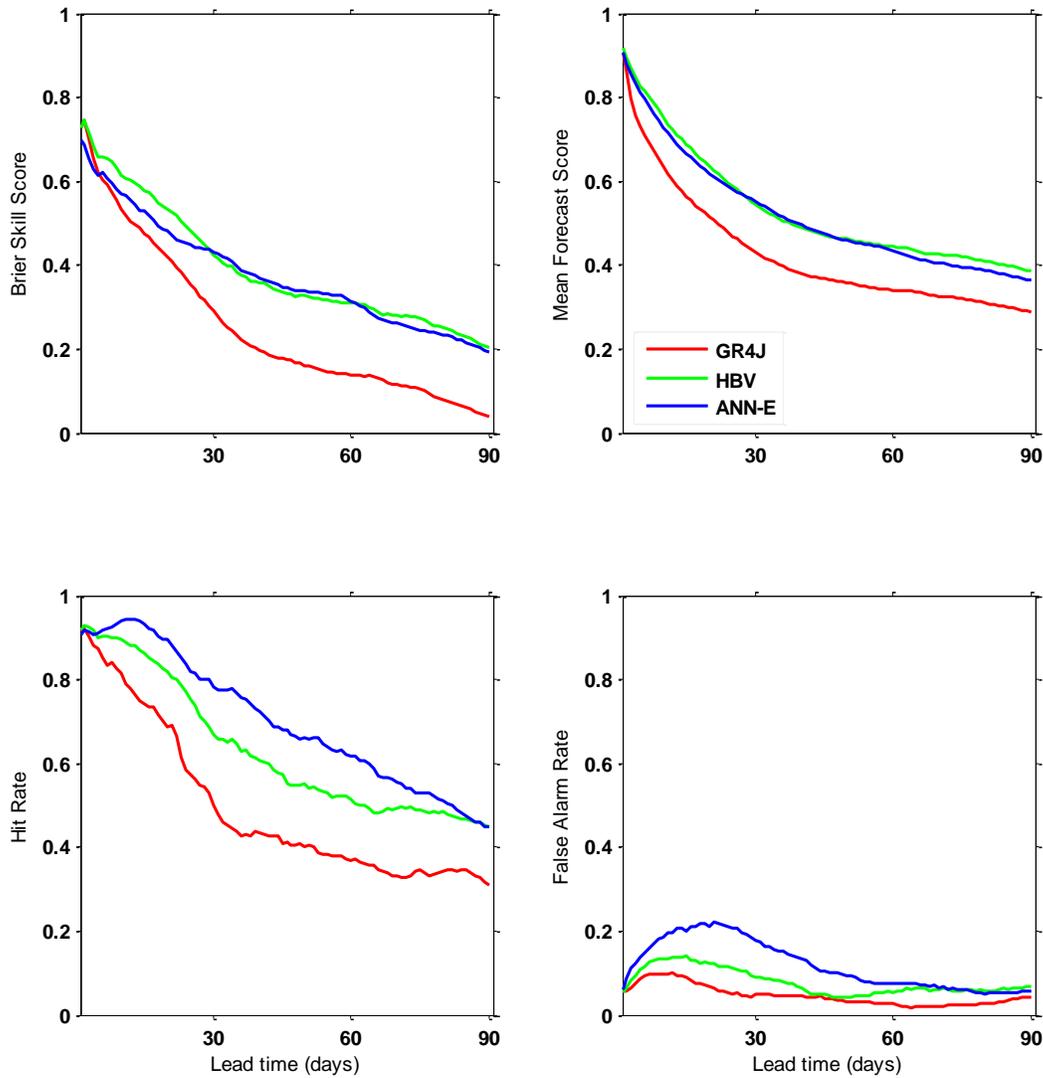
548
549
550   ==Figure 7 Skill scores for forecasting low flows at different lead times for three different==
551   ==hydrological models for the test period 2002-2005. Note that all forecasts (including high and low==
552   ==flow time steps) are used to estimate these skill scores.==
553

554   Figure 8 compares the reliability of probabilistic 90 day low flows forecasts below different

555   thresholds (i.e. Q75, Q90 and Q95) using ensemble P and PET as input for three models. The

556   figure shows that the Q75 and Q90 low flow forecasts issued by the HBV model are more

557   reliable compared to the other models. Moreover, all three models under-predict most of the

558   forecast intervals. It appears from Figure 8c that very critical low flows (i.e. Q99) are under-

559   predicted by the GR4J model.

