



## LJMU Research Online

**Broderick, C, Matthews, TR, Wilby, RL, Bastola, S and Murphy, C**

**Transferability of hydrological models and ensemble averaging methods between contrasting climatic periods**

<http://researchonline.ljmu.ac.uk/id/eprint/4690/>

### Article

**Citation** (please note it is advisable to refer to the publisher's version if you intend to cite from this work)

**Broderick, C, Matthews, TR, Wilby, RL, Bastola, S and Murphy, C (2016) Transferability of hydrological models and ensemble averaging methods between contrasting climatic periods. *Water Resources Research*, 52 (10). pp. 8343-8373. ISSN 1944-7973**

LJMU has developed [LJMU Research Online](#) for users to access the research output of the University more effectively. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LJMU Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain.

The version presented here may differ from the published version or from the version of the record. Please see the repository URL above for details on accessing the published version and note that access may require a subscription.

For more information please contact [researchonline@ljmu.ac.uk](mailto:researchonline@ljmu.ac.uk)

<http://researchonline.ljmu.ac.uk/>

1 **Transferability of hydrological models and ensemble averaging**  
2 **methods between contrasting climatic periods**

3 **Authors:**

4 Ciaran Broderick<sup>1</sup>, Tom Matthews<sup>2</sup>, Robert L. Wilby<sup>3</sup>, Satish Bastola<sup>4</sup>, Conor Murphy<sup>1</sup>

5 **Affiliation:**

6 Maynooth University<sup>1</sup>

7 Liverpool John Moores University<sup>2</sup>

8 University of Loughborough<sup>3</sup>

9 Georgia Institute of Technology<sup>4</sup>

10 **Key points:**

- 11     ▪ Differential split sample testing of hydrological models should include use of best  
12     available analogues of expected climate changes.
- 13     ▪ For climate impact assessment use a multi-model ensemble with an objective  
14     averaging technique to combine members.
- 15     ▪ Evaluate parameter and model transferability using a range of climate analogues,  
16     catchment types and performance criteria.

17 **Transferability of hydrological models and ensemble averaging methods between**  
18 **contrasting climatic periods**

19 *Ciaran Broderick, Tom Matthews, Robert L. Wilby, Satish Bastola, Conor Murphy*

20 **Abstract**

21 Understanding hydrological model predictive capabilities under contrasting climate  
22 conditions enables more robust decision making. Using Differential Split Sample Testing  
23 (DSST) we analyse the performance of six hydrological models for 37 Irish catchments under  
24 climate conditions unlike those used for model training. Additionally, we consider four  
25 ensemble averaging techniques when examining inter-period transferability. DSST is  
26 conducted using two/three-year non-continuous blocks of (i) the wettest/driest years on  
27 record based on precipitation totals, and (ii) years with a more/less pronounced seasonal  
28 precipitation regime. Model transferability between contrasting regimes was found to vary  
29 depending on the testing scenario, catchment and evaluation criteria considered. As expected,  
30 the ensemble average outperformed most individual ensemble members. However, averaging  
31 techniques differed considerably in the number of times they surpassed the best individual  
32 model-member. Bayesian Model Averaging (BMA) and the Granger-Ramanathan (GRA)  
33 method were found to outperform the simple arithmetic mean (SAM) and Akaike Information  
34 Criteria Averaging (AICA). Here, GRA performed better than the best individual model in  
35 51% to 86% of cases (according to the Nash-Sutcliffe criterion). When assessing model  
36 predictive skill under climate change conditions we recommend (i) setting up DSST to select  
37 the best available analogues of expected annual mean and seasonal climate conditions; (ii)  
38 applying multiple performance criteria; (iii) testing transferability using a diverse set of  
39 catchments and; (iv) using a multi-model ensemble in conjunction with an appropriate  
40 averaging technique. Given the computational efficiency and performance of GRA relative to  
41 BMA, the former is recommended as the preferred ensemble averaging technique for climate  
42 assessment.

43 **1. Introduction**

44 Evaluating hydrological responses to climate change is an important area of research.  
45 Conventional impact assessments typically involve: (i) projecting climate responses using  
46 General Circulation Model (GCM) simulations forced by greenhouse gas emission scenarios;  
47 (ii) post-processing/downscaling GCM output; and (iii) estimating catchment scale impacts  
48 using hydrological models. This top-down approach introduces uncertainties at each step  
49 which vary depending on factors including the catchment and regional climate characteristics.  
50 Even so-called ‘stress testing’ (or sensitivity-based) techniques – which move away from  
51 direct reliance on GCMs – are subject to uncertainties in hydrological model structures and  
52 parameter sets [Prudhomme *et al.*, 2010, 2015; Whateley *et al.*, 2014; Wilby *et al.*, 2014].

53 Hydrological model uncertainty stems from errors in input (e.g. precipitation) and output (e.g.  
54 streamflow) data, as well as from deficiencies in model structures and non-uniqueness of  
55 model parameters. Previous studies have encountered difficulties when addressing structural

56 uncertainty, particularly when trying to identify a single, optimum model for a given  
57 catchment type [Clark *et al.*, 2008; van Esse *et al.*, 2013; Coxon *et al.*, 2014]. Similarly,  
58 uncertainty relating to model calibration/training arises due to equifinality or the inability to  
59 determine a globally optimum parameter set [Beven, 2006]. For climate impact studies,  
60 additional uncertainties arise due to hydrological models being applied to conditions outside  
61 those used for model training. Hence, the assumption of parametric stationarity – whereby  
62 parameters provide realistic simulations when applied under hydroclimatological conditions  
63 dissimilar to those used for model development - has been widely questioned. A number of  
64 authors have called for a more rigorous and systematic approach to interrogating  
65 transferability and model robustness for climate impact studies [Hartmann and Bárdossy,  
66 2005; Wilby, 2005; Beven, 2006; Wilby and Harris, 2006; Andréassian *et al.*, 2009; Vaze *et*  
67 *al.*, 2010; Merz *et al.*, 2011; Coron *et al.*, 2012; Li *et al.*, 2012; Seiller *et al.*, 2012, 2015;  
68 Brigode *et al.*, 2013; Westra *et al.*, 2014; Thirel *et al.*, 2015a, 2015b].

69 Studies employing Differential Split Sample Testing [DSST; Klemeš, 1986] show  
70 dependence of model parameters on the climate and meteorological conditions dominating  
71 the training period and their role in activating different rainfall-runoff processes [Wagener,  
72 2003; Choi and Beven, 2007; Herman *et al.*, 2013]. One consequence is that identification of  
73 a ‘best’ hydrological model becomes intractable, as relative performances vary in time. This  
74 highlights the importance of employing a multiple rather than single model strategy and  
75 understanding potential deficiencies in model performance when extrapolated beyond  
76 training conditions. Such difficulties are further compounded by the absence of universally  
77 accepted metrics to benchmark performance [Krause *et al.*, 2005]. Model ensembles that  
78 better characterise the structural uncertainty space are one practical solution; the ensemble  
79 may reflect the strengths of individual models which may each omit or provide a biased  
80 representation of system processes. The importance of including model components which  
81 capture processes associated with particular catchment types - as a means to improving  
82 performance and physical realism in the structure - is demonstrated by previous multi-model  
83 studies [van Esse *et al.*, 2013; Coxon *et al.*, 2014]. Whilst previous research shows that using  
84 a multi-model ensemble is superior to relying on an individual model, the best way of  
85 combining ensemble members remains an area of active research [e.g. Shamseldin *et al.*,  
86 1997; Abraham and See, 2002; Ajami *et al.*, 2006; Hansen, 2008; Diks and Vrugt, 2010;  
87 Arsenault *et al.*, 2015].

88 Only when critical uncertainties have been addressed [Clark *et al.*, 2016], and sufficient  
89 testing has been conducted to establish performance under a range of conditions, can model  
90 projections be used to make well informed adaptation decisions (including under ‘stress test’  
91 conditions). To this end, the present study uses DSST to examine temporal transferability of  
92 a multi-model hydrological ensemble. The study has two aims. First, we analyse the  
93 performance of six lumped Conceptual Rainfall-Runoff (CRR) models applied under climate  
94 conditions that differ from those used for model training, for catchments across the Island of  
95 Ireland (IoI). Previous studies have assessed climate change impacts on Irish catchments  
96 [Steele-Dunne *et al.*, 2008; Bastola *et al.*, 2011, 2012], but systematic appraisal of model  
97 transferability has yet to be undertaken. In addition, there is limited information about which

98 model(s) perform best across catchments with contrasting hydrological and climate  
99 characteristics. Second, we examine through comparison of multiple methods, the extent to  
100 which an ensemble offers improved transferability beyond reliance on individual model  
101 structures. This study expands on existing research, [Vaze *et al.*, 2010; Merz *et al.*, 2011;  
102 Coron *et al.*, 2012; Li *et al.*, 2012] – and the work of Seiller *et al.*, [2012, 2015] in particular  
103 – by contributing to knowledge of model limitations under non-stationary conditions. In  
104 particular, we quantify how model performance may be diminished by transference and  
105 whether this is greater with respect to wetter/drier conditions and specific seasonal  
106 precipitation regimes. We also examine the suitability of using observed records as an  
107 analogue to determine predictive performance under possible future conditions, demonstrate  
108 an approach for training and unbiased model evaluation, and examine methods to improve  
109 model application in climate impact studies.

110 The following section describes the study catchments, hydrological models and averaging  
111 techniques employed. We also outline the criteria for selecting contrasting climate periods.  
112 Section 3 presents the results of the analyses. Section 4 discusses the new insights gained  
113 from the transferability and ensemble averaging assessment before suggesting priorities for  
114 further research.

## 115 **2. Methods**

### 116 **2.1 Study Catchments and Data**

117 The study was undertaken using 37 catchments from IoI (Figure 1; Table 1): 35 from the Irish  
118 Reference Network (IRN) [Murphy *et al.*, 2013]; two from the UK Benchmark Network  
119 [Hannaford and Marsh, 2008]. These catchments have near natural flow regimes, are  
120 minimally influenced by human activity and possess quality assured, long-term observational  
121 records. Catchments along the western seaboard are more exposed to Atlantic weather  
122 systems and subject to more pronounced orographic enhancement. As a result they tend to  
123 have higher annual precipitation totals.

124 Daily streamflow, precipitation and potential evapotranspiration (PET) data for the period  
125 1970-2010 were used. Observed streamflow data for the Republic of Ireland were provided  
126 by the Office of Public Works (OPW; <http://www.opw.ie/hydro/>) and the Environmental  
127 Protection Agency. Data for Northern Ireland (Gauge ID: 201008 and 201005) were obtained  
128 from the UK National River Flow Archive (<http://nrfa.ceh.ac.uk/>). Not all catchments have  
129 continuous records for the study period, hence model transferability was only assessed using  
130 periods with at least 90% data coverage.

131 Catchment average rainfall was estimated from a quality-assured 1km × 1km gridded dataset  
132 provided by Met Éireann [Walsh, 2012]. Daily PET, estimated via the Penman method [Allen  
133 *et al.*, 1998], was also provided by Met Éireann for the closest synoptic station to each  
134 catchment centroid (Figure 1). Gaps in the records were infilled through regression with  
135 highly correlated (Pearson's coefficient >0.7) neighbouring stations. Additionally, to ensure a

136 robust statistical relationship donor sites that provided an overlapping period of >5 years  
137 were selected.

138 No previous study has developed a typology of catchments for IoI [e.g. *Chiverton et al.*,  
139 2015]. Here, we use the Base Flow Index (BFI) to characterise differences in our catchment  
140 sample. The BFI is defined as the proportion of catchment outflow derived from saturated  
141 groundwater storage or baseflow as opposed to direct runoff [*Sear et al.*, 1999]. Generally,  
142 catchments with a high BFI have greater recharge and storage capacity, and thus potential to  
143 sustain flow during drier periods. Such catchments also tend to have a slower (i.e. time to  
144 peak) and more damped response to storm events [*Chiverton et al.*, 2015]. While the extent of  
145 surface/groundwater dominance and the associated BFI value is typically linked to catchment  
146 geology [*Coxon et al.*, 2014], it is associated with other characteristics including: vegetation,  
147 topography, climatic history, land cover and soil type [*Bloomfield et al.*, 2009; *Price*, 2011].  
148 Our focus on this index follows *Coxon et al.* [2014] who used the index as a key property  
149 when differentiating model performance for UK catchments. Similarly, *van Esse et al.* [2013]  
150 distinguish between groundwater and surface runoff dominated catchments when comparing  
151 model structures for 237 French catchments.

152 The hydrograph separation technique of *Gustard et al.* [1992] is used to estimate the BFI.  
153 This involves dividing the discharge series into non-overlapping, five-day blocks, then  
154 calculating the minimum for each block. Minima less than 0.9 times surrounding five-day  
155 blocks are taken as the base flow separation line. Daily base flow values are estimated using  
156 linear interpolation between the identified central minima. Values above observed daily flow  
157 are (re)set to the observed value. The index is estimated as the ratio between the total volume  
158 of flow and the volume of flow beneath the base flow line. The range of BFI values in our  
159 catchment network is shown in Table 1.

## 160 **2.2 Hydrological Models**

161 Six lumped CRR models (NAM, HyMod, Tank, HBV, GR4J and AWBM) are used to  
162 explore transferability under contrasting climate conditions. Developing a competent  
163 ensemble necessitates using models of sufficient diversity to ensure structural uncertainty is  
164 well represented and the ensemble has good performance potential under a range of  
165 hydroclimatological conditions [*Thiboult et al.*, 2016]. From a structural perspective, the  
166 inclusion of ‘quick’ flow pathways through upper layers and routing algorithms that regulate  
167 the volume and timing of peak flow events is important in ‘flashier’ catchments. Conversely,  
168 structures which provide a better representation of longer term storage components, with  
169 delayed outlet, inter-store routing and enhanced infiltration and exchange processes are  
170 needed for catchments with higher baseflow contributions [*van Esse et al.*, 2013]. Hence,  
171 selecting physically plausible structures which also provide contrasting conceptualizations  
172 and numerical descriptions of the main rainfall-runoff mechanisms were key criteria in model  
173 choice. Models were also selected on the basis that they have i) been used previously in  
174 similar intercomparison studies, ii) demonstrated performance as functional across diverse  
175 conditions, and iii) modest computational/data requirements that are amenable to climate  
176 impact assessment [*Bastola et al.*, 2011; *Seiller et al.*, 2012].

177 Our sample includes complex models with a relatively large number of empirically estimated  
178 (free) parameters alongside more parsimonious structures. All were applied in a lumped  
179 configuration at a daily time step using the same PET and precipitation inputs. Each model  
180 includes routines for evaporative losses and soil moisture accounting. The temperate IoI  
181 climate means snowfall occurs relatively infrequently and generally remains on the ground  
182 for only 1-2 days – although heavier snowfalls can persist for 10-12 days [Murphy, 2012;  
183 Sweeney, 2014]. Consequently, snowpack development is not a significant component of the  
184 hydrological regime and thus a snowmelt routine is not included. All models divide saturation  
185 excess between slower/quicker responding pathways and allow temporal distribution of  
186 individual and combined flow components. They differ in the number/type/configuration of  
187 stores (e.g. interception, root zone, series/parallel), the constituents of total flow included  
188 (e.g. interflow, overland flow), and the routing mechanisms employed (e.g. (non-) linear  
189 storage, unit hydrograph). Full model descriptions can be found in the literature so only a  
190 brief synopsis is provided for each below and in Table 2.

191 *NAM* (Nedbor-Afstromnings-Model [Madsen, 2000]) simulates runoff using three storage  
192 components: surface storage, root zone storage and a groundwater store. Stores are depleted  
193 through evaporative loss, lateral flow and infiltration. Overland flow is generated when  
194 capacity in the surface store is exceeded. A proportion of this excess also infiltrates to the  
195 root and lower groundwater zones. Surface and interflow contributions are routed through  
196 two linear reservoirs; base flow is routed through a single linear reservoir.

197 *HyMod* (HYdrologic MODel [Wagener et al., 2001]) has five reservoirs including a non-  
198 linear soil moisture store, three ‘quick’ flow linear reservoirs (in series) and a parallel  
199 groundwater reservoir. Actual evapotranspiration depends on saturation of the soil moisture  
200 store and evapotranspiration at the potential rate. It is noted that HBV and HyMod share a  
201 similar soil moisture accounting routine.

202 *Tank* [Sugawara, 1995], with 15 parameters, is the most complex model employed in the  
203 study. It has a hierarchy of four vertical non-linear storage reservoirs simulating, lateral flow,  
204 saturated flow and unsaturated moisture fluxes. Each tank discharges both vertically and  
205 horizontally. Parameters control the height of the horizontal outlet from each tank and their  
206 discharge rate; parameters also regulate the vertical infiltration rate. The lateral contribution  
207 from successive stores captures total runoff contributions from surface, intermediate, sub-  
208 base and base flow respectively.

209 *HBV* (Hydrologiska Byråns Vattenbalansavdelning [Seibert, 1996]) generates runoff using  
210 three storage reservoirs, including a soil moisture zone along with an upper and lower  
211 subsurface reservoir. It incorporates a set of runoff response algorithms and a function for  
212 streamflow routing. Within HBV groundwater recharge and actual evaporation are estimated  
213 as a function of water levels in the upper storage zone. Discharge occurs both laterally –  
214 through the lower (one linear outflow) and upper zone (two linear outflows) – and vertically  
215 from the upper zone only; a triangular weighting function is used to route their combined  
216 outflows.

217 *GR4J* (Génie Rural à 4 paramètres Journalier [*Perrin et al.*, 2003]) is the most parsimonious  
218 structure used, incorporating only four free parameters. Effective rainfall and soil moisture are  
219 estimated from net precipitation. Fluxes from the soil moisture zone along with effective rainfall  
220 are partitioned as a 10:90 split between two routing channels representing direct and delayed  
221 runoff respectively. The first routing applies a single unit hydrograph and the second a unit  
222 hydrograph and nonlinear storage function. Groundwater exchanges with deeper aquifers and/or  
223 adjoining catchments are represented using a gain/loss function applied to each routing channel.

224 *AWBM* (Australian Water Balance Model [*Boughton*, 2004]) uses three area-weighted surface  
225 reservoirs with different storage capacities to simulate partial areas of runoff. Water levels in  
226 each are iteratively adjusted according to daily rainfall and evaporative loss. The observed  
227 input evaporation series is subject to a multiplicative correction factor to adjust for any  
228 potential over estimation of PET. This factor is treated as an additional model parameter  
229 (sampling range 0.9-1.0) and estimated accordingly (Section 2.4). Saturation excess from the  
230 soil moisture routine is partitioned and routed between a base flow and surface runoff store;  
231 total runoff is taken as their combined outflows.

### 232 **2.3 Differential split sampling**

233 We adopted a modified version of the DSST approach of *Klemeš* [1986] involving an initial  
234 fitting or ‘*training*’ procedure, followed by performance evaluation for independent ‘*control*’  
235 conditions (similar to training) and ‘*testing*’ period (representing the opposing precipitation  
236 regime to the control). Using the period employed for model training as a benchmark to  
237 assess transferability precludes an unbiased estimate of how well models generalize across  
238 different climate regimes. Hence, to remove bias towards the training data an independent  
239 control period was used. Figure 2 describes the DSST procedure which is applied both for  
240 identification of model parameters (Section 2.4) and model averaging (Section 2.5).  
241 Differences in performance between the control (e.g. A in Figure 2) and testing (e.g. B in  
242 Figure 2) periods are indicative of transferability when trained under dissimilar conditions  
243 (e.g. use B to simulate regime type A in Figure 2).

244 Two sets of DSST were conducted. First, for each catchment we examined transferability  
245 between the ‘wettest’ and ‘driest’ years – identified from total annual precipitation statistics.  
246 Second, we examined transferability between years with contrasting annual precipitation  
247 patterns. In both cases, hydrological years (1st October to 30th September) were used. For the  
248 former, each CRR model was trained using the 1st, 3rd and 5th ranked wettest years. Model  
249 performance on the 2nd, 4th and 6th ranked wettest years (taken as the wet period control)  
250 provide a benchmark to test the transferability of models trained on the contrasting 1st, 3rd  
251 and 5th ranked driest years (Figure 3(a, b)). The opposing transferability assessment was also  
252 conducted using the 6 driest years. Differences in rainfall ( $\text{mm yr}^{-1}$ ) between DSST periods  
253 are smallest for Gauge ID 19001 (21/23 % drier/wetter) and greatest for Gauge ID 18006  
254 (33/50 % drier/wetter). Differences in wet/dry DSST periods relative to the 1976-2005  
255 climatological mean for each catchment are shown in Figure 4(a).

256 Climate model projections suggest wetter winters and drier summers for IoI [*Steele-Dunne et*  
257 *al.*, 2008; *Bastola et al.*, 2011, 2012; *Matthews et al.*, 2016], necessitating transferability of

258 models to an amplified seasonal regime. This is particularly important given how the  
259 dynamics of intra-seasonal processes during training (the rate, timing and distribution of  
260 storage recharge and reduction through the year) may affect the model response when used to  
261 simulate more extreme wetting-up and drying episodes [Wagener, 2003; Herman *et al.*,  
262 2013]. The type of seasonal regime is expected to influence the structural  
263 components/parameters for soil moisture accounting and the behaviour of longer term stores,  
264 as well as the threshold and time delay of different flow paths. Hence, under transference the  
265 training scenario used has particular implications for accurate simulation of baseflow and  
266 storm event dynamics.

