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

Authors:
Ciaran Broderick[1], Tom Matthews[2], Robert L. Wilby[3], Satish Bastola[4], Conor Murphy[1]

Affiliation:
Maynooth University[1]
Liverpool John Moores University[2]
University of Loughborough[3]
Georgia Institute of Technology[4]

Key points:

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

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

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

Abstract

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

1. Introduction

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

Hydrological model uncertainty stems from errors in input (e.g. precipitation) and output (e.g. streamflow) data, as well as from deficiencies in model structures and non-uniqueness of model parameters. Previous studies have encountered difficulties when addressing structural uncertainty, particularly when trying to identify a single, optimum model for a given catchment type [*Clark et al.*, 2008; *van Esse et al.*, 2013; *Coxon et al.*, 2014]. Similarly, uncertainty relating to model calibration/training arises due to equifinality or the inability to determine a globally optimum parameter set [*Beven*, 2006]. For climate impact studies, additional uncertainties arise due to hydrological models being applied to conditions outside those used for model training. Hence, the assumption of parametric stationarity – whereby parameters provide realistic simulations when applied under hydroclimatological conditions dissimilar to those used for model development - has been widely questioned. A number of authors have called for a more rigorous and systematic approach to interrogating transferability and model robustness for climate impact studies [*Hartmann and Bárdossy*, 2005; *Wilby*, 2005; *Beven*, 2006; *Wilby and Harris*, 2006; *Andréassian et al.*, 2009; *Vaze et al.*, 2010; *Merz et al.*, 2011; *Coron et al.*, 2012; *Li et al.*, 2012; *Seiller et al.*, 2012, 2015; *Brigode et al.*, 2013; *Westra et al.*, 2014; *Thirel et al.*, 2015a, 2015b].

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

Only when critical uncertainties have been addressed [*Clark et al.*, 2016], and sufficient testing has been conducted to establish performance under a range of conditions, can model projections be used to make well informed adaptation decisions (including under 'stress test' conditions). To this end, the present study uses DSST to examine temporal transferability of a multi-model hydrological ensemble. The study has two aims. First, we analyse the performance of six lumped Conceptual Rainfall-Runoff (CRR) models applied under climate conditions that differ from those used for model training, for catchments across the Island of Ireland (IoI). Previous studies have assessed climate change impacts on Irish catchments [*Steele-Dunne et al.*, 2008; *Bastola et al.*, 2011, 2012], but systematic appraisal of model transferability has yet to be undertaken. In addition, there is limited information about which model(s) perform best across catchments with contrasting hydrological and climate characteristics. Second, we examine through comparison of multiple methods, the extent to which an ensemble offers improved transferability beyond reliance on individual model structures. This study expands on existing research, [*Vaze et al.*, 2010; *Merz et al.*, 2011; *Coron et al.*, 2012; *Li et al.*, 2012] – and the work of *Seiller et al.*, [2012, 2015] in particular – by contributing to knowledge of model limitations under non-stationary conditions. In particular, we quantify how model performance may be diminished by transference and whether this is greater with respect to wetter/drier conditions and specific seasonal precipitation regimes. We also examine the suitability of using observed records as an analogue to determine predictive performance under possible future conditions, demonstrate an approach for training and unbiased model evaluation, and examine methods to improve model application in climate impact studies.

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

2. Methods

2.1 Study Catchments and Data

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

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

Catchment average rainfall was estimated from a quality-assured 1km × 1km gridded dataset provided by Met Éireann [*Walsh*, 2012]. Daily PET, estimated via the Penman method [*Allen et al*., 1998], was also provided by Met Éireann for the closest synoptic station to each catchment centroid (Figure 1). Gaps in the records were infilled through regression with highly correlated (Pearson's coefficient >0.7) neighbouring stations. Additionally, to ensure a robust statistical relationship donor sites that provided an overlapping period of >5 years were selected.

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

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

2.2 Hydrological Models

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

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

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

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

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

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

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

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

2.3 Differential split sampling

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

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

Climate model projections suggest wetter winters and drier summers for IoI [*Steele-Dunne et al.*, 2008; *Bastola et al.*, 2011, 2012; *Matthews et al.*, 2016], necessitating transferability of models to an amplified seasonal regime. This is particularly important given how the dynamics of intra-seasonal processes during training (the rate, timing and distribution of storage recharge and reduction through the year) may affect the model response when used to simulate more extreme wetting-up and drying episodes [*Wagener*, 2003; *Herman et al.*, 2013]. The type of seasonal regime is expected to influence the structural components/parameters for soil moisture accounting and the behaviour of longer term stores, as well as the threshold and time delay of different flow paths. Hence, under transference the training scenario used has particular implications for accurate simulation of baseflow and storm event dynamics.

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

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

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

Previous DSST studies have generally employed 5-10 year training/testing periods using both block sampling and non-continuous years [*Yapo et al.*, 1996; *Anctil et al.*, 2004; *Hartmann and Bárdossy*, 2005; *Merz et al.*, 2011; *Coron et al.*, 2012; *Li et al.*, 2012; *Seiller et al.*, 2012, 2015]. Assessing model suitability for climate impact assessment – for which models are applied under a projected climate that may diverge significantly from conditions experienced during observations – necessitates evaluating performance under as demanding a set of conditions as possible. This requires a compromise between maximizing difference in periods used to assess transferability *versus* achieving potentially more robust training. Given the short record length available (~30 years) and temperate nature of the IoI climate (which moderates the occurrence of extreme interannual/seasonal variability) DSST was undertaken using three/two-year non-continuous periods. This was considered sufficient to examine transferability under strict conditions yet provide sufficient training. Also, the shortened record lengths available for some catchments may omit years with more pronounced variability leading to a less strict DSST. However, based on relative differences in the rainfall regime between training/testing conditions for all IRN catchments, those with a shorter record length provide a similar level of diversity in precipitation (Figure 4).

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

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{T}(Q_o^t - Q_m^t)^2}{\sum_{t=1}^{T}\left(Q_o^t - \overline{Q_o^t}\right)^2} \tag{1}$$

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

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

$$\mathrm{PBIAS} = \frac{\sum_{\mathrm{t}=1}^{\mathrm{T}} Q_m^t - Q_o^t}{\sum_{\mathrm{t}=1}^{\mathrm{T}} Q_o^t} \times 100 \tag{3}$$

2.4 Parameter Selection

Parameter values sampled from different regions of parameter space can provide equally valid simulations of system behaviour [*Beven*, 2006]. This may, in part, be attributed to the over-parameterization of hydrological models, as well as to issues of parameter interdependence and identifiability. Although parameter sets may perform comparably well during training, their values are tuned to the training data used, meaning they can respond very differently when applied under dissimilar conditions [*Uhlenbrook et al.*, 1999]. Additionally, parameters may exhibit differing sensitivities depending on the climate conditions experienced during training; this has implications for identifiability and performance under contrasting conditions [*Merz et al*., 2011].

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

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

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

$$\mathrm{RMSE}_{\mathrm{sqrt}} = \sqrt{\frac{\sum_{t=1}^{T}\left(\sqrt{Q_m^t} - \sqrt{Q_o^t}\right)^2}{T}} \tag{4}$$

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

The top 10% parameter sets ranked according to $\mathrm{RMSE}_{\mathrm{sqrt}}$ for the training period were retained as behavioural and the associated $\mathrm{RMSE}_{\mathrm{sqrt}}$ values were used to estimate respective weights. Performance of the median simulation under control and opposing testing period(s) was used to examine model transferability. Here the median simulator refers to the combined 50th percentile of daily flow which is derived from the weighted flow series simulated by the retained parameter sets. As the likelihood measure does not conform to the properties of a formal objective function, and can return values greater than 1, a transformation function was required. Following *Blasone et al.* [2008] and *Mertens et al.* [2004] the posterior likelihood function for accepted parameter sets was calculated as the reciprocal of the returned efficiency criterion multiplied by a normalizing factor. In this case, the posterior likelihood function $L(\theta_i|Q)$ for each behavioural set (θ_i) was calculated using (equation 5):

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

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

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

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

2.5 Model Averaging

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

2.5.1 Bayesian Model Averaging (BMA)

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

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

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

$$p(M_k|D) \propto p(D|M_k)p(M_k) \tag{8}$$

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

$$E[\Delta|M_1, \ldots\ldots, M_k, D] = \sum_{k=1}^{K} w_k \hat{\Delta}_k \tag{9}$$

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

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

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

$$\alpha = \mu_k^2/\sigma_k^2;\ \beta_k = \sigma_k^2/\mu_k;\ \mu_k = M_k;\ \sigma_k^2 = b \cdot M_k + c \tag{12}$$

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

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

2.5.2 Akaike Information Criteria Averaging (AICA)

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

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

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

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

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

2.5.3 Granger-Ramanathan Averaging (GRA)

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

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

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

3. Results

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

3.1 Individual model performance – wettest/driest years

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

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

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

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

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

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

3.2 Individual model performance – seasonal assessment

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

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

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

Figure 12 shows the results of DSST applied to all behavioural parameter sets identified across the catchment sample. In terms of absolute model performance the highest NSE_{cubrt} control/testing values are generally returned for the WD/WD scenario. Based on the median estimate, GR4J performs well across the catchment sample, whereas AWBM generally returns the lowest scores. Difficulties in transference to a DW or DD regime are also highlighted by Figure 12. In contrast, parameters generally maintain performance when transferred to a WW regime irrespective of the training scenario.

3.3 Multimodel performance

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

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

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

For PBIAS, all averaging methods generally return a considerably lower proportion (<20%) of better performing estimates when benchmarked against the best model member. The results shown in Table 6 are reflected in Figure 14 which displays the best/worst ranked model averaging method for each catchment and seasonal DSST scenario; also considered is the best/worst performing model structure. Evident are the more favourable results for BMA/GRA according to the NSE/NSE_{cubrt} criterion. The ranking of methods is also largely consistent across individual catchments and for each DSST scenario. Figure 14 further highlights disparities in performance between the NSE and PBIAS measures. In the latter case, it is shown that the best individual model structure for each scenario typically performs better than the respective model averaging techniques. Figure 14 also highlights that the worst performing model is most often ranked lower than the worst performing averaging method.

