Climatic Change (2019) 157:407–428
https://doi.org/10.1007/s10584-019-02512-8

Evaluation of methods for selecting climate models to simulate future hydrological change

Andrew C. Ross[1,2] · **Raymond G. Najjar**[1]

Received: 7 December 2018 / Accepted: 24 July 2019 / Published online: 6 November 2019
© Springer Nature B.V. 2019

Abstract

A challenge for climate impact studies is the selection of a limited number of climate model projections among the dozens that are typically available. Here, we examine the impacts of methods for climate model selection on projections of runoff change for five different watersheds across the conterminous USA. The results from an ensemble of 29 global climate models and 29 corresponding hydrological model simulations are compared with the results that would have been obtained by applying six different selection methods to the climate model data and using only the selected models to drive the hydrological model. We evaluate each selection method based on whether the runoff projections produced by the method meet the method's objective and on whether the results are sensitive to the number of models chosen. The Katsavounidis–Kuo–Zhang (KKZ) method, which is intended to maximize the spread in the projected climate change, was the only method that met both criteria for suitability. Although the KKZ method generally performed well, the results from both it and the other methods varied somewhat unpredictably based on region and number of models chosen. This study shows that the methods and models used in similar top–down studies should be carefully chosen and that the results obtained should be interpreted with caution.

Keywords Climate model ensembles · Hydrological change · Model selection · Model uncertainty

1 Introduction

Climate change is expected to cause precipitation and evapotranspiration to change in most locations (Held and Soden 2006), leading to changes in runoff and streamflow (Nijssen et al.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10584-019-02512-8) contains supplementary material, which is available to authorized users.

✉ Andrew C. Ross
andrewross@princeton.edu

[1] Department of Meteorology and Atmospheric Science, The Pennsylvania State University, University Park, Pennsylvania, PA, USA

[2] Present address: Princeton University Program in Atmospheric and Oceanic Sciences, Princeton, NJ, USA

2001; Arnell and Gosling 2013). Changes in streamflow could have a number of environmental, social, and economic effects, including affecting the reliability of water resources (Vicuna et al. 2007; Schewe et al. 2014), altering estuarine salinity (Gibson and Najjar 2000), exacerbating eutrophication of coastal waters (Justić et al. 2005; Rabalais et al. 2009), and increasing the frequency of flooding (Milly et al. 2002; Hirabayashi et al. 2013). Therefore, it is critical to understand the impacts of the changing hydrological cycle and to begin to manage the associated risks.

To determine the impacts of climate change on hydrology and the potential implications for risk and adaptation, most studies have relied on hydrological models forced by projections of meteorological variables from global climate models (GCMs). Although this "top–down" approach has been criticized (e.g., Wilby and Dessai 2010), it has nevertheless remained the standard approach to studying the impacts of climate change on hydrology and the potential implications for water resource planning (Holman et al. 2012).

Conducting a top–down hydrological impact study often requires a great deal of computer and researcher time because several steps are typically necessary: (1) downloading global climate model data, (2) applying bias correction, (3) downscaling to finer spatial and temporal scales, (4) calibrating and evaluating one or more hydrological models, (5) running model simulations, (6) analyzing results, and (7) assessing alternatives and making decisions. Thus, although the Climate Model Intercomparison Project (CMIP) and other projects have created hundreds of climate model simulations that could be used to force hydrological models, it is not unusual for studies to use only a few models as forcing (e.g., Hagemann et al. (2013); Ott et al. (2013); Schewe et al. (2014)).

Many methods have been proposed to select a small number of climate models to use as hydrological model forcing. Al Aamery et al. (2016) used a large ensemble of climate model results to drive a hydrological model, and they showed that a smaller subset of the ensemble that contained balanced choices of climate model dataset (CMIP3 or CMIP5), climate model, emissions scenario, downscaling method, and bias correction technique yielded a similar streamflow projection mean and variance as the full ensemble. However, application of this technique still requires many model runs and access to or development of a range of modeling, downscaling, and bias correction methods. Al Aamery et al. (2016) and others (Teng et al. 2012; Vetter et al. 2017) have compared runoff and streamflow projections derived from various combinations of climate model forcings, emissions scenarios, downscaling methods, and hydrological models and have found that climate model structural uncertainty is typically the dominant source of uncertainty in future streamflow projections. As a result, it is reasonable to begin with a focus on uncertainty stemming only from different GCMs.

Some studies have proposed that focusing on models with good performance at simulating some aspects of historical climate may allow adequate coverage of GCM uncertainty and produce accurate projections without incurring significant computational costs. Weiland et al. (2012) showed that selecting models based on several criteria including the ability to reproduce observed mean runoff produced results that were reasonable and were comparable to results obtained by averaging models with weighting based on the same criteria (note that model selection could be considered as equivalent to model weighting with zero weights on rejected models and equal weights on selected models). On the other hand, Chen et al. (2017) found that weighting models equally, randomly, or based on performance produced little difference in projected hydrological change in a Canadian watershed. Selecting or weighting models based on historical performance has been criticized for several reasons, including the difficulty of identifying optimal metrics for measuring performance (Brekke et al. 2008; Knutti et al. 2010; Weigel et al. 2010) and the potential for models to obtain

reasonable historical performance for the wrong reasons (Tebaldi and Knutti 2007). However, Maraun et al. (2017) argue for the use of "emergent constraints," physically sound statistical relationships between model current climate and model-projected change, as a way of winnowing out implausible models.