560
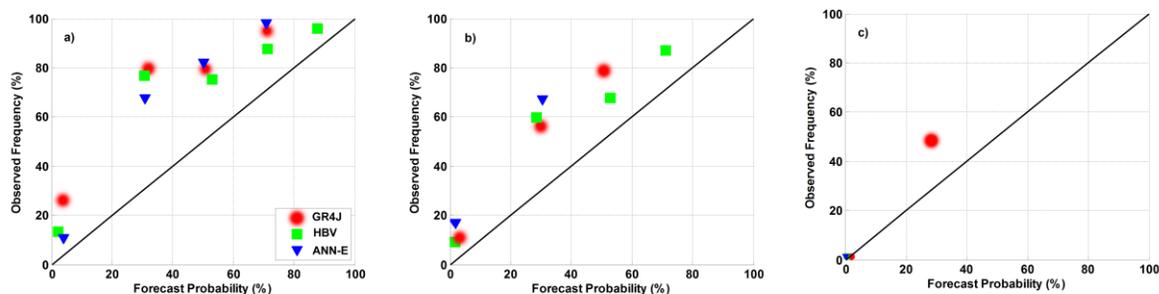
561
562
563   Figure 8 Reliability diagram for different low flow forecasts **a)** Low flows below Q75 threshold
564   (584 observed events in the test period 2002-2005) **b)** Low flows below Q90 threshold (250
565   observed events) **c)** Low flows below Q99 threshold (20 observed events). The forecasts are
566   issued for a lead time of 90 days for the test period 2002-2005 using ensemble P and PET as
567   input for GR4J, HBV and ANN-E models.
568

5   DISCUSSION

570   To compare data-driven and conceptual modelling approaches and to evaluate the effects of

571   seasonal meteorological forecasts on low flow forecasts, 40-member ensembles of ECMWF

572   seasonal meteorological forecasts were used as input for three low flow forecast models.

573   Different input combinations were compared to distinguish between the effects of ensemble P

574   and PET and model initial conditions on 90 day low flow forecasts. The models could reasonably

575   forecast low flows when ensemble P was introduced into the models. This result is in line with

576   that of Shukla and Lettenmaier (2011) who found that seasonal meteorological forecasts have a

577   greater influence than initial model conditions on the seasonal hydrological forecast skills. Two

578   other related studies also showed that the effect of a large spread in ensemble seasonal

579   meteorological forecasts is larger than the effect of initial conditions on hydrological forecasts

580   with lead times longer than 1-2 months (Li et al., 2009;Yossef et al., 2013). The encouraging

581   results of low flow forecasts using ensemble seasonal precipitation forecasts for the hydrological

582   models confirm the utility of seasonal meteorological forcing for low flow forecasts. Shukla et al.

583   (2012) also found useful forecast skills for both runoff and soil moisture forecasting at seasonal

584   lead times using the medium range weather forecasts.

585   In this study, we also assessed the effects of ensemble P and PET on the skill scores of low flow

586   forecasts with varying lead times up to 90 days. In general, the four skill scores show similar

587   results. Not surprisingly, all models under-predicted low flows without precipitation information

588   (zero P). The most evident two patterns in these scores are that first, the forecast skill drops

589   sharply until a lead time of 30 days and second, the skill of probabilistic low flow forecasts

590   issued by GR4J is the lowest, whereas the skill of forecasts issued by ANN-E is the highest

591   compared to the other two models. Further, our study showed that data-driven models can be

592   good alternatives to conceptual models for issuing seasonal low flow forecasts (e.g. Figure 6).

593   The two hydrological models used in this study have well defined surface and ground water

594   components. Therefore, they react to the weather inputs in a physically meaningful way.