4. Discussion

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

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

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

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

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

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

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

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

With respect to the BFI, it is worth noting how differences in the storage and routing configuration relate to infiltration processes and performance for groundwater/runoff dominated catchments. The influence of vertical soil heterogeneity and slope has on runoff generation is well documented [*Smith and Hebbert*, 1983; *Jackson*, 1992]. Typically for catchments with permeable homogeneous soils and a low anisotropy ratio (vertical conductivity/horizontal conductivity) movement through upper layers tends to occur vertically, with vertical increases in the saturated zone depth having a greater effect on runoff than lateral movements. Here catchments are likely to have a high BFI owing to better infiltration and delayed routing. In contrast, for catchments with a high anisotropy ratio where hillslope processes dominate, lateral flows are likely to be more significant. Hence models like HBV, which can better capture vertical variability in soil processes by using multiple vertical stores and a dedicated soil moisture routine, and which explicitly account for direct/lateral flows, may be more applicable to low BFI catchments. Furthermore the hillslope can be conceptualized as consisting of two soil layers, with the lower layer capable of retarding vertical flow at the boundary allowing development of subsurface stormflow. This corresponds well with the inclusion of an upper soil box in HBV from which two lateral outflows (one threshold based) are represented [*Smith and Hebbert*, 1983]. While GR4J also accounts for vertical variability, only two stores (production and routing) are included, and lateral flows are less well represented. In addition, the model has fewer free parameters to adjust in order to better capture horizontal/direct flows (e.g. the set 90:10 split between delayed and direct routing channels).

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

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

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

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

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

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

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

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

Future work will examine why the individual CRR models performed differently across the catchment sample used in this study. Exploring parameter sensitivity to time-varying hydroclimatic conditions would help link physical processes with model formulation and provide insight to the relative skill of ensemble members under different forcing scenarios (e.g. wet/dry and seasonal transitions). This would also help to establish the influence which information content in the training data and the associated activation frequency of key parameters have on transferability between contrasting regimes.

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

5. Conclusion

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

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

The results confirm that it is impossible to identify a single structure that performs optimally across all catchments, DSST scenarios and performance criteria. Moreover, the collective ensemble was shown to outperform the majority of individual ensemble members. However, averaging methods were found to differ considerably with respect to the frequency with which they surpass the best individual member, particularly for volumetric errors. Bayesian Model Averaging (BMA) and the Granger-Ramanathan method (GRA) were found to perform better under transference than using the Simple Arithmetic Mean (SAM) and Akaike Information Criteria Averaging (AICA). Further work could be done on the potential added value of using different variants of GRA including non-constrained weights and a bias correction step, as well as the transferability of averaging techniques that implement dynamic weighting [*See and Openshaw*, 2000; *Hu et al.*, 2001; *Wagener et al.*, 2003] .

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

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

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

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

Acknowledgements

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

References

Abrahart, R. J., and L. See (2002), Multi-model data fusion for river flow forecasting: an evaluation of six alternative methods based on two contrasting catchments, *Hydrol. Earth Syst. Sci.*, *6*(4), 655–670, doi: 10.5194/hess-6-655-2002.

Ajami, N. K., Q. Duan, X. Gao, and S. Sorooshian (2006), Multimodel combination techniques for analysis of hydrological simulations: Application to distributed model intercomparison project results, *J. Hydrometeorol*, *7*(4), 755–768, doi:10.1175/JHM519.1.

Akaike, H. (1974), A New Look at the Statistical Model Identification, in *Selected Papers of Hirotugu Akaike*, edited by E. Parzen, K. Tanabe, and G. Kitagawa, pp. 215–222, Springer New York.

Allen, R. G., L. S. Pereira, D. Raes, and M. Smith (1998), FAO Irrigation and drainage paper No. 56, *Rome: Food and Agriculture Organization of the United Nations*, *56*, 97–156.

Anctil, F., C. Perrin, and V. Andréassian (2004), Impact of the length of observed records on the performance of ANN and of conceptual parsimonious rainfall-runoff forecasting models, *Environ. Model. Softw.*, *19*(4), 357–368, doi:10.1016/S1364-8152(03)00135-X.

Andréassian, V., C. Perrin, L. Berthet, N. Le Moine, J. Lerat, C. Loumagne, L. Oudin, T. Mathevet, M. H. Ramos, and A. Valéry (2009), Crash tests for a standardized evaluation of hydrological models, *Hydrol. Earth Syst. Sci.*, *13*(10), 1757-1764, doi:10.5194/hess-13-1757-2009.

Arsenault, R., P. Gatien, B. Renaud, F. Brissette, and J.-L. Martel (2015), A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation, *J. Hydrol.*, *529*, 754–767, doi:10.1016/j.jhydrol.2015.09.001.

Atkinson, S. E., R. A.,Woods and M. Sivapalan (2002), Climate and landscape controls on water balance model complexity over changing timescales, *Water Resour. Res., 38*(12), 1314,doi:10.1029/2002WR001487.

Bastola, S., C. Murphy, and J. Sweeney (2011), The role of hydrological modelling uncertainties in climate change impact assessments of Irish river catchments, *Adv. Water Resour.*, *34*(5), 562–576, doi:10.1016/j.advwatres.2011.01.008.

Bastola, S., C. Murphy, and R. Fealy (2012), Generating probabilistic estimates of hydrological response for Irish catchments using a weather generator and probabilistic climate change scenarios: Probabilistic-based estimates of climate change effects, *Hydrol. Process.*, *26*(15), 2307–2321, doi:10.1002/hyp.8349.

Beven, K. (2006), A manifesto for the equifinality thesis, *J. Hydrol.*, *320*(1–2), 18–36, doi:10.1016/j.jhydrol.2005.07.007.

Beven, K., and A. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrol. Process.*, *6*(3), 279–298, doi:10.1002/hyp.3360060305.

Blasone, R.-S., J. A. Vrugt, H. Madsen, D. Rosbjerg, B. A. Robinson, and G. A. Zyvoloski (2008), Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Murphy, C., S. Harrigan, J. Hall, and R. L. Wilby (2013), Climate-driven trends in mean and high flows from a network of reference stations in Ireland, *Hydrol. Sci. J.*, *58*(4), 755–772, doi:10.1080/02626667.2013.782407.

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part I — A discussion of principles, *J. Hydrol.*, *10*(3), 282–290, doi:10.1016/0022-1694(70)90255-6.

Oudin, L., V. Andréassian, T. Mathevet, C. Perrin, and C. Michel (2006a), Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations: Dynamic averaging of rainfall-runoff models, *Water Resour. Res.*, *42*(7), W07410, doi:10.1029/2005WR004636.

Oudin, L., C. Perrin, T. Mathevet, V. Andréassian, and C. Michel (2006b), Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models, *J. Hydrol.*, *320*(1–2), 62–83, doi:10.1016/j.jhydrol.2005.07.016.

Oudin, L., V. Andréassian, C. Perrin, C. Michel, and N. Le Moine (2008), Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments, *Water Resour. Res.*, *44*(3), W03413, doi:10.1029/2007WR006240.

Parajka, J., A. Viglione, M. Rogger, J. L. Salinas, M. Sivapalan, and G. Blöschl (2013), Comparative assessment of predictions in ungauged basins – Part 1: Runoff-hydrograph studies, *Hydrol. Earth Syst. Sci.*, *17*(5), 1783–1795, doi:10.5194/hess-17-1783-2013.

Perrin, C., C. Michel, and V. Andréassian (2001), Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, *J. Hydrol.*, *242*(3–4), 275–301, doi:10.1016/S0022-1694(00)00393-0.

Perrin, C., C. Michel, and V. Andréassian (2003), Improvement of a parsimonious model for streamflow simulation, *J. Hydrol.*, *279*(1–4), 275–289, doi:10.1016/S0022-1694(03)00225-7.

Price, K. (2011), Effects of watershed topography, soils, land use, and climate on baseflow hydrology in humid regions: A review, *Prog. Phys. Geogr.*, *35*(4), 465-492, doi: 10.1177/0309133311402714.

Prudhomme, C., R. L. Wilby, S. Crooks, A. L. Kay, and N. S. Reynard (2010), Scenario-neutral approach to climate change impact studies: Application to flood risk, *J. Hydrol.*, *390*(3–4), 198–209, doi:10.1016/j.jhydrol.2010.06.043.

Prudhomme, C., E. Sauquet, and G. Watts (2015), Low Flow Response Surfaces for Drought Decision Support: A Case Study from the UK, *J. of Extr. Even.*, *2*(2), 1550005, doi:10.1142/S2345737615500050.

Pushpalatha, R., C. Perrin, N. Le Moine, T. Mathevet, and V. Andréassian, (2011), A downward structural sensitivity analysis of hydrological models to improve low-flow simulation, *J. Hydrol.*,*411*(1-2), 66–76, doi: 10.1016/j.jhydrol.2011.09.034

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev. 133*(5), 1155–1174, doi: 10.1175/MWR2906.1.

Rojas, R., L. Feyen, and A. Dassargues (2008), Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging, *Water Resour. Res.*, *44*(12), W12418, doi:10.1029/2008WR006908.

Samuel, J., P. Coulibaly, and R.A. Metcalfe (2012), Identification of rainfall–runoff model for improved baseflow estimation in ungauged basins. *Hydrol. Process.*, *26*(3), 356–366. doi:10.1002/hyp.8133

Sear, D. A., P. D., Armitage and F. H. Dawson (1999), Groundwater dominated rivers. *Hydrol. Process.*, *13*(3), 255–276. doi:10.1002/(SICI)1099-1085(19990228)

See, L., and S. Openshaw (2000), A hybrid multi-model approach to river level forecasting, *Hydrol. Sci. J.*, *45*(4), 523–536, doi:10.1080/02626660009492354.

Seibert, J. (1996), HBV light, User's manual, Uppsala University, Institute of Earth Science, Department of Hydrology, Uppsala.

Seiller, G., F. Anctil, and C. Perrin (2012), Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions, *Hydrol. Earth Syst. Sci.*, *16*(4), 1171–1189, doi:10.5194/hess-16-1171-2012.

Seiller, G., I. Hajji, and F. Anctil (2015), Improving the temporal transposability of lumped hydrological models on twenty diversified U.S. watersheds, *J. Hydrol.: Regional Studies*, *3*, 379–399, doi:10.1016/j.ejrh.2015.02.012.