Some studies have used clustering algorithms to objectively group and select climate models based on their similarity of historical climate or projected change. Houle et al. (2012) used the k-means algorithm to select five climate scenarios from 86 combinations of CMIP3 models and emissions scenarios. Casajus et al. (2016) used k-means clustering to select GCMs to drive a species distribution model and found that the model produced similar results whether driven by a six-GCM subset or by the full 27-GCM ensemble. Mendlik and Gobiet (2016) and Wilcke and Bärring (2016) described how to select models using agglomerative hierarchical clustering, including a method to determine the optimal number of clusters.

Rather than using clustering, Cannon (2015a) applied the Katsavounidis–Kuo–Zhang (KKZ) algorithm (Katsavounidis et al. 1994) to select models that have the largest possible spread from each other. Chen et al. (2016) and Wang et al. (2018) used climate models selected using both the k-means and KKZ algorithms to drive a hydrological model and to test whether properties of the selected climate model subensembles transferred to the hydrological model results. Although both algorithms produced model simulations that typically covered greater amounts of hydrological uncertainty than randomly selected subsets of equal size, Wang et al. (2018) found that the KKZ algorithm performed better than the k-means algorithm at covering hydrological uncertainty. Unlike k-means and other clustering algorithms, the KKZ algorithm also has the advantage of producing results that are ordered (Cannon 2015a); in other words, the KKZ algorithm produces a chain of subsets where each subset of size n is included in the subset of size $n + 1$.

Statistical dependence between models is a potential problem when using models from a multimodel ensemble like CMIP3 and CMIP5. Dependence between model results is likely to arise as some modeling groups contribute multiple models with minor differences to the CMIP archives, and even models from different groups often have shared model components or parameterizations (Masson and Knutti 2011; Knutti et al. 2013; Leduc et al. 2016). Dependence can cause an underestimation of the range of possible climate changes and a neglect of possible risks (Steinschneider et al. 2015) and can degrade the accuracy of ensemble predictions (Bishop and Abramowitz 2013; Abramowitz and Bishop 2015). Masson and Knutti (2011) and Knutti et al. (2013) used hierarchical clustering based on climatology, rather than on projected changes, to identify models with similar historical climates. Because Masson and Knutti (2011) and Knutti et al. (2013) found that many models with similar historical climates were developed by the same institution or at least shared similar model components and parameterizations, using clustering on climatology to select models could be a way to reduce the dependence of the selected subensemble.

The availability of multiple methods for model selection raises the question of which methods, if any, are suitable for use in impact studies. A few previous studies have compared the results from two similar methods (k-means vs. KKZ by Cannon (2015a), Chen et al. (2016), and Wang et al. (2018), and weighting and clustering based on historical performance by Zubler et al. (2016)). However, as far as we know, there has not been an objective comparison of multiple climate model selection methods that (1) compare both similar (e.g., k-means and KKZ) and different (e.g., performance) selection strategies; (2) use climate data for selection and evaluate the results after driving other models, such as hydrological models, with the selected climate input to account for the possibility that properties of the selected climate subensemble may not transfer to the results of the other models; and (3)

evaluate the results in multiple regions with different hydroclimates. An additional issue not considered in any of the previously cited evaluations of selection methods is whether the results are robust against the number of models selected. Although some sensitivity to the number of models is expected, particularly for very small subensembles, it would be undesirable if a selection method could cause a researcher to arrive at a different conclusion by choosing, say, ten models instead of five.

In this study, we evaluate six climate model selection methods: the KKZ method, two variants of two clustering methods, and a performance-based method. Although additional methods have been proposed in the literature (e.g., Sanderson et al. (2015), Herger et al. 2017)), we focus on these six methods in part because they could be easily used in studies by researchers who have expertise in fields other than climate science and climate model applications. We have excluded other methods that would require manual analysis of the climate model data (for example, to remove outlying models that poorly simulate some aspect of local climate) or that involve complicated mathematical and statistical methods. To account for potential lack of transferability of subensemble properties when using climate models to drive other models, we select models based on climate (precipitation and potential evapotranspiration) but evaluate the results using hydrological model simulations of runoff following a strategy similar to Chen et al. (2016). To determine whether the results are robust and how they may vary in different regions, we evaluate whether each selection method meets these criteria in five different large watersheds of the conterminous United States with different hydroclimates.

In each region, we test the hypothesis that each method is suitable. We propose that the suitability of a method should be defined based on two criteria. First, we define a suitable method as one in which the objective of the method is met with reasonably sized subensembles. For example, the objective of the KKZ method is to cover a wide range of model uncertainty, so for this method to be suitable, the results from small subensembles created using the KKZ method should cover most of the spread in the full-model ensemble. Second, to find methods that are robust against the number of models selected, we define a suitable method as one in which the properties of the model subensemble are not highly variable within a reasonable range of subensemble sizes. For both objectives, we primarily test the results using subensembles with between five and ten members, which we consider a reasonable range that is computationally feasible for most impact studies.