267 To explore the role of inter-seasonal precipitation differences, hydrological years were split  
268 into two six-month blocks representing summer (April to September, AMJJAS) and winter  
269 (October to March, ONDJFM) respectively. For each season, anomalies were calculated and  
270 a z-score transformation applied. Results were plotted with summer and winter anomalies  
271 located on the y- and x-axes respectively. Depending on location within each quadrant,  
272 individual hydrological years were classified as: Dry-Dry, Wet-Wet, Dry-Wet or Wet-Dry.  
273 The 1st and 3rd ranked years were used for model training; the 2nd and 4th ranked years  
274 were used both as the control and for assessing transferability from seasonal regimes in other  
275 quadrants.

276 Figure 3(c) shows the location of individual years within each quadrant. Note that seasonal  
277 totals are not plotted using z-score transformation. Instead, values were centred to give zero  
278 mean and scaled to have standard deviation equal to one. The experimental design recognizes  
279 that testing based on annual precipitation totals alone can mask significant variations *within*  
280 years with similar totals [Wilby *et al.*, 2015a; 2015b]. Here only two years are used for  
281 training/testing due to some catchments having few occurrences of the four seasonal regime  
282 types. Figure 4 (b-e) presents differences in rainfall seasonality used for DSST. Differences  
283 in summer precipitation for DSST periods, estimated relative to the long-term seasonal mean,  
284 range from +44% (Dry-Wet; 39006) to -40% (Wet-Dry; 19001). The winter period  
285 differences vary between -34% (Dry-Dry; 19001) and +25% (Wet-Wet; 14007).

286 We use the coding system X/Y to identify which scenario of temporal transference is  
287 examined. Here X and Y identify which independent training and evaluation period was used.  
288 Identification codes with the same first and second letter indicate training and evaluation  
289 under two similar regimes selected from the observed record. An independent 'control' is  
290 used to remove inherent bias towards the training period. Different first and second letters  
291 denote training and testing under an opposing set of conditions. For example, D/W (W/D)  
292 identifies the scenario of training on the driest (wettest) and testing on the wettest (driest)  
293 years respectively. The same applies to the seasonal experiment (e.g. DD/DD), whereby the  
294 first and last two letters indicate the seasonal precipitation regime (e.g. DD indicates Dry-  
295 Dry) used for training and testing/control respectively.

296 Previous DSST studies have generally employed 5-10 year training/testing periods using both  
297 block sampling and non-continuous years [Yapo *et al.*, 1996; Anctil *et al.*, 2004; Hartmann  
298 and Bárdossy, 2005; Merz *et al.*, 2011; Coron *et al.*, 2012; Li *et al.*, 2012; Seiller *et al.*, 2012,

299 2015]. Assessing model suitability for climate impact assessment – for which models are  
 300 applied under a projected climate that may diverge significantly from conditions experienced  
 301 during observations – necessitates evaluating performance under as demanding a set of  
 302 conditions as possible. This requires a compromise between maximizing difference in periods  
 303 used to assess transferability *versus* achieving potentially more robust training. Given the  
 304 short record length available (~30 years) and temperate nature of the IoI climate (which  
 305 moderates the occurrence of extreme interannual/seasonal variability) DSST was undertaken  
 306 using three/two-year non-continuous periods. This was considered sufficient to examine  
 307 transferability under strict conditions yet provide sufficient training. Also, the shortened  
 308 record lengths available for some catchments may omit years with more pronounced  
 309 variability leading to a less strict DSST. However, based on relative differences in the rainfall  
 310 regime between training/testing conditions for all IRN catchments, those with a shorter  
 311 record length provide a similar level of diversity in precipitation (Figure 4).

312 The Nash-Sutcliffe efficiency (NSE [*Nash and Sutcliffe, 1970*]) criterion and a volumetric  
 313 error measure (PBIAS) were used to assess performance when transferring models between  
 314 control and testing periods. NSE is known to be biased towards higher flows. To provide a  
 315 more balanced measure of performance across the hydrograph,  $NSE^{1/3}$  ( $NSE_{cubrt}$ ) was also  
 316 used. PBIAS provides a measure of the models' systematic error, as squared or absolute value  
 317 terms are absent. In contrast, the Nash Sutcliffe criterion squares the deviation thereby  
 318 weighting positive and negative outliers equally, thus providing a measure of performance in  
 319 reproducing patterns of variability in the observed series [*Gupta et al., 2009*]. The NSE and  
 320  $NSE_{cubrt}$  are defined as equation 1 and 2 respectively:

$$NSE = 1 - \frac{\sum_{t=1}^T (Q_o^t - Q_m^t)^2}{\sum_{t=1}^T (Q_o^t - \overline{Q_o})^2} \quad (1)$$

$$NSE_{cubrt} = 1 - \frac{\sum_{t=1}^T (\sqrt[3]{Q_o^t} - \sqrt[3]{Q_m^t})^2}{\sum_{t=1}^T (\sqrt[3]{Q_o^t} - \overline{\sqrt[3]{Q_o}})^2} \quad (2)$$

321 where  $Q_m$  and  $Q_o$  represent simulated and observed daily runoff respectively;  $\overline{Q_o}$  is the mean  
 322 observed streamflow for the estimation period,  $t$  is the time step, and  $T$  is the number of data  
 323 points. Similarly  $\sqrt[3]{Q_m}$  and  $\sqrt[3]{Q_o}$  represent simulated and observed daily runoff with a cube  
 324 root transformation applied;  $\overline{\sqrt[3]{Q_o}}$  is the mean observed cube root transformed streamflow.  
 325 The PBIAS measure (equation 3) is described by:

$$PBIAS = \frac{\sum_{t=1}^T Q_m^t - Q_o^t}{\sum_{t=1}^T Q_o^t} \times 100 \quad (3)$$

## 326 2.4 Parameter Selection

327 Parameter values sampled from different regions of parameter space can provide equally  
 328 valid simulations of system behaviour [*Beven, 2006*]. This may, in part, be attributed to the

329 over-parameterization of hydrological models, as well as to issues of parameter  
330 interdependence and identifiability. Although parameter sets may perform comparably well  
331 during training, their values are tuned to the training data used, meaning they can respond  
332 very differently when applied under dissimilar conditions [Uhlenbrook *et al.*, 1999].  
333 Additionally, parameters may exhibit differing sensitivities depending on the climate  
334 conditions experienced during training; this has implications for identifiability and  
335 performance under contrasting conditions [Merz *et al.*, 2011].

336 To address parameter uncertainty we employ the Generalized Likelihood Uncertainty  
337 Estimation procedure (GLUE [Beven and Binley, 1992]), a Monte Carlo based approach to  
338 model training and uncertainty assessment which is employed extensively in hydrological  
339 and environmental modelling [Blasone *et al.*, 2008; Bastola *et al.*, 2011; Shafii and Tolson,  
340 2015]. The GLUE procedure is applied to the training data (Figure 2); evaluation was  
341 undertaken using the control and testing data.

342 For each model, 10,000 simulations were conducted for the period 1970-2010 using  
343 parameter sets drawn randomly from a uniform (non-informative) prior distribution using  
344 Latin Hypercube Sampling [McKay *et al.*, 1979]. We use the period 1970-1973 as a spin-up  
345 period to equalize model stores, the proceeding years (up to 2010) are used for DSST (Figure  
346 2). The GLUE procedure was applied using identified non-continuous two/three-year DSST  
347 training scenarios. By simulating the full series and then extracting non-sequential 2/3 years  
348 periods for training/testing, the temporal dynamics and internal consistency of catchment  
349 stores are maintained.

350 A likelihood measure was used to distinguish between behavioural and non-behavioural  
351 parameter sets conditional on the input data and observations. In this case, the Root Mean  
352 Squared Error (RMSE) was applied to square root transformed streamflow series (equation  
353 4):

$$\text{RMSE}_{\text{sqrt}} = \sqrt{\frac{\sum_{t=1}^T (\sqrt{Q_m^t} - \sqrt{Q_o^t})^2}{T}} \quad (4)$$

354 where  $\sqrt{Q_o^t}$  and  $\sqrt{Q_m^t}$  represent the square root of observed and simulated runoff at time step  
355  $t$  respectively;  $T$  is the total number of observations. This measure reduces bias towards  
356 higher flows associated with the standard RMSE and is a general purpose criterion for  
357 hydrograph fitting [Oudin *et al.*, 2006a, 2006b]. Using a set of performance measures  
358 different to the likelihood function above removes potential bias towards the training  
359 criterion, allowing more equitable assessment of transferability.

360 The top 10% parameter sets ranked according to  $\text{RMSE}_{\text{sqrt}}$  for the training period were  
361 retained as behavioural and the associated  $\text{RMSE}_{\text{sqrt}}$  values were used to estimate respective  
362 weights. Performance of the median simulation under control and opposing testing period(s)  
363 was used to examine model transferability. Here the median simulator refers to the combined  
364 50th percentile of daily flow which is derived from the weighted flow series simulated by the  
365 retained parameter sets. As the likelihood measure does not conform to the properties of a

366 formal objective function, and can return values greater than 1, a transformation function was  
 367 required. Following *Blasone et al.* [2008] and *Mertens et al.* [2004] the posterior likelihood  
 368 function for accepted parameter sets was calculated as the reciprocal of the returned  
 369 efficiency criterion multiplied by a normalizing factor. In this case, the posterior likelihood  
 370 function  $L(\theta_i|Q)$  for each behavioural set ( $\theta_i$ ) was calculated using (equation 5):

$$L(\theta_i|Q) = \frac{1}{F_i} \cdot \frac{1}{C} \quad (5)$$

371  
 372 where  $Q$  represents the observed runoff series and  $C$  is a scaling constant such that the sum of  
 373  $L(\theta_i|Q)$  over the accepted simulations equals unity; here  $F_i$  is the  $RMSE_{sqrt}$  for  $\theta_i$  divided by  
 374 the minima of the likelihood measure returned for the retained set. These Rescaled  
 375 Likelihoods ( $RL$ ) were used to assign a weight to the behavioural simulations. The prediction  
 376 quantiles at each time step were empirically derived according to (equation 6):

$$P[\hat{Z}_t < z] = \sum_{i=1}^N RL[f(\theta_i)|\hat{Z}_{t,i}, z] \quad (6)$$

378  
 379 where  $P$  is the selected quantile,  $\theta_i$  is the  $i$ -th parameter set and  $N$  is the number of  
 380 behavioural parameters. The value of the discharge series at time  $t$  by model  $f(\theta_i)$  is  
 381 represented by  $\hat{Z}$ . The median was taken as the most likely estimate and used as input for  
 382 model averaging.

### 383 **2.5 Model Averaging**

384 Numerous averaging techniques have been proposed. These range from simple averaging –  
 385 where all outcomes are considered equally probable – to more sophisticated weight-based  
 386 methods which may be static or dynamically tuned to system behaviour [*See and Openshaw,*  
 387 *2000; Hu et al., 2001*]. Here, four averaging techniques were considered, namely: Bayesian  
 388 model averaging (BMA), Akaike information criterion averaging (AICA), a variant of the  
 389 Granger-Ramanathan method (GRA) and simple arithmetic mean (SAM). Methods were  
 390 selected on the basis that they have achieved good results in previous inter-comparison  
 391 studies [*Diks and Vrugt, 2010; Arsenault et al., 2015*], differ in complexity, and are  
 392 representative of contrasting methodological approaches. In cases where weights were  
 393 applied, their values were estimated over the training period (Figure 2), with transferability of  
 394 the ensemble average to each opposing testing period being assessed. SAM is the least  
 395 sophisticated method considered, and assigns equal weight to each ensemble member  
 396 irrespective of past performance. While simplistic, previous studies have demonstrated that  
 397 SAM can improve performance over individual model structures [*Seiller et al., 2012, 2015*].  
 398 Additionally, SAM provides a benchmark against which to compare more complex averaging  
 399 methods. The median prediction from the GLUE method as applied above to each model and  
 400 DSST scenario was taken as the input for averaging.

401 **2.5.1 Bayesian Model Averaging (BMA)**

402 BMA is a statistical framework for combining output from competing members of an  
 403 ensemble to give a more realistic description of predictive uncertainty [Hoeting *et al.*, 1999;  
 404 Raftery *et al.*, 2005; Rojas *et al.*, 2008]. A comprehensive description of the technique is  
 405 provided by Hoeting *et al.* [1999] and Bastola *et al.* [2011]. BMA weights simulations from  
 406 individual model members based on their relative skill estimated over a training period.  
 407 According to BMA the full predictive distribution for the quantity of interest ( $\Delta$ ) is described  
 408 by (equation 7):

$$p(\Delta|M_1, \dots, M_K, D) = \sum_{k=1}^K p(\Delta|M_k, D)p(M_k|D) \quad (7)$$

409

410 The above is estimated as the mean of the posterior predictive distribution for  $\Delta$  predicted by  
 411 each individual model  $p(\Delta|M_k, D)$  weighted by the associated posterior model  
 412 probability  $p(M_k|D)$ . The posterior probability of model  $M_k$  is given by (equation 8):

$$p(M_k|D) \propto p(D|M_k)p(M_k) \quad (8)$$

413 where  $p(D|M_k)$  is the integrated likelihood of model ( $M_k$ ). A distribution for the prior  
 414 probability of each model  $p(M_k)$  must be specified. In this case, as no prior assumptions  
 415 regarding the likely performance or suitability of individual model structures were made, a  
 416 uniform (non-informative) distribution was selected. This ensured model weights  
 417 (likelihoods) were estimated conditional only on observed data used for training. The mean  
 418 and variance of the predictive distribution for  $\Delta$  were estimated using (equation 9 and  
 419 equation 10):

$$E[\Delta|M_1, \dots, M_K, D] = \sum_{k=1}^K w_k \hat{\Delta}_k \quad (9)$$

$$Var[\Delta|M_1, \dots, M_K, D] = \sum_{k=1}^K (Var(\Delta|D, M_k) + \hat{\Delta}_k) w_k - E(\Delta|D)^2 \quad (10)$$

420 where  $\hat{\Delta}_k = E(\Delta|D, M_k)$ . The weighting for models in the ensemble ( $w_k$ ) varies between zero  
 421 and one with the cumulative sum equal to unity. The total variance or predictive uncertainty  
 422 is estimated as a combination of inter- and intra-model variance. Streamflow is non-zero,  
 423 strictly positive and highly skewed meaning it does not conform to a Gaussian distribution.  
 424 Thus the probability density function of the model output at time step  $t$  was modelled using a  
 425 gamma distribution (equation 11) with heteroscedastic variance (equation 12).

$$p(\Delta|M_k) = \Delta^{\alpha_k-1} e^{(\Delta/\beta_k)} / (\Gamma(\alpha_k)\theta^{\alpha_k}) \quad (11)$$

$$\alpha = \mu_k^2/\sigma_k^2; \beta_k = \sigma_k^2/\mu_k; \mu_k = M_k; \sigma_k^2 = b \cdot M_k + c \quad (12)$$

$$l(w_1, \dots, w_k | \sigma_1^2 \dots \sigma_k^2, \Delta) = \sum_{t=1}^n \log(w_1 p(\Delta | M_1) + \dots + w_k p(\Delta | M_k)) \quad (13)$$

426 Here  $b$  and  $c$  are the coefficients which relate the model simulated series with the respective  
 427 variances. Over each training period the BMA weights and variances were estimated from  
 428 observed streamflow data through Markov Chain Monte Carlo (MCMC) sampling. This was  
 429 undertaken using the Differential Evolution Adaptive Metropolis (DREAM) algorithm [Vrugt  
 430 *et al.*, 2008]. The maximum a-posteriori probability estimate of the weights - as determined  
 431 over the training period - were used to average model simulations. Performance of the model  
 432 average when temporally transferred to each testing period was then assessed using the  
 433 adopted set of performance criteria.

### 434 2.5.2 Akaike Information Criteria Averaging (AICA)

435 AICA [Akaike, 1974] is a method for combining ensemble members based on both  
 436 performance and model parsimony. Weights represent a trade-off between reducing the  
 437 overall prediction bias while tending towards less complex models. Such a measure is  
 438 important when considering model transferability, where increasing the number of  
 439 parameters could increase the likelihood of over-fitting, thus limiting a model's ability to  
 440 generalize to unseen conditions. As specified by Buckland *et al.* [1997] and Burnham and  
 441 Anderson [2003] the weights are calculated by (equation 14):

$$\beta_{AICA,k} = \frac{\exp\left(-\frac{1}{2}I_k\right)}{\sum_{k=1}^K \exp\left(-\frac{1}{2}I_k\right)} \quad (14)$$

442 where  $I_k$  (equation 15) is an information criterion estimated based on the mean of the  
 443 logarithm of the model variances.

$$I_k = -2\log(L_k) + q(p_k) \quad (15)$$

444 In the above  $L_k$  is the maximum likelihood of model  $k$  and  $q(p_k)$  is its associated penalty  
 445 term which, in this case, is taken for each ensemble member as double the number of  
 446 calibration parameters or  $q(p_k) = 2p$ .

### 447 2.5.3 Granger-Ramanathan Averaging (GRA)

448 GRA simulations are combined using Ordinary Least Squares (OLS) optimized by  
 449 minimizing the root mean squared difference between simulated and observed series.  
 450 Previous studies have employed different variants of the method including applying a bias  
 451 correction and using (non)constrained linear coefficients [Diks and Vrugt, 2010; Arsenault *et*  
 452 *al.*, 2015]. In this study the OLS algorithm is constrained so that weights are positive and sum  
 453 to unity – a prior bias correction was not applied. The model weighting vector ( $\beta_{GRA}$ ) was  
 454 estimated according to (equation 16):

$$\beta_{GRA} = (X^T X)^{-1} X^T Y \quad (16)$$

455 where  $Y$  is a vector representing the observed discharge series for the training period and  $X$  is  
456 an  $n \times m$  matrix whose columns ( $m$ ) correspond to the daily ( $n$  rows) simulated flow series  
457 from each model member.

### 458 3. Results

459 This section presents results from the DSST undertaken to assess the performance of a six  
460 member CRR model ensemble under contrasting climate conditions. For each of the 37  
461 catchments DSST was conducted using the wettest/driest three year non-continuous periods  
462 on record. Similarly, performance when models were transferred between contrasting wet/dry  
463 seasonal scenarios was examined. Note that while DSST analysis is conducted using non-  
464 continuous periods, all model simulations are run continuously using the entire period for  
465 which input data (rainfall and PET) are available (~1970-2010). DSST was conducted for  
466 individual model structures and for the ensemble collectively, using the four different model  
467 averaging techniques.

#### 468 3.1 Individual model performance – wettest/driest years

469 Figure 5 shows individual model structures ranked according to performance when tested for  
470 each wet/dry scenario (W/D, D/W), catchment and evaluation criterion. Performance is  
471 examined using median GLUE simulations. According to the NSE criterion, HBV and GR4J  
472 generally perform best. HBV is typically ranked higher for catchments with a low BFI; GR4J  
473 performs better on catchments with a higher BFI. While both models perform well for  
474  $NSE_{cubrt}$ , NAM is also ranked among the best models for this criterion, most notably for the  
475 W/D scenario. Tank and AWBM typically return the lowest NSE and  $NSE_{cubrt}$  values across  
476 catchments. Much less consistency is evident amongst the results for PBIAS: in some  
477 instances Tank is ranked among the best performing models with GR4J amongst the worst.  
478 The favourable results for GR4J – particularly under NSE for high BFI catchments  
479 corroborate the findings of previous model intercomparison studies [*Pushpalatha et al.*, 2011;  
480 *van Esse et al.*, 2013]. Given the lack of convergence in results across catchments, testing  
481 criteria and DSST scenarios, there is considerable uncertainty when identifying a preferred  
482 model structure (albeit that a combination of GR4J and HBV appears a good compromise,  
483 with either model ranked first for 118 out of the 148 tests according to the NSE criterion).

484 Figure 6 plots scores for the evaluation criteria by comparing performance for the same three  
485 year control period when trained using (dis)similar wet/dry annual regimes (Figure 2).  
486 Differences are examined using median GLUE simulations. Distances from the diagonal  
487 ( $x=y$ ) indicate differences in performance under transference. Based on results for both DSST  
488 scenarios, NSE values vary between 0.51 (GR4J; D/W; Gauge ID 26029) and 0.97 (GR4J;  
489 D/W; Gauge ID 27002). Gauge 26029(27002) has a BFI of 0.23(0.70), a mean elevation of  
490 217(73) m, and an area of 117(511) km<sup>2</sup>. While runoff is approximately twice as much for  
491 26029 (1308 mm yr<sup>-1</sup>) as 27002 (651 mm yr<sup>-1</sup>), annual precipitation is relatively similar (1569  
492 – 1319 mm yr<sup>-1</sup>). In other words, skill is least for small, higher elevation, hydrologically  
493 responsive catchments.

494 PBIAS values range from 29% (AWBM; W/D; Gauge ID 7009; BFI 0.70) to -36.0% (NAM;  
495 W/D; Gauge ID 18003; BFI 0.54). With respect to the BFI, catchment elevation, runoff (mm  
496  $\text{yr}^{-1}$ ) and precipitation receipts (mm  $\text{yr}^{-1}$ ) are generally of (lesser) importance in  
497 differentiating model performance. Each is also negatively correlated with the BFI (Pearson's  
498 coefficient of -0.76, -0.72 and -0.70 respectively), indicating some redundancy in using the  
499 full suite of characterises to differentiate performance. Catchment area is more poorly  
500 correlated both with model performance and BFI across catchments (Pearson's coefficient  
501 =0.54). Broadly speaking, groundwater dominated catchments tend to have lower  
502 precipitation receipts, yield less runoff and are located in lower lying areas; the converse  
503 generally holds for catchments dominated by surface runoff.