595   However, in black box models, the step functions (transfer functions or activation functions) may

596   limit model sensitivity after the training. The ANN model will then react to a certain range of

597   inputs based on the objective function. This feature of an ANN is the main reason for the small

598   (and uniform) uncertainty range in the figures. The over prediction of the models is closely

599   related to the over prediction of the P by the ensembles. Low flows are usually over predicted by

600   the models for the entire period. However, there are under-predictions of low flows for some days

601   in November-December as well. Before June, none of the low flows are captured by the ensemble

602   members. The best performing period is the fall and the worst performing period is the spring

603   period for the models. The poor performance of the models during the spring period can be

604   explained by the high precipitation amount in this period. Since the first part of the objective

605   function used in this study solely focuses on low flows, the high flow period is less important in

606   the calibration. The low flows occurring in the spring period are, therefore, missed in the

607   forecasts. The simulation of snow cover during winter and snow melt during the spring can both

608   have effects on the forecasts too.

609

610

611   **6   CONCLUSIONS**

612   Three hydrological models have been compared regarding their performance in the calibration,

613   validation and forecast periods, and the effect of seasonal meteorological forecasts on the skill of

614   low flow forecasts has been assessed for varying lead times. The comparison of three different

615   models help us to contrast data-driven and conceptual models in low flow forecasts, whereas

616   running the models with different input combinations, e.g. climate mean precipitation and

617   ensemble potential evapotranspiration, help us to identify which input source led to the largest

618   range in the forecasts. A new hybrid low flow objective function, comprising the mean absolute

619   error of low flows and the mean absolute error of inverse discharges, is used for comparing low

620   flow simulations, whereas the skill of the probabilistic seasonal low flow forecasts has been

621   evaluated based on the ensemble forecast range, Brier Skill Score, reliability, hit/false alarm rates

622   and Mean Forecast Score. The latter skill score (MFS) focusing on low flows is firstly introduced

623   in this study. In general our results showed that;

624   •   Based on the results of the calibration and validation, one hidden neuron in ANN was

625   found to be enough for seasonal forecasts as additional hidden neurons did not increase

626   the simulation performance. The difference between calibration and validation

627   performances was smallest for the HBV model, i.e. the most sophisticated model used in

628   this study.

629   •   Based on the results of the comparison of different model inputs for two years (i.e. 2002

630   and 2003), the largest range for 90 day low flow forecasts is found for the GR4J model

631   when using ensemble seasonal meteorological forecasts as input. Moreover, the

632   uncertainty arising from ensemble precipitation has a larger effect on seasonal low flow

633   forecasts than the effects of ensemble potential evapotranspiration. All models are prone

634   to over-predict low flows using ensemble seasonal meteorological forecasts. However, the

635   precipitation forecasts in the forecast period are crucial for improving the low flow

636   forecasts. As expected, all three models, i.e. GR4J, HBV and ANN-E under-predicted 90

637   day ahead low flows in 2003 without rainfall data.

638   •   Based on the results of the comparison of forecast skills with varying lead times, the false

639   alarm rate of GR4J is the lowest indicating the ability of the model of forecasting non-

640   occurrence of low flow days. The low flow forecasts issued by HBV are more reliable

641   compared to the other models. The hit rate of ANN-E is higher than that of the two

642   conceptual models used in this study. Overall, the ANN-E and HBV models are the best

643   performing two of the three models using ensemble P and PET.

644

645   Further work should examine the effect of model parameters and initial conditions on the

646   seasonal low flow forecasts as the values of the maximum soil moisture and percolation related

647   parameters of conceptual models can result in over- or under-prediction of low flows. It is

648   noteworthy to mention that the data-driven model developed in this study, i.e. ANN-E, can be

649   applied to other large river basins elsewhere in the world. Surprisingly, ANN-E and HBV showed

650   a similar skill for seasonal forecasts, where a priori we expected that the two conceptual models,

