Water Resources Management (2021) 35:113–133
https://doi.org/10.1007/s11269-020-02714-1

Evaluation of Sub-Selection Methods for Assessing Climate Change Impacts on Low-Flow and Hydrological Drought Conditions

Saeed Golian[1] · Conor Murphy[1]

Received: 1 May 2020 / Accepted: 8 November 2020 /
Published online: 21 November 2020

Abstract

A challenge for climate impact studies is the identification of a sub-set of climate model projections from the many typically available. Sub-selection has potential benefits, including making large datasets more meaningful and uncovering underlying relationships. We examine the ability of seven sub-selection methods to capture low flow and drought characteristics simulated from a large ensemble of climate models for two catchments. Methods include Multi-Cluster Feature Selection (MCFS), Unsupervised Discriminative Features Selection (UDFS), Diversity-Induced Self-Representation (DISR), Laplacian score (LScore), Structure Preserving Unsupervised Feature Selection (SPUFS), Non-convex Regularized Self-Representation (NRSR) and Katsavounidis–Kuo–Zhang (KKZ). We find that sub-selection methods perform differently in capturing varying aspects of the parent ensemble, i.e. median, lower or upper bounds. They also vary in their effectiveness by catchment, flow metric and season, making it very difficult to identify a best sub-selection method for widespread application. Rather, researchers need to carefully judge sub-selection performance based on the aims of their study, the needs of adaptation decision making and flow metrics of interest, on a catchment by catchment basis.

Keywords Climate change · General circulation models (GCMs) · Subs-selection · Uncertainty · Drought

✉ Saeed Golian
saeed.golian@mu.ie

[1] Irish Climate Analysis and Research UnitS (ICARUS), Department of Geography, Maynooth University, Maynooth, Co. Kildare, Ireland

1 Introduction

General Circulation Models (GCMs) are essential for studying changes in the climate system and informing adaptation. However, the unknown trajectory of future emissions, differences in the sensitivity of GCMs to anthropogenic forcing and the chaotic nature of the climate system, mean model projections are subject to much uncertainty (Knutti et al., 2010). Best practice dictates that impact analyses adequately account for uncertainty (Clark et al., 2016). Hence output from large-scale model experiments – comprising simulations from multi-model and/or perturbed physics ensembles – e.g. CMIP (*Coupled Model Intercomparison Project;* Taylor et al., 2012)*,* CORDEX (Giorgi et al., 2009) and *climateprediction.net* (Stainforth et al., 2005), represent an invaluable resource. While large ensembles allow better investigation of climate risk, such datasets are assembled on the basis of opportunity; consequently suffering from a lack of model independence and biased representation of constituent uncertainties (Pirtle et al. 2010; Knutti et al., 2010; Masson and Knutti, 2011; Mendlik and Gobiet, 2016). Furthermore, in integrated assessments where additional stressors (e.g. land-use change, socio-economic scenarios) must be considered, identifying a reduced set of EMs helps integrate climate into already complex decision processes. Similarly, for impact assessment, while large ensembles can provide a richer picture of the range of possible changes (Fung et al., 2013), they may also limit the complexity of impact models used and depth of analysis permissible (Christierson et al., 2012), including of additional uncertainties (e.g. impact model, data uncertainties) (Broderick et al., 2016; Broderick et al., 2019). Thus, sub-selection offers a potential means of efficiently navigating the uncertainty cascade (Wilby and Dessai, 2010; Smith et al., 2018).

Many sub-selection methods have been developed and used in different hydroclimatological studies. Mendlik and Gobiet (2016) developed a method for sub-selection based on principal component and cluster analysis and showed that their method reduced computational costs for climate impact modeling. Wang et al. (2018) applied two selection methods, K means clustering and the Katsavounidis–Kuo–Zhang (KKZ) method, to select subsets of 50 climate simulations over two sub-catchments and found that KKZ performs better than K-means. Seo et al. (2019) also used the KKZ algorithm for sub-selecting climate scenarios, demonstrating that KKZ could reduce the number of GCMs in an ensemble while maintaining the ranges of multiple climate extremes indices. Ross and Najjar (2019) compared the performance of sub-selection methods including hierarchical clustering, K-means and KKZ on projections of runoff change for five US watersheds. They found the KKZ model performs satisfactorily over all watersheds and number of selected models.

Given their potential utility, we examine the ability of seven widely-used unsupervised sub-selection methods to capture low flow and drought characteristics simulated from a large ensemble of climate models for two Irish catchments. Sub-selection methods are applied to identify reduced-size ensembles from monthly change factors of precipitation and temperature from the CMIP5 ensemble. We focus on the ability to sub-selection methods to capture the median and range (90% confidence intervals) simulated for each catchment by the CMIP5 ensemble. Finally, we test the sensitivity of sub-selection methods to the number of members selected and discuss the value of sub-selection approaches for climate impact assessment and adaptation decision-making.

2 Data and Methods

2.1 Study Catchments and Hydrological Model

Study design is displayed in Fig. 1a. Both catchments are located in eastern Ireland (Fig. 1b) and differ in terms of key characteristics, i.e. size, average annual precipitation and runoff generation processes (as indicted by the baseflow index (BFI)). BFI measures the proportion of streamflow derived from baseflow or saturated groundwater storage as opposed to direct runoff (Gustard et al. 1992). For more details on how BFI is calculated, readers are referred to Mills et al. (2014). Catchment 06030 is a small catchment (10.2 Km^2) with average precipitation of 1150 mm and mean annual temperature of 8.5 °C. Hydrological response is dominated by surface runoff with a low BFI of 0.37. By contrast, catchment 14,019 has an area of 1697 Km^2, is located in the drier southeast with annual average precipitation of 869 mm, average annual temperature of 9.3 °C and a larger groundwater contribution to streamflow with a BFI of 0.63. For each catchment daily discharge observations were obtained from the Office of Public Works (OPW) for the years 1961–2017 for catchment 14,019 and 1975–2017 for catchment 6030. Observed daily precipitation and temperature data were derived from 1 × 1 km grids (Walsh 2012) and area-averaged for each catchment for the period concurrent with discharge observations. To calculate potential evapotranspiration we used the formula proposed by Oudin et al. (2005):

$$\begin{array}{ll} PET = \dfrac{R_e}{\lambda\rho}\dfrac{T_a+5}{100} & if\ T_a > 0 \\ PET = 0 & Otherwise \end{array} \tag{1}$$

where PET is the rate of evapotranspiration (mm/day), R_e is extraterrestrial radiation (MJ kh^{-1}), λ is latent heat flux in (MJ kg-1), ρ is density of water (kg m-3) and T_a is mean daily air temperature (C^o). Extraterrestrial radiation is a highly regular variable and thus temperature is the key component explaining fluctuations of PET (Koutsoyiannis 2013). Consequently, as we just consider the effects of precipitation and temperature as the main drivers for climate change in our study, we neglect projections of extraterrestrial radiation.

We employ the GR4J hydrological model to simulate observed and projected changes in discharge. The ability of this model to capture multiple hydrological signatures for Irish catchments, including groundwater dominated catchments has been shown by Broderick et al. (2016) and Broderick et al. (2019). GR4J is a simple lumped two-storage conceptual rainfall-runoff model developed as part of the airGR R hydrological modelling package (Coron et al. 2017). The model takes precipitation (P) and potential evapotranspiration (PET) as inputs and is based on four parameters: X1, the maximum soil moisture storage (mm); X2, the groundwater exchange coefficient (mm); X3, the maximum capacity of the routing storage (mm); and X4, the time peak ordinate of hydrograph unit UH1 or flow delay (day). Fig. S1 shows the structure of the GR4J model.