2 Methods

2.1 Study regions

We evaluated the selection methods in five different watersheds (Fig. 1). The watersheds are essentially the same as the pilot sites used by Johnson et al. (2012), which were selected based on several criteria including representing a diverse range of climate, geography, and hydrology. In terms of land cover and vegetation, the watersheds predominantly contain a mix of evergreen forest and cropland (Willamette River); cropland and woodland (Minnesota River); deciduous forest, woodland, and mixed (Susquehanna River); wooded grassland (Theodore Roosevelt Lake, hereafter Roosevelt); and a combination of most categories (Apalachicola River). On average, annual total precipitation is greater than total potential evapotranspiration in three of the five watersheds (Willamette, Susquehanna, and Apalachicola). The Roosevelt is the most arid of the five watersheds, with a precipitation to potential evapotranspiration ratio of 0.69.

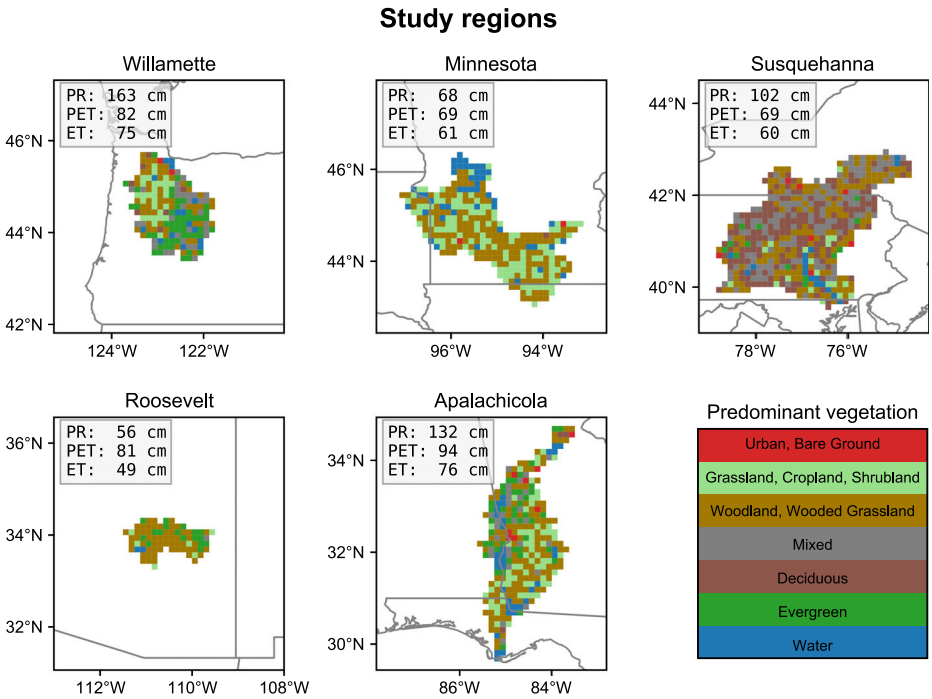

Fig. 1 The five study regions and the corresponding 1/8 degree model grid cells. Grid cell color indicates predominant vegetation type (obtained from the NLDAS dataset). The text in the upper left corner of each subfigure gives the mean annual total precipitation (PR), potential evapotranspiration (PET), and actual evapotranspiration (ET) in centimeters (obtained from the Maurer et al. (2002) dataset). The subfigures are arranged geographically, so that the upper right subfigure is in the Northeast US, etc.

2.2 Climate and hydrology models

A broad overview of the data used and model selection methods is provided in Fig. 2. Climate and hydrological data were extracted from the Downscaled CMIP5 Climate and Hydrology Projections dataset produced by the U.S. Bureau of Reclamation (Brekke et al. 2013, 2014). This dataset contains output from CMIP5 GCMs that has been bias-corrected and statistically downscaled to daily data at 1/8 degree resolution using the bias correction and spatial disaggregation methods developed by Wood et al. (2004) and described in Brekke et al. (2013). A list of the models contained in the dataset is provided in Table S1 in the supporting information. The dataset also provides hydrological output produced by driving the Variable Infiltration Capacity (VIC) model (Liang et al. 1994, 1996; Nijssen et al. 1997) with the downscaled GCM data. Brekke et al. (2014) provides additional details on the model configuration and calibration.

From the climate and hydrology model dataset, we obtained data for precipitation, potential evapotranspiration, and total runoff from 29 different climate models. Because runoff was not routed through a river network over the eastern half of the USA, we instead focused on long-term watershed averages of runoff (as a proxy for river discharge) and precipitation and PET. We used model data from 1950 to 1999 forced with historical greenhouse gas concentrations ("historical") and from 2050 to 2099 forced using the RCP 8.5 scenario of