651   GR4J and HBV, would show similar results up to a lead time of 90 days.

652

653   **ACKNOWLEDGEMENTS**

669

670

671   **REFERENCES**

672   Adamowski, J., Chan, H. F., Prasher, S. O., Ozga-Zielinski, B., and Sliusarieva, A.: Comparison
673   of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial
674   neural network, and wavelet artificial neural network methods for urban water demand
675   forecasting in Montreal, Canada, Water Resour. Res., 48, W01528, 10.1029/2010wr009945,
676   2012.
677   Archer, D. R., and Fowler, H. J.: Using meteorological data to forecast seasonal runoff on the
678   River Jhelum, Pakistan, J. Hydrol., 361, 10-23, 10.1016/j.jhydrol.2008.07.017, 2008.
679   ATV-DVWK: Verdunstung in Bezug zu Landnutzung, Bewuchs und Boden, Merkblatt ATV-
680   DVWK-M 504, Hennef, 2002.
681   Bell, V. A., Davies, H. N., Kay, A. L., Marsh, T. J., Brookshaw, A., and Jenkins, A.: Developing
682   a large-scale water-balance approach to seasonal forecasting: application to the 2012 drought in
683   Britain, Hydrol. Processes, 10.1002/hyp.9863, 2013.
684   Bierkens, M. F. P., and van Beek, L. P. H.: Seasonal Predictability of European Discharge: NAO
685   and Hydrological Response Time, J. Hydrometeorol., 10, 953-968, Doi 10.1175/2009jhm1034.1,
686   2009.
687   Booij, M. J.: Impact of climate change on river flooding assessed with different spatial model
688   resolutions, J. Hydrol., 303, 176-198, DOI 10.1016/j.jhydrol.2004.07.013, 2005.
689   Chiew, F. H. S., Zhou, S. L., and McMahon, T. A.: Use of seasonal streamflow forecasts in water
690   resources management, J. Hydrol., 270, 135-144, 2003.
691   Chowdhury, S., and Sharma, A.: Multisite seasonal forecast of arid river flows using a dynamic
692   model combination approach, Water Resour. Res., 45, W10428, 10.1029/2008wr007510, 2009.
693   Coley, D. M., and Waylen, P. R.: Forecasting dry season streamflow on the Peace River at
694   Arcadia, Florida, USA, J. Am. Water Resour. Assoc., 42, 851-862, 2006.
695   De Vos, N. J., and Rientjes, T. H. M.: Multiobjective training of artificial neural networks for
696   rainfall-runoff modeling, Water Resour. Res., 44, W08434, 10.1029/2007wr006734, 2008.
697   Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: Effect of different uncertainty sources on the
698   skill of 10 day ensemble low flow forecasts for two hydrological models, Water Resour. Res., 49,
699   4035-4053, 10.1002/wrcr.20294, 2013a.
700   Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: Identification of appropriate lags and temporal
701   resolutions for low flow indicators in the River Rhine to forecast low flows with different lead
702   times, Hydrol. Processes, 27, 2742-2758, 10.1002/hyp.9402, 2013b.
703   Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: Effect of different uncertainty sources on the
704   skill of 10 day ensemble low flow forecasts for two hydrological models, Water Resour. Res., 49,
705   10.1002/wrcr.20294, 2013c.
706   Devineni, N., Sankarasubramanian, A., and Ghosh, S.: Multimodel ensembles of streamflow
707   forecasts: Role of predictor state in developing optimal combinations, Water Resour. Res., 44,
708   W09404, 10.1029/2006wr005855, 2008.
709   Doblas-Reyes, F. J., Weisheimer, A., Déqué, M., Keenlyside, N., McVean, M., Murphy, J. M.,
710   Rogel, P., Smith, D., and Palmer, T. N.: Addressing model uncertainty in seasonal and annual
711   dynamical ensemble forecasts, Q. J. R. Meteorol. Soc., 135, 1538-1559, 10.1002/qj.464, 2009.
712   Dutra, E., Di Giuseppe, F., Wetterhall, F., and Pappenberger, F.: Seasonal forecasts of droughts
713   in African basins using the Standardized Precipitation Index, Hydrol. Earth Syst. Sci., 17, 2359-
714   2373, 10.5194/hess-17-2359-2013, 2013.