Observations prior to 2002 were used for model calibration and post 2002 used for verification. The first year of observations was used as a warmup period. To calibrate parameters, we applied Memetic Algorithms with Local Search Chains (MA-LS-Chains). These are hybridizations of genetic algorithms with local search methods (Bergmeir et al., 2016). The non-parametric Kling-Gupta efficiency (KGE) (Pool et al. 2018) was considered as the objective function:

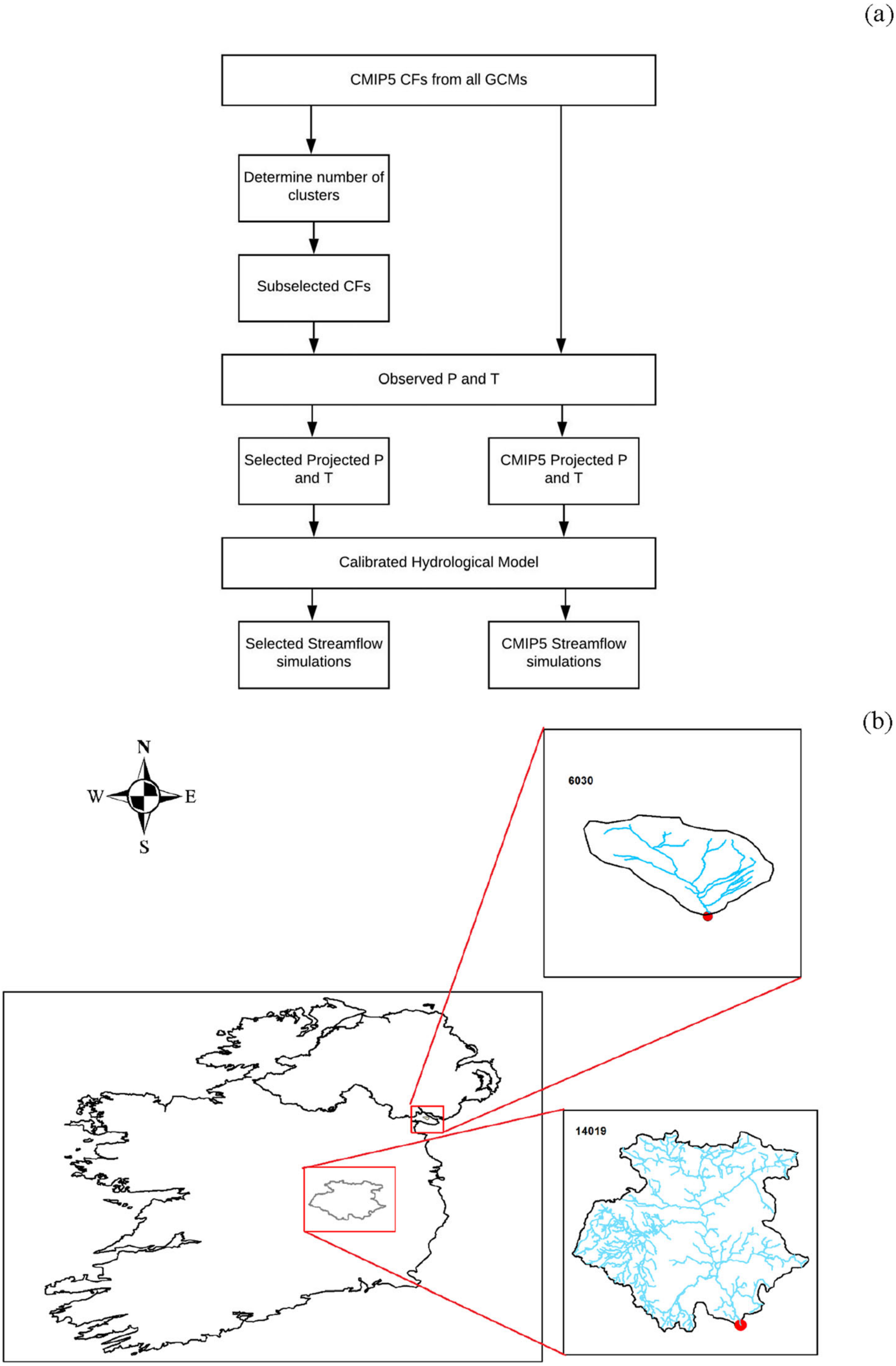
(a)
CMIP5 CFs from all GCMs
Determine number of clusters
Subselected CFs
Observed P and T
Selected Projected P and T
CMIP5 Projected P and T
Calibrated Hydrological Model
Selected Streamflow simulations
CMIP5 Streamflow simulations
(b)
N
W
E
S
6030
14019

Fig. 1 a) A diagram of the modeling framework of this study and b) Location of the selected catchments In Ireland

$$KGE = 1 - \sqrt{(r_s - 1)^2 + (\beta - 1)^2 + (\alpha_{NP} - 1)^2} \tag{2}$$

$$\beta = \frac{\overline{Q}sim}{\overline{Q}obs} \tag{3}$$

$$\alpha_{NP} = 1 - 0.5\sum\left|\frac{Q_{sim}(I(k))}{n\overline{Q}_{sim}} - \frac{Q_{obs}(J(k))}{n\overline{Q}_{obs}}\right| \tag{4}$$

Where r_s is Spearman's rank correlation, (k) and J(k) are the time steps when the kth largest flow occurs within the simulated and observed times series, respectively. $\overline{Q}_{obs}$ and $\overline{Q}_{sim}$ are average flow for observed and simulated datasets, respectively. The non-parametric Kling-Gupta efficiencies range from −∞ to 1. The closer to 1, the more accurate the model is. In comparison to ordinary KGE (Gupta et al. 2009), non-parametric KGE doesn't require assumptions about data linearity, normality or outliers.

2.2 Design of Model Experiment

For both catchments we extract change factors for precipitation (P) and temperature (T) from the ensemble of climate models within the CMIP5 archive. To establish the central estimate and range of change projected for low flow and drought metrics from the full ensemble, we use all derived change factors to force GR4J to examine changes in low flow and drought indicators for a period representing mid-century. To compare the effectiveness of sub-selection methods to capture the central estimate and range of change in flow indicators derived from the full ensemble we take the following steps. First, to identify the size of sub-selected ensembles, cluster analysis is performed on change factors for P and T from the full ensemble using two different approaches (Gap Statsitics and Dedrograms). Next, we examine seven different sub-selection routines for their ability to capture key features of the full ensemble. For each sub-selection method, change factors representing the selected climate model members are used to force GR4J to examine changes in flow indicators for the 2050s. Lastly, we compare changes in flow indicators derived from both the full CMIP5 ensemble and sub-selected ensembles to evaluate differences.

2.3 Climate Change Projections

To quantify future climate change we employ the large ensemble of climate model projections contained within the Coupled Model Intercomparison Project Phase 5 (henceforth CMIP5) (Taylor et al., 2011). Only GCMs forced with Representative Concentration Pathway (RCP) 8.5 are employed to maximize distinction between simulated and observed values. Table S1 of supplementary information details the CMIP5 members used. We employ a simple downscaling technique using deriving monthly change factors for temperature and precipitation based on differences between our baseline (1976–2005) and the 2050s (2040–2069). The CF approach assumes that while GCMs may be affected by biases in simulating baseline climatology, their response to altered forcing and associated climate change signal is well represented. While it has a number of limitations, the method allows rapid assessment of climate impacts and can be expanded to include other (higher order) statistics.

Monthly series for precipitation and temperature were extracted for each CMIP5 ensemble member for the grid(s) overlying the respective catchments. Once derived, CFs based on relative (precipitation) and absolute (temperature) differences between the baseline (1976–2005) and 2050s (2040–2069) were estimated and applied to baseline observations. Equations 1 and 2, outline the scaling used for temperature and precipitation respectively.

$$T_{scal,fut,s} = T_{obs,s} + \left(\overline{T}_{GCM,fut,s} - \overline{T}_{GCM,\ \ base,s}\right) \quad (5)$$

$$P_{scal,fut,s} = P_{obs,s} \times \left(\overline{P}_{GCM,fut,s} / P_{GCM,\ \ base,s}\right) \quad (6)$$

Here observed temperature ($T_{obs,\ s}$) for the baseline period is adjusted to obtain a seasonal mean (s) value representative of future conditions ($T_{scal,\ fut,\ s}$) by adding the difference in temperature projected by each EM $\left(\overline{T}_{GCM,fut,s} - \overline{T}_{GCM,\ \ base,s}\right)$. For precipitation ($P_{scal,\ fut,\ s}$) the observed data ($P_{obs,\ s}$) is multiplied using the ratio of future to baseline simulated periods $\left(\overline{P}_{GCM,fut,s} / P_{GCM,\ \ base,s}\right)$. The adjusted P and T (PET) were then used to force the calibrated conceptual hydrological model, GR4J for the selected catchments.