504 Given that the NSE criterion is based on the sum of squared errors, irrespective of the model  
505 structure catchments with a high BFI also return higher NSE and  $\text{NSE}_{\text{cubrt}}$  values. This is due  
506 to catchments with greater storage capacity (higher BFI) tending to be less responsive to  
507 storm events, and thus producing a less variable flow series. For example, using HBV Gauge  
508 ID 21002 with BFI of 0.21 returns a NSE value of 0.55 for the D/W testing scenario. In  
509 contrast Gauge ID: 26021 (BFI 0.82) returns a NSE of 0.77 for the same model and testing  
510 scenario.

511 As shown by Figure 6, in some cases models experience a slight improvement in  
512 performance under transference. Overall, however, the greatest deviations from the diagonals  
513 are due to declining performance. Based on the greater variability and spread of the  $\text{NSE}_{\text{cubrt}}$   
514 values, models tend to experience the largest reductions in performance when trained on a  
515 wet period and transferred to a dry (i.e. W/D *versus* D/D) [Seiller *et al.*, 2012, 2015]. Figure 6  
516 is supplemented by Table 3 which lists for each catchment the DSST scenario and model  
517 associated with the greatest singular decline in performance. Decreases under transference are  
518 estimated in relative (NSE and  $\text{NSE}_{\text{cubrt}}$ ) and absolute (PBIAS) terms using performance for  
519 the control (Figure 2) as a benchmark, and represents a 'worst-case' scenario for each  
520 catchment. Greater relative decreases are associated with  $\text{NSE}_{\text{cubrt}}$  as opposed to the NSE  
521 measure; in some cases up to a 21% decrease in this criterion is observed.

522 Figure 7 shows NSE,  $\text{NSE}_{\text{cubrt}}$  and PBIAS estimates for individual model structures across all  
523 catchments when transferability between the wettest/driest years is examined. Boxplots are  
524 calculated using behavioural parameter sets identified over the training period; performance  
525 under control and testing conditions is examined. Parameter sets generally perform well  
526 across all catchments, with median NSE and  $\text{NSE}_{\text{cubrt}}$  values  $\geq 0.7$ . Only HBV, GR4J and  
527 NAM have a median NSE value greater than 0.75 for both control periods (D/D and W/W);  
528 AWBM returns the lowest median NSE and  $\text{NSE}_{\text{cubrt}}$  values respectively. Despite GR4J and  
529 HBV performing well across catchments, they exhibit a relatively large range under temporal  
530 transference. This suggests that the weighting applied through the GLUE procedure offsets  
531 the poor performance of some parameters within the behavioural set.

### 532 **3.2 Individual model performance – seasonal assessment**

533 In addition to examining transferability between the wettest and driest hydrological years,  
534 assessment was also undertaken between years with contrasting seasonal regimes. Testing

535 was performed based on sample sizes of two years using the median GLUE simulation.  
536 Figure 8 shows highest to lowest ranked model structures according to performance over  
537 each testing scenario for the NSE,  $NSE_{cubrt}$  and PBIAS criterion respectively. AWBM, along  
538 with HyMod and Tank (to a lesser extent) are the lowest ranked models for the NSE measure.  
539 HBV is generally ranked highest for catchments with lower base flow contributions; GR4J  
540 tends to be ranked higher for catchments with a larger BFI. Either HBV (52.2% of cases) or  
541 GR4J (27.2% of cases) are ranked first for 354 of 444 transference tests according to the NSE  
542 criterion. For  $NSE_{cubrt}$  both models are similarly dominant, with GR4J (50.2% of cases) or  
543 HBV (29.0% of cases) being ranked first for 344 testing scenarios. Lowest NSE and  $NSE_{cubrt}$   
544 values are generally given by AWBM which is ranked first/last for 10/503 cases of the same  
545 888 transference tests. In contrast to the NSE criteria, there is much greater uncertainty in  
546 results for PBIAS. AWBM tends to be highest ranked for catchments with a low BFI,  
547 however this is reversed as the BFI increases. Additional weaker patterns in results emerge,  
548 including the poor ranking for Tank (NSE and Abs PBIAS) and NAM ( $NSE_{cubrt}$ ) under  
549 transference to a Dry-Dry (DD) seasonal regime. Similarly AWBM performs poorly for  
550 transference to a Wet-Wet (WW) and Dry-Wet (DW) scenario according to all criteria.  
551 However, the degree of inconsistency highlights the complexity of model transference, with  
552 performance being related to the individual model structure, catchment and climate regime  
553 type.

554 Figures 9 (NSE), 10 ( $NSE_{cubrt}$ ) and 11 (PBIAS) present results of the DSST scenarios, whilst  
555 Table 4 lists for each catchment the scenario of seasonal transference and associated model  
556 structure that yields the greatest decrease in performance relative to the control for each  
557 evaluation criterion. For 29 of the 37 catchments transference to a DW (Dry-Wet; 14 cases)  
558 or DD (Dry-Dry; 15) seasonal regime returns the largest reductions in the NSE criterion.  
559 Within this the DD/DW (11 cases) and DW/DD (8 cases) scenarios are notable for returning  
560 the greatest number of poor performances. These range from a decrease in NSE of -46.4%  
561 (WD/DD; Gauge ID: 25006; Tank) to -3.2% (DD/DW; Gauge ID: 18003; HBV). In contrast,  
562 the decline in performance when transferred to a WW or WD scenario is much less, while the  
563 DW/WW or WW/DW tests do not lead to the greatest singular decrease for any catchment.

564 A similar and more pronounced pattern is evident in the results for  $NSE_{cubrt}$  and PBIAS. For  
565 the  $NSE_{cubrt}$  criterion transference to a DW or DD regime is found for 33 catchments, with  
566 seven registering reductions of 20-30% relative to the control. Poor transference to a DD and  
567 WD is similarly evident for the PBIAS criterion. As shown in Table 4, deficiencies in  
568 performance across catchments are generally associated with a more pronounced  
569 underestimation of flow volumes (WD/DD; Gauge ID: 18005; GR4J). Although there is a  
570 degree of variation between models, GR4J (NSE; PBIAS), HyMod and AWBM ( $NSE_{cubrt}$ )  
571 yield greatest reductions relative to the control.

572 Figure 12 shows the results of DSST applied to all behavioural parameter sets identified  
573 across the catchment sample. In terms of absolute model performance the highest  $NSE_{cubrt}$   
574 control/testing values are generally returned for the WD/WD scenario. Based on the median  
575 estimate, GR4J performs well across the catchment sample, whereas AWBM generally  
576 returns the lowest scores. Difficulties in transference to a DW or DD regime are also

577 highlighted by Figure 12. In contrast, parameters generally maintain performance when  
578 transferred to a WW regime irrespective of the training scenario.

### 579 **3.3 Multimodel performance**

580 Attention is now given to how use of the four different averaging methods over our multi-  
581 model ensembles may improve transferability. Figure 13 plots NSE values for individual  
582 models against corresponding values returned when model averaging is applied. Plots are  
583 based on the results of DSST conducted using contrasting wet/dry annual regimes for each  
584 catchment. Table 5 lists the frequency with which each method outperforms the individual  
585 ensemble members. In the majority of cases, model averaging surpasses performance of any  
586 single structure, even for SAM where the application of equal weights returns  $NSE_{cubrt}$  values  
587 better than individual models in more than 79% of cases. Model averaging performs better  
588 for the NSE criteria than for PBIAS. With respect to volumetric error SAM returns similar  
589 values to the more complex averaging methods employing objective weighting criteria. Both  
590 BMA and GRA perform similarly across DSST scenarios, exhibiting only a slight difference  
591 in performance under transference to each testing period(s).

592 Despite the ensemble average clearly being better than individual model members (Figure 13  
593 and Table 5), differences are evident not just in how well each averaging method performs  
594 but also in the evaluation measure used. For both Nash Sutcliffe measures, GRA and BMA  
595 are most consistent in exceeding the best ensemble member and perform considerably better  
596 than simple averaging. AICA fails under all DSSTs to provide encouraging results.  
597 Considering all DSST scenarios AICA assigns the largest weight to HBV and GR4J in 50%  
598 and 31% of cases respectively. In contrast, AWBM is never assigned a weight above zero. As  
599 would be expected, the objective methods perform well over the period used for estimation of  
600 model weights, highlighting an inherent bias to the training data. This is particularly evident  
601 for GRA according to the NSE and  $NSE_{cubrt}$  criterion. In both cases this method achieves  
602 almost perfect results (Table 5).

603 Table 6 lists the frequency with which each model averaging technique outperforms the best  
604 performing individual model from the ensemble. In the majority of cases GRA and BMA are  
605 better under transference (and for the control) than the best performing model member  
606 according to both the NSE and  $NSE_{cubrt}$  measures. In general, GRA performs better than  
607 BMA for the NSE criterion, particularly with respect to the best performing model member.  
608 However, the opposite applies for  $NSE_{cubrt}$  – albeit that returned differences are of a lesser  
609 magnitude. As is demonstrated by differences between the control and testing periods, neither  
610 GRA nor BMA experience a significant drop in performance under transference. Generally,  
611 the averaging methods perform similarly across each opposing DSST period. Overall, GRA  
612 emerges as the most consistent technique, returning high NSE and  $NSE_{cubrt}$  values across all  
613 DSST scenarios.

614 For PBIAS, all averaging methods generally return a considerably lower proportion (<20%)  
615 of better performing estimates when benchmarked against the best model member. The  
616 results shown in Table 6 are reflected in Figure 14 which displays the best/worst ranked  
617 model averaging method for each catchment and seasonal DSST scenario; also considered is

618 the best/worst performing model structure. Evident are the more favourable results for  
619 BMA/GRA according to the  $NSE/NSE_{cubrt}$  criterion. The ranking of methods is also largely  
620 consistent across individual catchments and for each DSST scenario. Figure 14 further  
621 highlights disparities in performance between the NSE and PBIAS measures. In the latter  
622 case, it is shown that the best individual model structure for each scenario typically performs  
623 better than the respective model averaging techniques. Figure 14 also highlights that the  
624 worst performing model is most often ranked lower than the worst performing averaging  
625 method.

#### 626 **4. Discussion**

627 While in some cases model performance was shown to improve relative to the control when  
628 trained under a contrasting set of conditions, in general there was a degradation in  
629 performance. The extent of this degradation depends on model structure, catchment, DSST  
630 scenario, performance criterion and averaging technique. For all catchments, no clear  
631 relationship could be identified between decline in performance under transference and  
632 relative differences in precipitation between DSST periods. This may be due to variations in  
633 training/control and testing conditions being broadly similar across the catchment sample  
634 (Figure 4(a)). In addition, despite using a two/three year period to maximize  
635 interannual/seasonal differences, the dissimilarity between training/testing conditions varies  
636 only within a limited range. Furthermore, when considering results for the catchment sample  
637 collectively, there are a number of interacting factors external to the driving climate regime.  
638 These include differences in the catchment properties and model/data uncertainties which  
639 may preclude or complicate a simple quantitative (linear or otherwise) relationship between  
640 differences in performance and differences in the associated annual/seasonal precipitation  
641 regime. As a result, no generally applicable quantitative threshold for transferability –  
642 indicating when models may become inaccurate or non-functional – can be identified. This  
643 underlines the necessity of conducting DSST on a catchment-by-catchment and model-  
644 specific basis.

645 Generally, models were challenged when transferring between wetter and drier periods.  
646 Overall, the greatest performance declines were associated with transference from wet to dry  
647 conditions. This is evident both in terms of transference between wetter/drier years and  
648 between contrasting seasonal precipitation regimes. For the latter, models struggled when  
649 simulating years with a dry winter followed by dry summer, particularly with respect to the  
650 (low flow)  $NSE_{cubrt}$  criterion. In contrast, models were less affected by transference to a wet-  
651 dry or wet-wet seasonal regime. This finding applies both to the median estimate derived  
652 using GLUE and behavioural parameter sets across the catchment sample. Hence, if climate  
653 change tends towards drier conditions, then we would expect models calibrated on a wetter  
654 present to be less accurate under future forcing. Conversely, for a more pronounced seasonal  
655 regime (wetter winters and drier summers) models may maintain performance. Difficulties in  
656 transference to a ‘drier’ regime may be related to nonlinearities in the hydrological processes  
657 being more pronounced and poorly conditioned under a ‘wetter’ regime [Atkinson *et al.*,  
658 2002, van Esse *et al.*, 2013]. Sensitivity to training using wet or dry periods is highlighted by

659 *Li et al.* [2012], who indicate that models intended to simulate a wet/dry climate scenario  
660 should be trained using a similar period from the observed record.

661 While our findings support previous research [*Li et al.*, 2012; *Seiller et al.*, 2012, 2015], they  
662 contradict Wilby and Harris [2006] who found greater transferability from wet to dry  
663 conditions in the Thames basin (SE England). Here it is highlighted that data information  
664 content, in terms of threshold parameter activation, is higher during wet periods, thereby  
665 improving transference to dry (as opposed to wet) conditions. However, as applies to all  
666 previous studies a direct comparison is complicated by differences in the hydroclimatological  
667 regime and the degree of dissimilarity between DSST conditions [*Brigode et al.*, 2013]. For  
668 example differences between ‘wet’ and ‘dry’ are more pronounced in SE England than the  
669 IoI.

670 Typically, the structures that performed well under control conditions also performed well  
671 under transference, with the model rankings generally unchanged. Overall declines in  
672 performance were not sufficient to conclude that the models may be inaccurate or non-  
673 functional under altered climate conditions. However, it is acknowledged that the historical  
674 record may only provide limited analogues to represent plausible ranges of future changes.  
675 For instance, there is no three year period that is >20% wetter or drier than the climatology  
676 mean (1976-2005) to stress test operational limitations under the full range of possible future  
677 climates [*Matthews et al.*, 2016]. Consequently, we emphasise that caution be exercised in  
678 assuming model reliability under input forcing that differs markedly from the data available  
679 for model development. This concurs with *Bastola et al.*, [2011] who found substantial  
680 divergence between individual CRR model structures when driven using the same  
681 downscaled climate change projections, even though the models performed similarly under  
682 observed conditions. Difficulties encountered in temporal transferability mirror those of  
683 spatial transferability, whereby rainfall-runoff models are developed for ungauged  
684 catchments using parameters calibrated at suitable donor sites identified based on physical  
685 similarity and/or spatial proximity [*Oudin et al.*, 2008; *Parajka et al.*, 2013]. The DSST  
686 method used here would provide a suitable approach for interrogating the performance of  
687 different regionalization techniques under contrasted conditions.

688 Our results confirm that it is impossible to identify a single optimum model structure across  
689 all catchments and all DSST scenarios. In addition, performance was found to vary  
690 considerably depending on the evaluation criteria used, with differences being most apparent  
691 when comparing the NSE and PBIAS. However, under transference for the NSE criteria, a  
692 number of models can be identified that are likely to be more/less robust for climate  
693 assessment. Overall, HBV, GR4J and to a lesser extent NAM were consistently the best  
694 performing models, with HBV (GR4J) generally ranked the highest for catchments with a  
695 lower (higher) groundwater contribution. For climate impact studies the case for GR4J is  
696 further strengthened by its relatively parsimonious structure. In contrast, AWBM generally  
697 performed poorly across DSST periods for the majority of catchments. This may be due to its  
698 relatively large number of parameters (i.e. low parsimony) or the fact that, despite its  
699 plausible structure it was conceived for a different (Australian) hydro-climate regime. It is

700 noted that, contrary to other models AWBM requires that surface stores are satisfied before  
701 excess moisture required to sustain baseflow and surface runoff is generated.

702 The favourable results for HBV and GR4J are consistent with previous studies [*Perrin et al.*,  
703 2001; *Seiller et al.*, 2012, 2015]. The good performance of GR4J may, in part, be attributed to  
704 its inclusion of a water exchange function alongside two independent parallel routing paths,  
705 which *van Esse et al.*, [2013] cite as important both for ground water-dominated catchments  
706 and successful transference between contrasting wet/dry periods. Conversely high BFI  
707 catchments with less dynamic flow behaviour may be better represented using linear-models.  
708 In our case the higher performance of HBV for responsive catchments may be due to its use  
709 of two linear outflows from the upper reservoir (one of which is threshold activated) allowing  
710 better representation of lateral and direct flow dynamics during storm events. This is  
711 supported by the better performance of HBV (GR4J) for the NSE ( $NSE_{cubrt}$ ) criterion which is  
712 more representative of high (low) flow dynamics. *Fenicia et al.*, [2014] note the importance  
713 of storage elements connected in series (versus a parallel configuration) for catchments with  
714 impermeable bedrock dominated by lateral flows. Such catchments may also favour non-  
715 linear models where threshold exceedance activates more direct flow paths. As shown by  
716 others, improvements in HBV simulation of groundwater catchments may be gained  
717 (particularly for recession dynamics) if reservoir discharges were modelled using a power  
718 function [*Samuel et al.* 2012; *van Esse et al.*, 2013].

719 The number of model parameters is an important factor that can directly affect model  
720 performance. In baseflow dominated catchments parsimonious models with less complexity  
721 (e.g. GR4J) may be sufficient. However, in catchments with a low BFI and thus higher  
722 variability in runoff a more complex model (more parameters; e.g. HBV) may be required.  
723 When comparing HBV and HyMod – which share similar soil moisture accounting routines –  
724 our results suggest that the greater parametric complexity of HBV and use of a parallel rather  
725 than serial routing/storage structure is more successful. Based on the differing number of free  
726 parameters (Table 3), the performance of AWBM and Tank indicates that a greater degree of  
727 freedom in terms of fitting does not necessarily lead to superior performance. In fact, this  
728 may increase the risk of over-fitting during training, and hence a lesser ability to generalize  
729 across diverse conditions.

730 With respect to the BFI, it is worth noting how differences in the storage and routing  
731 configuration relate to infiltration processes and performance for groundwater/runoff  
732 dominated catchments. The influence of vertical soil heterogeneity and slope has on runoff  
733 generation is well documented [*Smith and Hebbert*, 1983; *Jackson*, 1992]. Typically for  
734 catchments with permeable homogeneous soils and a low anisotropy ratio (vertical  
735 conductivity/horizontal conductivity) movement through upper layers tends to occur  
736 vertically, with vertical increases in the saturated zone depth having a greater effect on runoff  
737 than lateral movements. Here catchments are likely to have a high BFI owing to better  
738 infiltration and delayed routing. In contrast, for catchments with a high anisotropy ratio  
739 where hillslope processes dominate, lateral flows are likely to be more significant. Hence  
740 models like HBV, which can better capture vertical variability in soil processes by using  
741 multiple vertical stores and a dedicated soil moisture routine, and which explicitly account for

742 direct/lateral flows, may be more applicable to low BFI catchments. Furthermore the hillslope  
743 can be conceptualized as consisting of two soil layers, with the lower layer capable of  
744 retarding vertical flow at the boundary allowing development of subsurface stormflow. This  
745 corresponds well with the inclusion of an upper soil box in HBV from which two lateral  
746 outflows (one threshold based) are represented [*Smith and Hebbert, 1983*]. While GR4J also  
747 accounts for vertical variability, only two stores (production and routing) are included, and  
748 lateral flows are less well represented. In addition, the model has fewer free parameters to  
749 adjust in order to better capture horizontal/direct flows (e.g. the set 90:10 split between  
750 delayed and direct routing channels).

751 Relative to other criteria, model performance for PBIAS was more varied: notably, in some  
752 cases, AWBM was returned as the best performing model. Performance in simulating the  
753 long-term water balance is related to how precipitation is partitioned between evaporation  
754 and streamflow. Hence, performance hinges on those model parameters relating to  
755 evaporation influence on the water balance [*Herman et al., 2013*]. The more favourable  
756 performance of AWBM may be due to it being the only model that incorporates an  
757 adjustment factor for PET. However, determining which parameters influence the overall  
758 water balance would require an in-depth and systematic sensitivity assessment that is beyond  
759 the scope of this study. In addition, as noted by *Herman et al. [2013]* selecting behavioural  
760 parameter sets using RMSE alone (as in this study) is no guarantee of achieving an accurate  
761 water balance. Thus, differences between the NSE and PBIAS criteria may also reflect the  
762 choice of likelihood function.

763 Differences in the performance criteria suggest that model selection should give due  
764 consideration to those components of the flow regime that are most relevant to the study  
765 objectives. For example, AWBM may be more appropriate for assessing climate driven  
766 changes in the long-term water balance, as opposed to assessing changes in dynamic  
767 behaviour (e.g. timing and magnitude of flood peaks). However, given that it only provides a  
768 measure of systematic error, and is thus a less comprehensive indicator of overall  
769 performance, selecting a model on the basis of mean bias alone lacks rigor. Hence, to inform  
770 robust model selection for climate studies, modellers should examine temporal transferability  
771 giving weight to multiple performance criteria. Here each criterion can be treated equally, or  
772 based on the study objective weights can be used to place greater emphasis on performance  
773 for particular parts of the hydrological regime.

774 When benchmarked against a single model structure, the ensemble average provides a better  
775 overall estimator. The performance of averaging techniques was shown to remain relatively  
776 consistent under transference. Additionally, methods based on objective weighting are  
777 recommended over simple averaging. The results confirm findings from previous studies  
778 which stress the value of a multi-model strategy [e.g. *Shamseldin et al., 1997; Velázquez et*  
779 *al., 2010, 2011, Seiller et al., 2012, 2015; Arsenault et al., 2015*]. When benchmarked against  
780 the best individual model structure, greater variation in the averaging methods emerged.  
781 These differences are related primarily to the choice of evaluation criteria rather than the  
782 DSST scenario or catchment selected. All methods performed considerably better for the

783 NSE as opposed to PBIAS measure. This suggests that any potential bias towards certain  
784 error types should be considered when selecting an averaging technique.