715   Dutra, E., Pozzi, W., Wetterhall, F., Di Giuseppe, F., Magnusson, L., Naumann, G., Barbosa, P.,
716   Vogt, J., and Pappenberger, F.: Global meteorological drought – Part 2: Seasonal forecasts,
717   Hydrol. Earth Syst. Sci., 18, 2669-2678, 10.5194/hess-18-2669-2014, 2014.
718   Eberle, M.: Hydrological Modelling in the River Rhine Basin Part III - Daily HBV Model for the
719   Rhine Basin BfG-1451, Institute for Inland Water Management and Waste Water Treatment
720   (RIZA) and Federal Institute of Hydrology (BfG) Koblenz, Germany 2005.
721   ECMWF: Describing ECMWF's forecasts and forecasting system, ECMWF newsletter 133,
722   Available from: http://old.ecmwf.int/publications/manuals/mars/ (last access: 26/07/2014), 2012.
723   Elshorbagy, A., Corzo, G., Srinivasulu, S., and Solomatine, D. P.: Experimental investigation of
724   the predictive capabilities of data driven modeling techniques in hydrology - Part 1: Concepts and
725   methodology, Hydrol. Earth Syst. Sci., 14, 1931-1941, 10.5194/hess-14-1931-2010, 2010.
726   Engeland, K., Renard, B., Steinsland, I., and Kolberg, S.: Evaluation of statistical models for
727   forecast errors from the HBV model, J. Hydrol., 384, 142-155, 2010.
728   EU: Horizon 2020 – Work Programme 2014-2015: Water 7_2015: Increasing confidence in
729   seasonal-to-decadal   predictions   of   the   water   cycle.
730   http://www.aber.ac.uk/en/media/departmental/researchoffice/funding/UKRO-Horizon-
731   2020_climate_change_draft_wp.pdf, (Accessed September 4, 2013).
732   , 2013.
733   Felipe, P.-S., and Nelson, O.-N.: Forecasting of Monthly Streamflows Based on Artificial Neural
734   Networks, J. Hydrol. Eng., 14, 1390-1395, 2009.
735   Fundel, F., Jörg-Hess, S., and Zappa, M.: Monthly hydrometeorological ensemble prediction of
736   streamflow droughts and corresponding drought indices, Hydrol. Earth Syst. Sci., 17, 395-407,
737   10.5194/hess-17-395-2013, 2013.
738   Ganguli, P., and Reddy, M. J.: Ensemble prediction of regional droughts using climate inputs and
739   SVM-copula approach, Hydrol. Processes, (accepted), 10.1002/hyp.9966, 2013.
740   Gaume, E., and Gosset, R.: Over-parameterisation, a major obstacle to the use of artificial neural
741   networks in hydrology?, Hydrol. Earth Syst. Sci., 7, 693-706, 10.5194/hess-7-693-2003, 1999.
742   Giuntoli, I., Renard, B., Vidal, J. P., and Bard, A.: Low flows in France and their relationship to
743   large-scale climate indices, J. Hydrol., 482, 105-118, 10.1016/j.jhydrol.2012.12.038, 2013.
744   Gobena, A. K., and Gan, T. Y.: Incorporation of seasonal climate forecasts in the ensemble
745   streamflow prediction system, J. Hydrol., 385, 336-352, 10.1016/j.jhydrol.2010.03.002, 2010.
746   Görgen, K., Beersma, J., Brahmer, G., Buiteveld, H., Carambia, M., de Keizer, O., Krahe, P.,
747   Nilson, E., Lammersen, R., Perrin, C., and Volken, D.: Assessment of Climate Change Impacts
748   on Discharge in the Rhine River Basin: Results of the RheinBlick 2050 Project, Lelystad, CHR,
749   ISBN   978-90-70980-35-1,   211p.   Available   from:
750   http://www.news.admin.ch/NSBSubscriber/message/attachments/20770.pdf   (last   access:
751   30/10/2014), 2010.
752   Govindaraju, R. S., and Rao, A. R.: Artificial Neural Networks in Hydrology, Kluwer Academic
753   Publishers Norwell, MA, USA, 329 pp., 2000.
754   Hartmann, H. C., Pagano, T. C., Sorooshian, S., and Bales, R.: Confidence builders: Evaluating
755   seasonal climate forecasts from user perspectives, Bulletin of the American Meteorological
756   Society, 83, 683-698, 2002.
757   Jaun, S., and Ahrens, B.: Evaluation of a probabilistic hydrometeorological forecast system,
758   Hydrol. Earth Syst. Sci., 13, 1031-1043, 2009.
759   Kahya, E., and Dracup, J. A.: U.S. streamflow patterns in relation to the El Niño/Southern
760   Oscillation, Water Resour. Res., 29, 2491-2503, 10.1029/93wr00744, 1993.