Shafii, M., and B. A. Tolson (2015), Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives: Hydrological consistecy optimization, *Water Resour. Res.*, *51*(5), 3796–3814, doi:10.1002/2014WR016520.

Shamseldin, A. Y., K. M. O'Connor, and G. C. Liang (1997), Methods for combining the outputs of different rainfall–runoff models, *J. Hydrol.*, *197*(1–4), 203–229, doi:10.1016/S0022-1694(96)03259-3.

Smith, R. E., and R. H. B. Hebbert (1983), Mathematical simulation of interdependent surface and subsurface hydrologic processes, *Water Resour. Res.*, *19*(4), 987–1001, doi:10.1029/WR019i004p00987.

Staudinger, M., K. Stahl, J. Seibert, M. P. Clark, and L. M. Tallaksen (2011), Comparison of hydrological model structures based on recession and low flow simulations, *Hydrol. Earth Syst. Sci.*, *15*(11), 3447–3459, doi:10.5194/hess-15-3447-2011.

Steele-Dunne, S., P. Lynch, R. McGrath, T. Semmler, S. Wang, J. Hanafin, and P. Nolan (2008), The impacts of climate change on hydrology in Ireland, *J. Hydrol.*, *356*(1–2), 28–45, doi:10.1016/j.jhydrol.2008.03.025.

Sugawara, M. (1995), Tank Model, in *Computer models of watershed hydrology*, edited by V. P. Singh, pp. 165–214, Water Resources Publications, Littleton, Colorado.

Sweeney, J. (2014), Regional weather and climates of the British Isles – Part 6: Ireland, *Weather*, *69*(1), 20–27, doi:10.1002/wea.2230.

Thiboult, A., F. Anctil, and M. A. Boucher (2016), Accounting for three sources of uncertainty in ensemble hydrological forecasting, *Hydrol. Earth Syst. Sci.*, *20*(5), 1809-1825, doi:10.5194/hess-20-1809-2016.

Thirel, G. et al. (2015a), Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments, *Hydrol. Sci. J.*, *60*(7–8), 1184–1199, doi:10.1080/02626667.2014.967248.

Thirel, G., V. Andréassian, and C. Perrin (2015b), On the need to test hydrological models under changing conditions, *Hydrol. Sci. J.*, *60*(7–8), 1165–1173, doi:10.1080/02626667.2015.1050027.

Uhlenbrook, S., J. Seibert, C. Leibundgut, and A. Rodhe (1999), Prediction uncertainty of conceptual rainfall-runoff models caused by problems in identifying model parameters and structure, *Hydrol. Sci. J.*, *44*(5), 779–797, doi:10.1080/02626669909492273.

Vaze, J., D. A. Post, F. H. S. Chiew, J.-M. Perraud, N. R. Viney, and J. Teng (2010), Climate non-stationarity – Validity of calibrated rainfall–runoff models for use in climate change studies, *J. Hydrol.*, *394*(3–4), 447–457, doi:10.1016/j.jhydrol.2010.09.018.

Velázquez, J. A., F. Anctil, and C. Perrin (2010), Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments, *Hydrol. Earth Syst. Sci.*, *14*(11), 2303–2317, doi:10.5194/hess-14-2303-2010.

Velázquez, J. A., F. Anctil, M. H. Ramos, and C. Perrin (2011), Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures, *Adv. Geosci.*, *29*, 33–42, doi:10.5194/adgeo-29-33-2011.

Vrugt, J. A., C. G. H. Diks, and M. P. Clark (2008), Ensemble Bayesian model averaging using Markov Chain Monte Carlo sampling, *Environ. Fluid. Mech.*, *8*(5–6), 579–595, doi:10.1007/s10652-008-9106-3.

Wagener, T. (2003), Evaluation of catchment models, *Hydrol. Process.*, *17*(16), 3375–3378, doi:10.1002/hyp.5158.

Wagener, T., D. P. Boyle, M. J. Lees, H. S. Wheater, H. V. Gupta, and S. Sorooshian (2001), A framework for development and application of hydrological models, *Hydrol. Earth Syst. Sci. Discussions*, *5*(1), 13–26.

Wagener, T., N. McIntyre, M. J. Lees, H. S. Wheater, and H. V. Gupta (2003), Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis, *Hydrol. Process.*, *17*(2), 455–476, doi:10.1002/hyp.1135.

Walsh, S. (2012), Long term rainfall averages for Ireland, in *National Hydrology Seminar 2012*, Office of Public Works, Tullamore, Ireland.

Westra, S., M. Thyer, M. Leonard, D. Kavetski, and M. Lambert (2014), A strategy for diagnosing and interpreting hydrological model nonstationarity, Water Resour. Res., 50(6), 5090–5113, doi:10.1002/2013WR014719.

Whateley, S., S. Steinschneider, and C. Brown (2014), A climate change range-based method for estimating robustness for water resources supply, *Water Resour. Res.*, *50*(11), 8944–8961, doi:10.1002/2014WR015956.

Wilby, R. L. (2005), Uncertainty in water resource model parameters used for climate change impact assessment, *Hydrol. Process.*, *19*(16), 3201–3219, doi:10.1002/hyp.5819.

Wilby, R. L., and I. Harris (2006), A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the River Thames, UK, *Water Resour. Res.*, *42*(2), W02419, doi:10.1029/2005WR004065.

Wilby, R. L., C. W. Dawson, C. Murphy, P. OConnor, and E. Hawkins (2014), The Statistical DownScaling Model - Decision Centric (SDSM-DC): conceptual basis and applications, *Clim Res*, *61*(3), 259–276, doi:10.3354/cr01254.

Wilby, R. L., S. Noone, C. Murphy, T. Matthews, S. Harrigan, and C. Broderick (2015a), An evaluation of persistent meteorological drought using a homogeneous Island of Ireland precipitation network, *Int. J. Climatol.*, *36*(8), 2854–2865, doi:10.1002/joc.4523.

Wilby, R. L., C. Prudhomme, S. Parry, and K. G. L. Muchan (2015b), Persistence of Hydrometeorological Droughts in the United Kingdom: A Regional Analysis of Multi-Season Rainfall and River Flow Anomalies, *J. of Extr. Even.*, 1550006, doi:10.1142/S2345737615500062.

Yapo, P. O., H. V. Gupta, and S. Sorooshian (1996), Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data, *J. Hydrol.*, *181*(1–4), 23–48, doi:10.1016/0022-1694(95)02918-4.

Table 1. Hydroclimatic and physical descriptors for the 37 selected catchments. Flow indices are estimated from 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.

						Precipitation (mm) 1976-2005		
Gauge ID	**Area (km^2)**	**Mean Elevation (m)**	**BFI (-)**	**Runoff (mm yr^{-1})**	**Start date**	**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

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)

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

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 (%Δ; NSE, NSE_{cubrt}) and absolute (Δ; 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	%Δ	Scenario	Model	%Δ	Scenario	Model	Δ
		NSE			NSE_{cubrt}			PBIAS		
21002	0.21	DD/DW	GR4J	-5.19	WW/DW	AWBM	-2.42	DD/DW	GR4J	**-5.6**
26029	0.23	DD/WW	HBV	-6.91	WW/WD	AWBM	-5.58	WD/DD	HBV	**-7.0**
38001	0.26	WD/WW	GR4J	-8.26	WW/DW	AWBM	-13.37	DW/WW	GR4J	**-7.4**
23002	0.28	DD/DW	HyMod	-25.33	WW/DD	AWBM	-28.24	DD/DW	HBV	**-11.8**
201008	0.32	DW/DD	GR4J	-16.31	DW/DD	AWBM	-13.40	DW/DD	GR4J	**-16.0**
15003	0.38	DW/DD	Tank	-14.03	DD/DW	Tank	-14.50	DD/DW	GR4J	**-7.5**
18050	0.38	DW/WD	NAM	-5.45	DW/WD	NAM	-11.39	WD/DW	GR4J	**-11.1**
35002	0.4	DD/DW	HyMod	-6.04	DW/DD	HyMod	-5.24	WW/WD	GR4J	**-7.6**
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	**-9.3**
201005	0.47	DD/DW	HyMod	-10.43	WD/DD	Tank	-13.39	DD/DW	GR4J	**-8.8**
25002	0.48	DD/WW	HyMod	-8.96	DD/WW	Tank	-6.89	DW/DD	GR4J	**-10.3**
18006	0.5	DD/WW	HBV	-5.07	DD/DW	GR4J	-7.08	DW/DD	GR4J	**-13.4**
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	**-7.3**
25030	0.54	WD/DD	GR4J	-27.62	WW/DD	AWBM	-22.82	WW/DD	GR4J	**-18.5**
18003	0.54	DD/DW	HBV	-3.23	DW/DD	AWBM	-10.49	WW/WD	GR4J	**-4.2**
32012	0.56	WD/DD	GR4J	-5.35	DW/DD	AWBM	-4.82	DW/DD	GR4J	**-7.1**
19001	0.59	DW/DD	HBV	-18.49	DD/DW	HBV	-16.03	DD/DW	GR4J	**-11.9**
6013	0.6	WW/DW	GR4J	-15.55	WD/DW	NAM	-14.64	WW/DD	HBV	**-18.9**
36010	0.6	DD/DW	GR4J	-14.22	DW/DD	HyMod	-17.89	DD/DW	GR4J	**-11.6**
6014	0.61	DD/DW	GR4J	-10.52	WW/DW	HyMod	-11.92	DD/DW	GR4J	**-14.4**
14007	0.62	DD/DW	HBV	-16.75	WW/DD	AWBM	-9.72	WD/DD	HyMod	**-14.7**
15006	0.62	DW/DD	Tank	-14.36	WD/DW	Tank	-13.29	DW/DD	HyMod	**-10.8**
18002	0.62	WW/DD	GR4J	-4.58	DW/DD	AWBM	-6.61	WW/WD	GR4J	**-7.2**
16008	0.63	DD/DW	GR4J	-13.74	WD/DW	NAM	-18.62	DD/DW	GR4J	**-18.5**
35005	0.63	DD/WD	NAM	-2.57	WD/WW	NAM	-3.56	DD/DW	GR4J	**-3.1**
16009	0.64	DD/WD	NAM	-8.03	DW/DD	AWBM	-20.08	DD/WW	GR4J	**-5.4**
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	**-33.3**
27002	0.7	WD/DW	AWBM	-15.88	WD/DD	GR4J	-5.44	WD/DW	GR4J	-4.6
7009	0.7	WW/DW	GR4J	-11.35	DW/DD	HyMod	-6.05	WW/DD	HBV	**-7.2**
18005	0.71	WD/DD	GR4J	-36.39	WD/DD	GR4J	-29.16	WD/DD	GR4J	**-36.0**
34001	0.77	WD/DD	AWBM	-6.04	WD/DW	AWBM	-5.66	DD/WD	GR4J	**-5.9**
26021	0.82	DW/DD	GR4J	-27.16	DD/DW	HBV	-19.19	WD/DD	HBV	**-11.7**

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

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

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.