785 As reported by previous studies, the AICA method was found to perform relatively poorly  
786 [Diks and Vrugt, 2010; Arsenault *et al.*, 2015] due to a tendency to heavily weight a single  
787 member, thereby discounting additional information provided by the ensemble. As  
788 implemented here, AICA is strictly a model averaging technique. This is generally not the  
789 case with conventional information criterion methods which seek to identify the single ‘best’  
790 model based on parsimony and performance. This suggests that, although it can be used as a  
791 model averaging technique, there are better alternatives. But the method does have value if  
792 there are any concerns about over-fitting models with a large number of parameters.

793 Overall, GRA produced the most consistent results across catchments and DSST periods.  
794 Whilst BMA was found to perform comparably, this method is computational demanding and  
795 requires considerable run time to achieve convergence. However, it is acknowledged that the  
796 deterministic nature of this study ignores the importance of uncertainty in model averaging.  
797 For this purpose, BMA provides a coherent framework which allows explicit quantification  
798 of both within and between model uncertainties. Given its importance for robust decision  
799 making, the benefit of selecting an averaging method like BMA which provides a  
800 comprehensive and statistically robust framework for uncertainty assessment should receive  
801 due consideration.

802 It could be argued that a more carefully selected model may provide a better tool for impact  
803 assessment. Whilst this may be appealing, particularly given the additional resources required  
804 to develop a multimodel ensemble, it ignores the fact that structural uncertainties make this a  
805 particularly risky strategy. This will always be the case because of our inability to fully  
806 explore model behaviour under (unknowable) future climate forcing using historical data. It  
807 is also noted that the process of parameter selection (whether using an optimization routine or  
808 a method such as GLUE), and the training data used, limit model ability to produce accurate  
809 simulations when extrapolated beyond this context.

810 Our results demonstrate that the best model varies depending on the DSST scenario,  
811 performance measure and catchment considered, thus making optimal model identification  
812 unlikely. Such an approach would also require tuning the selection for each catchment, which  
813 an adequate averaging technique should achieve without necessitating prior screening. An  
814 alternative strategy might be to select an optimum model subset. However, this process is  
815 subject to the same uncertainties outlined above, and is complicated by the optimal subset not  
816 always being comprised of the best individual models [Velázquez *et al.*, 2011; Seiller *et al.*,  
817 2012, 2015]. This approach further runs the risk of pooling insufficient information to  
818 provide a good measure of structural uncertainty, with too few members resulting in  
819 diminished predictive power and the added benefit of the ensemble ultimately being lost.

820 Future work will examine why the individual CRR models performed differently across the  
821 catchment sample used in this study. Exploring parameter sensitivity to time-varying  
822 hydroclimatic conditions would help link physical processes with model formulation and

823 provide insight to the relative skill of ensemble members under different forcing scenarios  
824 (e.g. wet/dry and seasonal transitions). This would also help to establish the influence which  
825 information content in the training data and the associated activation frequency of key  
826 parameters have on transferability between contrasting regimes.

827 Whilst the current study considers six dissimilar CRR models, each has a fixed structure  
828 which, it is assumed, will generalize across a variety of catchment types. However, there is  
829 scope for exploring temporal transferability using a flexible modelling framework such as  
830 SUPERFLEX [*Fenicia et al.*, 2011] or FUSE [*Clark et al.*, 2008]. Previous studies have  
831 highlighted the benefits of moving away from the ‘one-size-fits-all’ approach to one based on  
832 developing a structure commensurate with the hydrological complexity of the study  
833 catchment [*Staudinger et al.*, 2011; *Euser et al.*, 2013]. Although potentially allowing for  
834 more appropriate structure selection this would still require DSST to evaluate capabilities  
835 beyond the training set(s). Similarly using a flexible framework, whereby the effect of  
836 individual components can be isolated allows a more tenable link between physical  
837 catchment properties/processes and the model structure. Parametric uncertainty  
838 notwithstanding, it facilitates attributing differences in performance to specific structural  
839 configurations.

## 840 **5. Conclusion**

841 This study employed Differential Split Sample Testing (DSST) to scrutinize the temporal  
842 transferability of six conceptual rainfall runoff models based on contrasting two/three year  
843 non-continuous periods. Using 37 Irish catchments with diverse hydrological regimes, model  
844 performance was assessed when transferred between the wettest/driest years on record and  
845 between contrasting wet/dry seasonal combinations. The study also considered the benefits of  
846 employing combined model estimates derived from four different ensemble averaging  
847 techniques.

848 Overall, HBV, GR4J and to a lesser extent NAM were consistently the best performing  
849 models, with HBV (GR4J) generally ranking highest for catchments with a lower (higher)  
850 groundwater contribution. Transferability of individual structures was found to vary  
851 depending on the DSST scenario, catchment and testing criteria used. The greatest declines in  
852 performance were associated with transference to drier conditions, with the extent of decline  
853 dependent on the performance criterion used.

854 The results confirm that it is impossible to identify a single structure that performs optimally  
855 across all catchments, DSST scenarios and performance criteria. Moreover, the collective  
856 ensemble was shown to outperform the majority of individual ensemble members. However,  
857 averaging methods were found to differ considerably with respect to the frequency with  
858 which they surpass the best individual member, particularly for volumetric errors. Bayesian  
859 Model Averaging (BMA) and the Granger-Ramanathan method (GRA) were found to  
860 perform better under transference than using the Simple Arithmetic Mean (SAM) and Akaike  
861 Information Criteria Averaging (AICA). Further work could be done on the potential added  
862 value of using different variants of GRA including non-constrained weights and a bias

863 correction step, as well as the transferability of averaging techniques that implement dynamic  
864 weighting [*See and Openshaw, 2000; Hu et al., 2001; Wagener et al., 2003*].

865 Given that the historical record may not provide sufficient analogues to represent the  
866 plausible range of projected climate changes, it is likely that the predictive errors from DSST  
867 will be underestimated and the demand for models to offer functional simulations under  
868 increasingly different conditions will almost certainly be greater than can be captured here. It  
869 is noted that we only examined performance based on mean seasonal/annual conditions.  
870 Other objective functions could be used to test model performance under extreme high or low  
871 flows (which may be of greater interest to decision-makers than average flow conditions).

872 Moreover, there is scope to develop an expanded DSST methodology that incorporates an  
873 assessment of extremes, particularly as transferability at seasonal/annual timescales may  
874 mask performance with respect to exact non-stationarities in the intensity and occurrence of  
875 extreme events. Similarly, while we focus on precipitation, it may be helpful to consider  
876 using other climate variables (e.g. temperature, evaporation, wind speed, cloud cover) when  
877 selecting contrasting periods of record for model training and transference testing [e.g. *Seiller*  
878 *et al., 2012; 2015*]. This may be particularly pertinent in regions where evapotranspiration  
879 and/or snow-melt presently play a greater role, or where climate scenarios suggest that such  
880 drivers are likely to become more/less significant in the future.

881 In closing, we emphasise that the predictive skill of hydrological models under different  
882 climate conditions should be considered routinely, particularly when results are used to  
883 inform adaptation decision making. Thus, it is important that codes of good practice are  
884 established to ensure models are applied in consistent and appropriate ways. On the basis of  
885 our findings, we offer the following five recommendations:

- 886 1. Clearly articulate the objectives of the climate assessment; these will define the  
887 options in the next four choices (below).
- 888 2. Set up the DSST to select the best available analogues of expected annual mean,  
889 seasonal mean, or sub-seasonal (extreme) climate conditions for model training and  
890 evaluation, depending on the study objectives.
- 891 3. Apply multiple performance criteria that are pertinent to the study objectives when  
892 assessing the transferability of model parameters between contrasting climate  
893 conditions; do not rely on a single performance metric.
- 894 4. Test parameter transferability using a range of catchment types to better appreciate the  
895 form(s) of hydroclimatic regime that are simulated with more or less reliability by a  
896 given model, and for the specified objective function(s).
- 897 5. Use a multi-model ensemble in conjunction with an objectively based averaging  
898 technique – ideally BMA or GRA – to obtain the most reliable estimate of future river  
899 flow under a changing climate.

900  
901  
902  
903

## 904 Acknowledgements

905 We thank each of the data providers for access to precipitation, PET and river flow data. CM  
906 and CB acknowledge funding provided by the Irish Environmental Protection Agency under  
907 project 2014-CCRP-MS.16. We thank Katie Smith and Christel Prudhomme of CEH  
908 Wallingford for valuable feedback. The thoughtful comments of three reviewers improved  
909 the paper considerably.

## 910 References

- 911 Abrahart, R. J., and L. See (2002), Multi-model data fusion for river flow forecasting: an  
912 evaluation of six alternative methods based on two contrasting catchments, *Hydrol.*  
913 *Earth Syst. Sci.*, 6(4), 655–670, doi: 10.5194/hess-6-655-2002.
- 914 Ajami, N. K., Q. Duan, X. Gao, and S. Sorooshian (2006), Multimodel combination  
915 techniques for analysis of hydrological simulations: Application to distributed model  
916 intercomparison project results, *J. Hydrometeorol.*, 7(4), 755–768,  
917 doi:10.1175/JHM519.1.
- 918 Akaike, H. (1974), A New Look at the Statistical Model Identification, in *Selected Papers of*  
919 *Hirotsugu Akaike*, edited by E. Parzen, K. Tanabe, and G. Kitagawa, pp. 215–222,  
920 Springer New York.
- 921 Allen, R. G., L. S. Pereira, D. Raes, and M. Smith (1998), FAO Irrigation and drainage paper  
922 No. 56, *Rome: Food and Agriculture Organization of the United Nations*, 56, 97–156.
- 923 Anctil, F., C. Perrin, and V. Andréassian (2004), Impact of the length of observed records on  
924 the performance of ANN and of conceptual parsimonious rainfall-runoff forecasting  
925 models, *Environ. Model. Softw.*, 19(4), 357–368, doi:10.1016/S1364-8152(03)00135-  
926 X.
- 927 Andréassian, V., C. Perrin, L. Berthet, N. Le Moine, J. Lerat, C. Loumagne, L. Oudin, T.  
928 Mathevet, M. H. Ramos, and A. Valéry (2009), Crash tests for a standardized  
929 evaluation of hydrological models, *Hydrol. Earth Syst. Sci.*, 13(10), 1757–1764,  
930 doi:10.5194/hess-13-1757-2009.
- 931 Arsenault, R., P. Gatien, B. Renaud, F. Brissette, and J.-L. Martel (2015), A comparative  
932 analysis of 9 multi-model averaging approaches in hydrological continuous  
933 streamflow simulation, *J. Hydrol.*, 529, 754–767, doi:10.1016/j.jhydrol.2015.09.001.
- 934 Atkinson, S. E., R. A. Woods and M. Sivapalan (2002), Climate and landscape controls on  
935 water balance model complexity over changing timescales, *Water Resour. Res.*,  
936 38(12), 1314, doi:10.1029/2002WR001487.
- 937 Bastola, S., C. Murphy, and J. Sweeney (2011), The role of hydrological modelling  
938 uncertainties in climate change impact assessments of Irish river catchments, *Adv.*  
939 *Water Resour.*, 34(5), 562–576, doi:10.1016/j.advwatres.2011.01.008.
- 940 Bastola, S., C. Murphy, and R. Fealy (2012), Generating probabilistic estimates of  
941 hydrological response for Irish catchments using a weather generator and probabilistic  
942 climate change scenarios: Probabilistic-based estimates of climate change effects,  
943 *Hydrol. Process.*, 26(15), 2307–2321, doi:10.1002/hyp.8349.
- 944 Beven, K. (2006), A manifesto for the equifinality thesis, *J. Hydrol.*, 320(1–2), 18–36,  
945 doi:10.1016/j.jhydrol.2005.07.007.
- 946 Beven, K., and A. Binley (1992), The future of distributed models: Model calibration and  
947 uncertainty prediction, *Hydrol. Process.*, 6(3), 279–298,  
948 doi:10.1002/hyp.3360060305.
- 949 Blasone, R.-S., J. A. Vrugt, H. Madsen, D. Rosbjerg, B. A. Robinson, and G. A. Zyvoloski  
950 (2008), Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov

951 Chain Monte Carlo sampling, *Adv. Water Resour.*, 31(4), 630–648,  
952 doi:10.1016/j.advwatres.2007.12.003.

953 Bloomfield, J.P., D.J. Allen, and K.J. Griffiths (2009), Examining geological controls on  
954 baseflow index (BFI) using regression analysis: An illustration from the Thames  
955 Basin, UK, *J. Hydrol.*, 373(1-2), 164–176, doi:10.1016/j.jhydrol.2009.04.025

956 Boughton, W. (2004), The Australian water balance model, *Environ. Model. Softw.*, 19(10),  
957 943–956, doi:10.1016/j.envsoft.2003.10.007.

958 Brigode, P., L. Oudin, and C. Perrin (2013), Hydrological model parameter instability: A  
959 source of additional uncertainty in estimating the hydrological impacts of climate  
960 change?, *J. Hydrol.*, 476, 410–425, doi:10.1016/j.jhydrol.2012.11.012.

961 Buckland, S. T., K. P. Burnham, and N. H. Augustin (1997), Model Selection: An Integral  
962 Part of Inference, *Biometrics*, 53(2), 603–618, doi:10.2307/2533961.

963 Burnham, K. P., and D. R. Anderson (2003), *Model Selection and Multimodel Inference: A*  
964 *Practical Information-Theoretic Approach*, Springer Science & Business Media,  
965 Springer, New York.

966 Chiverton, A., J. Hannaford, I. Holman, R. Corstanje, C. Prudhomme, J. Bloomfield, and T.  
967 M. Hess (2015), Which catchment characteristics control the temporal dependence  
968 structure of daily river flows?, *Hydrol. Process.*, 29(6), 1353–1369,  
969 doi:10.1002/hyp.10252.

970 Choi, H. T., and K. Beven (2007), Multi-period and multi-criteria model conditioning to  
971 reduce prediction uncertainty in an application of TOPMODEL within the GLUE  
972 framework, *J. Hydrol.*, 332(3–4), 316–336, doi:10.1016/j.jhydrol.2006.07.012.

973 Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener,  
974 and L. E. Hay (2008), Framework for Understanding Structural Errors (FUSE): A  
975 modular framework to diagnose differences between hydrological models, *Water*  
976 *Resour. Res.*, 44(12), W00B02, doi:10.1029/2007WR006735.

977 Clark, M.P., R.L., Wilby, E.D., Gutmann, et al. (2016) Characterizing Uncertainty of the  
978 Hydrologic Impacts of Climate Change, *Curr. Clim. Change. Rep.* 2(2), 55-64,  
979 doi:10.1007/s40641-016-0034-x.

980 Coron, L., V. Andréassian, C. Perrin, J. Lerat, J. Vaze, M. Bourqui, and F. Hendrickx (2012),  
981 Crash testing hydrological models in contrasted climate conditions: An experiment on  
982 216 Australian catchments, *Water Resour. Res.*, 48(5), W05552,  
983 doi:10.1029/2011WR011721.

984 Coxon, G., J. Freer, T. Wagener, N. A. Odoni, and M. Clark (2014), Diagnostic evaluation of  
985 multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework  
986 for 24 UK catchments, *Hydrol. Process.*, 28(25), 6135–6150, doi:10.1002/hyp.10096.

987 Diks, C. G. H., and J. A. Vrugt (2010), Comparison of point forecast accuracy of model  
988 averaging methods in hydrologic applications, *Stoch. Environ. Res. Risk Assess.*,  
989 24(6), 809–820, doi:10.1007/s00477-010-0378-z.

990 van Esse, W. R., C. Perrin, M. J. Booij, D. C. M. Augustijn, F. Fenicia, D. Kavetski, and F.  
991 Lobligeois (2013), The influence of conceptual model structure on model  
992 performance: a comparative study for 237 French catchments, *Hydrol. Earth Syst.*  
993 *Sci.*, 17(10), 4227–4239, doi:10.5194/hess-17-4227-2013.

994 Euser, T., H. C. Winsemius, M. Hrachowitz, F. Fenicia, S. Uhlenbrook, and H. H. G.  
995 Savenije (2013), A framework to assess the realism of model structures using  
996 hydrological signatures, *Hydrol. Earth Syst. Sci.*, 17(5), 1893–1912, doi:10.5194/hess-  
997 17-1893-2013.

998 Fenicia, F., D. Kavetski, and H. H. G. Savenije (2011), Elements of a flexible approach for  
999 conceptual hydrological modeling: 1. Motivation and theoretical development, *Water*  
1000 *Resour. Res.*, 47(11), W11510, doi:10.1029/2010WR010174.

1001 Fenicia, F., D. Kavetski, H.H.G. Savenije, M.P. Clark, G. Schoups, L. Pfister, and J. Freer  
1002 (2014), Catchment properties, function, and conceptual model representation: is there  
1003 a correspondence?, *Hydrol. Process.*, 28(4): 2451–2467. doi:10.1002/hyp.9726.

1004 Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez (2009), Decomposition of the mean  
1005 squared error and NSE performance criteria: Implications for improving hydrological  
1006 modelling, *J. Hydrol.*, 377(1–2), 80–91, doi:10.1016/j.jhydrol.2009.08.003.

1007 Gustard, A., A. Bullock, and J.M., Dixon (1992), Low flow estimation in the United  
1008 Kingdom. Report No. 108., Institute of Hydrology, Wallingford, United Kingdom.

1009 Hannaford, J., and T. J. Marsh (2008), High-flow and flood trends in a network of  
1010 undisturbed catchments in the UK, *Int. J. Climatol.*, 28(10), 1325–1338,  
1011 doi:10.1002/joc.1643.

1012 Hansen, B. E. (2008), Least-squares forecast averaging, *J. Econometrics*, 146(2), 342–350,  
1013 doi:10.1016/j.jeconom.2008.08.022.

1014 Hartmann, G., and A. Bárdossy (2005), Investigation of the transferability of hydrological  
1015 models and a method to improve model calibration, *Adv. Geosci.*, 5, 83–87.

1016 Herman, J. D., P. M. Reed, and T. Wagener (2013), Time-varying sensitivity analysis  
1017 clarifies the effects of watershed model formulation on model behavior, *Water*  
1018 *Resour. Res.*, 49(3), 1400–1414, doi:10.1002/wrcr.20124.

1019 Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999), Bayesian model  
1020 averaging: a tutorial, *Statistical science*, 14(4), 382–401, doi:  
1021 doi:10.1214/ss/1009212519.

1022 Hu, T. S., K. C. Lam, and S. T. Ng (2001), River flow time series prediction with a range-  
1023 dependent neural network, *Hydrol. Sci. J.*, 46(5), 729–745, doi:  
1024 10.1080/02626660109492867.

1025 Jackson, C. R. (1992), Hillslope infiltration and lateral downslope unsaturated flow, *Water*  
1026 *Resour. Res.*, 28(9), 2533–2539, doi:10.1029/92WR00664.

1027 Klemeš, V. (1986), Operational testing of hydrological simulation models, *Hydrol. Sci. J.*,  
1028 31(1), 13–24, doi:10.1080/02626668609491024.

1029 Krause, P., D. P. Boyle, and F. Bäse (2005), Comparison of different efficiency criteria for  
1030 hydrological model assessment, *Adv. Geosci.*, 5, 89–97, doi: 10.5194/adgeo-5-89-  
1031 2005.

1032 Li, C. Z., L. Zhang, H. Wang, Y. Q. Zhang, F. L. Yu, and D. H. Yan (2012), The  
1033 transferability of hydrological models under nonstationary climatic conditions,  
1034 *Hydrol. Earth Syst. Sci.*, 16(4), 1239–1254, doi:10.5194/hess-16-1239-2012.

1035 Madsen, H. (2000), Automatic calibration of a conceptual rainfall–runoff model using  
1036 multiple objectives, *J. Hydrol.*, 235(3), 276–288, doi: 10.1016/S0022-1694(00)00279-  
1037 1.

1038 Matthews, T., D. Mullan, R. L. Wilby, C. Broderick, and C. Murphy (2016), Past and future  
1039 climate change in the context of memorable seasonal extremes, 11, 37–52, *Clim. Risk*  
1040 *Man.*, doi:10.1016/j.crm.2016.01.004.

1041 McKay, M., R. Beckman, and W. Conover (1979), A Comparison of Three Methods for  
1042 Selecting Values of Input Variables in the Analysis of Output from a Computer Code.  
1043 *Technometrics*, 21(2), 239–245, doi: 10.2307/1268522.

1044 Mertens, J., H. Madsen, L. Feyen, D. Jacques, and J. Feyen (2004), Including prior  
1045 information in the estimation of effective soil parameters in unsaturated zone  
1046 modelling, *J. Hydrol.*, 294(4), 251–269, doi:10.1016/j.jhydrol.2004.02.011.

1047 Merz, R., J. Parajka, and G. Blöschl (2011), Time stability of catchment model parameters:  
1048 Implications for climate impact analyses, *Water Resour. Res.*, 47(2), W02531,  
1049 doi:10.1029/2010WR009505.