2.4 Flow Indicators Evaluated

The following indicators were used to evaluate the ability of sub-selection methods to capture the median and 90% uncertainty ranges of projected changes in hydrological droughts and low flows.

• *Average monthly discharge.*

• *Seasonal minimum 7-day low flows*: the minimum cumulative discharge for running 7-days totals for each season. Seasons are winter [DJF], spring [MAM], summer [JJA] and autumn [SON].

• *Standardised River Flow Index (SRI):* used to evaluate drought events and derived using a non-parametric standardization approach (Hao el al. 2014, Farahmand and AghaKouchak 2015). First, the empirical probabilities of the simulated/observed discharge values are computed. By applying an empirical approach, we avoid assumptions about the underlying distribution of discharge data (Farahmand and AghaKouchak 2015). To capture seasonal variation, standardisation is applied to 3 month accumulated runoff (SRI3). A drought event was identified when SRI3 drops below a threshold (zero). The ability of various sub-selection methods to capture changes in drought characteristics from the parent ensemble is evaluated using the percent error in number of drought events, drought duration and average drought intensity. Percent relative error in the number of drought events for different sub-selection methods ($NoEv_{sun-selection}$) compared to the case with all GCMs ($NoEv_{All\_GCM}$) is calculated as follows:

$$Percent\ Error\ (\%) = \frac{\left(NoEv_{All\_GCM} - NoEv_{sub-selection}\right)}{NoEv_{All\_GCM}} \times 100 \quad (7)$$

2.5 Selecting the Number of Sub-Ensemble Members

The primary aim of sub-selection is to reduce the effective size of the parent ensemble by removing redundant model simulations, while retaining as much spread in the ensemble distribution as possible. We employ cluster analysis (gap statistics and dendrograms) to inform the number of independent members to be retained in sub-selected ensembles. We also evaluate the sensitivity of results to number of clusters selected, varying the optimum number by ± 50%.

2.5.1 Gap Statistics

The Gap statistic is a standard method for determining the number of clusters in a dataset. The Gap statistic standardizes the graph of $\log W_k$, where W_k is the intra-cluster dispersion, by comparing it to its expectation under an appropriate null reference distribution of the data. The following measure represents the sum of intra-cluster distances between points in a given cluster C_k containing n_k points:

$$D_k = \sum_{x_i \in C_k} \sum_{x_j \in C_k} \left\| x_i - x_j \right\|^2 = 2n_k \sum_{x_i \in C_k} \left\| x_i - \mu_k \right\|^2 \tag{8}$$

Adding the normalized intra-cluster sum of squares gives a measure of the compactness of our clustering:

$$W_k = \sum_{x_i \in C_k} \frac{1}{2n_k} D_k \tag{9}$$

The Gap Statistic approach aims to standardize the comparison of $\log W_k$ with a null reference distribution of the data, i.e. a distribution with no obvious clustering (Tibshirani et al., 2001). The estimate for the optimal number of clusters K is the value for which $\log W_k$ falls the farthest below this reference curve. This information is contained in the following formula for the gap statistic:

$$\mathrm{Gap}_n(k) = E_n^*\{\log W_k\} - \log W_k \tag{10}$$

Where E_n^* denotes expectation under a sample of size n from the reference distribution. Our estimate $\widehat{k}$ will be the value maximizing $\mathrm{Gap}_n(k)$ after accounting for the sampling distribution. For more details see Tibshirani et al. (2001).

2.5.2 Dendrograms

Agglomerative hierarchical clustering algorithms build a cluster hierarchy commonly displayed as a dendrogram. The algorithm begins with each object in a separate cluster. At each step, the two clusters most similar are joined into a single new cluster. The height of the dendrogram represents the distance or dissimilarity between clusters (Koga et al., 2007). The larger the height difference between two agglomerations (or clusters), the larger their dissimilarity. To measure dissimilarity between two clusters (cluster centers), different distance measures, e.g. Euclidean can be used.

2.6 Sub-Selection Methods

We consider seven unsupervised sub-selection methods.

2.6.1 Diversity-Induced Self-Representation (DISR)

Diversity-Induced Self-Representation (DISR) is an unsupervised feature selection method (Liu et al. 2017) that effectively selects features with both representativeness and diversity. The similarity between the *i*th and *j*th models can be calculated using their dot product weight as:

$$S_{ij} = f_i^T . f_j, i, j = 1, 2, \ldots, m \tag{11}$$

where m is the number of models. Each S_{ij} indicates how well f_i differentiatesf_j. Readers are referred to Liu et al. (2017) for more details.

2.6.2 Laplacian Score (LScore)

Laplacian Score (LScore) is an unsupervised linear feature extraction method. For each feature/variable, it computes the Laplacian score based on an observation that data from the same class are often close to each other. Let Lr denote the Laplacian Score of the r-th feature (model). Let f_{ri} denote the i-th sample of the r-th feature, $i = 1, \ldots, p$. The LScore algorithm can be stated as follows (He et al. 2006):

1. Construct a nearest neighbor graph $G$with p nodes. The i-th node corresponds to x_i . An edge can be put between nodes i and j if x_i and x_j are "close", i.e. x_i is among k nearest neighbors of x_j or x_j is among k nearest neighbors of x_i. When the label information is available, one can put an edge between two nodes sharing the same label.
2. If nodes i and j are connected (or are neighbors), put $S_{ij} = e^{-\|x_i - x_j\|}\,\frac{2}{t}$, where t is a suitable constant. Otherwise, put $S_{ij} = 0$. The weight matrix S of the graph models the local structure of the data space.
3. For the r-th feature (model), we define:

$$\mathbf{f}_r = \left[f_{r1}, f_{r2}, \ldots, f_{rp}\right]^T, D = diag(S\mathbf{1}), \mathbf{1} = [1, \ldots, 1]^T, L = D{-}S \tag{12}$$

where the matrix L is often called the graph Laplacian. Let

$$\widetilde{\mathbf{f}}_r = \mathbf{f}_r - \frac{\mathbf{f}_r^T D\mathbf{1}}{\mathbf{1}^T D\mathbf{1}} \tag{13}$$

4. Compute the Laplacian Score of the r-th feature as follows:

$$L_r = \frac{\widetilde{\mathbf{f}}_r^T L \widetilde{\mathbf{f}}_r}{\widetilde{\mathbf{f}}_r^T D \widetilde{\mathbf{f}}_r} \tag{14}$$

2.6.3 Multi-Cluster Feature Selection (MCFS)

By using spectral analysis techniques, Multi-Cluster Feature Selection (MCFS) suggests a principled way to measure the correlations between different features without label information (Cai et al., 2010). MCFS has been shown to preserve the cluster structure of data. This method utilizes spectral regression with $l1$- norm regularization to select the features. For more details, readers are referred to Cai et al. (2010).

2.6.4 Non-convex Regularized Self-Representation (NRSR)

In the Non-convex Regularized Self-Representation (NRSR) algorithm, features can be represented by a linear combination of other features, and propose to impose $L_{2,\,p}$ -norm ($0 < p < 1$) regularization on self-representation coefficients for unsupervised feature selection. This method is more efficient in selecting salient features compared to methods using conventional $L_{2,\,1}$ norm regularization. For more details, readers are referred to Wang et al. (2015) and Zhu et al. (2017).

2.6.5 Structure Preserving Unsupervised Feature Selection (SPUFS)

Structure Preserving Unsupervised Feature Selection (SPUFS) is an unsupervised feature selection method based on a self-expression model to capture the relationships between the features without learning the cluster labels. In this method, there is a cost function which has two penalties, sparsity and preservation of local structure. Each feature is reconstructed through a linear combination of all features in the original feature space and considering the local manifold structure of the data using an object similarity matrix. More details and formulation can be found in Lu et al. (2018).