	NSE				NSE_{cubrt}				Absolute PBIAS			
DSST	BMA	AICA	GRA	SAM	BMA	AICA	GRA	SAM	BMA	AICA	GRA	SAM
D/D	41	5	65	14	**86**	**0**	**70**	**49**	**20**	**0**	**15**	**8**
W/D	49	0	68	16	86	5	70	51	17	0	16	14
W/W	46	0	81	27	**95**	**3**	**86**	**51**	**15**	**0**	**18**	**16**
D/W	32	3	70	16	84	0	81	32	14	0	16	3
DD/DD	44	3	60	16	**75**	**3**	**72**	**43**	**15**	**0**	**19**	**5**
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
WD/WD	46	10	70	16	**72**	**8**	**73**	**43**	**12**	**0**	**15**	**15**
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
WD/WD	46	8	71	19	**89**	**5**	**84**	**27**	**11**	**0**	**12**	**12**
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/WD	51	0	76	27	86	3	76	51	14	0	10	8
WW/WW	54	5	68	8	**80**	**3**	**81**	**30**	**19**	**0**	**17**	**11**
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

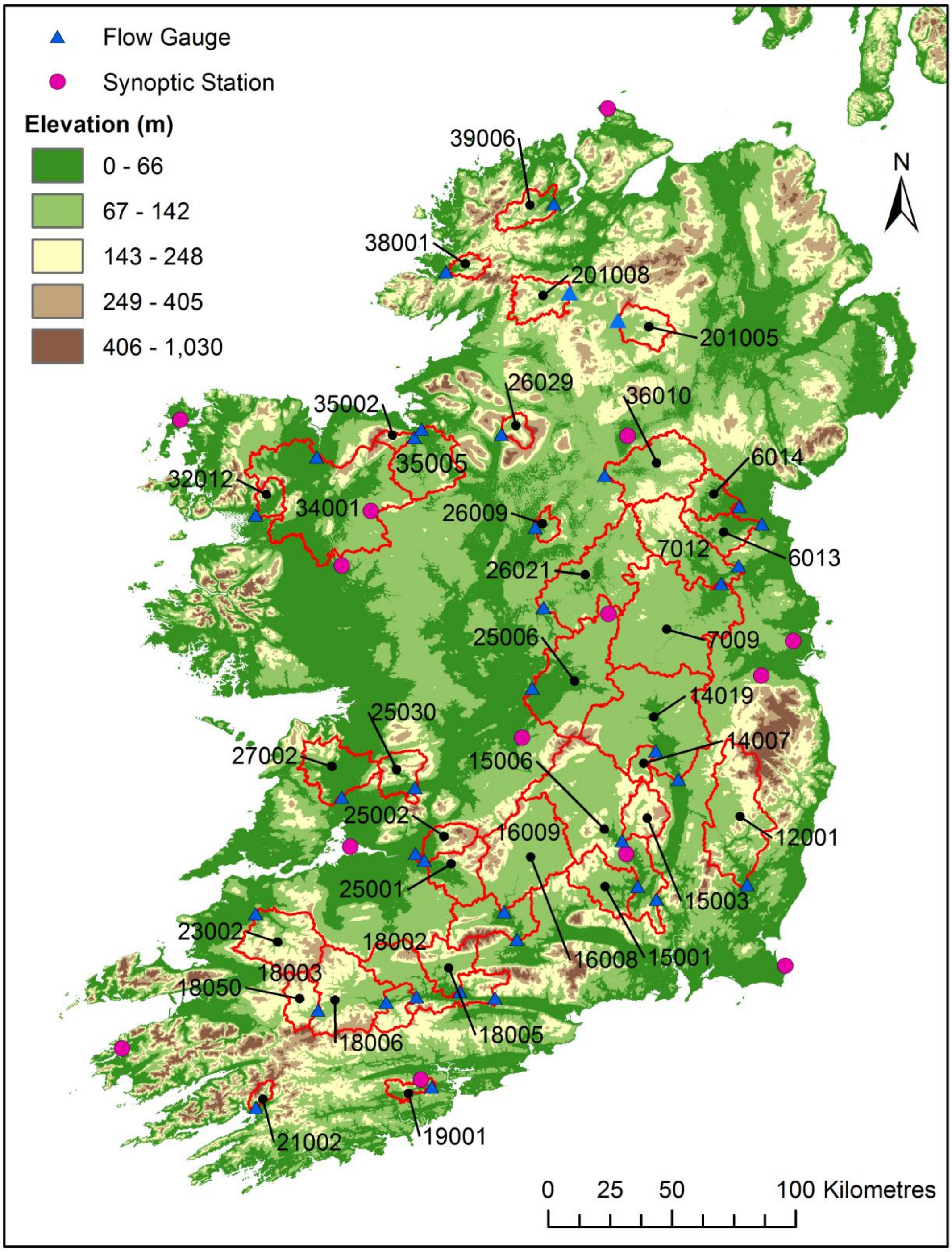

Figure 1. Study catchments and Met Éireann synoptic stations. Catchment identification codes are shown; red lines denote the respective catchment boundaries.

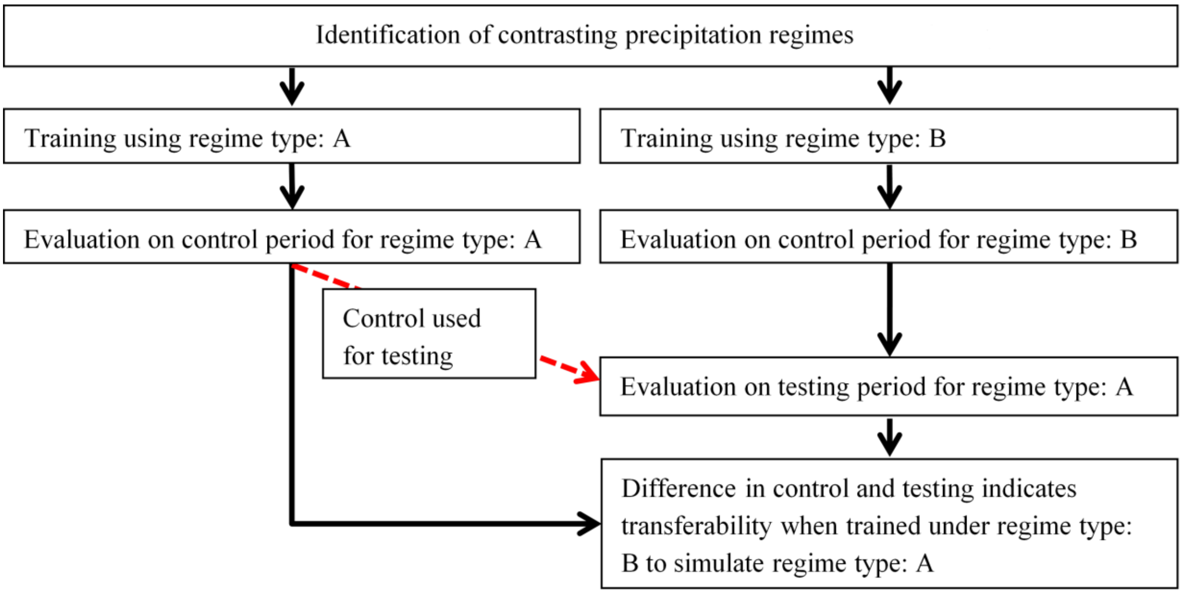

Figure 2. Flow diagram of the Differential Split Sample Testing (DSST) procedure used - incorporating training and performance assessment for an independent control and testing period respectively. This DSST procedure is used for estimation of weights in the Generalised Likelihood Uncertainty Estimation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

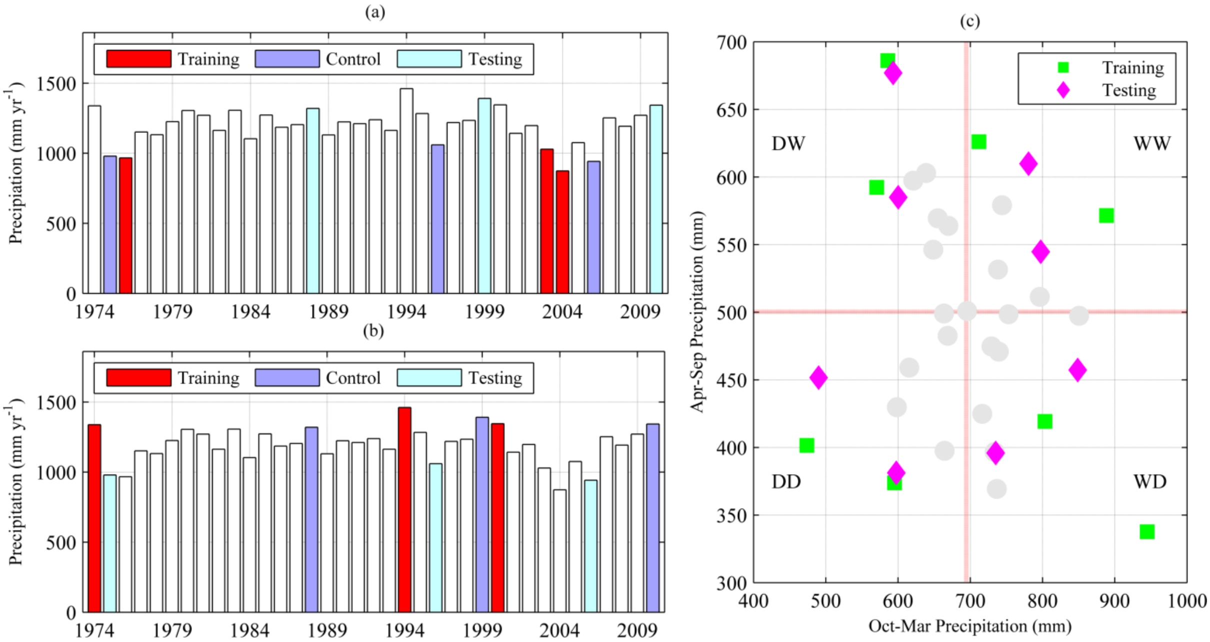

Figure 3. Panel (a) and (b): precipitation totals (1974-2010) for the hydrological year (1st October - 30th September; catchment ID 15006). Panel (c): winter (ONDJFM; *x-axis*) and summer (AMJJAS; *y-axis*) seasonal precipitation for six month periods of the hydrological year. Training and testing periods 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.