1050 Murphy, A. (2012), Snowfall in Ireland, Met Éireann, Glasnevin Hill, Dublin 9, Ireland.

- 1051 Murphy, C., S. Harrigan, J. Hall, and R. L. Wilby (2013), Climate-driven trends in mean and  
1052 high flows from a network of reference stations in Ireland, *Hydrol. Sci. J.*, 58(4), 755–  
1053 772, doi:10.1080/02626667.2013.782407.
- 1054 Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part  
1055 I — A discussion of principles, *J. Hydrol.*, 10(3), 282–290, doi:10.1016/0022-  
1056 1694(70)90255-6.
- 1057 Oudin, L., V. Andréassian, T. Mathevet, C. Perrin, and C. Michel (2006a), Dynamic  
1058 averaging of rainfall-runoff model simulations from complementary model  
1059 parameterizations: Dynamic averaging of rainfall-runoff models, *Water Resour. Res.*,  
1060 42(7), W07410, doi:10.1029/2005WR004636.
- 1061 Oudin, L., C. Perrin, T. Mathevet, V. Andréassian, and C. Michel (2006b), Impact of biased  
1062 and randomly corrupted inputs on the efficiency and the parameters of watershed  
1063 models, *J. Hydrol.*, 320(1–2), 62–83, doi:10.1016/j.jhydrol.2005.07.016.
- 1064 Oudin, L., V. Andréassian, C. Perrin, C. Michel, and N. Le Moine (2008), Spatial proximity,  
1065 physical similarity, regression and ungauged catchments: A comparison of  
1066 regionalization approaches based on 913 French catchments, *Water Resour. Res.*,  
1067 44(3), W03413, doi:10.1029/2007WR006240.
- 1068 Parajka, J., A. Viglione, M. Rogger, J. L. Salinas, M. Sivapalan, and G. Blöschl (2013),  
1069 Comparative assessment of predictions in ungauged basins – Part 1: Runoff-  
1070 hydrograph studies, *Hydrol. Earth Syst. Sci.*, 17(5), 1783–1795, doi:10.5194/hess-17-  
1071 1783-2013.
- 1072 Perrin, C., C. Michel, and V. Andréassian (2001), Does a large number of parameters  
1073 enhance model performance? Comparative assessment of common catchment model  
1074 structures on 429 catchments, *J. Hydrol.*, 242(3–4), 275–301, doi:10.1016/S0022-  
1075 1694(00)00393-0.
- 1076 Perrin, C., C. Michel, and V. Andréassian (2003), Improvement of a parsimonious model for  
1077 streamflow simulation, *J. Hydrol.*, 279(1–4), 275–289, doi:10.1016/S0022-  
1078 1694(03)00225-7.
- 1079 Price, K. (2011), Effects of watershed topography, soils, land use, and climate on baseflow  
1080 hydrology in humid regions: A review, *Prog. Phys. Geogr.*, 35(4), 465–492, doi:  
1081 10.1177/0309133311402714.
- 1082 Prudhomme, C., R. L. Wilby, S. Crooks, A. L. Kay, and N. S. Reynard (2010), Scenario-  
1083 neutral approach to climate change impact studies: Application to flood risk, *J.*  
1084 *Hydrol.*, 390(3–4), 198–209, doi:10.1016/j.jhydrol.2010.06.043.
- 1085 Prudhomme, C., E. Sauquet, and G. Watts (2015), Low Flow Response Surfaces for Drought  
1086 Decision Support: A Case Study from the UK, *J. of Extr. Even.*, 2(2), 1550005,  
1087 doi:10.1142/S2345737615500050.
- 1088 Pushpalatha, R., C. Perrin, N. Le Moine, T. Mathevet, and V. Andréassian, (2011), A  
1089 downward structural sensitivity analysis of hydrological models to improve low-flow  
1090 simulation, *J. Hydrol.*, 411(1-2), 66–76, doi: 10.1016/j.jhydrol.2011.09.034
- 1091 Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model  
1092 averaging to calibrate forecast ensembles, *Mon. Weather Rev.* 133(5), 1155–1174,  
1093 doi: 10.1175/MWR2906.1.
- 1094 Rojas, R., L. Feyen, and A. Dassargues (2008), Conceptual model uncertainty in groundwater  
1095 modeling: Combining generalized likelihood uncertainty estimation and Bayesian  
1096 model averaging, *Water Resour. Res.*, 44(12), W12418, doi:10.1029/2008WR006908.
- 1097 Samuel, J., P. Coulibaly, and R.A. Metcalfe (2012), Identification of rainfall–runoff model  
1098 for improved baseflow estimation in ungauged basins. *Hydrol. Process.*, 26(3), 356–  
1099 366. doi:10.1002/hyp.8133

- 1100 Sear, D. A., P. D., Armitage and F. H. Dawson (1999), Groundwater dominated rivers.  
1101 *Hydrol. Process.*, 13(3), 255–276. doi:10.1002/(SICI)1099-1085(19990228)
- 1102 See, L., and S. Openshaw (2000), A hybrid multi-model approach to river level forecasting,  
1103 *Hydrol. Sci. J.*, 45(4), 523–536, doi:10.1080/02626660009492354.
- 1104 Seibert, J. (1996), HBV light, User’s manual, Uppsala University, Institute of Earth Science,  
1105 Department of Hydrology, Uppsala.
- 1106 Seiller, G., F. Anctil, and C. Perrin (2012), Multimodel evaluation of twenty lumped  
1107 hydrological models under contrasted climate conditions, *Hydrol. Earth Syst. Sci.*,  
1108 16(4), 1171–1189, doi:10.5194/hess-16-1171-2012.
- 1109 Seiller, G., I. Hajji, and F. Anctil (2015), Improving the temporal transposability of lumped  
1110 hydrological models on twenty diversified U.S. watersheds, *J. Hydrol.: Regional  
1111 Studies*, 3, 379–399, doi:10.1016/j.ejrh.2015.02.012.
- 1112 Shafii, M., and B. A. Tolson (2015), Optimizing hydrological consistency by incorporating  
1113 hydrological signatures into model calibration objectives: Hydrological consistency  
1114 optimization, *Water Resour. Res.*, 51(5), 3796–3814, doi:10.1002/2014WR016520.
- 1115 Shamseldin, A. Y., K. M. O’Connor, and G. C. Liang (1997), Methods for combining the  
1116 outputs of different rainfall–runoff models, *J. Hydrol.*, 197(1–4), 203–229,  
1117 doi:10.1016/S0022-1694(96)03259-3.
- 1118 Smith, R. E., and R. H. B. Hebbert (1983), Mathematical simulation of interdependent  
1119 surface and subsurface hydrologic processes, *Water Resour. Res.*, 19(4), 987–1001,  
1120 doi:10.1029/WR019i004p00987.
- 1121 Staudinger, M., K. Stahl, J. Seibert, M. P. Clark, and L. M. Tallaksen (2011), Comparison of  
1122 hydrological model structures based on recession and low flow simulations, *Hydrol.  
1123 Earth Syst. Sci.*, 15(11), 3447–3459, doi:10.5194/hess-15-3447-2011.
- 1124 Steele-Dunne, S., P. Lynch, R. McGrath, T. Semmler, S. Wang, J. Hanafin, and P. Nolan  
1125 (2008), The impacts of climate change on hydrology in Ireland, *J. Hydrol.*, 356(1–2),  
1126 28–45, doi:10.1016/j.jhydrol.2008.03.025.
- 1127 Sugawara, M. (1995), Tank Model, in *Computer models of watershed hydrology*, edited by  
1128 V. P. Singh, pp. 165–214, Water Resources Publications, Littleton, Colorado.
- 1129 Sweeney, J. (2014), Regional weather and climates of the British Isles – Part 6: Ireland,  
1130 *Weather*, 69(1), 20–27, doi:10.1002/wea.2230.
- 1131 Thiboult, A., F. Anctil, and M. A. Boucher (2016), Accounting for three sources of  
1132 uncertainty in ensemble hydrological forecasting, *Hydrol. Earth Syst. Sci.*, 20(5),  
1133 1809–1825, doi:10.5194/hess-20-1809-2016.
- 1134 Thirel, G. et al. (2015a), Hydrology under change: an evaluation protocol to investigate how  
1135 hydrological models deal with changing catchments, *Hydrol. Sci. J.*, 60(7–8), 1184–  
1136 1199, doi:10.1080/02626667.2014.967248.
- 1137 Thirel, G., V. Andréassian, and C. Perrin (2015b), On the need to test hydrological models  
1138 under changing conditions, *Hydrol. Sci. J.*, 60(7–8), 1165–1173,  
1139 doi:10.1080/02626667.2015.1050027.
- 1140 Uhlenbrook, S., J. Seibert, C. Leibundgut, and A. Rodhe (1999), Prediction uncertainty of  
1141 conceptual rainfall–runoff models caused by problems in identifying model  
1142 parameters and structure, *Hydrol. Sci. J.*, 44(5), 779–797,  
1143 doi:10.1080/02626669909492273.
- 1144 Vaze, J., D. A. Post, F. H. S. Chiew, J.-M. Perraud, N. R. Viney, and J. Teng (2010), Climate  
1145 non-stationarity – Validity of calibrated rainfall–runoff models for use in climate  
1146 change studies, *J. Hydrol.*, 394(3–4), 447–457, doi:10.1016/j.jhydrol.2010.09.018.
- 1147 Velázquez, J. A., F. Anctil, and C. Perrin (2010), Performance and reliability of multimodel  
1148 hydrological ensemble simulations based on seventeen lumped models and a thousand

- 1149 catchments, *Hydrol. Earth Syst. Sci.*, 14(11), 2303–2317, doi:10.5194/hess-14-2303-  
1150 2010.
- 1151 Velázquez, J. A., F. Anctil, M. H. Ramos, and C. Perrin (2011), Can a multi-model approach  
1152 improve hydrological ensemble forecasting? A study on 29 French catchments using  
1153 16 hydrological model structures, *Adv. Geosci.*, 29, 33–42, doi:10.5194/adgeo-29-33-  
1154 2011.
- 1155 Vrugt, J. A., C. G. H. Diks, and M. P. Clark (2008), Ensemble Bayesian model averaging  
1156 using Markov Chain Monte Carlo sampling, *Environ. Fluid. Mech.*, 8(5–6), 579–595,  
1157 doi:10.1007/s10652-008-9106-3.
- 1158 Wagener, T. (2003), Evaluation of catchment models, *Hydrol. Process.*, 17(16), 3375–3378,  
1159 doi:10.1002/hyp.5158.
- 1160 Wagener, T., D. P. Boyle, M. J. Lees, H. S. Wheater, H. V. Gupta, and S. Sorooshian (2001),  
1161 A framework for development and application of hydrological models, *Hydrol. Earth*  
1162 *Syst. Sci. Discussions*, 5(1), 13–26.
- 1163 Wagener, T., N. McIntyre, M. J. Lees, H. S. Wheater, and H. V. Gupta (2003), Towards  
1164 reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability  
1165 analysis, *Hydrol. Process.*, 17(2), 455–476, doi:10.1002/hyp.1135.
- 1166 Walsh, S. (2012), Long term rainfall averages for Ireland, in *National Hydrology Seminar*  
1167 *2012*, Office of Public Works, Tullamore, Ireland.
- 1168 Westra, S., M. Thyer, M. Leonard, D. Kavetski, and M. Lambert (2014), A strategy for  
1169 diagnosing and interpreting hydrological model nonstationarity, *Water Resour. Res.*,  
1170 50(6), 5090–5113, doi:10.1002/2013WR014719.
- 1171 Whateley, S., S. Steinschneider, and C. Brown (2014), A climate change range-based method  
1172 for estimating robustness for water resources supply, *Water Resour. Res.*, 50(11),  
1173 8944–8961, doi:10.1002/2014WR015956.
- 1174 Wilby, R. L. (2005), Uncertainty in water resource model parameters used for climate change  
1175 impact assessment, *Hydrol. Process.*, 19(16), 3201–3219, doi:10.1002/hyp.5819.
- 1176 Wilby, R. L., and I. Harris (2006), A framework for assessing uncertainties in climate change  
1177 impacts: Low-flow scenarios for the River Thames, UK, *Water Resour. Res.*, 42(2),  
1178 W02419, doi:10.1029/2005WR004065.
- 1179 Wilby, R. L., C. W. Dawson, C. Murphy, P. OConnor, and E. Hawkins (2014), The Statistical  
1180 DownScaling Model - Decision Centric (SDSM-DC): conceptual basis and  
1181 applications, *Clim Res*, 61(3), 259–276, doi:10.3354/cr01254.
- 1182 Wilby, R. L., S. Noone, C. Murphy, T. Matthews, S. Harrigan, and C. Broderick (2015a), An  
1183 evaluation of persistent meteorological drought using a homogeneous Island of  
1184 Ireland precipitation network, *Int. J. Climatol.*, 36(8), 2854–2865,  
1185 doi:10.1002/joc.4523.
- 1186 Wilby, R. L., C. Prudhomme, S. Parry, and K. G. L. Muchan (2015b), Persistence of  
1187 Hydrometeorological Droughts in the United Kingdom: A Regional Analysis of  
1188 Multi-Season Rainfall and River Flow Anomalies, *J. of Extr. Even.*, 1550006,  
1189 doi:10.1142/S2345737615500062.
- 1190 Yapo, P. O., H. V. Gupta, and S. Sorooshian (1996), Automatic calibration of conceptual  
1191 rainfall-runoff models: sensitivity to calibration data, *J. Hydrol.*, 181(1–4), 23–48,  
1192 doi:10.1016/0022-1694(95)02918-4.

1194

1195

1196 Table 1. Hydroclimatic and physical descriptors for the 37 selected catchments. Flow indices are estimated from  
 1197 daily data for the period 1974-2010. The Base Flow Index (BFI) is calculated according to *Gustard et al.*, [1992].  
 Mean annual (hydrological year) and six-month winter/summer (ONDJFM/AMJJAS) precipitation totals for the  
 period 1976-2005 are shown.

Gauge ID	Area (km <sup>2</sup> )	Mean Elevation (m)	BFI (-)	Runoff (mm yr <sup>-1</sup> )	Start date	Precipitation (mm) 1976-2005		
						Annual	Winter	Summer
6013	308	84	0.60	432	Jul-75	881	497	384
6014	270	84	0.61	510	Jun-75	919	526	393
7009	1683	85	0.70	471	Jan-73	890	496	393
7012	2460	91	0.68	491	Jan-73	908	508	400
12001	1031	161	0.69	650	Jan-73	1095	632	463
14007	114	136	0.62	538	Jan-73	915	520	395
14019	1702	94	0.65	417	Oct-81	868	486	382
15001	444	118	0.52	500	Jan-73	971	559	413
15003	297	209	0.38	634	Oct-73	1027	584	443
15006	2417	137	0.62	528	Dec-76	975	558	417
16008	1091	138	0.63	702	May-72	1037	606	431
16009	1583	139	0.64	656	Jan-73	1078	632	445
18002	2329	165	0.62	807	Jul-77	1267	773	495
18003	1257	181	0.54	873	Jan-73	1357	845	511
18005	378	158	0.71	725	Jan-73	1189	699	491
18006	1055	188	0.50	975	Jan-73	1379	862	517
18050	250	210	0.38	1073	Jan-72	1588	999	589
19001	103	100	0.59	744	May-81	1236	753	483
21002	66	247	0.21	2031	Jan-73	2277	1422	855
23002	647	196	0.28	1082	Oct-75	1443	880	563
25001	647	153	0.53	758	Jan-73	1185	679	505
25002	222	190	0.48	854	Oct-75	1291	742	550
25006	1188	89	0.69	460	Jan-73	922	515	406
25030	278	136	0.54	918	Feb-80	1196	703	493
26009	90	91	0.43	570	Jan-73	1065	609	456
26021	1072	90	0.82	559	Jan-73	967	547	420
26029	117	217	0.23	1308	Jan-73	1569	923	646
27002	511	73	0.70	651	Jan-73	1319	787	532
32012	145	131	0.56	1285	Jan-73	1690	1027	663
34001	1971	81	0.77	907	Jan-73	1334	811	523
35002	76	198	0.40	1352	Jan-73	1631	984	647
35005	639	100	0.63	820	Jan-73	1268	747	521
36010	771	124	0.60	580	Jan-73	1028	584	444
38001	111	186	0.26	1528	Nov-76	1899	1140	759
39006	245	131	0.46	1129	Jan-73	1530	929	601
201005	277	163	0.47	793	Jan-74	1141	649	492
201008	335	172	0.32	1340	Jan-73	1676	1007	668

1198

1199

1200

1201

Table 2. Structural components of the six lumped conceptual rainfall-runoff models. Routing mechanisms are abbreviated as unit hydrograph (uh), non-linear store (nls) and linear store (ls) respectively.

Model	Number of free parameters	Represented catchment stores	Represented flow component / routing mechanism
NAM	9	surface; root zone; groundwater	overland (ls); interflow (ls); baseflow (ls)
HyMod	5	soil; 'quick' flow reservoirs (×3); 'slow' groundwater	overland (three ls in series); baseflow (single ls in parallel)
Tank	15	soil; intermediate (upper and lower); groundwater	sum of lateral outflow from each model store
HBV	9	soil; lower soil; groundwater	triangular weighting of combined lateral outflow from the lower soil and groundwater store
GR4J	4	production; routing	10:90 split between direct (uh) and delayed (using a uh and single routing nls) routing
AWBM	10	variable soil surface stores (×3); surface runoff; groundwater store	overland (ls); baseflow (ls)

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

Table 3. The DSST scenario and model associated with the greatest singular decrease in performance under transference between ‘wet/dry’ annual regimes. Differences are estimated using performance under control conditions as a benchmark (i.e. control *versus* testing). Percent (% $\Delta$ ; NSE, NSE<sub>cubrt</sub>) and absolute ( $\Delta$ ; PBIAS) differences are given. PBIAS values in bold denote an underestimation of the total observed flow under transference (e.g. W/D). Values underlined indicate that models trained under dissimilar conditions both (under/over) estimate the total volume.

ID	BFI	Scenario	Model	% $\Delta$	Scenario	Model	% $\Delta$	Scenario	Model	$\Delta$
		NSE		NSE <sub>cubrt</sub>		PBIAS				
6013	0.60	D/W	HyMod	-2.8	W/D	AWBM	-1.5	W/D	AWBM	<b>-4.4</b>
6014	0.61	D/W	HBV	-5.0	W/D	AWBM	-4.8	W/D	AWBM	<b>-4.8</b>
7009	0.70	D/W	Tank	-3.8	W/D	AWBM	-6.6	W/D	AWBM	<b>-4.6</b>
7012	0.68	D/W	HBV	-14.0	W/D	AWBM	-21.6	W/D	GR4J	<b>-11.3</b>
12001	0.69	D/W	NAM	-3.8	W/D	AWBM	-6.1	W/D	HBV	<b>-7.9</b>
14007	0.62	W/D	Tank	-5.0	D/W	Tank	-5.6	D/W	GR4J	<b>-10.1</b>
14019	0.65	D/W	Tank	-1.0	D/W	Tank	-0.9	D/W	GR4J	<b>-4.1</b>
15001	0.52	D/W	HyMod	-3.6	W/D	AWBM	-5.2	W/D	GR4J	<b>-7.3</b>
15003	0.38	W/D	GR4J	-5.3	W/D	AWBM	-7.5	D/W	AWBM	<u>-10.5</u>
15006	0.62	W/D	GR4J	-3.6	W/D	AWBM	-9.4	W/D	GR4J	<b>-9.9</b>
16008	0.63	D/W	HyMod	-8.7	W/D	HyMod	-7.0	D/W	GR4J	<b>-10.7</b>
16009	0.64	D/W	HyMod	-6.6	D/W	HyMod	-4.1	W/D	GR4J	<b>-9.5</b>
18002	0.62	D/W	HBV	-1.6	D/W	HyMod	-1.2	D/W	GR4J	<b>-4.6</b>
18003	0.54	W/D	Tank	-2.8	W/D	AWBM	-7.6	D/W	GR4J	<b>-9.6</b>
18005	0.71	D/W	NAM	-4.1	W/D	HyMod	-6.9	W/D	GR4J	<b>-8.4</b>
18006	0.50	W/D	GR4J	-14.6	W/D	AWBM	-20.6	W/D	AWBM	<b>-18.4</b>
18050	0.38	D/W	HBV	-4.3	W/D	AWBM	-6.3	W/D	HyMod	<u>-3.9</u>
19001	0.59	D/W	HyMod	-2.4	W/D	AWBM	-5.4	W/D	HBV	<b>-5.9</b>
21002	0.21	W/D	GR4J	-13.3	D/W	HyMod	-5.3	D/W	HyMod	-5.8
23002	0.28	W/D	GR4J	-6.1	D/W	NAM	-6.1	W/D	NAM	<b>-12.0</b>
25001	0.53	D/W	HyMod	-5.8	W/D	Tank	-10.3	D/W	GR4J	<b>-10.8</b>
25002	0.48	D/W	GR4J	-6.4	W/D	GR4J	-5.6	D/W	GR4J	<b>-13.3</b>
25006	0.69	D/W	NAM	-3.8	W/D	HyMod	-5.0	D/W	AWBM	<u>-5.3</u>
25030	0.54	D/W	HBV	-9.4	W/D	HyMod	-5.1	D/W	GR4J	<b>-7.6</b>
26009	0.43	W/D	GR4J	-5.5	W/D	AWBM	-6.8	W/D	GR4J	<b>-8.6</b>
26021	0.82	D/W	NAM	-4.0	W/D	AWBM	-5.3	D/W	GR4J	<b>-11.2</b>
26029	0.23	D/W	HyMod	-3.2	W/D	NAM	-2.7	W/D	Tank	<u>-3.5</u>
27002	0.70	D/W	NAM	-5.1	W/D	AWBM	-10.1	D/W	GR4J	<b>-11.9</b>
32012	0.56	W/D	AWBM	-5.4	W/D	HyMod	-18.0	W/D	HBV	<u>-10.2</u>
34001	0.77	W/D	Tank	-14.9	W/D	Tank	-5.5	D/W	GR4J	<b>-16.2</b>
35002	0.40	D/W	HyMod	-2.5	W/D	HyMod	-17.7	W/D	HBV	<b>-9.3</b>
35005	0.63	D/W	NAM	-7.1	W/D	HyMod	-12.5	W/D	HBV	<b>-4.2</b>
36010	0.60	D/W	Tank	-3.0	W/D	Tank	-2.5	W/D	HyMod	<u>-4.3</u>
38001	0.26	D/W	HyMod	-4.1	W/D	AWBM	-2.4	D/W	GR4J	<b>-5.6</b>
39006	0.46	D/W	NAM	-2.7	W/D	HBV	-7.3	D/W	GR4J	<b>-5.3</b>
201005	0.47	D/W	HBV	-1.5	W/D	HyMod	-1.4	D/W	GR4J	<b>-4.1</b>
201008	0.32	W/D	HBV	-10.9	D/W	AWBM	-7.4	W/D	HBV	<b>-12.2</b>

1224

1225

1226

Table 4. The DSST scenario and model associated with the greatest singular decrease in performance under transference between seasonal (DD, WW, DW and WD) precipitation regimes. Differences are estimated using performance under control conditions as a benchmark (i.e. control *versus* testing). Percent (% $\Delta$ ; NSE, NSE<sub>cubrt</sub>) and absolute ( $\Delta$ ; PBIAS) differences are given. PBIAS values in bold denote an underestimation of the total observed flow under transference (e.g. WD/DD). Values underlined indicate that models trained under dissimilar conditions both (under/over) estimate the total volume.