2.6.6 Unsupervised Discriminative Features Selection (UDFS)

UDFS aims to find discriminative features under an unsupervised learning framework. The class label can be predicted by a linear classifier and iteratively updates its discriminative nature using $\ell_{2,\,1}$-norm minimization while attaining row-sparsity scores for selecting features. It has been shown that this algorithm can outperform other unsupervised algorithms to select discriminative features for data representation (Yang et al., 2011). However, its orthogonal constraint on the feature selection projection matrix is unreasonable since feature weight vectors are not necessarily orthogonal with each other in nature (Qian and Zhai, 2013).

2.6.7 Katsavounidis–Kuo–Zhang (KKZ)

The Katsavounidis–Kuo–Zhang (KKZ) method (Katsavounidis et al. 1994) identifies a set of optimal seed cases as initial centroids in K means clustering, and was introduced by Cannon (2015) for selection of subsets of climate simulations. The KKZ algorithm selects members in a recursive manner that cover a spread of multivariate space comprehensively.

The specific procedure is as follows:

1. The climate simulation closest to the centroid of the whole ensemble is selected as the first simulation.

2. The simulation farthest from the first selected simulation is selected as the second simulation. The Euclidean distance is applied to calculate the distance, $d(i, j)$, between two models (the ith and jth models):

$$d(i,j) = \sqrt{\sum_{p=1}^{P} \left(y_{ip} - y_{jp}\right)^2} \tag{15}$$

Where y_{ip} and y_{jp} represents the value of the pth variable for the ith and jth models, respectively.

3. Subsequent simulations are selected as follows.

1 For each remaining simulation, its distances to every previously selected simulation are calculated.
2 Each remaining simulation is designated with the minimum distance among all distances calculated in step 3(i).
3 The simulation with the largest minimum distance, which is designated in step 3(ii), is selected as the next selected simulation.

3 Results and Discussion

3.1 Model Calibration and Verification

For catchments 6030 and 14,019, the non-parametric KGE scores derived are 0.93 and 0.92 for calibration and 0.85 and 0.88 for validation, respectively. Calibrated X_1, X_2, X_3 and X_4 parameters are 100(mm), 0.63(mm), 33.38(mm) and 1.4(day) for catchment 6030 and 523.04(mm), 0.07(mm), 32.84(mm) and 2.82(day) for catchment 14,019. Fig. S2 shows the scatterplot of simulated-observed flow for training and verification periods. The correlation coefficient equaled to 0.86 and 0.70 for catchment 6030 and 0.92 and 0.91 for catchment 14,019 over training and verification. RMSE of 1.57 and 0.37 for training and 3.62 and 0.46 for verification were returned for catchment 6030 and 14,019, respectively. Hydrological model performance for the groundwater-dominated catchment (14019) is better, compared to the surface water dominated catchment (6030).

3.2 Cluster Analysis

Based on the Gap-statistics method (Fig. S3) 11 clusters/members are proposed for catchment 6030 and 10 clusters for catchment 14,019. Using the dendrogram method, the approximate number of clusters was obtained by cutting the dendrogram tree with a horizontal line at a height where the line can traverse the maximum distance up and down without intersecting the merging point. For example in our case (Fig. S4), using the Gap-statistics method, the cutting height should be between 0.75 and 0.80 which suggests 10 clusters for both catchments. Given the similar results of both methods, 10 clusters were selected for both study catchments.

3.3 Sub-Selection and Evaluation of Low Flows

Table S2 (supplementary material) provides the resultant 10-member sub-ensembles selected using each of the seven sub-selection methods. There is considerable difference between the selected members identified using each sub-selection algorithm. Some methods, i.e. KKZ, SPUFS and UDFS, selected different versions of same GCMs while others, e.g. DISR and NRSR, have more variability among selected GCMs to represent the whole ensemble range.

To examine how well sub-selection methods represent the entire CMIP5 ensemble, we compare simulated flow metrics from each sub-selected ensemble with those forced using the full CMIP5 ensemble. Figure 2 shows results for average monthly discharge. For catchment 6030, riverflow is not well represented in low flow months, i.e. July, August and September using LScore and MCFS methods, but all other sub-ensembles performed satisfactory and similarly to the CMIP5 ensemble in other months. This may be attributed to the challenge of capturing low-flows. It has been shown by many researchers that low-flow simulation highly depends on the study region, season and also performance criteria the hydrological models are calibrated based on (Nicolle et al., 2014). For catchment 14,019, DISR and UDFS didn't preserve the upper and lower ranges in low flow months (July, August and September) while

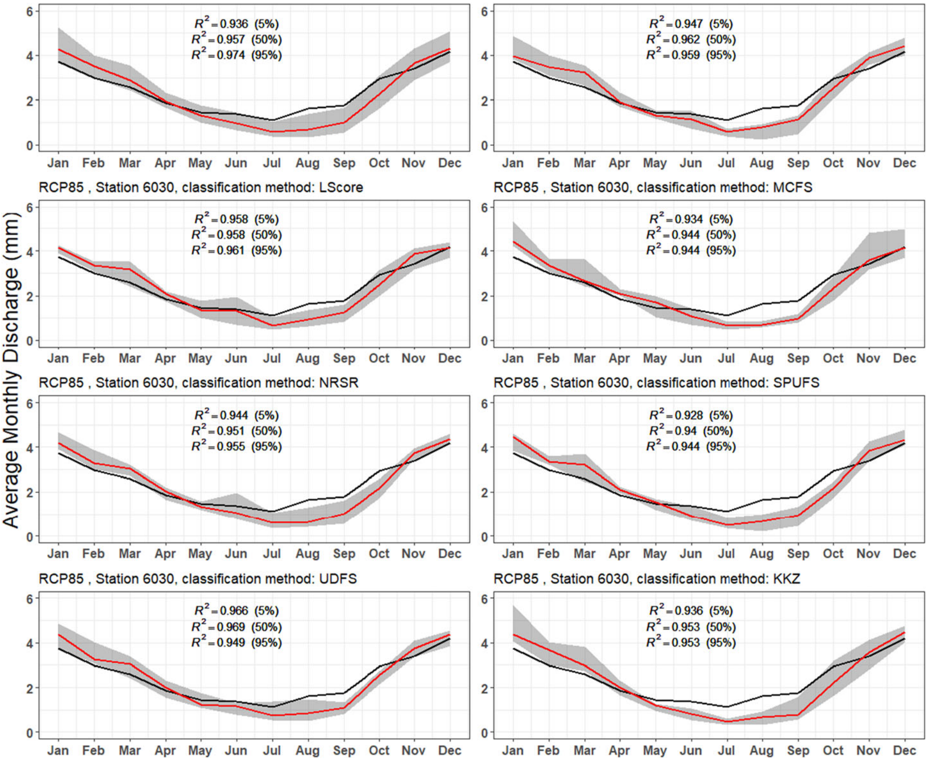

Fig. 2 Average monthly streamflow for catchment 6030 using CFs from the full CMIP5 ensemble and those from sub-selection methods for 2050s under RCP8.5. Black line shows the observed streamflow, red line shows the median of the simulated streamflow and the grey area shows the 90% uncertainty band for Catchments a) 6030 and b) 14,019. Grey region: the 95% uncertainty band, red line: 50% quantile flow, black line: Average observed monthly flow (Correlation between observed flow and simulated with 5, 50 and 95% probabilities are included for further comparison)