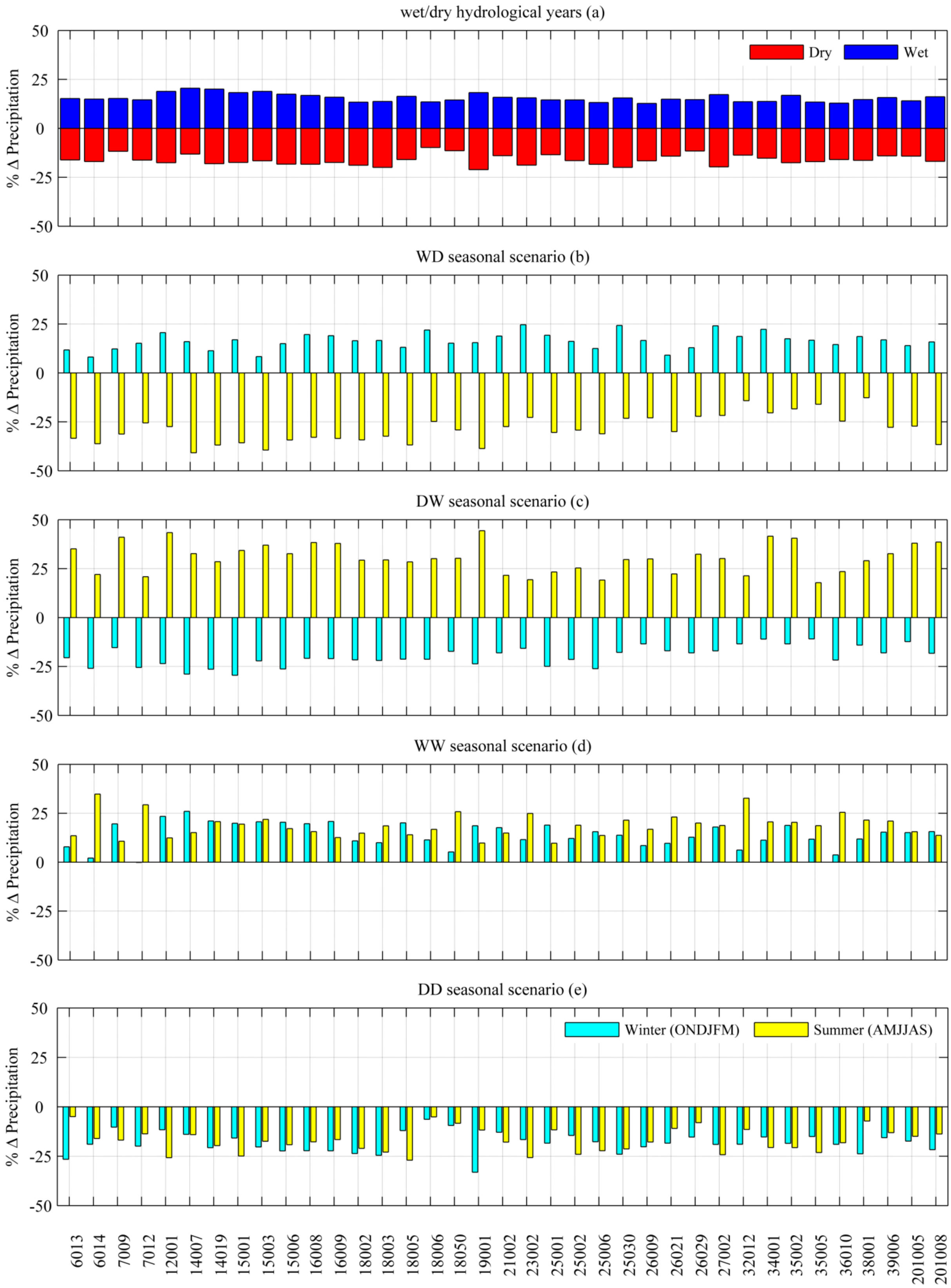

Figure 4. Percent differences in total seasonal/annual precipitation relative to 1976-2005 (Table 1) for DSST testing/control periods. Differences in contrasting 'wet'/'dry' hydrological years (1st October - 30th September) are shown (a). Relative differences for six-month winter (ONDJFM) and summer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

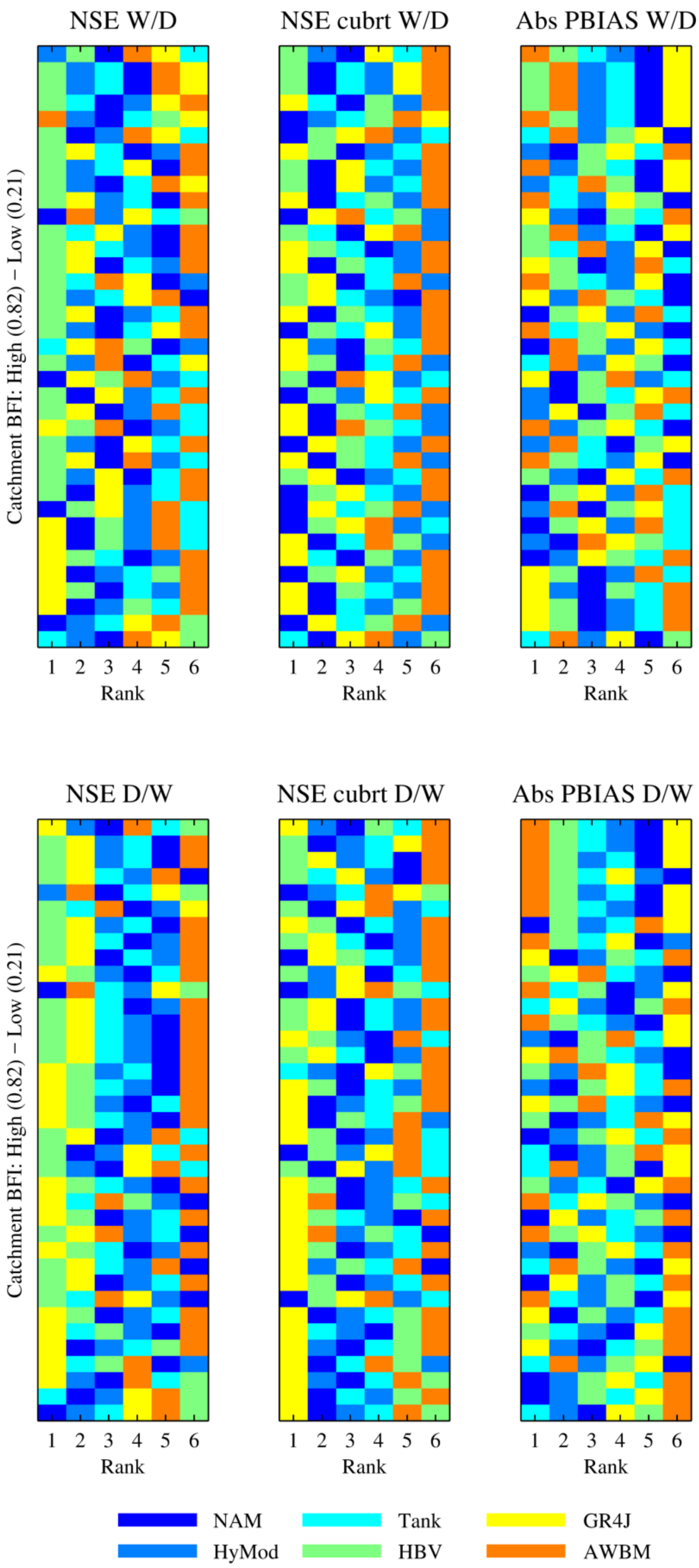

Figure 5. Individual model structures ranked (*x-axis*; best (1) to worst (6)) according to performance when tested under transference between 'wet'/'dry' annual regimes. Catchments (*y-axis*) are sorted according to their BFI in ascending order. Models are ranked according to the absolute (Abs) PBIAS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

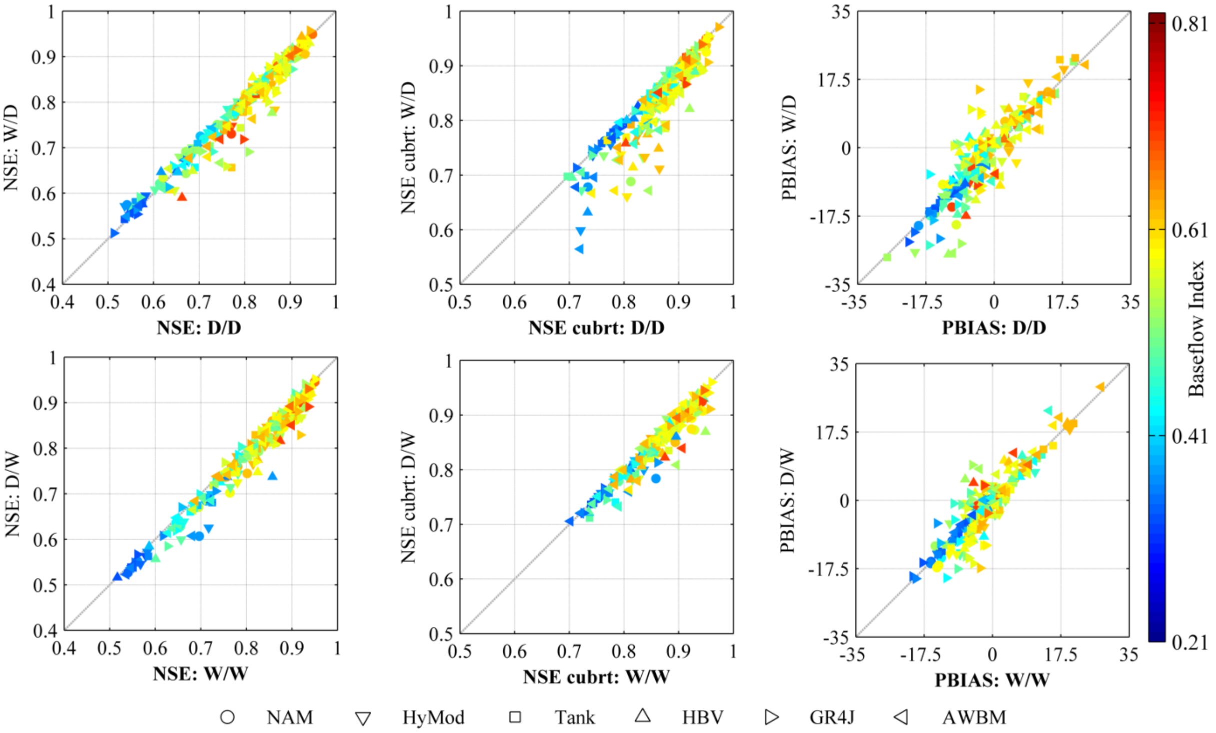

Figure 6. Testing (*y-axis*) and control (*x-axis*; shown in bold) results for two ('wet'/'dry') annual precipitation regimes. Models producing similar results for each DSST fall closer to the 45° line. Marker type corresponds to an individual model structure; markers are also coded using graduated shading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