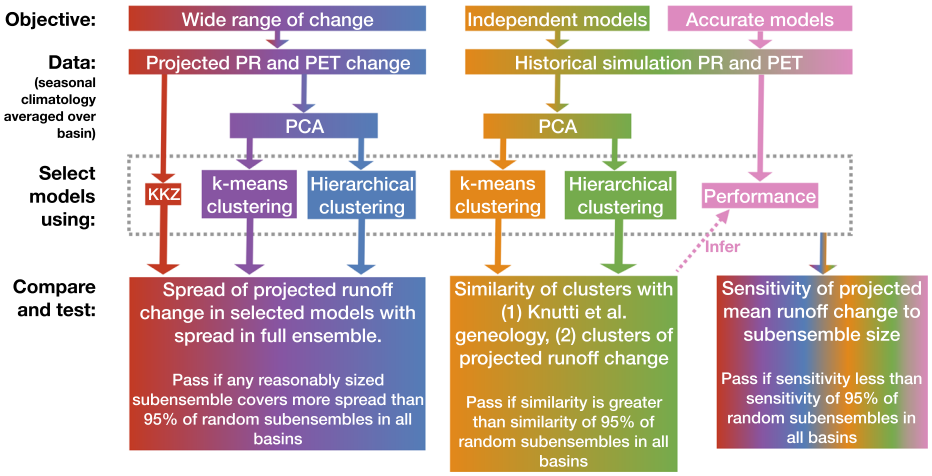

Fig. 2 Basic flowchart showing the general objective of each model selection method, the model data used, and the criteria by which the methods were evaluated. Note that data processing and model selection were applied separately for each watershed, and the final evaluation of whether a method was successful was based on results from all watersheds

high greenhouse gas emissions (Riahi et al. 2011) ("future"). Values calculated as changes were based on the change in mean seasonal totals between the historical and the future periods.

Bias correction and downscaling routines, such as those used to prepare the GCM output for use with the hydrological model, may alter model-projected changes (e.g., Maraun 2013, 2014) and potentially reduce the effectiveness of model selection methods that use projected changes. Therefore, to isolate the impact of the selection methods, we applied most of the selection methods to the bias-corrected and downscaled GCM data. Note that applying bias correction to a CMIP-like ensemble could be computationally demanding, but there are several alternatives including using a database where bias correction has already been applied to a large ensemble (as we do in this study) or applying model selection to raw climate model data and using a bias correction routine that preserves projected changes (such as Li et al. (2010), Pierce et al. (2015), and Cannon et al. (2015b)). Some methods do require historical data without bias correction in order to identify differences in simulated historical climate, and these data were acquired directly from the CMIP5 database. The CMIP5 database does not contain the necessary potential evapotranspiration data. Rather than attempting to re-run the hydrological model with non-bias-corrected boundary conditions to estimate PET, we built ordinary linear regression models to predict seasonal total PET from seasonal mean temperature. The models were fit to the bias-corrected mean temperature and PET data from the hydrology dataset and applied to non-bias-corrected mean temperature from the CMIP5 database (bilinearly interpolated to the same 1/8 degree resolution as the bias-corrected data). A separate model was fit for each season and GCM. Overall, the estimation worked well, with a 10-fold cross-validation root mean square error of 9.1 mm and a mean absolute percentage error of 5.0%.

2.3 Climate model selection methods

We followed a strategy similar to Chen et al. (2016) and performed most of this study as a thought experiment where we have only the climate model data (precipitation and potential evapotranspiration) from the 29 climate models and are seeking to select only a few models to downscale runoff projections with a hydrological model. Model selection was performed separately for each watershed in the study using six methods proposed by recent studies (Cannon 2015a; Chen et al. 2016; Mendlik and Gobiet 2016; Wilcke and Bärring 2016) which are presented in detail next. To further support the use of these methods by other scientists, we have provided complete code for the six methods and a reproducible example using the open-source programming language R at https://doi.org/10.6084/m9.figshare.6816851.

2.3.1 Katsavounidis–Kuo–Zhang algorithm

The Katsavounidis–Kuo–Zhang (KKZ) algorithm (Katsavounidis et al. 1994) selects models that span a wide range of projected changes. Given a set of variables from a multimodel ensemble, the algorithm selects the model closest to the centroid (in Euclidean distance) as the first model and the model farthest away from the first model as the second. Each subsequent selection is based on the model with the maximum distance from the nearest of any previously selected model. In this study, the algorithm was applied to the difference between the historical and future seasonal climatologies of potential evapotranspiration (PET) and precipitation (PR) totals averaged over each watershed (a total of 8 variables; 4 seasons $\times$ 2 climate variables). The KKZ method could also be applied to the principal components of projected change used in the clustering methods (Section 2.3.2). However, we chose to follow Cannon (2015a) and Chen et al. (2016) and use projected change itself. We did compare the spread coverage metric used to evaluate the KKZ method (Section 2.4) with and without the inclusion of a principal components analysis (PCA) step and found that including PCA produced negligible differences in the results. The Willamette watershed was the only exception; including PCA produced considerably worse performance in the spread coverage metric for this watershed.

2.3.2 Clustering on projected change

Hierarchical clustering In agglomerative hierarchical clustering, each model begins as a member of its own cluster. The two clusters that are the most similar, based on a chosen linkage method, are merged together to form a new cluster. The merging process continues, with the number of clusters decreasing, until a stopping criterion is reached or all models are members of the same cluster.

The clustering methods used in this study are similar to those suggested by Mendlik and Gobiet (2016) and Wilcke and Bärring (2016). Ward's method was used as the linkage method; it merges the two clusters that result in the smallest increase in the error sum of squares (the sum over all clusters of the squared Euclidean distances between the cluster members and their centroid) (Kaufmann and Rousseeuw 1990). For model selection, the model closest to the cluster centroid was selected. For clusters with two models, which are both equally close to the centroid, the model with the greatest total distance from all other models was selected.