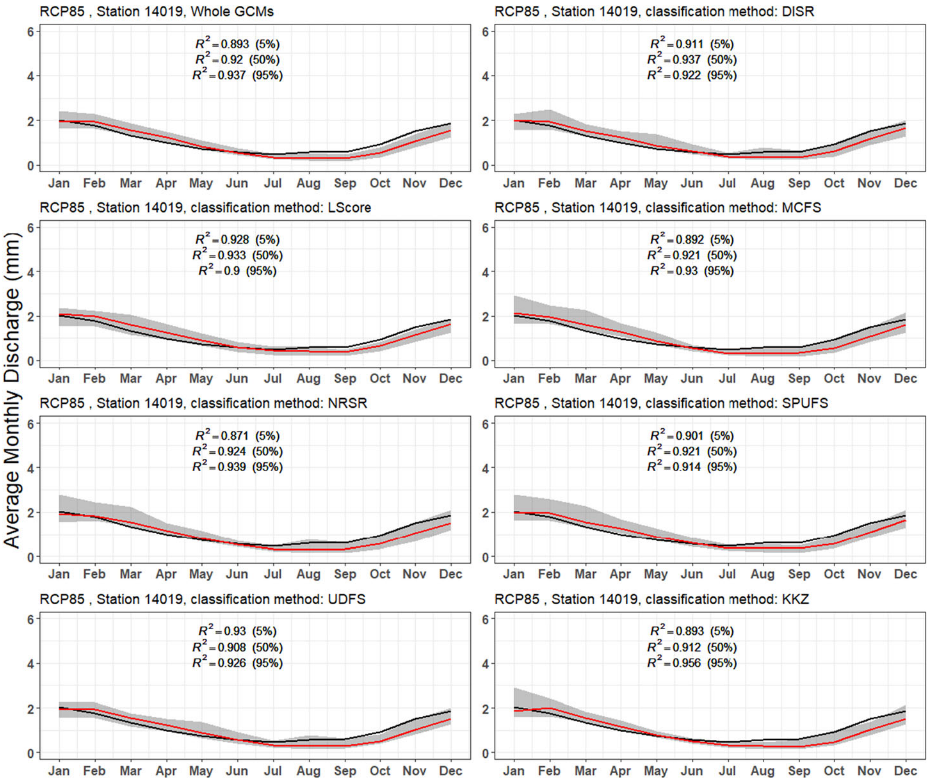

Fig. 2 (continued)

NRSR, SPUFS KKZ and MCFS show considerable difference for high flow months (i.e. January, February and March), relative to the full ensemble.

Using the entire CMIP5 ensemble as the reference, Taylor diagrams in Fig. S5 show the performance of different sub-selection methods in capturing the upper (95%) and lower (5%) uncertainty ranges, together with the median simulated changes in average monthly discharge. Different sub-selection models perform well and very similarly in replicating the CMIP5 ensemble median. However, greater differences are evident in capturing the upper/ lower bounds of the parent ensemble. For the lower bound (5%), all methods except DISR and UPFS for catchment 6030 and DISR and LScore for catchment 14,019 performed well. For the upper bound (95%), more dispersion is evident, while the performance of all sub-selection methods are worse compared to their ability to replicate the CMIP5 median and lower bound. DISR, MCFS and KKZ for catchment 6030 and DISR, LScore and UDFS for catchment 14,019 show the best performance in capturing the upper bound of simulations. In general, it can be deduced that the sub-selection methods evaluated performed best at capturing the characteristics of CMIP5 ensemble for the median, than for lower bound and worst for upper bound.

Figure 3 and Fig. S6 (Supplementary Information) examine the ability of each 10 member sub-selected ensemble to capture key characteristics of the CMIP5 ensemble for summer and autumn 7-day low flows, respectively. In summer, SPUFS and NRSR performed better over

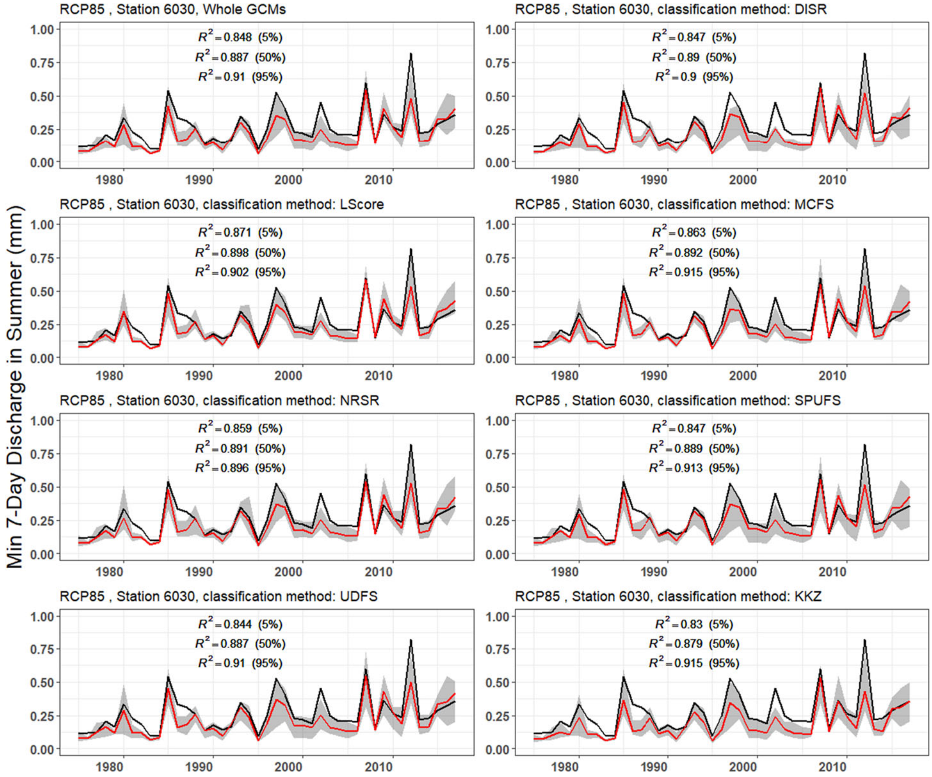

Fig. 3 Time series of 7-day lowflow values in summer simulated for the full CMIP5 ensemble and different sub-selection methods for Catchments a) 6030 and b) 14,019. Grey region: the 95% uncertainty band, red line: 50% quantile flow, black line: Average observed monthly flow (Correlation between observed flow and simulated with 5, 50 and 95% probabilities are included for further comparison)

both catchments, while LScore and MCFS for catchment 6030 and UDFS and DISR for catchment 14,019, had the worst performance. In autumn, KKZ and NRSR for catchment 6010 and KKZ, SPUFS and MCFS for catchment 14,019 show better skill in preserving the 90% uncertainty bounds of the entire CMIP5 ensemble. DISR and LScore had the worst performance over both catchments.

Figure 4 examines the ability of different sub-selection methods to capture the median, upper and lower uncertainty bounds of the CMIP5 ensemble for summer and autumn minimum 7-day low flows in each catchment. For summer in catchment 14,019, UDFS and KKZ followed by LScore and NRSR performed best in capturing the upper bound, while DISR, MCFS, LScore and SPUFS followed by NRSR performed best in capturing the median and lower bound from the full ensemble. For autumn in catchment 6030, DISR and NRSR followed by SPUFS do best at replicating the upper bound. For the median, LScore and NRSR followed by UDFS perform best, while KKZ and NRSR followed by UDFS have slightly better performance in capturing the lower bound of the full ensemble.