761   Kalra, A., Ahmad, S., and Nayak, A.: Increasing streamflow forecast lead time for snowmelt-
762   driven catchment based on large-scale climate patterns, Adv. Water Resour., 53, 150-162,
763   10.1016/j.advwatres.2012.11.003, 2013.
764   Kasiviswanathan, K. S., Raj, C., Sudheer, K. P., and Chaubey, I.: Constructing prediction interval
765   for artificial neural network rainfall runoff models based on ensemble simulations, J. Hydrol.,
766   499, 275-288, 10.1016/j.jhydrol.2013.06.043, 2013.
767   Kuo, C.-C., Gan, T. Y., and Yu, P.-S.: Seasonal streamflow prediction by a combined climate-
768   hydrologic system for river basins of Taiwan, J. Hydrol., 387, 292-303, 2010.
769   Li, H., Luo, L., and Wood, E. F.: Seasonal hydrologic predictions of low-flow conditions over
770   eastern USA during the 2007 drought, Atmospheric Science Letters, 9, 61-66, 2008.
771   Li, H., Luo, L., Wood, E. F., and Schaake, J.: The role of initial conditions and forcing
772   uncertainties in seasonal hydrologic forecasting, J. Geophys. Res., 114, D04114,
773   10.1029/2008jd010969, 2009.
774   Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergstrom, S.: Development and
775   test of the distributed HBV-96 hydrological model, J. Hydrol., 201, 272-288, 1997.
776   Luo, L., Wood, E. F., and Pan, M.: Bayesian merging of multiple climate model forecasts for
777   seasonal hydrological predictions, J. Geophys. Res., 112, D10102, 10.1029/2006jd007655, 2007.
778   Madadgar, S., and Moradkhani, H.: A Bayesian Framework for Probabilistic Seasonal Drought
779   Forecasting, J. Hydrometeorol., 14, 1685-1705, 10.1175/JHM-D-13-010.1, 2013.
780   Martina, M. L. V., Todini, E., and Libralon, A.: A Bayesian decision approach to rainfall
781   thresholds based flood warning, Hydrol. Earth Syst. Sci., 10, 413-426, 10.5194/hess-10-413-
782   2006, 2006.
783   Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., Le Lay, M.,
784   Besson, F., Soubeyroux, J. M., Viel, C., Regimbeau, F., Andréassian, V., Maugis, P., Augeard,
785   B., and Morice, E.: Benchmarking hydrological models for low-flow simulation and forecasting
786   on French catchments, Hydrol. Earth Syst. Sci. Discuss., 10, 13979-14040, 10.5194/hessd-10-
787   13979-2013, 2013.
788   Olsson, J., and Lindström, G.: Evaluation and calibration of operational hydrological ensemble
789   forecasts in Sweden, J. Hydrol., 350, 14-24, 2008.
790   Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for
791   streamflow simulation, J. Hydrol., 279, 275-289, 2003.
792   Pokhrel, P., Wang, Q. J., and Robertson, D. E.: The value of model averaging and dynamical
793   climate model predictions for improving statistical seasonal streamflow forecasts over Australia,
794   Water Resour. Res., 49, 6671-6687, 10.1002/wrcr.20449, 2013.
795   Pushpalatha, R., Perrin, C., Moine, N. L., Mathevet, T., and Andréassian, V.: A downward
796   structural sensitivity analysis of hydrological models to improve low-flow simulation, J. Hydrol.,
797   411, 66–76, 2011.
798   Pushpalatha, R., Perrin, C., Moine, N. L., and Andréassian, V.: A review of efficiency criteria
799   suitable for evaluating low-flow simulations, J. Hydrol., 420–421, 171-182,
800   10.1016/j.jhydrol.2011.11.055, 2012.
801   Renner, M., Werner, M. G. F., Rademacher, S., and Sprokkereef, E.: Verification of ensemble
802   flow forecasts for the River Rhine, J. Hydrol., 376, 463-475, 2009.
803   Robertson, D. E., Pokhrel, P., and Wang, Q. J.: Improving statistical forecasts of seasonal
804   streamflows using hydrological model output, Hydrol. Earth Syst. Sci., 17, 579-593,
805   10.5194/hess-17-579-2013, 2013.
806   Roulin, E.: Skill and relative economic value of medium-range hydrological ensemble
807   predictions, Hydrol. Earth Syst. Sci., 11, 725-737, 2007.