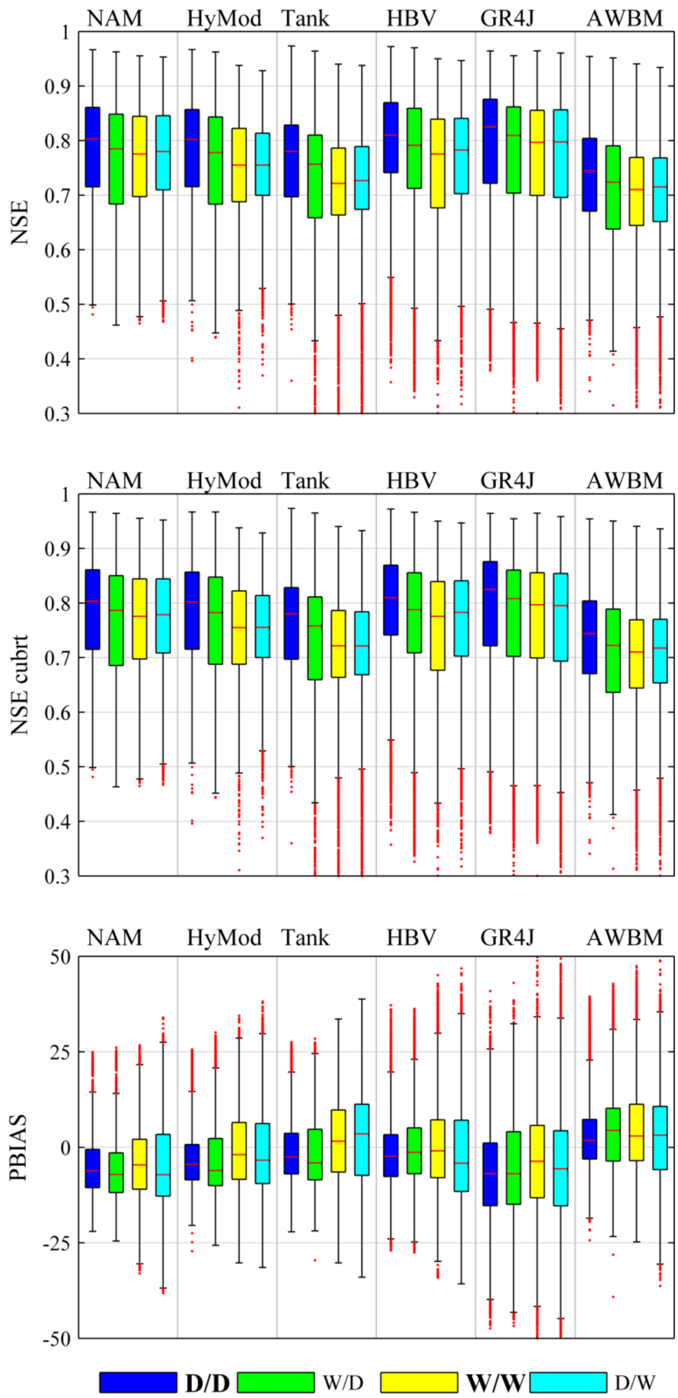

Figure 7. The combined performance of behavioural parameter sets for all catchments and rainfall-runoff models. DSST results are for two ('dry'/'wet') annual precipitation regimes are shown. The red line represents the median estimate; box edges denote the 25th and 75th percentiles. Whiskers are located at Q3+1.5×(Q3-Q1) and Q1-1.5×(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 <0.3 are not shown.

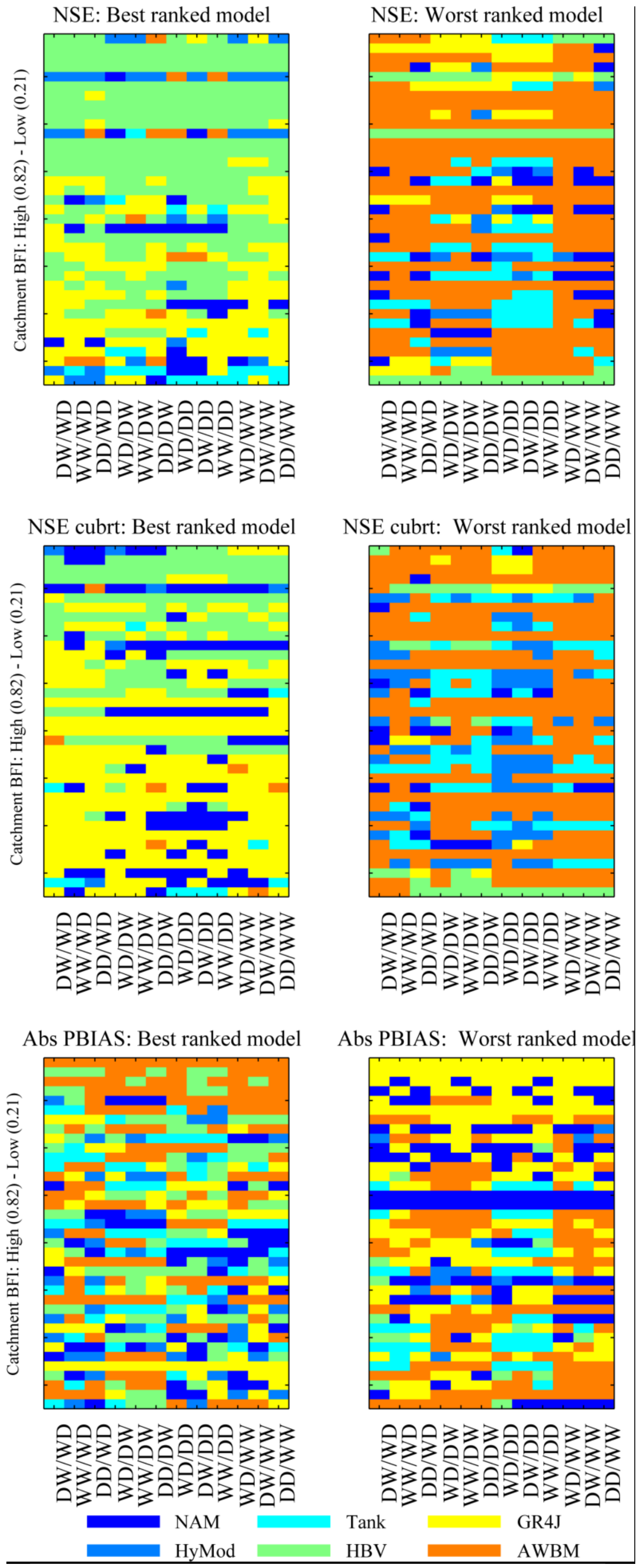

Figure 8. Best and worst ranked hydrological model according to DSST results for four (DD, WW, DW, WD) seasonal precipitation regimes (*x-axis*). Catchments (*y-axis*) are sorted according to their BFI in ascending order.