Using the Standardized Streamflow Index (SSI), we extracted statistics on the number of drought events, drought duration and intensities for all sub-selected ensembles and compared results with the case considering all CMIP5 GCMs. The relative errors (%) in replicating the CMIP5 ensemble median, upper and lower bounds for average drought intensity, average

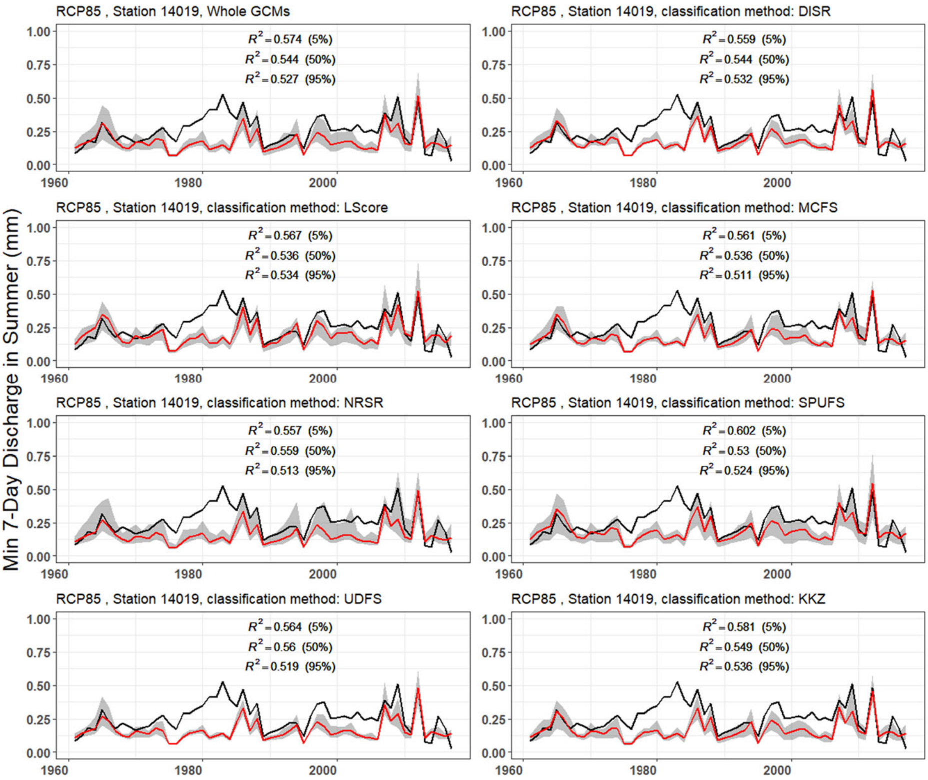

Fig. 3 (continued)

drought duration and number of drought events are presented in Table 1. For both catchments, all sub-selected ensembles perform very satisfactory with percent relative errors below 5% in most cases. For catchment 6030, MCFS and LScore followed by KKZ perform better for the CMIP5 upper bound. For the median, UDFS and DISR followed by NRSR and MCFS and for the lower bound KKZ followed by LScore and UDFS performs slightly better. For catchment 14,019, DISR and UDFS performed better for the upper bound, while NRSR and MCFS/UDFS and KKZ performed better for the median/lower bound, respectively.

Based on the above, it is very difficult to identify a best sub-selection method for widespread application. Our results show that sub-selection methods perform differently in capturing aspects of the parent ensemble, i.e. median, lower or upper bounds. They also vary in their effectiveness by catchment, flow metric and season. For catchment 14,019 while DISR is one of the weakest methods for 7-day low flows, it performs the best in drought analysis using SSI-based characteristics Other research also confirmed that there were not any single best sub-selection methods for their application data (e.g. Afzal and Torkar, 2016). Among sub-selection methods evaluated, the most consistent performance was found for KKZ, SPUFS and NRSR in terms of preserving 7-day low flow uncertainty bounds and SSI-based characteristics, however they were not always the best, i.e. their performance varied unpredictably based on metrics/criteria assessed. Others have reported similar results, e.g. Ross and Najjar (2019) and Cannon (2015).

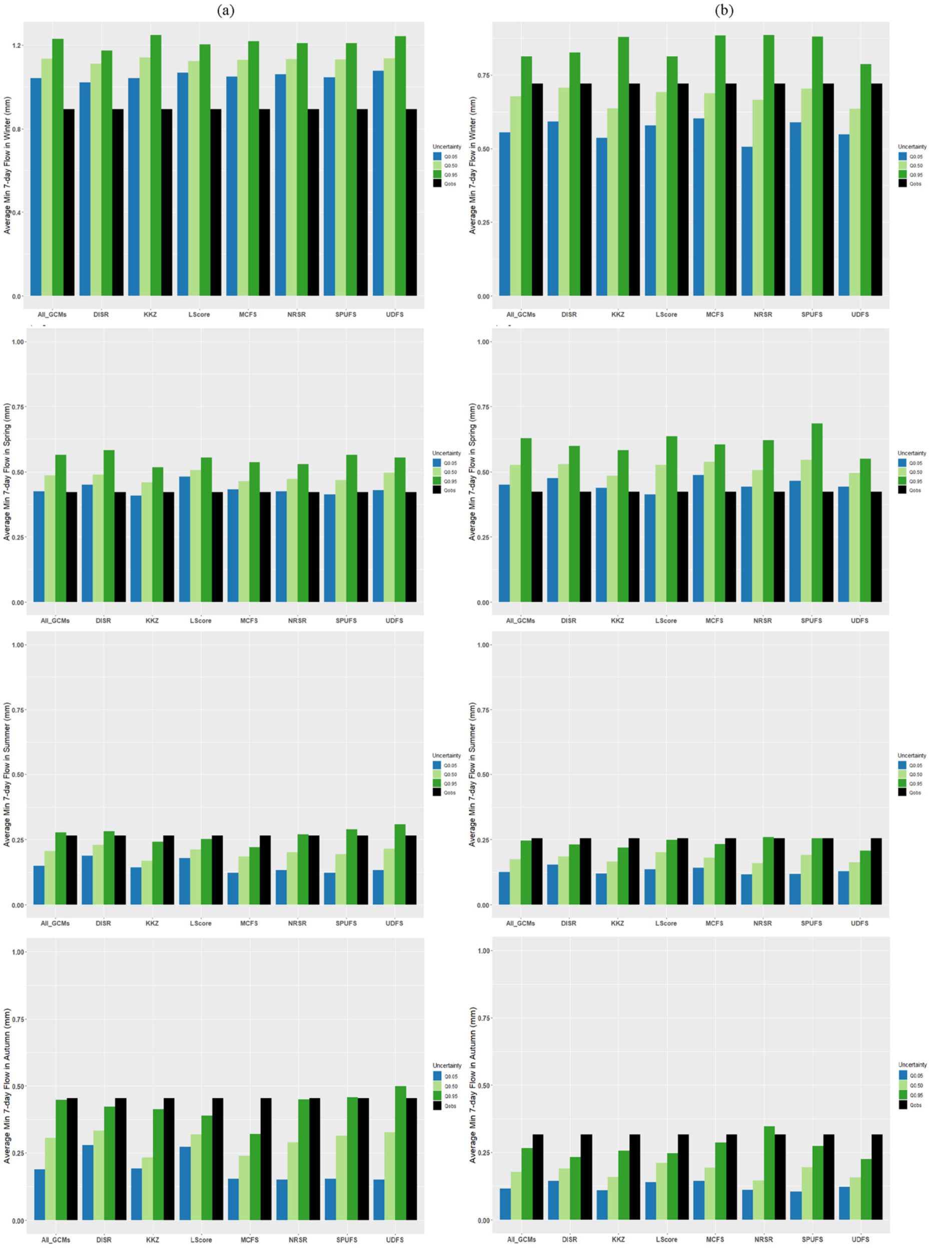

Fig. 4 Barplot of the average seasonal 7-day lowflow simulated based on different sub-selection methods for Catchments a) 6030 and b) 14,019 (blue = 5% lower bound, light green = median and dark green = 95% upper bound)

Our results further show that whether ensemble members are selected from entirely different GCM families (e.g. NRSR with 8 different GCM families for both catchments) or are allowed to contain greater membership from the same family of GCMs (e.g. KKZ with 5 and 6 and SPUFS with 4 and 7 GCM families for catchment 6030 and 14,019, respectively),

Table 1 Percent relative error of each sub-selection method for drought indices extracted from standardized streamflow index (SDI)