Prior to applying the clustering algorithm, principal components analysis was used to reduce the dimensionality of the climate data and to convert it to uncorrelated variables. PCA was applied to the same matrices of projected PET and PR change used in the KKZ method. PCA was performed using the covariance matrix (i.e., with the data centered but not scaled) since both PET and PR change have the same units and reasonably similar variance. Following Mendlik and Gobiet (2016), the appropriate number of components to retain was selected using the broken-stick method (Frontier 1976), which selects components that have larger eigenvalues than would be expected if the total variance in the dataset was randomly distributed among all possible components (Jackson 1993). The number of components selected for each watershed is provided in Table S2 in the Supporting Information.

Like the KKZ method, the use of PCA prior to the application of the clustering algorithm is not strictly necessary. We compared the spread coverage metric for the clustering results (Section 2.4) with and without the inclusion of the PCA step and found that including PCA produced slightly better (higher) spread coverage in nearly all cases. Therefore, we chose to include the PCA step.

***k*-means clustering** The *k*-means method (Houle et al. 2012; Cannon 2015a; Casajus et al. 2016) is similar to hierarchical clustering with Ward's method in that the objective is to minimize the squared distance between cluster members and their centroids. However, *k*-means does so by iteratively determining cluster membership and updating the positions of cluster centroids rather than by iteratively merging clusters together. In this study, we used the implementation of the *k*-means method provided as the default in the R programming language, which is based on the algorithm of Hartigan and Wong (1979). The algorithm requires a set of initial cluster centers; we ran the algorithm 1000 times with randomly chosen initial centers and selected the resulting clustering that minimized the sum of the squared distances between cluster members and their centroids. All other methods were identical to those used for hierarchical clustering on projected change (Section 2.3.2), including model selection based on distance to cluster centroid and application of principal components analysis to projected PET and PR change.

2.3.3 Hierarchical and *k*-means clustering on climatology

Both hierarchical and *k*-means clustering were also applied to the leading principal components of the modeled historical seasonal climatology of precipitation and potential evapotranspiration (without bias correction). Similar clustering methods have been used to identify models with similar structure and code base (Masson and Knutti 2011; Knutti et al. 2013), and this method could potentially reduce the interdependence of model projections. The clustering methods were the same as for the previous clustering of projected changes.

2.3.4 Historical performance

To measure historical performance, we use the mean absolute error (MAE), a measure of average model error that was argued for by Willmott and Matsuura (2005, 2009). We ranked the climate models based on the MAE of the seasonal climatology of precipitation. The MAE of a model was defined as an average over the four seasons of the absolute value of the difference between the model mean and the observed (Maurer et al. 2002) mean for the season. For each watershed, both the models and the observations were averaged over the watershed and over the historical period (1950 to 1999). A subensemble of size n was defined as the n models with the lowest error. Using the root mean square error of the

precipitation climatology, as suggested by Chai and Draxler (2014), produced essentially the same results.

2.4 Evaluation methods

Each of the six model selection methods was evaluated based on two criteria: first, whether the objective of the method was met with a small number of models; and second, whether the results obtained from the method were sensitive to the number of models chosen. To determine whether the method objective was met, we used different metrics for the methods seeking to maximize the range of projected change included (KKZ and clustering on change) and the metrics seeking to reduce model dependence (clustering on climatology); however, we used the same metric for each method to evaluate the sensitivity to the number of models chosen. For both of the objective metrics and the sensitivity metric, we compared the results with a naive baseline that a useful model selection method should improve upon by randomly choosing models without replacement or randomly assigning models to clusters. We set a threshold at the 95th percentile of 10,000 random simulations, so that an evaluation score was considered acceptable if it was better than could easily be obtained by chance from random selection. Results from the subensembles could also be compared with the full-model ensemble (for example, to test whether the complete range of change in the full ensemble is included in a subensemble), but we think random simulations are a more realistic benchmark as model selection inherently involves a trade-off between reduced computational costs and reduced representation of the full-model ensemble.

To evaluate whether the clustering-on-change and KKZ methods met their objective of covering the majority of the spread in projected change in the full ensemble, the fraction of the total ensemble range covered by the models selected by these methods was calculated (similar to the fractional range coverage metric in McSweeney et al. (2015)). Given an ensemble of projections of runoff change R and a subset of the ensemble R_{sub}, the fraction of the range covered is:

$$\frac{\max R_{sub} - \min R_{sub}}{\max R - \min R} \tag{1}$$

The fraction of the range covered was calculated for each selection method, season, and subensemble size. Then, to summarize these results, we calculated the fraction of the four-season range covered, which is simply the product of the fraction of the range covered in each season s:

$$\prod_s \left(\frac{\max R_{sub} - \min R_{sub}}{\max R - \min R} \right)_s \tag{2}$$

This is similar in concept to the percent spread area calculated by Chen et al. (2016). When compared with the random simulation baseline, we considered a method to be successful if the range coverage metric for any subensemble with between five and ten members exceeded the 95th percentile.