ID	BFI	Scenario	Model	% $\Delta$	Scenario	Model	% $\Delta$	Scenario	Model	$\Delta$
		NSE		NSE <sub>cubrt</sub>		PBIAS				
21002	0.21	DD/DW	GR4J	-5.19	WW/DW	AWBM	-2.42	DD/DW	GR4J	<b>-5.6</b>
26029	0.23	DD/WW	HBV	-6.91	WW/WD	AWBM	-5.58	WD/DD	HBV	<b>-7.0</b>
38001	0.26	WD/WW	GR4J	-8.26	WW/DW	AWBM	-13.37	DW/WW	GR4J	<b>-7.4</b>
23002	0.28	DD/DW	HyMod	-25.33	WW/DD	AWBM	-28.24	DD/DW	HBV	<b>-11.8</b>
201008	0.32	DW/DD	GR4J	-16.31	DW/DD	AWBM	-13.40	DW/DD	GR4J	<b>-16.0</b>
15003	0.38	DW/DD	Tank	-14.03	DD/DW	Tank	-14.50	DD/DW	GR4J	<b>-7.5</b>
18050	0.38	DW/WD	NAM	-5.45	DW/WD	NAM	-11.39	WD/DW	GR4J	<b>-11.1</b>
35002	0.4	DD/DW	HyMod	-6.04	DW/DD	HyMod	-5.24	WW/WD	GR4J	<b>-7.6</b>
26009	0.43	DD/DW	HyMod	-13.51	DW/DD	HyMod	-11.81	DD/DW	AWBM	-6.9
39006	0.46	WW/DD	GR4J	-4.72	WW/DD	AWBM	-12.59	WW/DD	GR4J	<b>-9.3</b>
201005	0.47	DD/DW	HyMod	-10.43	WD/DD	Tank	-13.39	DD/DW	GR4J	<b>-8.8</b>
25002	0.48	DD/WW	HyMod	-8.96	DD/WW	Tank	-6.89	DW/DD	GR4J	<b>-10.3</b>
18006	0.5	DD/WW	HBV	-5.07	DD/DW	GR4J	-7.08	DW/DD	GR4J	<b>-13.4</b>
15001	0.52	DW/DD	Tank	-19.84	DW/DD	HyMod	-16.03	DW/DD	HyMod	-24.2
25001	0.53	DW/WD	NAM	-6.98	DD/DW	Tank	-10.51	WW/DD	GR4J	<b>-7.3</b>
25030	0.54	WD/DD	GR4J	-27.62	WW/DD	AWBM	-22.82	WW/DD	GR4J	<b>-18.5</b>
18003	0.54	DD/DW	HBV	-3.23	DW/DD	AWBM	-10.49	WW/WD	GR4J	<b>-4.2</b>
32012	0.56	WD/DD	GR4J	-5.35	DW/DD	AWBM	-4.82	DW/DD	GR4J	<b>-7.1</b>
19001	0.59	DW/DD	HBV	-18.49	DD/DW	HBV	-16.03	DD/DW	GR4J	<b>-11.9</b>
6013	0.6	WW/DW	GR4J	-15.55	WD/DW	NAM	-14.64	WW/DD	HBV	<b>-18.9</b>
36010	0.6	DD/DW	GR4J	-14.22	DW/DD	HyMod	-17.89	DD/DW	GR4J	<b>-11.6</b>
6014	0.61	DD/DW	GR4J	-10.52	WW/DW	HyMod	-11.92	DD/DW	GR4J	<b>-14.4</b>
14007	0.62	DD/DW	HBV	-16.75	WW/DD	AWBM	-9.72	WD/DD	HyMod	<b>-14.7</b>
15006	0.62	DW/DD	Tank	-14.36	WD/DW	Tank	-13.29	DW/DD	HyMod	<b>-10.8</b>
18002	0.62	WW/DD	GR4J	-4.58	DW/DD	AWBM	-6.61	WW/WD	GR4J	<b>-7.2</b>
16008	0.63	DD/DW	GR4J	-13.74	WD/DW	NAM	-18.62	DD/DW	GR4J	<b>-18.5</b>
35005	0.63	DD/WD	NAM	-2.57	WD/WW	NAM	-3.56	DD/DW	GR4J	<b>-3.1</b>
16009	0.64	DD/WD	NAM	-8.03	DW/DD	AWBM	-20.08	DD/WW	GR4J	<b>-5.4</b>
14019	0.65	WD/DD	GR4J	-14.37	WW/DD	HyMod	-20.51	DW/WD	HyMod	-18.8
7012	0.68	DW/DD	Tank	-45.25	DW/DD	HyMod	-16.23	DW/DD	HyMod	-15.5
25006	0.69	DW/DD	Tank	-46.42	DW/DD	HyMod	-33.43	DW/WD	HyMod	-12.0
12001	0.69	DD/DW	GR4J	-30.05	DD/DW	GR4J	-31.64	DW/DD	GR4J	<b><u>-33.3</u></b>
27002	0.7	WD/DW	AWBM	-15.88	WD/DD	GR4J	-5.44	WD/DW	GR4J	<b><u>-4.6</u></b>
7009	0.7	WW/DW	GR4J	-11.35	DW/DD	HyMod	-6.05	WW/DD	HBV	<b>-7.2</b>
18005	0.71	WD/DD	GR4J	-36.39	WD/DD	GR4J	-29.16	WD/DD	GR4J	<b><u>-36.0</u></b>
34001	0.77	WD/DD	AWBM	-6.04	WD/DW	AWBM	-5.66	DD/WD	GR4J	<b>-5.9</b>
26021	0.82	DW/DD	GR4J	-27.16	DD/DW	HBV	-19.19	WD/DD	HBV	<b>-11.7</b>

1227

1228 Table 5. Frequency (%) with which each model averaging technique outperforms individual members of the  
 1229 model ensemble calculated for each DSST and training period. Results for the training and control periods are  
 listed in bold.

DSST	NSE				NSE <sub>cubrt</sub>				Absolute PBIAS			
	BMA	AICA	GRA	SAM	BMA	AICA	GRA	SAM	BMA	AICA	GRA	SAM
<b>D (training)</b>	<b>80</b>	<b>80</b>	<b>100</b>	<b>72</b>	<b>99</b>	<b>70</b>	<b>99</b>	<b>85</b>	<b>75</b>	<b>50</b>	<b>66</b>	<b>60</b>
<b>D/D</b>	<b>87</b>	<b>82</b>	<b>94</b>	<b>78</b>	<b>98</b>	<b>71</b>	<b>95</b>	<b>87</b>	<b>57</b>	<b>56</b>	<b>60</b>	<b>57</b>
W/D	89	74	94	81	97	63	92	89	60	54	66	55
<b>W (training)</b>	<b>85</b>	<b>72</b>	<b>100</b>	<b>85</b>	<b>100</b>	<b>75</b>	<b>99</b>	<b>91</b>	<b>58</b>	<b>51</b>	<b>77</b>	<b>60</b>
<b>W/W</b>	<b>89</b>	<b>76</b>	<b>96</b>	<b>82</b>	<b>99</b>	<b>70</b>	<b>97</b>	<b>90</b>	<b>55</b>	<b>54</b>	<b>67</b>	<b>64</b>
D/W	86	77	95	76	97	68	95	86	58	58	70	60
<b>DD (training)</b>	<b>80</b>	<b>68</b>	<b>100</b>	<b>82</b>	<b>99</b>	<b>70</b>	<b>98</b>	<b>85</b>	<b>68</b>	<b>52</b>	<b>65</b>	<b>55</b>
<b>DD/DD</b>	<b>82</b>	<b>70</b>	<b>90</b>	<b>81</b>	<b>90</b>	<b>65</b>	<b>90</b>	<b>82</b>	<b>64</b>	<b>87</b>	<b>68</b>	<b>52</b>
WD/DD	86	69	89	83	95	63	89	91	60	55	60	58
DW/DD	86	67	87	77	91	61	85	86	57	50	63	53
WW/DD	91	68	93	84	95	65	90	92	54	52	64	55
<b>WD (training)</b>	<b>84</b>	<b>82</b>	<b>100</b>	<b>80</b>	<b>99</b>	<b>69</b>	<b>97</b>	<b>79</b>	<b>57</b>	<b>49</b>	<b>75</b>	<b>65</b>
<b>WD/WD</b>	<b>89</b>	<b>86</b>	<b>95</b>	<b>77</b>	<b>80</b>	<b>71</b>	<b>95</b>	<b>80</b>	<b>55</b>	<b>50</b>	<b>69</b>	<b>61</b>
DD/WD	77	71	91	77	91	67	92	88	50	51	64	60
DW/WD	86	76	91	74	96	74	92	85	58	50	63	58
WW/WD	88	77	92	76	96	71	92	89	57	46	61	64
<b>WD (training)</b>	<b>85</b>	<b>80</b>	<b>100</b>	<b>78</b>	<b>100</b>	<b>75</b>	<b>98</b>	<b>85</b>	<b>57</b>	<b>52</b>	<b>80</b>	<b>62</b>
<b>WD/WD</b>	<b>87</b>	<b>82</b>	<b>90</b>	<b>79</b>	<b>98</b>	<b>75</b>	<b>97</b>	<b>86</b>	<b>66</b>	<b>58</b>	<b>76</b>	<b>69</b>
WD/DW	88	77	95	85	96	72	95	90	60	54	66	64
DD/DW	82	71	91	82	92	64	91	88	55	51	64	62
WW/DW	89	73	94	86	96	71	95	91	51	44	59	64
<b>WW (training)</b>	<b>90</b>	<b>81</b>	<b>100</b>	<b>75</b>	<b>100</b>	<b>80</b>	<b>99</b>	<b>82</b>	<b>65</b>	<b>55</b>	<b>78</b>	<b>59</b>
<b>WW/WW</b>	<b>92</b>	<b>84</b>	<b>91</b>	<b>77</b>	<b>92</b>	<b>75</b>	<b>99</b>	<b>86</b>	<b>69</b>	<b>57</b>	<b>76</b>	<b>62</b>
DW/WW	89	79	92	76	95	72	92	85	64	55	69	60
WD/WW	89	76	95	80	96	73	95	89	63	52	68	59
DD/WW	84	73	95	77	93	67	91	86	61	55	66	62

1230

1231

1232

1233

1234

1235

1236

1237

1238

Table 6. Frequency (%) with which each model averaging technique outperforms the best individual model member of the ensemble for each DSST. Results for the control are listed in bold.

1239

DSST	NSE				NSE <sub>subrt</sub>				Absolute PBIAS			
	BMA	AICA	GRA	SAM	BMA	AICA	GRA	SAM	BMA	AICA	GRA	SAM
<b>D/D</b>	41	5	65	14	<b>86</b>	<b>0</b>	<b>70</b>	<b>49</b>	<b>20</b>	<b>0</b>	<b>15</b>	<b>8</b>
W/D	49	0	68	16	86	5	70	51	17	0	16	14
<b>W/W</b>	46	0	81	27	<b>95</b>	<b>3</b>	<b>86</b>	<b>51</b>	<b>15</b>	<b>0</b>	<b>18</b>	<b>16</b>
D/W	32	3	70	16	84	0	81	32	14	0	16	3
<b>DD/DD</b>	44	3	60	16	<b>75</b>	<b>3</b>	<b>72</b>	<b>43</b>	<b>15</b>	<b>0</b>	<b>19</b>	<b>5</b>
WD/DD	41	0	57	22	70	11	53	57	18	0	18	5
DW/DD	41	0	51	16	62	3	51	41	15	0	14	3
WW/DD	51	3	62	24	76	3	54	62	17	0	13	5
<b>WD/WD</b>	46	10	70	16	<b>72</b>	<b>8</b>	<b>73</b>	<b>43</b>	<b>12</b>	<b>0</b>	<b>15</b>	<b>15</b>
DD/WD	30	0	54	16	57	5	62	41	13	0	15	5
DW/WD	35	5	52	14	78	3	59	35	18	0	16	11
WW/WD	41	5	68	16	84	3	68	43	16	0	12	11
<b>WD/WD</b>	46	8	71	19	<b>89</b>	<b>5</b>	<b>84</b>	<b>27</b>	<b>11</b>	<b>0</b>	<b>12</b>	<b>12</b>
WD/DW	41	5	73	27	81	8	78	46	12	0	15	14
DD/DW	32	0	68	27	68	3	65	46	13	0	11	5
WW/DW	51	0	76	27	86	3	76	51	14	0	10	8
<b>WW/WW</b>	54	5	68	8	<b>80</b>	<b>3</b>	<b>81</b>	<b>30</b>	<b>19</b>	<b>0</b>	<b>17</b>	<b>11</b>
DW/WW	43	3	57	11	78	0	65	35	17	0	15	5
WD/WW	46	8	73	16	78	5	76	46	20	0	18	8
DD/WW	30	3	70	14	73	3	68	32	21	0	11	11

1240

1241

1242

1243

1244

1245

1246

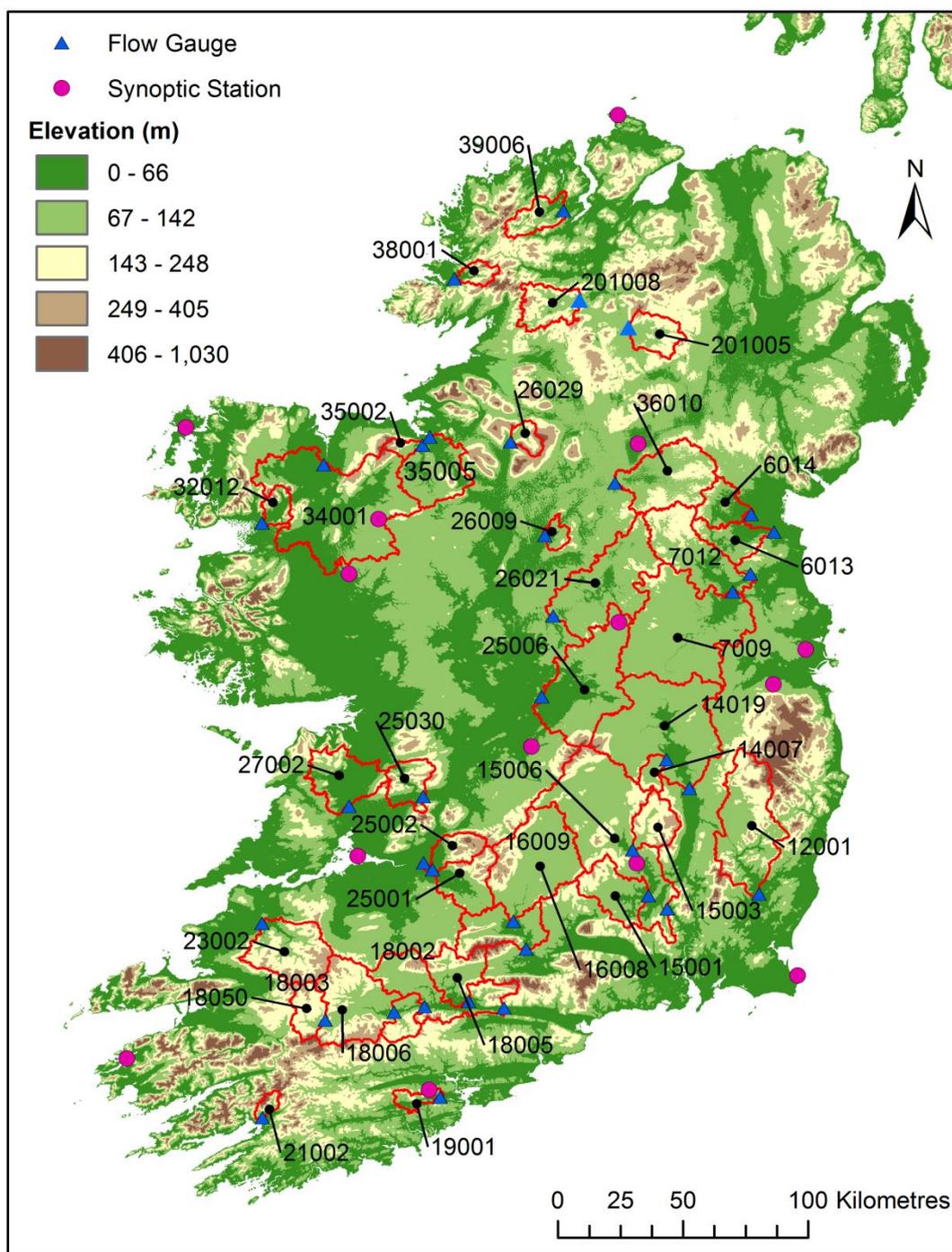
1247

1248

1249

1250

1251



1253

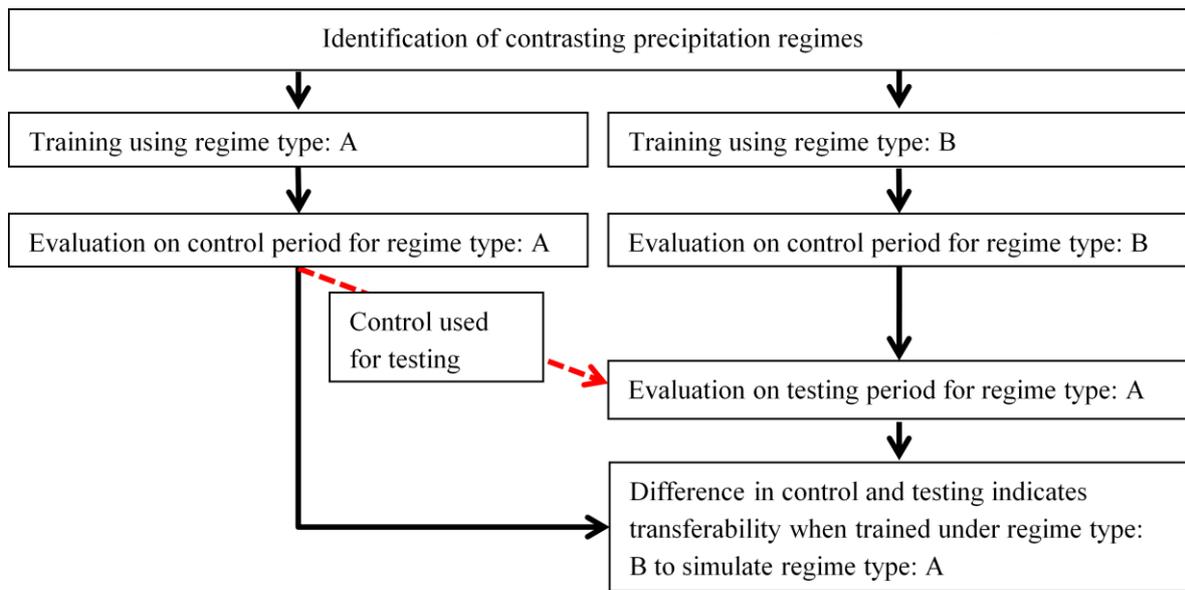
1254 Figure 1. Study catchments and Met Éireann synoptic stations. Catchment identification codes  
1255 are shown; red lines denote the respective catchment boundaries.

1256

1257

1258

1259



1260

1261 Figure 2. Flow diagram of the Differential Split Sample Testing (DSST) procedure used - incorporating  
1262 training and performance assessment for an independent control and testing period respectively. This  
1263 DSST procedure is used for estimation of weights in the Generalised Likelihood Uncertainty Estimation  
1264 procedure (GLUE; Section 2.4) and for model averaging (Section 2.5).