Catchment	Uncertainty	Drought Index	DISR	LScore	MCFS	NRSR	SPUFS	UDFS	KKZ
6030	P0.05*	Percent Error in No Drought Events	−3.7	−1.85	1.85	5.56	−3.7	5.56	3.7
		Percent Error in Average Intensity	4.2	1.96	−1.84	−5.14	3.86	−4.92	−3.6
		Percent Error in Average Duration	3.85	1.89	−1.82	−5.26	3.85	−5.26	−3.57
	P0.50*	Percent Error in No Drought Events	0	−1.82	−1.82	−1.82	−3.64	0	1.82
		Percent Error in Average Intensity	0.21	1.79	1.67	1.54	3.65	−0.05	−1.78
		Percent Error in Average Duration	0	1.85	1.43	1.85	3.77	0	−2.19
	P0.95*	Percent Error in No Drought Events	3.57	1.79	−3.57	−5.36	−3.57	−1.79	0
		Percent Error in Average Intensity	−3.59	−1.93	3.46	5.65	3.4	1.98	0.07
		Percent Error in Average Duration	−3.45	−1.75	3.7	5.66	3.7	1.82	0
14,019	P0.05*	Percent Error in No Drought Events	0.00	4.65	4.65	−9.30	2.33	0.00	4.65
		Percent Error in Average Intensity	0.05	−4.38	−4.50	10.19	−2.22	−0.06	−4.35
		Percent Error in Average Duration	0.00	−4.44	−4.44	10.26	−2.27	0.00	−4.44
	P0.50*	Percent Error in No Drought Events	−2.08	−6.25	0.00	0.00	2.08	−4.17	−4.17
		Percent Error in Average Intensity	2.13	6.74	−0.02	0.00	−2.06	4.42	4.25
		Percent Error in Average Duration	2.13	6.67	−0.40	0.00	−2.44	4.35	4.35
	P0.95*	Percent Error in No Drought Events	−2.04	−4.08	−8.16	0.00	4.08	0.00	0.00
		Percent Error in Average Intensity	2.30	4.48	8.78	−0.33	−3.89	−0.09	0.03
		Percent Error in Average Duration	2.08	4.26	8.45	−0.40	−3.92	0.00	0.00

*- P0.05, P0.50 and P0.95 are riverflow with 5%, 50% and 95% probability of occurrence

Table 2 Percent relative error in drought characteristics compared to full ensemble case based on different number of clusters

Catchment	K	Drought Index	DISR	LScore	MCFS	NRSR	SPUFS	UDFS	KKZ
6030	K = 5	Percent Error in No Drought Events	3.64	1.82	−1.82	−3.64	0.00	0.00	1.82
		Percent Error in Average Intensity	−3.10	−1.48	1.52	3.95	0.04	0.17	−2.04
		Percent Error in Average Duration	−3.51	−1.79	1.85	3.77	−0.42	0.00	−2.19
	K = 10	Percent Error in No Drought Events	0.00	−1.82	−1.82	−1.82	−3.64	0.00	1.82
		Percent Error in Average Intensity	0.21	1.79	1.67	1.54	3.65	−0.05	−1.78
		Percent Error in Average Duration	0.00	1.85	1.43	1.85	3.77	0.00	−2.19
	K = 15	Percent Error in No Drought Events	1.82	−1.82	−3.64	−3.64	−3.64	1.82	1.82
		Percent Error in Average Intensity	−1.48	1.88	3.54	3.68	3.70	−1.79	−2.01
		Percent Error in Average Duration	−1.79	1.85	3.77	3.77	3.77	−1.79	−2.19
14,019	K = 5	Percent Error in No Drought Events	−2.10	−8.33	0.00	−4.17	−2.08	−6.25	−2.08
		Percent Error in Average Intensity	2.27	9.17	−0.26	4.54	2.20	6.92	2.26
		Percent Error in Average Duration	2.13	9.09	−0.40	4.35	2.13	6.67	2.13
	K = 10	Percent Error in No Drought Events	−2.08	−6.25	0.00	0.00	2.08	−4.17	−4.17
		Percent Error in Average Intensity	2.13	6.74	−0.02	0.00	−2.06	4.42	4.25
		Percent Error in Average Duration	2.13	6.67	−0.40	0.00	−2.44	4.35	4.35
	K = 15	Percent Error in No Drought Events	−4.17	−6.25	2.08	−2.08	0.00	−8.33	0.00
		Percent Error in Average Intensity	4.42	6.74	−2.14	2.13	−0.09	9.17	−0.09
		Percent Error in Average Duration	4.35	6.67	−2.44	2.13	0.00	9.09	0.00

the results can be very similar in terms of preserving the uncertainty bounds of the entire CMIP5 archive. Furthermore, while the same CMIP5 ensemble members are often represented in the majority of sub-selection methods for a specific catchment, the overlap of members for both catchment and sub-selection methods is very limited. Therefore, it is not practical to extend a subset of CMIP5 ensembles that are selected based on a particular sub-selection method over a single catchment/region to another catchment/region within the same geographical domain. Other studies also shown that the results of climate impact studies depend greatly on sub-selection method, available GCM models and the performance criteria (Evans et al., 2013; Kiesel et al., 2020).

3.4 Sensitivity of Results to Number of Ensemble Members

To examine the sensitivity of results to number of clusters/members identified to construct our sub-selected ensembles, we extended the study to consider 5 and 15 members. Results, relative to the full CMIP5 ensemble are compared with those for the 10-member ensemble for seasonal 7-day low flow values and drought characteristics derived from the SSI series (Fig. S7). Differences in the sensitivity of sub-selection methods to ensemble members are evident between catchments. The groundwater-dominated catchment (14019) tends to be less sensitive to sub-ensemble size in comparison to the runoff dominated catchment (6030). However, this finding is also sensitive to the specific sub-selection method. DISR and KKZ tend to be most robust to ensemble size in simulating summer and autumn 7-day low flows, with errors being similar for 10 and 15 members. There is no consistent increase or decrease in the performance of sub-selection methods as the sub-ensemble member size increases or decreases. Sub-selection methods performed differently based on different seasons and number of clusters over each catchment. For example, in winter DISR, LScore, MCFS and NRSR performed better in catchment 14,019 while KKZ, LScore and MCFS have the best performance in catchment 6030. Table 2 summarizes the percent relative error of varying sub-ensemble size in replicating drought characteristics from the CMIP5 ensemble. In both catchments the results are sensitive to number of clusters/ensemble members, however it is notable that for SSI drought metrics the smallest errors are typically found for an ensemble size of 10 for almost all sub-selection methods.

4 Conclusion

In this study, we presented a methodology for evaluating the effectiveness of sub-selection methods at replicating the characteristics (median, upper and lower bound) of the CMIP5 ensemble in simulating monthly mean discharge, low-flow and drought characteristics for two Irish catchments. Our results show that sub-selection methods perform differently in capturing aspects of the parent ensemble, i.e. median, lower or upper bounds. They also vary in their effectiveness by catchment, flow metric and season. Therefore, researchers need to carefully evaluate sub-selection performance based on the aims of their study and the flow metrics of interest, on a catchment by catchment basis.

Moreover, results from sub-selection methods are sensitive to the number of clusters/ensemble size. This was particularly the case for the runoff dominated catchment, with errors in replicating the parent ensemble showing little consistency with ensemble size. Interestingly, the simulation of drought characteristics was less sensitive to sub-ensemble size in comparison to absolute low flows, as characterized by the minimum 7-day low flows. Given these findings we recommend that before choosing sub-selection methods that thought is given to which aspects of the uncertainty are most important for decision making. For drought and low flow analyses, it is perhaps most important to preserve estimation of the median and lower bound for stress-testing adaptation responses in water systems. Finally, while sub-selection methods hold promise for reducing the complexity of larger ensembles they are subject to considerable challenges in their application with no clear guidance possible on which methods to use where, or for which metrics. This is even more complex when dealing with low-flow processes which have more inherent complexity and uncertainty due to surface-groundwater interactions, as evident in this study. Application of these approaches should therefore be undertaken with care, with the aims of the modelling activity clearly in mind.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11269-020-02714-1.

Acknowledgments The authors thank two reviewers for helpful comments in improving the manuscript.
Availability of Data and Material Available from corresponding author on request.

Author's Contributions Both authors conceived and designed the study. Material preparation, data collection and analysis were performed by Saeed Golian. Both authors wrote the manuscript and approved the final draft.

Funding SG and CM acknowledge funding from Science Foundation Ireland Career Development Award, grant number SFI/17/CDA/4783.

Compliance with Ethical Standards

Conflict of Interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Consent for Publication Herewith, we declare our consent for our manuscript to be published in "Water Resources Management" journal.