We applied two main tests to the results from the clustering-on-climatology methods. First, we tested whether the methods successfully chose models that were developed by different institutions. Given the findings of Masson and Knutti (2011), Knutti et al. (2013), and Leduc et al. (2016), the developing institution should have a large role in determining whether one climate model is independent of another. To apply the test, we used the Hubert and Arabie (1985) adjusted Rand index to compare the clusterings produced by the two clustering-on-climatology methods with an adaptation of the clustering based on institution and code similarity developed by Knutti et al. (2013) (the colors in their Fig. 1a). Given two

different partitions, or clusterings, of the data, the Rand index is based on how all possible pairs of points in the data are grouped:

$$RI = \frac{a + d}{N}, \tag{3}$$

where a and d are the number of pairs that the two partitions agree about and N is the total number of pairs that can be formed from the data (Milligan and Cooper 1986). Specifically, a is the number of pairs that are grouped together in both partitions; for example, if the two partitions both grouped the HadGEM2-ES and HadGEM2-AO models together, this would contribute 1 to the value of a. d is the number of pairs that are in different clusters in both partitions. A Rand index of 0 indicates complete disagreement between the two partitions, and a Rand index of 1 indicates complete agreement. The adjusted Rand index (ARI) modifies the Rand index to have an expected value of 0, so that an ARI of 0 is equal to the level of agreement that would be obtained on average by randomly creating subsets from the data. The maximum possible ARI, indicating complete agreement, is still 1. To compare with a random baseline, we calculated the 95th percentile of the ARIs when comparing the random clusters with the Knutti et al. (2013) reference clustering. Variability in the 95th percentile between watersheds and selection methods was small, so we set the threshold at 0.087, the average over all watersheds and selection methods. A method with maximum ARIs above 0.087 in every watershed was considered to have passed the test. To be consistent with the evaluation for clustering on change, which considers a method successful if any reasonably sized subensemble (between five and ten models) exceeds the random threshold, we also created clusterings on climatology with between five and ten clusters and selected the maximum ARI for each watershed and clustering method.

The second test applied to the results from the clustering-on-climatology methods used the ARI to compare whether clusters of historical climatology correspond to clusters of projected runoff change. This test evaluates the hypothesis underlying the clustering-on-climatology methods—that a subensemble of models with unique or independent historical climatologies will produce a subensemble of results with unique or independent projected changes—regardless of whether or not the clustering on climatology identifies models that are known to be independent. Like the other datasets, principal components analysis was applied to the seasonal runoff change projections prior to application of the clustering algorithms. Also as before, we used random simulations to set a threshold value of 0.080 for the ARI. Subensembles with between five and ten clusters were compared with the optimal clustering of projected runoff change, which was determined using the average silhouette width (Rousseeuw 1987) following Wilcke and Bärring (2016), and the maximum ARI for each watershed and clustering method was chosen. The optimal number of clusters for each watershed is provided in Table S3 in the Supporting Information.

As an alternative test of the clustering-on-climatology methods, we compared the Knutti et al. (2013) clustering based on institution and code history with the clusters of projected runoff change obtained in the previous test. Knutti et al. (2013) show that this clustering corresponds well to clusters derived from full global data with more sophisticated methods, and the clustering could provide a better picture of model dependence that translates to a better identification of unique changes at the regional scale. Melsen et al. (2018) is an example of a study that used this method to select models. As before, we used the ARI for comparison and used random simulations to set a threshold value of 0.085 for the ARI. In this experiment, we varied the number of clusters of projected runoff change between five and ten, and chose the maximum adjusted Rand index.

We do not directly test whether the performance method meets its objective of selecting models that accurately predict future change. Instead, we use the clustering-on-climatology results to determine whether there is any similarity between historical climate and projected runoff change. If there is little similarity between historical climate and projected runoff change, selecting models with accurate historical climate simulations is unlikely to result in a subensemble of models that also has accurate projected runoff change.

Finally, to determine how results from all of the methods were sensitive to the number of models selected, we (1) used each selection method to create subensembles containing between five and ten models; (2) calculated the subensemble mean annual total runoff change projected for each watershed, selection method, and subensemble size; and (3) calculated the standard deviation of the mean runoff change for each selection method and watershed. A higher standard deviation indicates that for the given method and watershed, the subensemble mean runoff change is more sensitive to the number of models chosen. As a baseline metric to compare with, we also created 10,000 subensembles using random selection without replacement and computed the mean and the 5th and 95th percentiles of the standard deviation as the sizes of the subensembles varied between 5 and 10 models. The random subensembles were ordered; in other words, to create one of the 10,000 subensembles, we randomly chose 5 models, then iteratively increased the subensemble size by randomly choosing one additional model at a time, rather than by completely redrawing a new random subensemble each time. We considered a method to have failed the sensitivity test if the subensemble mean runoff change was more sensitive to the number of models selected than the 95th percentile of the random simulations in any watershed.

3 Results

3.1 Full-model ensemble

Substantial uncertainty about runoff in the second half of the twenty-first century is present in all five of the study watersheds (Fig. S1 in the supporting information). Only in two watersheds (the Minnesota and Susquehanna) do at least three-fourths of the models agree about the sign of the projected annual runoff change, and only in the Susquehanna watershed do all 29 models agree about the sign. Unanimous agreement about the sign of seasonal runoff changes is only found for winter in Susquehanna and Minnesota (presumably from decreased snowfall and increased snowmelt) and for summer in the Willamette, while at least three-fourths of the models also agree about the sign of change for one or two seasons in each watershed (autumn in Apalachicola, spring in Susquehanna and Roosevelt, spring and autumn in Minnesota, and winter and spring in Willamette).