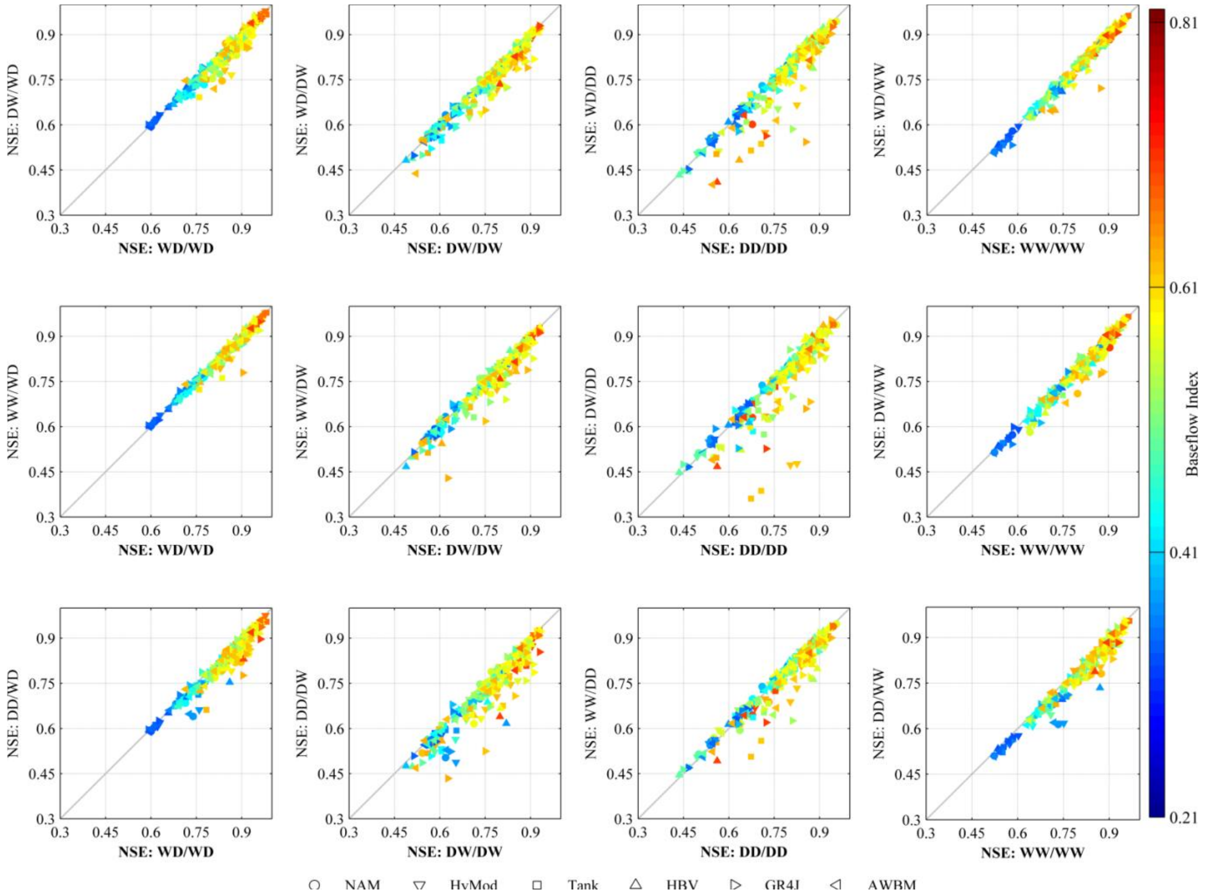

Figure 9. NSE testing (*y-axis*) and control (*x-axis*; shown in bold) results for four (DD, WW, DW, WD) seasonal precipitation regimes. Models producing similar results for each DSST fall closer to the 45° line. Marker type corresponds to an individual model structure; markers are also coded using graduated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

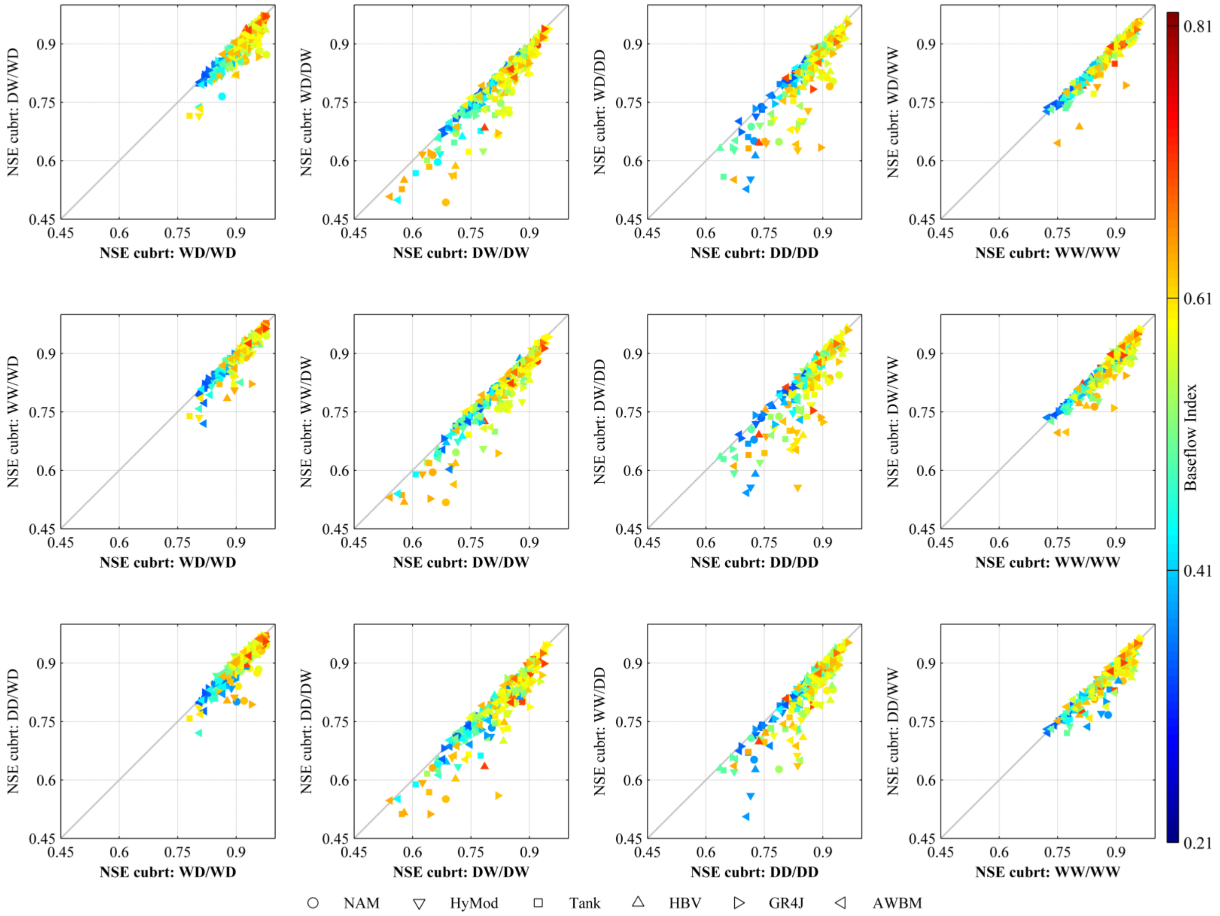

Figure 10. NSE_{cubrt} testing (*y-axis*) and control (*x-axis*; shown in bold) results for four (DD, WW, DW, WD) seasonal precipitation regimes. Models producing similar results for each DSST fall closer to the 45° line. Marker type corresponds to an individual model structure; markers are also coded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

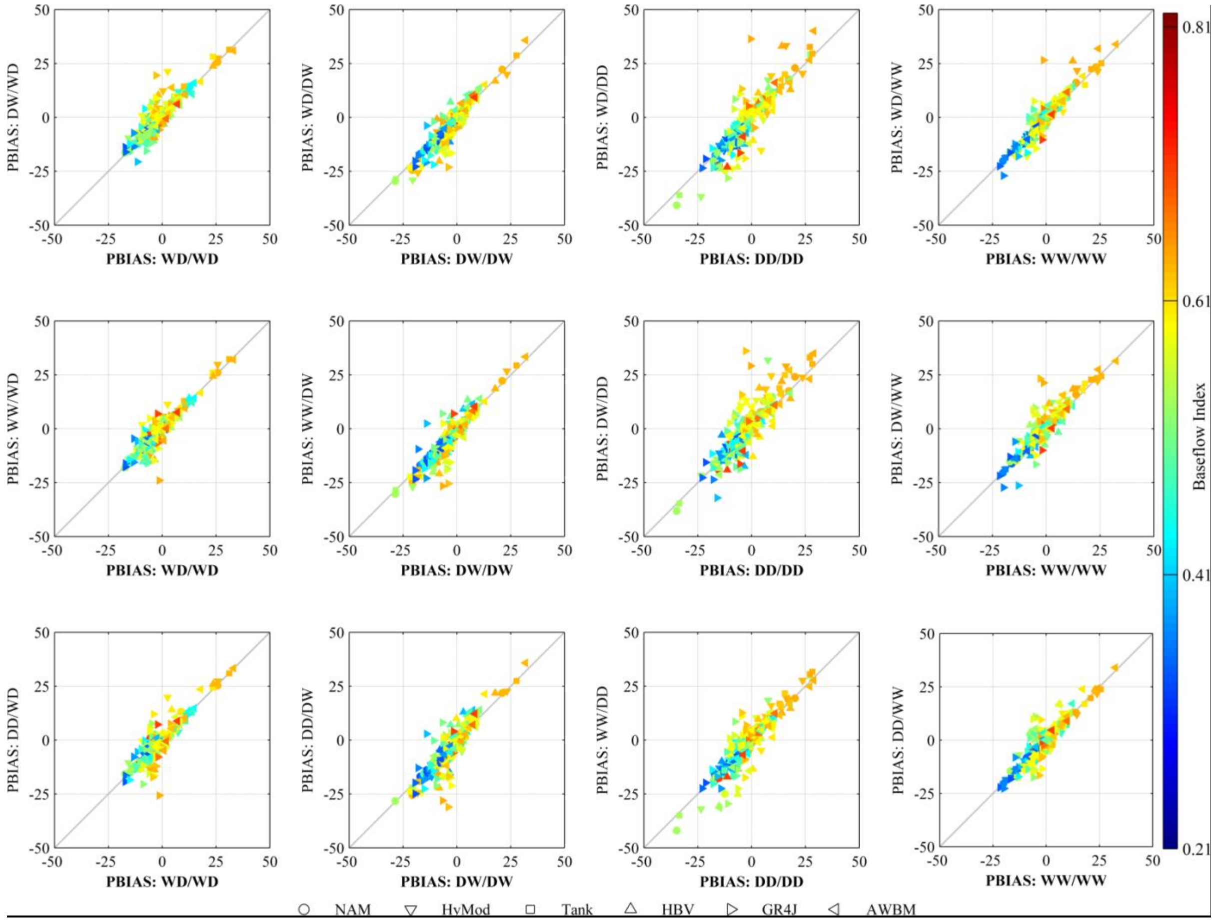

Figure 11. PBIAS testing (*y-axis*) and control (*x-axis*; shown in bold) results for four (DD, WW, DW, WD) seasonal precipitation regimes. Models producing similar results for each DSST fall closer to the 45° line. Marker type corresponds to an individual model structure; markers are also coded using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