808   Rutten, M., van de Giesen, N., Baptist, M., Icke, J., and Uijttewaal, W.: Seasonal forecast of
809   cooling water problems in the River Rhine, Hydrol. Processes, 22, 1037-1045, 2008.
810   Saadat, S., Khalili, D., Kamgar-Haghighi, A., and Zand-Parsa, S.: Investigation of spatio-
811   temporal patterns of seasonal streamflow droughts in a semi-arid region, Natural Hazards, 1-24,
812   10.1007/s11069-013-0783-y, 2013.
813   Sauquet, E., Lerat, J., and Prudhomme, C.: La prévision hydro-météorologique à 3-6 mois. Etat
814   des connaissances et applications, La Houille Blanche, 77-84, 2008.
815   Schubert, S., Koster, R., Hoerling, M., Seager, R., Lettenmaier, D., Kumar, A., and Gutzler, D.:
816   Predicting Drought on Seasonal-to-Decadal Time Scales, Bulletin of the American
817   Meteorological Society, 88, 1625-1630, 10.1175/bams-88-10-1625, 2007.
818   Shamseldin, A. Y.: Application of a neural network technique to rainfall-runoff modelling, J.
819   Hydrol., 199, 272-294, 10.1016/s0022-1694(96)03330-6, 1997.
820   Shukla, S., and Lettenmaier, D. P.: Seasonal hydrologic prediction in the United States:
821   understanding the role of initial hydrologic conditions and seasonal climate forecast skill, Hydrol.
822   Earth Syst. Sci., 15, 3529-3538, DOI 10.5194/hess-15-3529-2011, 2011.
823   Shukla, S., Voisin, N., and Lettenmaier, D. P.: Value of medium range weather forecasts in the
824   improvement of seasonal hydrologic prediction skill, Hydrol. Earth Syst. Sci., 16, 2825-2838,
825   10.5194/hess-16-2825-2012, 2012.
826   Shukla, S., Sheffield, J., Wood, E. F., and Lettenmaier, D. P.: On the sources of global land
827   surface hydrologic predictability, Hydrol. Earth Syst. Sci., 17, 2781-2796, 10.5194/hess-17-2781-
828   2013, 2013.
829   Soukup, T. L., Aziz, O. A., Tootle, G. A., Piechota, T. C., and Wulff, S. S.: Long lead-time
830   streamflow forecasting of the North Platte River incorporating oceanic-atmospheric climate
831   variability, J. Hydrol., 368, 131-142, 2009.
832   Thirel, G., Rousset-Regimbeau, F., Martin, E., and Habets, F.: On the Impact of Short-Range
833   Meteorological Forecasts for Ensemble Streamflow Predictions, J. Hydrometeorol., 9, 1301-
834   1317, 10.1175/2008jhm959.1, 2008.
835   Tian, Y., Booij, M. J., and Xu, Y.-P.: Uncertainty in high and low flows due to model structure
836   and parameter errors, Stoch. Environ. Res. Risk Assess., 28, 319-332, 10.1007/s00477-013-0751-
837   9, 2014.
838   Tootle, G. A., and Piechota, T. C.: Suwannee River Long Range Streamflow Forecasts Based On
839   Seasonal Climate Predictors, JAWRA Journal of the American Water Resources Association, 40,
840   523-532, 2004.
841   Towler, E., Roberts, M., Rajagopalan, B., and Sojda, R. S.: Incorporating probabilistic seasonal
842   climate forecasts into river management using a risk-based framework, Water Resour. Res., 49,
843   4997–5008, 10.1002/wrcr.20378, 2013.
844   Van den Tillaart, S. P. M., Booij, M. J., and Krol, M. S.: Impact of uncertainties in discharge
845   determination on the parameter estimation and performance of a hydrological model, Hydrology
846   Research, 44, 454–466 2013.
847   Van Dijk, A. I. J. M., Peña-Arancibia, J. L., Wood, E. F., Sheffield, J., and Beck, H. E.: Global
848   analysis of seasonal streamflow predictability using an ensemble prediction system and
849   observations from 6192 small catchments worldwide, Water Resour. Res., 49, 2729–2746,
850   10.1002/wrcr.20251, 2013.
851   Van Ogtrop, F. F., Vervoort, R. W., Heller, G. Z., Stasinopoulos, D. M., and Rigby, R. A.: Long-
852   range forecasting of intermittent streamflow, Hydrol. Earth Syst. Sci., 15, 3343-3354,
853   10.5194/hess-15-3343-2011, 2011.