3.2 Model subensembles

3.2.1 Method objectives

We begin the model subensemble analysis with the methods that have a goal of achieving a wide range of change: the KKZ and clustering-on-change methods. The KKZ method meets the goal of including extreme changes in model subensembles (Fig. 3). The KKZ method works particularly well in the Apalachicola and Willamette watersheds, where only five models are necessary for subensembles created using the method to cover over 75% of the range of projected change in the full ensemble (2). Although the method requires more

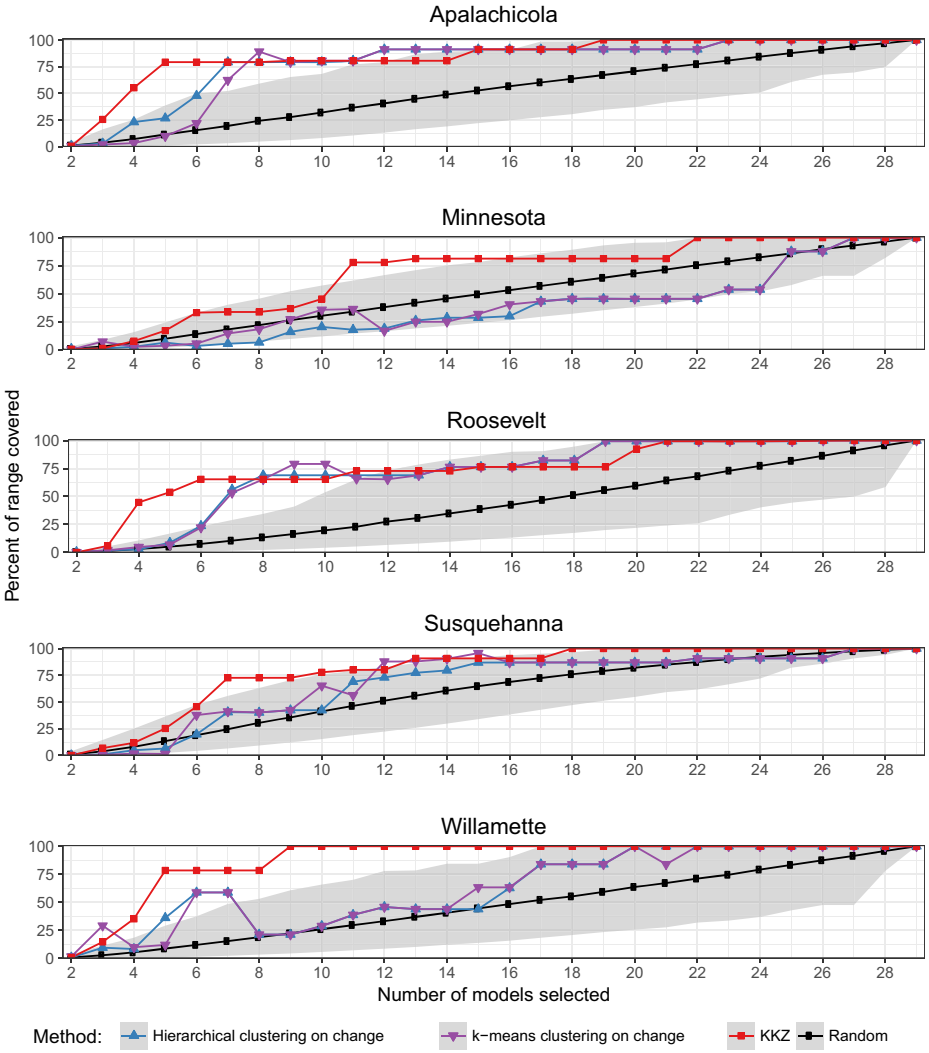

Fig. 3 Comparison of the percent of the range of seasonal runoff change covered by the models chosen by the KKZ and clustering-on-change methods. The shaded area shows the 5–95 percentile range of 10,000 subensembles created randomly without replacement, and the black line shows the mean of these random subensembles

models to achieve similar spread coverage in the Minnesota and Susquehanna watersheds, in all five watersheds at least one subensemble with between five and ten members (the range we consider reasonable) exceeds the 95th percentile of the random selection simulations.

Compared with subensembles created with the KKZ method, subensembles produced using both of the clustering-on-change methods (hierarchical and *k*-means) typically cover a smaller percentage of the full-ensemble spread for a given subensemble size. In addition, for subensembles within the five to ten model range, both clustering methods fail to exceed the 95th percentile of random selections in the Minnesota and Susquehanna watersheds, and many subensembles had less spread coverage than the average of the random selections. The

results from the two clustering methods are not notably different from each other, except in the Apalachicola watershed, where hierarchical clustering obtains higher spread coverage with smaller subensemble sizes than k-means does, and in the Minnesota watershed, where the opposite occurs.