1265

1266

1267

1268

1269

1270

1271

1272

1273

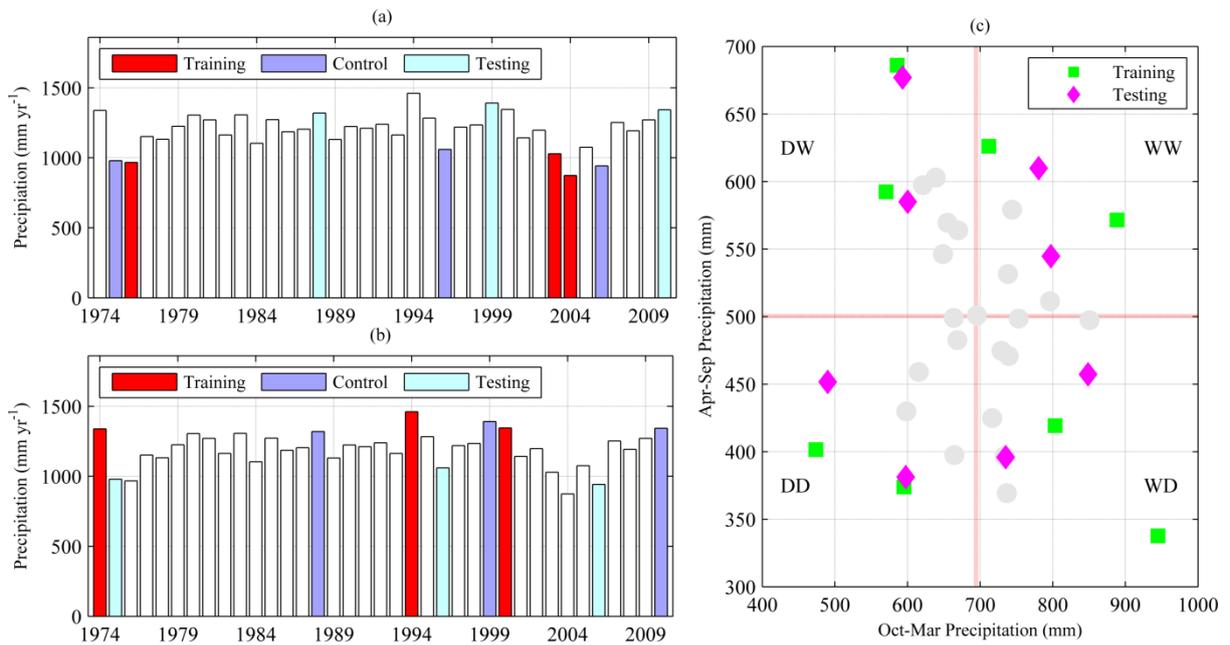
1274

1275

1276

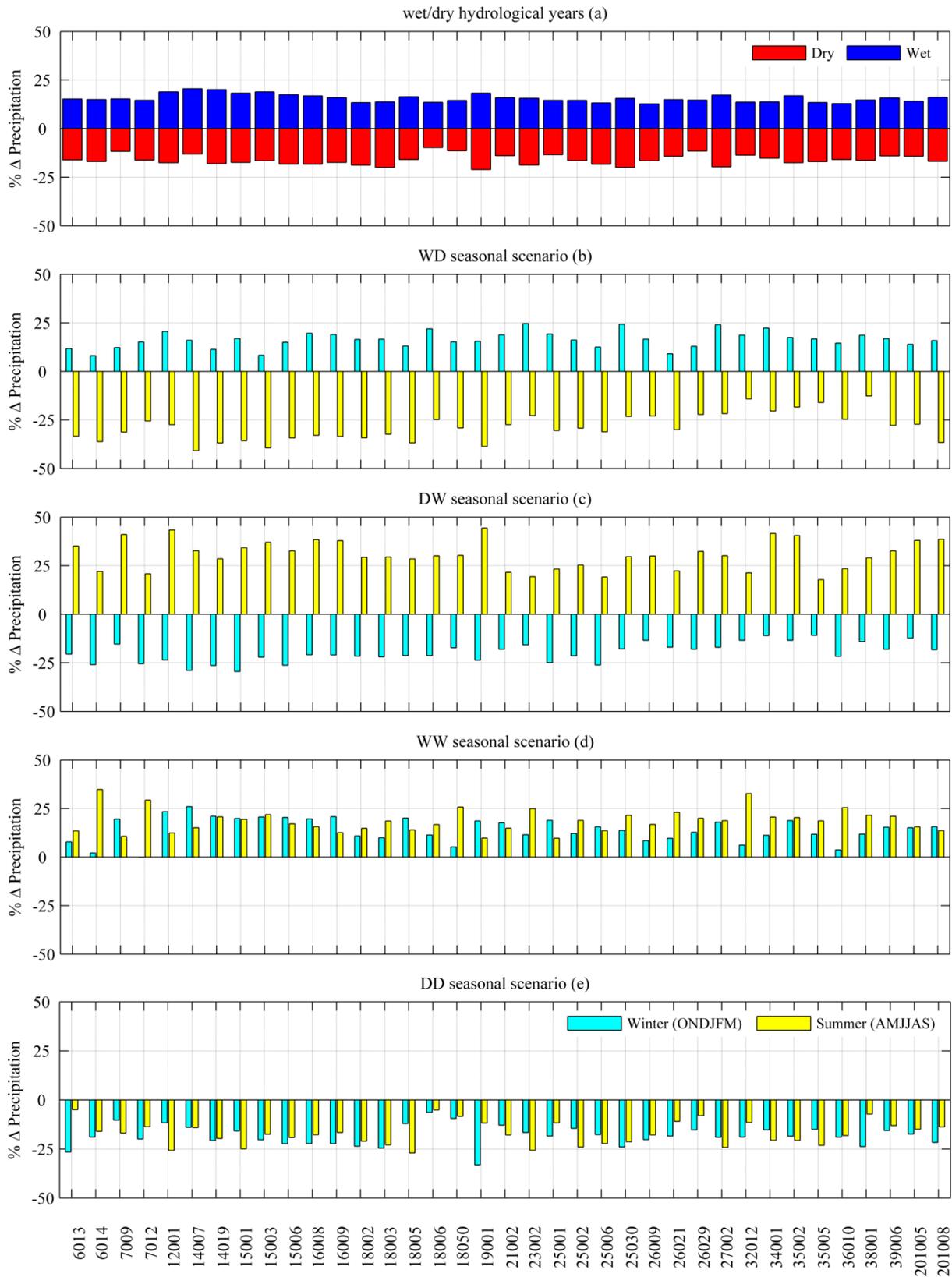
1277

1278  
1279  
1280  
1281  
1282  
1283  
1284



1285  
1286  
1287  
1288

Figure 3. Panel (a) and (b): precipitation totals (1974-2010) for the hydrological year (1<sup>st</sup> October - 30<sup>th</sup> September; catchment ID 15006). Panel (c): winter (ONDJFM; *x-axis*) and summer (AMJJAS; *y-axis*) seasonal precipitation for six month periods of the hydrological year. Training and testing periods used to assess transferability between ‘wet’/‘dry’ (D, W) years (a and b) are highlighted, as are periods (c) used to examine transferability between each of four (DD, WW, DW, WD) seasonal precipitation regimes.



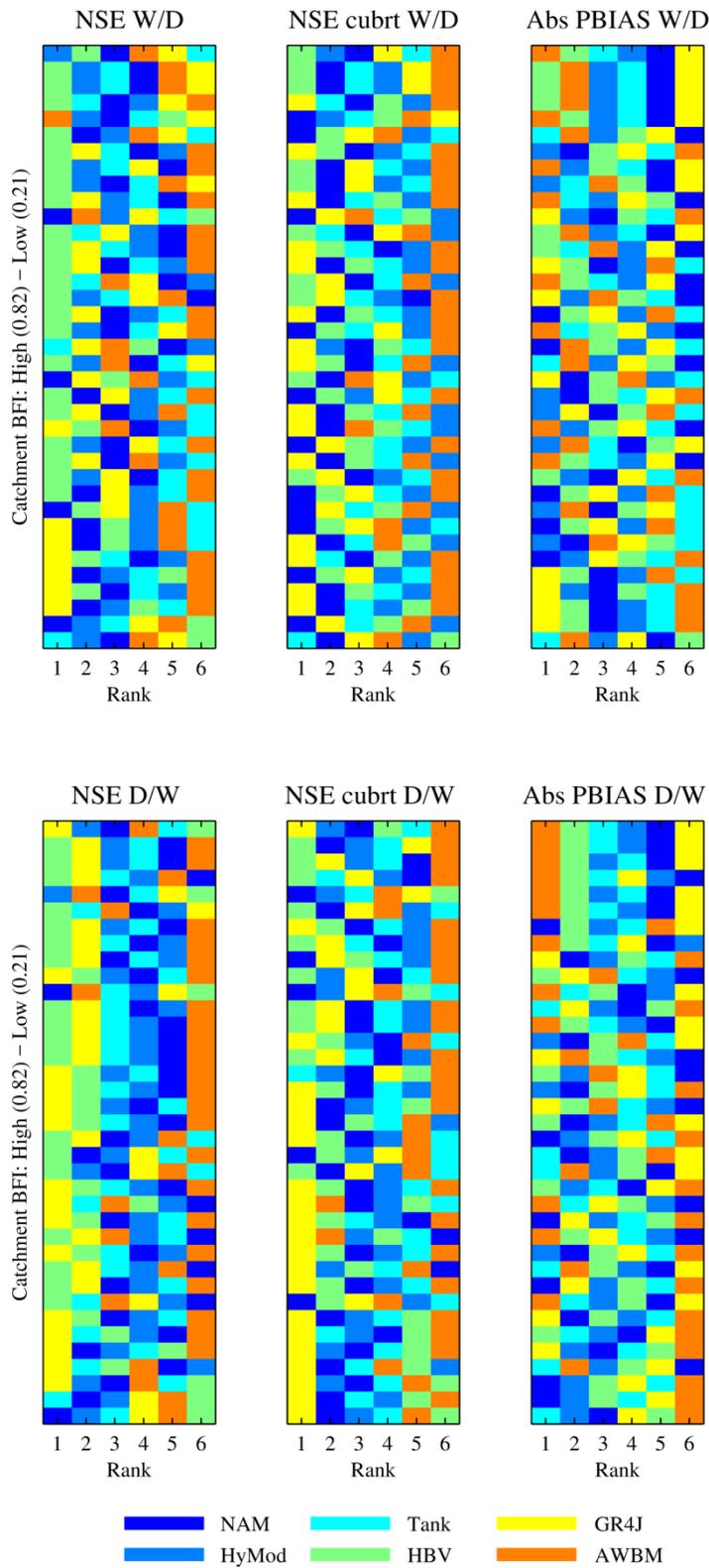
1289

1290

1291

1292

Figure 4. Percent differences in total seasonal/annual precipitation relative to 1976-2005 (Table 1) for DSST testing/control periods. Differences in contrasting 'wet'/'dry' hydrological years (1<sup>st</sup> October - 30<sup>th</sup> September) are shown (a). Relative differences for six-month winter (ONDJFM) and summer (AMJJAS) periods are shown for each seasonal (Wet-Dry, Dry-Wet, Wet-Wet and Dry-Dry) DSST scenario (b-e).



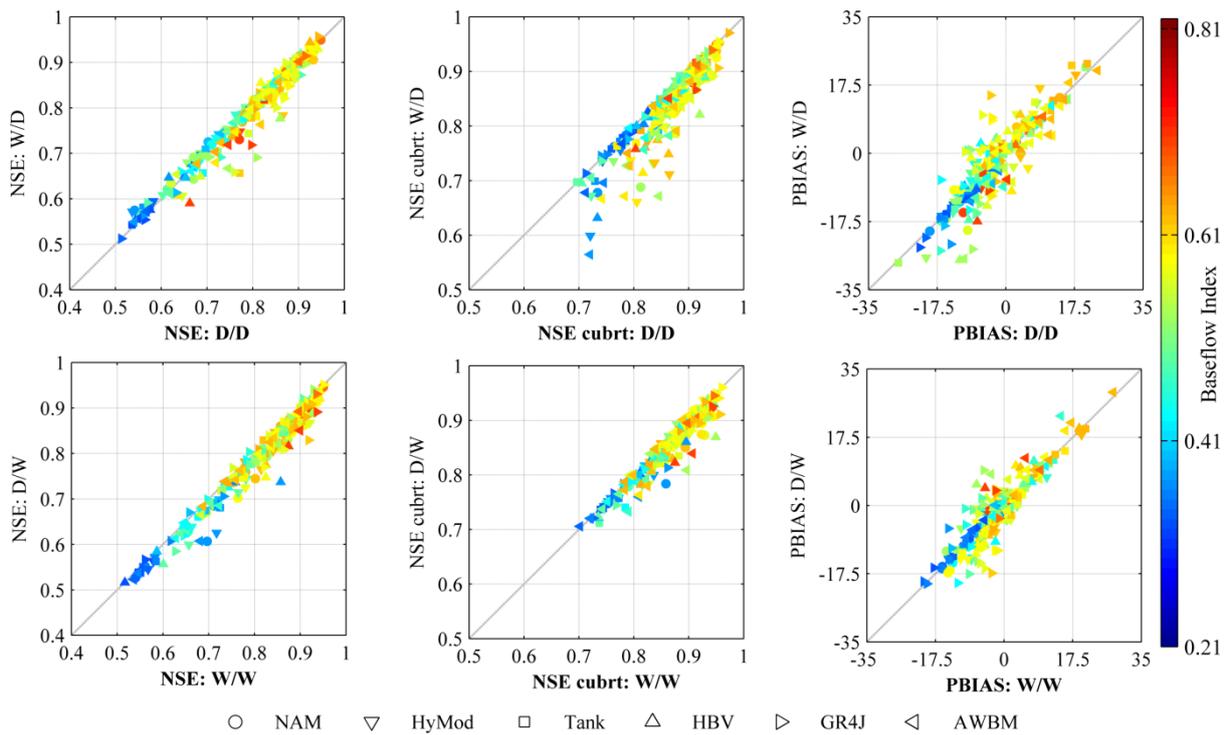
1294

1295

1296

1297

Figure 5. Individual model structures ranked (*x-axis*; best (1) to worst (6)) according to performance when tested under transference between 'wet'/'dry' annual regimes. Catchments (*y-axis*) are sorted according to their BFI in ascending order. Models are ranked according to the absolute (Abs) PBIAS value.



1299

1300

1301

Figure 6. Testing (*y-axis*) and control (*x-axis*; shown in bold) results for two ('wet'/'dry') annual precipitation regimes. Models producing similar results for each DSST fall closer to the 45° line. Marker type corresponds to an individual model structure; markers are also coded using graduated shading for Base Flow Index (BFI).

1302

1303

1304

1305

1306

1307

1308

1309

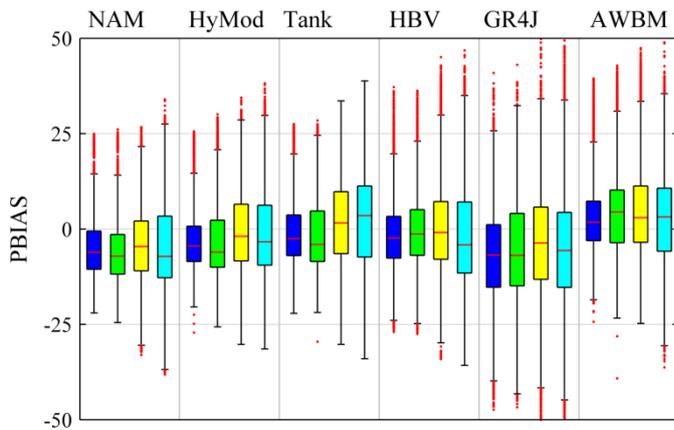
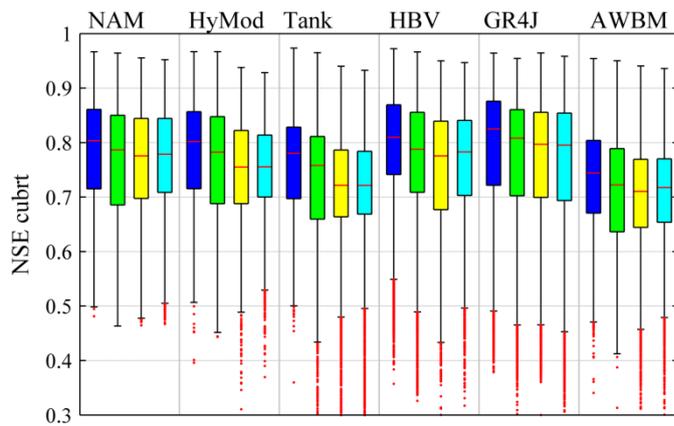
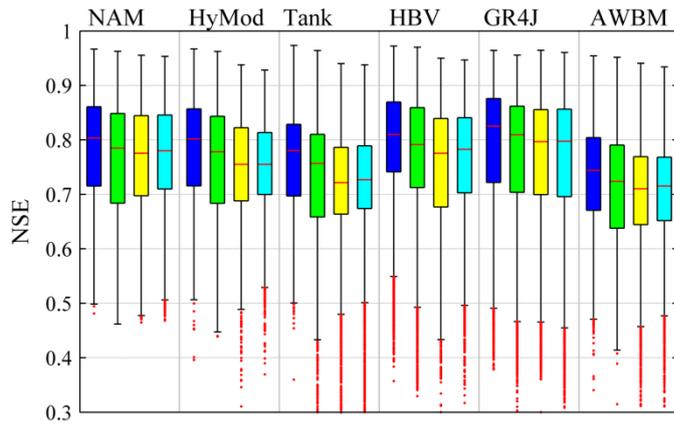
1310

1311

1312

1313

1314



■ **D/D**
■ W/D
 ■ W/W
 ■ D/W

1315

1316

1317

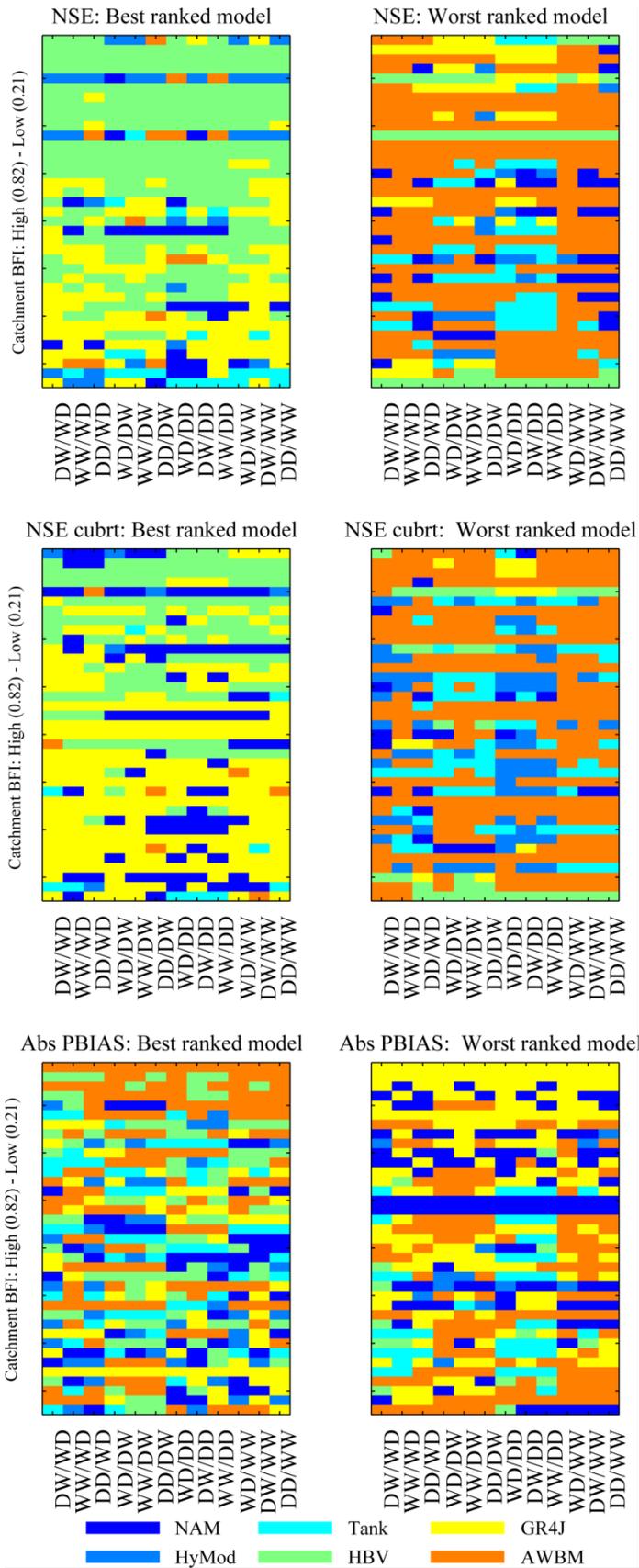
1318

1319

1320

1321

Figure 7. The combined performance of behavioural parameter sets for all catchments and rainfall-runoff models. DSST results are for two ('dry'/'wet') annual precipitation regimes are shown. The red line represents the median estimate; box edges denote the 25th and 75th percentiles. Whiskers are located at  $Q3+1.5 \times (Q3-Q1)$  and  $Q1-1.5 \times (Q3-Q1)$ , where  $Q1$  and  $Q3$  are the 25th and 75th percentiles respectively. Values beyond this are identified with red dots. Control scenarios are highlighted in bold. NSE/NSE<sub>cubrt</sub> values <0.3 are not shown.