854   Velázquez, J. A., Anctil, F., and Perrin, C.: Performance and reliability of multimodel
855   hydrological ensemble simulations based on seventeen lumped models and a thousand
856   catchments, Hydrol. Earth Syst. Sci., 14, 2303-2317, 10.5194/hess-14-2303-2010, 2010.
857   Vidal, J. P., Martin, E., Franchistéguy, L., Habets, F., Soubeyroux, J. M., Blanchard, M., and
858   Baillon, M.: Multilevel and multiscale drought reanalysis over France with the Safran-Isba-
859   Modcou hydrometeorological suite, Hydrol. Earth Syst. Sci., 14, 459-478, 2010.
860   Wang, E., Zhang, Y., Luo, J., Chiew, F. H. S., and Wang, Q. J.: Monthly and seasonal
861   streamflow forecasts using rainfall-runoff modeling and historical weather data, Water Resour.
862   Res., 47, W05516, 10.1029/2010wr009922, 2011.
863   Wang, W., Gelder, P. H. A. J. M. V., Vrijling, J. K., and Ma, J.: Forecasting daily streamflow
864   using hybrid ANN models, J. Hydrol., 324, 383-399, 10.1016/j.jhydrol.2005.09.032, 2006.
865   Wedgbrow, C. S., Wilby, R. L., Fox, H. R., and O'Hare, G.: Prospects for seasonal forecasting of
866   summer drought and low river flow anomalies in England and Wales, Int. J. Climatol., 22, 219-
867   236, 10.1002/joc.735, 2002.
868   Wedgbrow, C. S., Wilby, R. L., and Fox, H. R.: Experimental seasonal forecasts of low summer
869   flows in the River Thames, UK, using Expert Systems, Clim. Res., 28, 133-141, 2005.
870   Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Elsevier, New York., 1995.
871   Winsemius, H. C., Dutra, E., Engelbrecht, F. A., Archer Van Garderen, E., Wetterhall, F.,
872   Pappenberger, F., and Werner, M. G. F.: The potential value of seasonal forecasts in a changing
873   climate in southern Africa, Hydrol. Earth Syst. Sci., 18, 1525-1538, 10.5194/hess-18-1525-2014,
874   2014.
875   Wood, A. W., Maurer, E. P., Kumar, A., and Lettenmaier, D. P.: Long-range experimental
876   hydrologic forecasting for the eastern United States, J. Geophys. Res, 107, 4429,
877   10.1029/2001JD000659, 2002.
878   Wood, A. W., and Lettenmaier, D. P.: A Test Bed for New Seasonal Hydrologic Forecasting
879   Approaches in the Western United States, Bulletin of the American Meteorological Society, 87,
880   1699-1712, 10.1175/bams-87-12-1699, 2006.
881   Yossef, N. C., van Beek, L. P. H., Kwadijk, J. C. J., and Bierkens, M. F. P.: Assessment of the
882   potential forecasting skill of a global hydrological model in reproducing the occurrence of
883   monthly flow extremes, Hydrol. Earth Syst. Sci., 16, 4233-4246, 10.5194/hess-16-4233-2012,
884   2012.
885   Yossef, N. C., Winsemius, H., Weerts, A., van Beek, R., and Bierkens, M. F. P.: Skill of a global
886   seasonal streamflow forecasting system, relative roles of initial conditions and meteorological
887   forcing, Water Resour. Res., 49, 4687–4699, 10.1002/wrcr.20350, 2013.