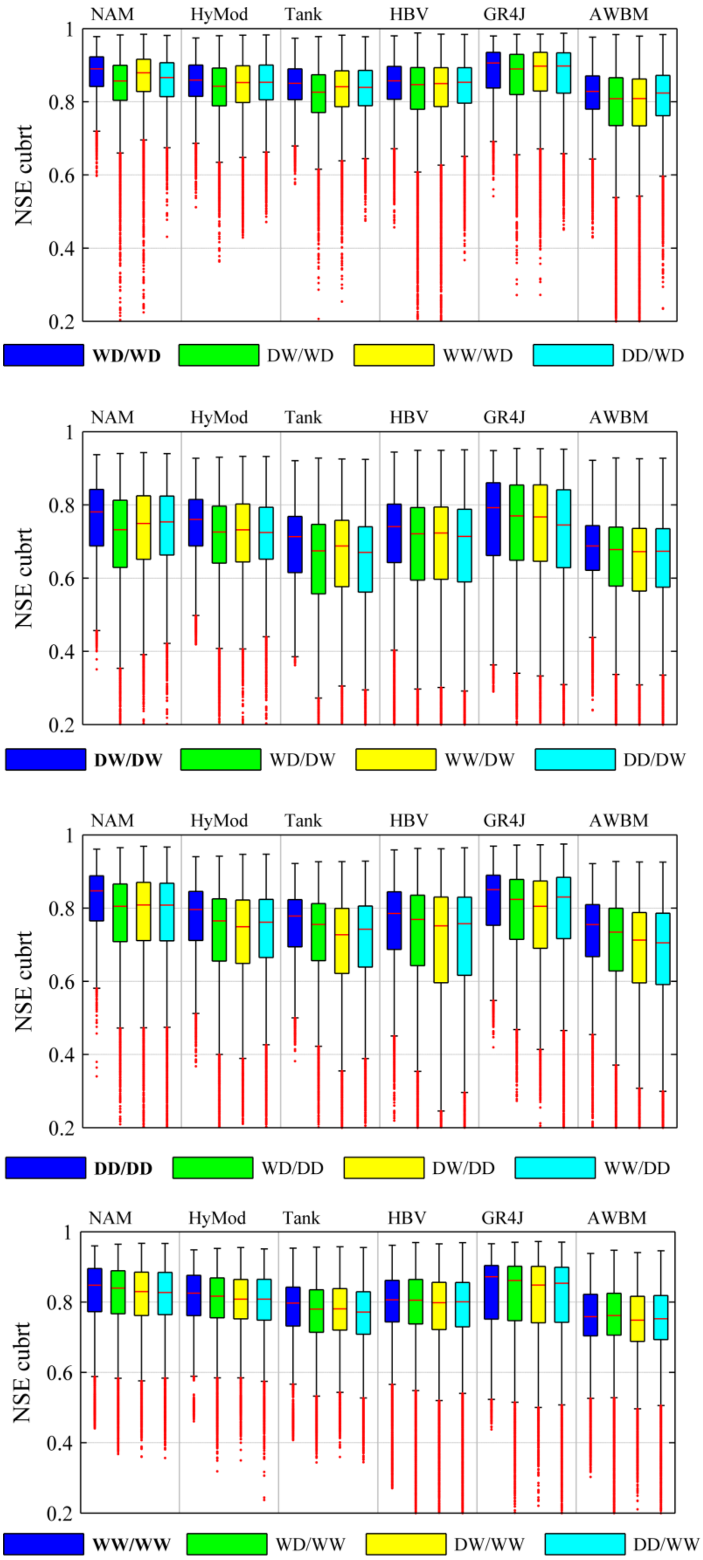

Figure 12. NSE_{cubrt} boxplots developed using the combined behavioural parameter sets of all six rainfall-runoff models for 37 catchments and four (DD, WW, DW, WD) seasonal precipitation regimes. The red line represents the median estimate; box edges denote the 25th and 75th percentiles. Whiskers are located at Q3+1.5×(Q3-Q1) and Q1-1.5×(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.

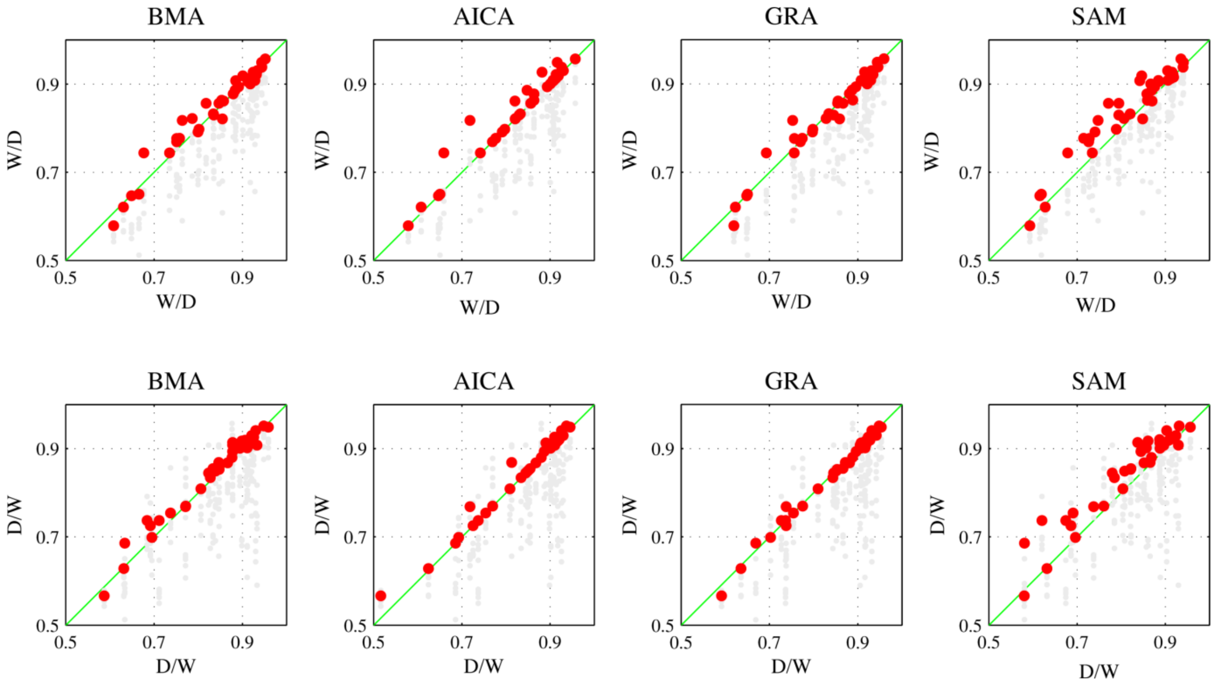

Figure 13. NSE scores for ‘wet’/‘dry’ DSST period obtained from four different model averaging techniques plotted against the corresponding NSE value from each model structure (grey dots). NSE values showing transference between the wettest/driest years for each catchment is plotted; red dots denote 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. $x=y$).

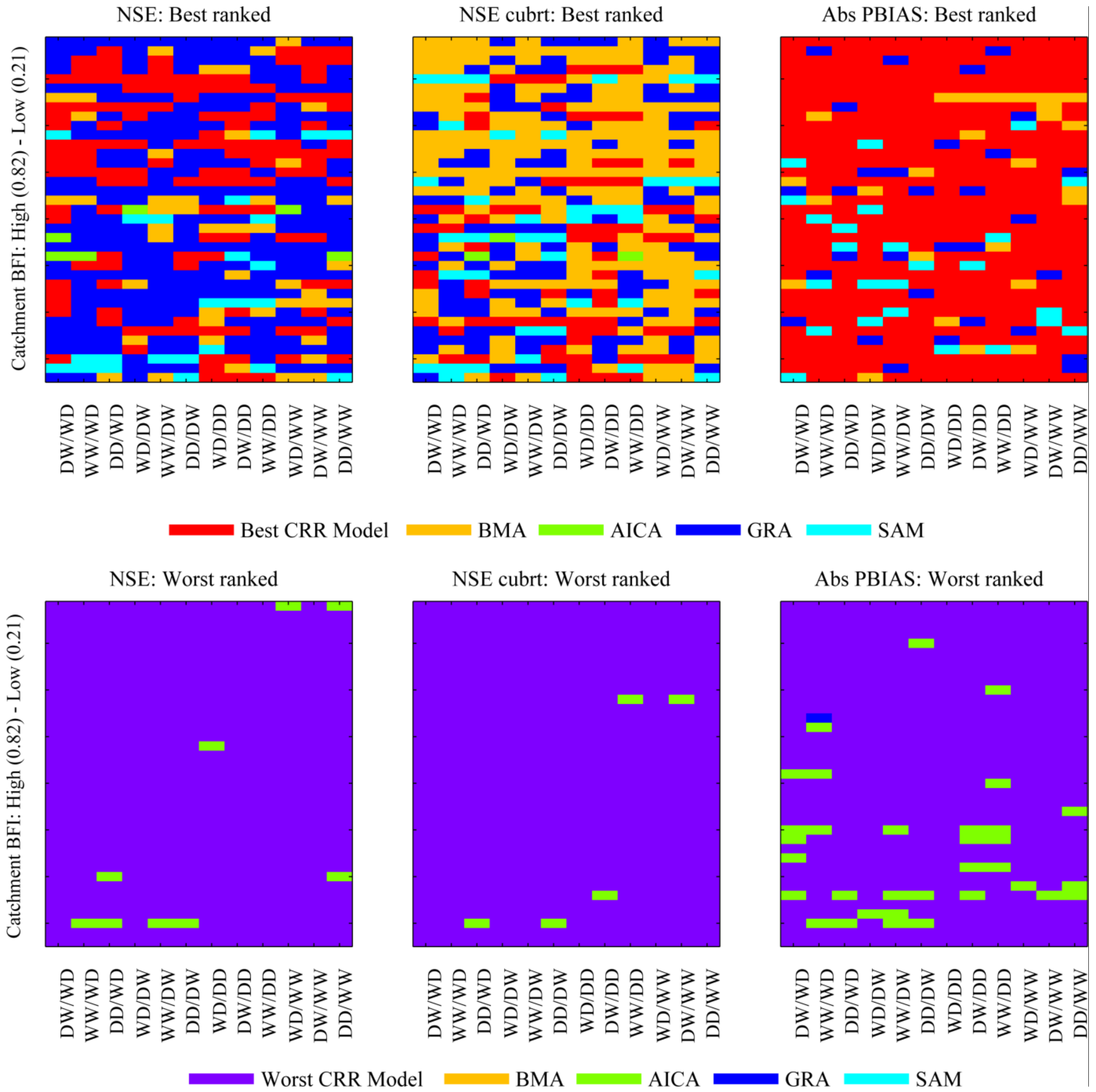

Figure 14. Best and worst ranked model averaging technique according to DSST results for four (DD, WW, DW, WD) seasonal precipitation regimes (*x-axis*). Also considered is the best and worst performing conceptual rainfall-runoff (CRR) model for each scenario. Catchments (*y-axis*) are sorted according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