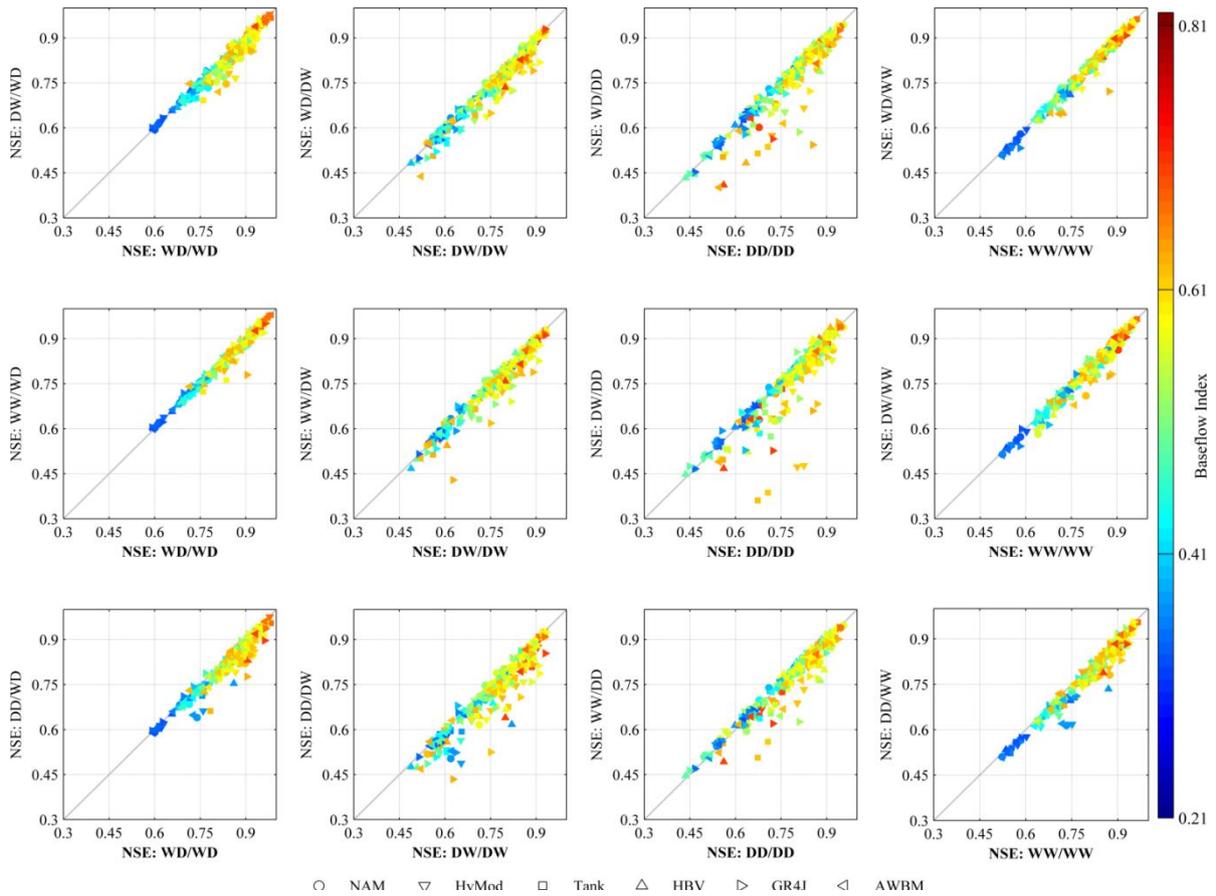
1322

1323

1324

Figure 8. Best and worst ranked hydrological model according to DSST results for four (DD, WW, DW, WD) seasonal precipitation regimes (*x-axis*). Catchments (*y-axis*) are sorted according to their BFI in ascending order.

1325

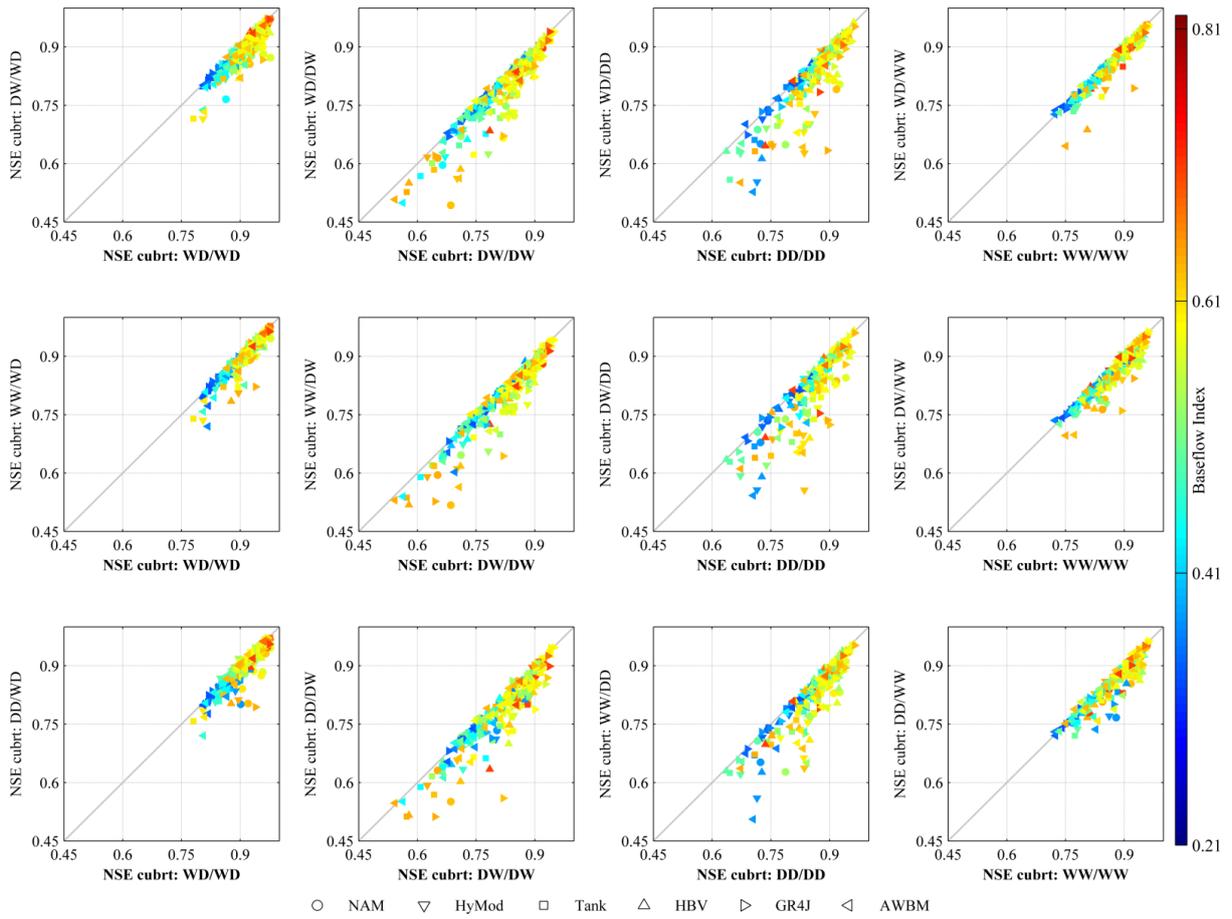


1326

1327

Figure 9. NSE testing (*y-axis*) and control (*x-axis*; shown in bold) results for four (DD, WW, DW, WD) seasonal precipitation regimes. Models producing similar results for each DSST fall closer to the 45° line. Marker type corresponds to an individual model structure; markers are also coded using graduated shading for Base Flow Index (BFI).

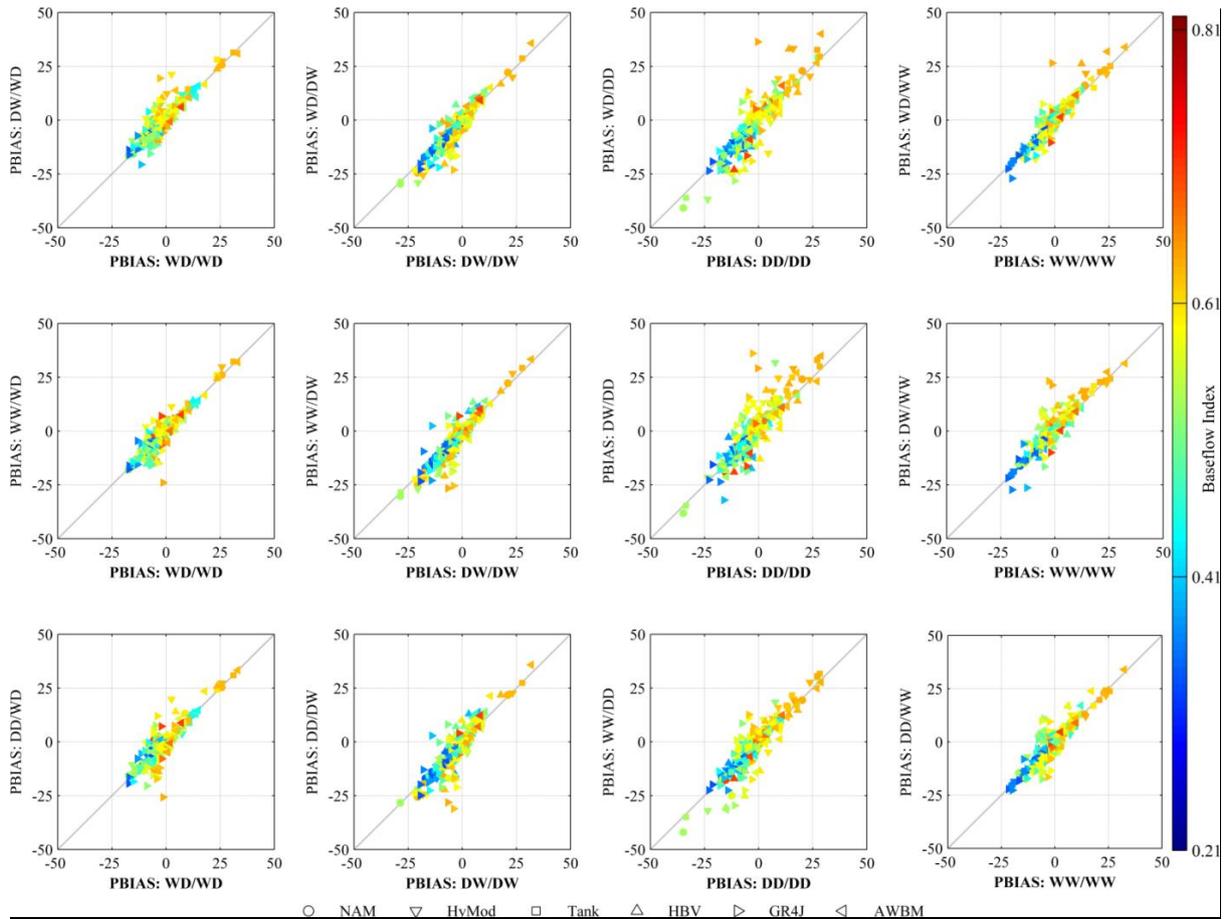
1328



1329  
1330  
1331

Figure 10.  $NSE_{cubrt}$  testing (*y-axis*) and control (*x-axis*; shown in bold) results for four (DD, WW, DW, WD) seasonal precipitation regimes. Models producing similar results for each DSST fall closer to the 45° line. Marker type corresponds to an individual model structure; markers are also coded using graduated shading for Base Flow Index (BFI).

1332



1333

1334

Figure 11. PBIAS testing (*y-axis*) and control (*x-axis*; shown in bold) results for four (DD, WW, DW, WD) seasonal precipitation regimes. Models producing similar results for each DSST fall closer to the 45° line. Marker type corresponds to an individual model structure; markers are also coded using graduated shading for Base Flow Index (BFI).

1335

1336

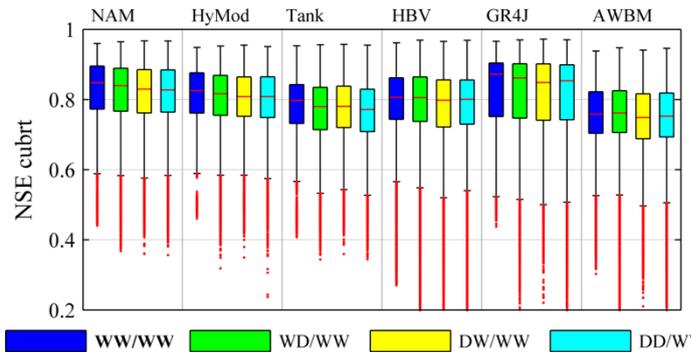
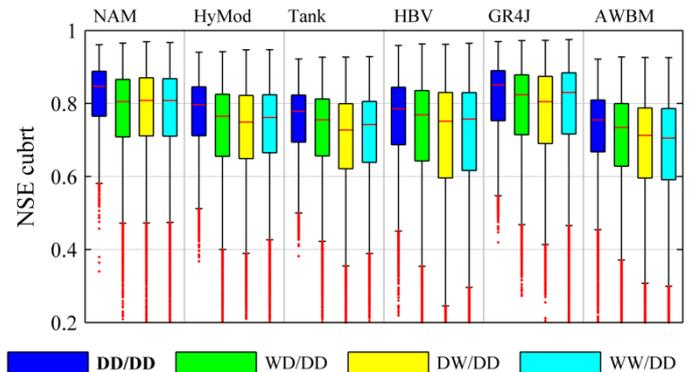
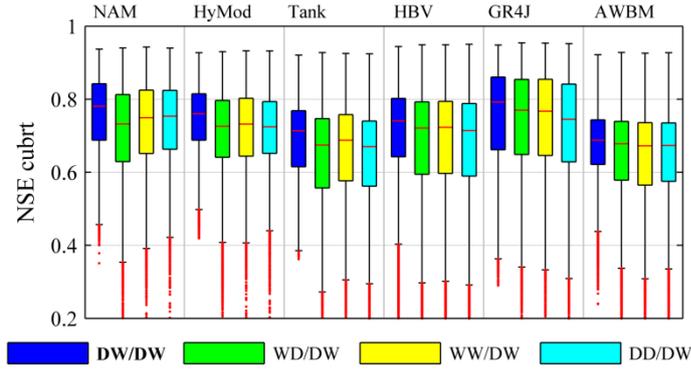
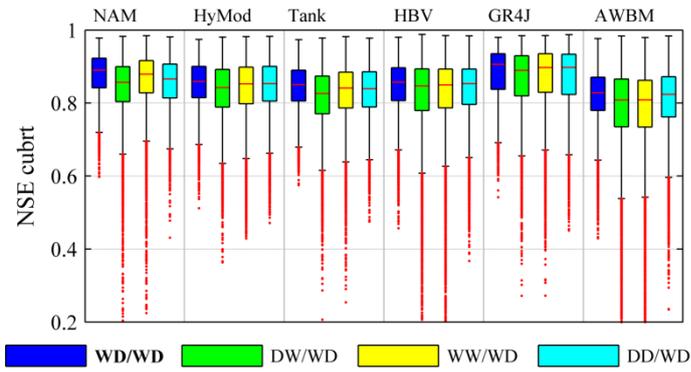
1337

1338

1339

1340

1341



1342

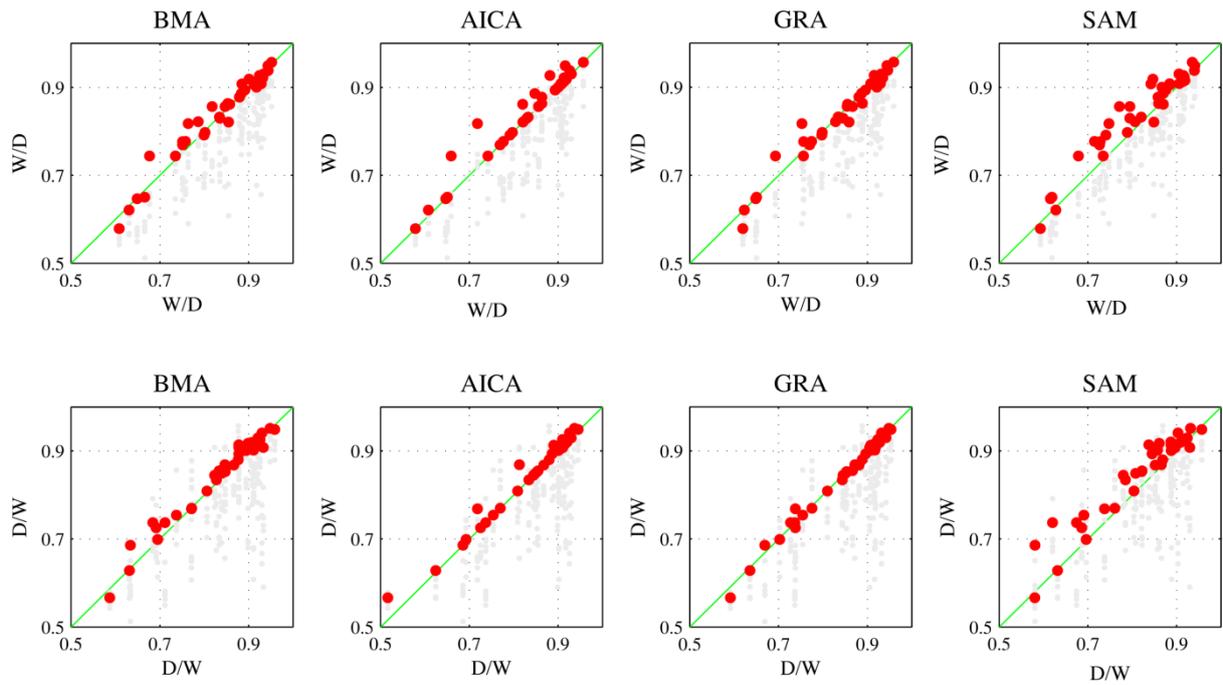
1343

1344

1345

1346

Figure 12.  $NSE_{cubrt}$  boxplots developed using the combined behavioural parameter sets of all six rainfall-runoff models for 37 catchments and four (DD, WW, DW, WD) seasonal precipitation regimes. The red line represents the median estimate; box edges denote the 25th and 75th percentiles. Whiskers are located at  $Q3+1.5 \times (Q3-Q1)$  and  $Q1-1.5 \times (Q3-Q1)$ , where  $Q1$  and  $Q3$  are the 25th and 75th percentiles respectively. Values beyond this are identified with red dots. Control scenarios are highlighted in bold.  $NSE/NSE_{cubrt}$  values  $<0.2$  are not shown.



1347

1348

1349

Figure 13. NSE scores for 'wet'/'dry' DSST period obtained from four different model averaging techniques plotted against the corresponding NSE value from each model structure (grey dots). NSE values showing transference between the wettest/driest years for each catchment is plotted; red dots denote the best performing individual ensemble member. Model averaging improves relative to a single structure where points are plotted below the 45° continuous green line (i.e.  $x=y$ ).

1350

1351

1352

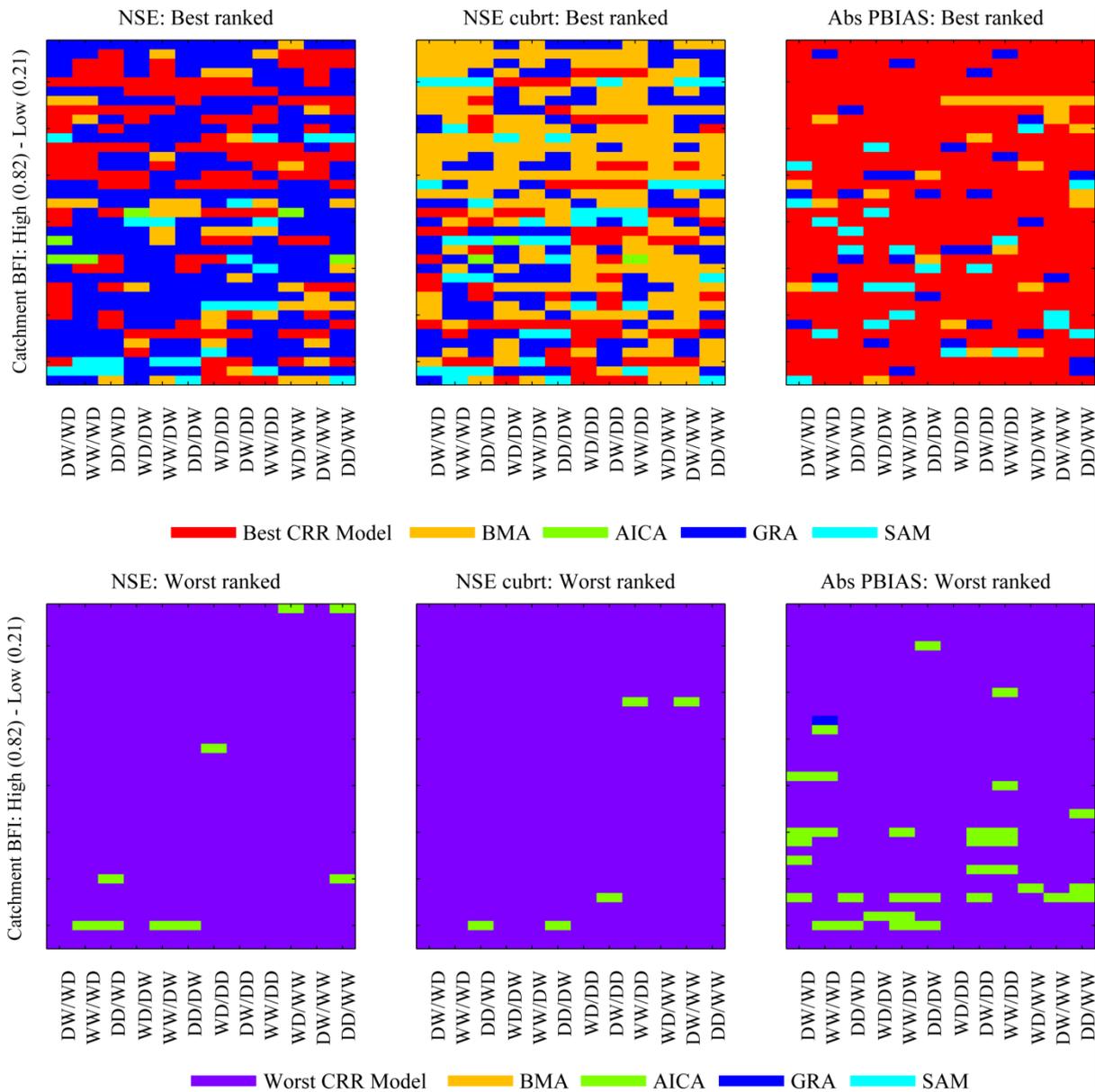
1353

1354

1355

1356

1357



1358

1359

1360

1361

1362

1363

1364

1365

1366

Figure 14. Best and worst ranked model averaging technique according to DSST results for four (DD, WW, DW, WD) seasonal precipitation regimes (*x-axis*). Also considered is the best and worst performing conceptual rainfall-runoff (CRR) model for each scenario. Catchments (*y-axis*) are sorted according to their BFI in ascending order.