References

Afzal W, Torkar R (2016) Towards benchmarking feature subset selection methods for software fault prediction. In Computational intelligence and quantitative software engineering (pp. 33–58). Springer, Cham

Bergmeir C, Molina D, Benítez JM Memetic Algorithms with Local Search Chains in R: The Rmalschains Package (2016) Journal of Statistical Software, 75(4), 1–33., doi:https://doi.org/10.18637/jss.v075.i04

Broderick C, Matthews T, Wilby RL, Bastola S, Murphy C (2016) Transferability of hydrological models and ensemble averaging methods between contrasting climatic periods. Water Resour Res 52(10):8343–8373

Broderick C, Murphy C, Wilby RL, Matthews T, Prudhomme C, Adamson M (2019) Using a scenario-neutral framework to avoid potential maladaptation to future flood risk. Water Resour Res 55(2):1079–1104

Cai D, Zhang C, He X (2010) Unsupervised feature selection for multi-cluster data. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 333–342)

Cannon AJ (2015) Selecting GCM scenarios that span the range of changes in a multimodel ensemble: application to CMIP5 climate extremes indices. J Clim 28:1260–1267. https://doi.org/10.1175/JCLI-D-14-00636.1

Christierson B v, Vidal J-P, Wade SD (2012) Using UKCP09 probabilistic climate information for UK water resource planning. J Hydrol 424:48–67

Clark MP, Wilby RL, Gutmann ED, Vano JA, Gangopadhyay S, Wood AW, Fowler HJ, Prudhomme C, Arnold JR, Brekke LD (2016) Characterizing uncertainty of the hydrologic impacts of climate change. Curr Clim Change Rep 2:55–64. https://doi.org/10.1007/s40641-016-0034-x

Coron L, Perrin C, Delaigue O, Thirel G, Michel C (2017) airGR: suite of GR hydrological models for precipitation-runoff modelling. R package version 1(9.64)

Evans JP, Ji F, Abramowitz G, Ekström M (2013) Optimally choosing small ensemble members to produce robust climate simulations. Environ Res Lett 8(4):044050

Farahmand A, AghaKouchak A (2015) A generalized framework for deriving nonparametric standardized drought indicators. Adv Water Resour 76:140–145

Fung F, Watts G, Lopez A, Orr HG, New M, Extence C (2013) Using large climate ensembles to plan for the hydrological impact of climate change in the freshwater environment. Water Resour Manag 27(4):1063–1084

Giorgi F, Jones C, Asrar GR et al (2009) Addressing climate information needs at the regional level: the CORDEX framework. World Meteorol Organ WMO Bull 58:175

Gupta HV, Kling H, Yilmaz KK, Martinez GF (2009) Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J Hydrol 377(1–2):80–91

Gustard, A., Bullock, A., and Dixon, J. M, 1992. Low flow estimation in the United Kingdom. Wallingford: Institute of Hydrology

Hao Z, AghaKouchak A, Nakhjiri N, Farahmand A (2014) Global integrated drought monitoring and prediction system. Scientific data 1(1):1–10

He X, Cai D, Niyogi P (2006) Laplacian score for feature selection. In Advances in neural information processing systems (pp. 507–514)

Katsavounidis I, Kuo CCJ, Zhang Z (1994) A new initialization technique for generalized Lloyd iteration. IEEE Signal Process Lett 1:144–146. https://doi.org/10.1109/97.329844

Kiesel J, Stanzel P, Kling H, Fohrer N, Jähnig SC, Pechlivanidis I (2020) Streamflow-based evaluation of climate model sub-selection methods. Climatic Change, pp.1–19

Koga H, Ishibashi T, Watanabe T (2007) Fast agglomerative hierarchical clustering algorithm using locality-sensitive hashing. Knowl Inf Syst 12(1):25–53

Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010) Challenges in combining projections from multiple climate models. J Clim 23:2739–2758. https://doi.org/10.1175/2009JCLI3361.1

Koutsoyiannis D (2013) Hydrology and change. Hydrol Sci J 58(6):1177–1197

Liu Y, Liu K, Zhang C, Wang J, Wang X (2017) Unsupervised feature selection via diversity-induced self-representation. Neurocomputing 219:350–363

Lu Q, Li X, Dong Y (2018) Structure preserving unsupervised feature selection. Neurocomputing 301:36–45. ISSN 09252312. https://doi.org/10.1016/j.neucom.2018.04.001

Masson D, Knutti R (2011) Climate model genealogy: CLIMATE MODEL GENEALOGY. Geophys. Res. Lett. 38, n/a-n/a. doi:https://doi.org/10.1029/2011GL046864

Mendlik T, Gobiet A (2016) Selecting climate simulations for impact studies based on multivariate patterns of climate change. Clim Chang 135(3–4):381–393

Mills P, Nicholson O, Reed D (2014) Physical catchment descriptors. Volume IV, flood studies update technical research report. Office of Public Works, Dublin https://opw.hydronet.com/data/files/Technical%20Research%20Report%20-%20Volume%20IV%20-%20Physical%20Catchment%20Descriptors.pdf

Nicolle P, Pushpalatha R, Perrin C, François D, Thiéry D, Mathevet T, Le Lay M, Besson F, Soubeyroux JM, Viel C, Regimbeau F (2014) Benchmarking hydrological models for low-flow simulation and forecasting on French catchments. Hydrol Earth Syst Sci 18:2829–2857

Oudin L, Hervieu F, Michel C, Perrin C, Andréassian V, Anctil F, Loumagne C (2005) Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. J Hydrol 303(1–4):290–306

Pirtle Z, Meyer R, Hamilton A (2010) What does it mean when climate models agree? A case for assessing independence among general circulation models. Environ Sci Policy 13(5):351–361

Pool S, Vis M, Seibert J (2018) Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency. Hydrol Sci J 63(13–14):1941–1953

Stainforth DA, Aina T, Christensen C, Collins M, Faull N, Frame DJ, Kettleborough JA, Knight S, Martin A, Murphy JM et al (2005) Uncertainty in predictions of the climate response to rising levels of greenhouse gases. Nature 433:403–406

Qian M, Zhai C (2013) Robust unsupervised feature selection. In Twenty-Third International Joint Conference on Artificial Intelligence

Ross AC, Najjar RG (2019) Evaluation of methods for selecting climate models to simulate future hydrological change. Clim Chang 157(3–4):407–428

Seo SB, Kim YO, Kim Y, Eum HI (2019) Selecting climate change scenarios for regional hydrologic impact studies based on climate extremes indices. Clim Dyn 52(3–4):1595–1611

Smith KA, Wilby RL, Broderick C, Prudhomme C, Matthews T, Harrigan S, Murphy C (2018) Navigating cascades of uncertainty—as easy as ABC? Not quite…. Journal of Extreme Events 5(01):1850007

Taylor KE, Stouffer RJ, Meehl GA (2011) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93(4):485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Taylor KE, Stouffer RJ, Meehl GA (2012) An overview of CMIP5 and the experiment design. Bull Am Meteorol Soc 93:485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63(2):411–423

Walsh S (2012) A Summary of climate averages 1981–2010 for Ireland. Climatological Note, 14

Wang HM, Chen J, Cannon AJ, Xu CY, Chen H (2018) Transferability of climate simulation uncertainty to hydrological impacts. Hydrol Earth Syst Sci 22(7):3739–3759

Wang W, Zhang H, Zhu P, Zhang D, Zuo W (2015) Non-convex regularized self-representation for unsupervised feature selection. In International Conference on Intelligent Science and Big Data Engineering (pp. 55–65). Springer, Cham

Wilby RL, Dessai S (2010) Robust adaptation to climate change. Weather 65:180–185

Yang Y, Shen HT, Ma Z, Huang Z, Zhou X (2011) L2, 1-norm regularized discriminative feature selection for unsupervised. In Twenty-Second International Joint Conference on Artificial Intelligence

Zhu P, Zhu W, Wang W, Zuo W, Hu Q (2017) Non-convex regularized self-representation for unsupervised feature selection. Image Vis Comput 60:22–29. ISSN 02628856. https://doi.org/10.1016/j.imavis.2016.11.014