The clustering-on-climatology methods are modestly capable of selecting models that have unique developing institutions and code bases as identified by Knutti et al. (2013) (Fig. 4a). All of the adjusted Rand indexes (ARIs) are above 0, which indicates that the similarity between climatology and developing institution is at least higher than expected by chance. However, the ARIs obtained for the Minnesota and Willamette watersheds are not

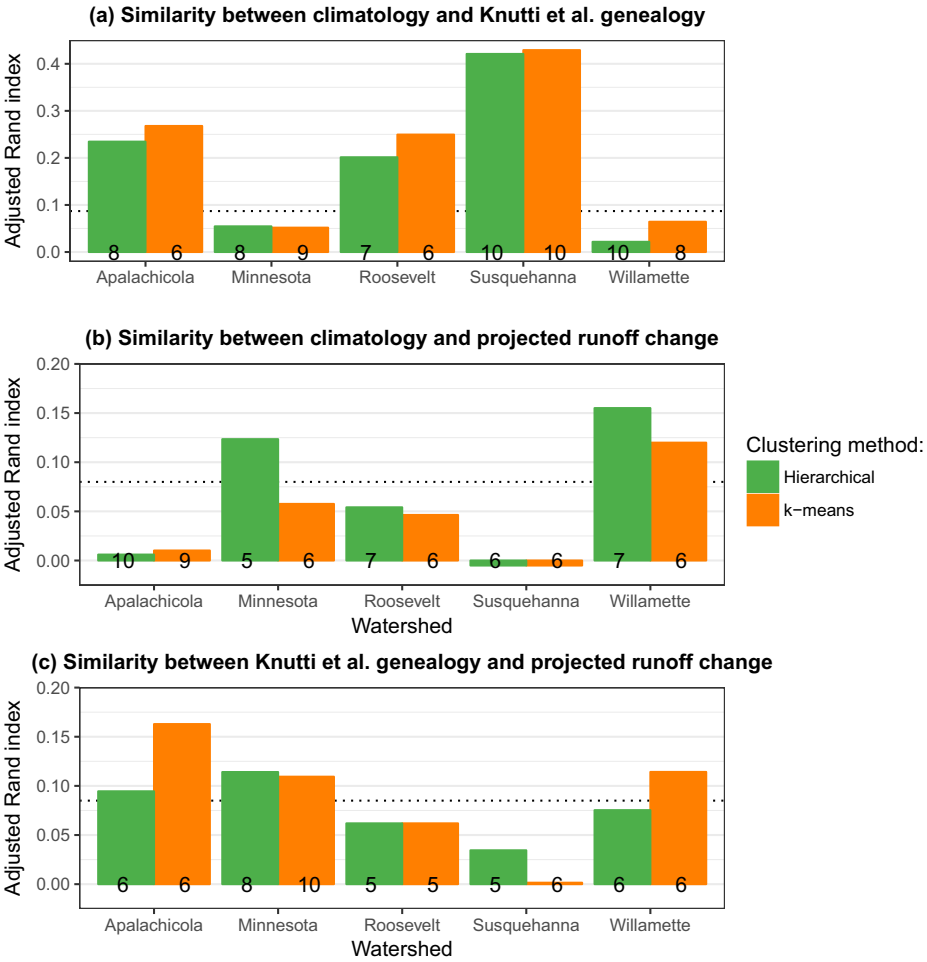

Fig. 4 Comparison between (**a**) clusters of precipitation and PET climatology for each watershed and the clustering derived from Knutti et al. (2013), (**b**) clusters of climatology and projected runoff change for each watershed, and (**c**) the clustering derived from Knutti et al. (2013) and the clustering of projected runoff change for each watershed. Dashed lines indicate the mean 95th percentile of adjusted Rand indices for random clusters. Integers at the bottom of each bar indicate the number of clusters between five and ten at which the maximum adjusted Rand index is obtained. Note that the range of the y-axis in (**a**) is different from the remaining subfigures

greater than the 95th percentile of random simulations, and the values for the Apalachicola and Roosevelt watersheds indicate only weak similarity. The similarity between clusters of climatology and projected runoff change is even more modest and is less than expected by chance in the Susquehanna watershed (Fig. 4b). Opposite of the results in Fig. 4a, only hierarchical clustering in the Minnesota watershed and both methods in the Willamette watershed exceed the 95th percentile of ARIs for random clusterings. Because the clustering methods using regional climatology were only modestly capable of identifying models that are considered independent and unique (Fig. 4a), we instead clustered models that were determined to be similar by Knutti et al. (2013) and compared this clustering with our clustering of projected runoff change (Fig. 4c). This results in an improvement in the similarity in most watersheds, although the similarity remains weak in all watersheds and the similarity in the Roosevelt and Susquehanna watersheds is still less than the 95th percentile of random subensembles.

Overall, the differences between using hierarchical or k-means clustering on climatology are not notable. Most differences are small, and there is no clear pattern of whether hierarchical or k-means clustering is better. The number of clusters at which the maximum adjusted Rand index is obtained also varies considerably, although in many cases six clusters of climatology compare well with the optimal clustering of projected runoff change and six clusters of projected runoff change compare well with the Knutti et al. (2013) genealogy.

We did not directly test whether selection based on model performance results in subensembles with accurate future projections. However, because the previous results show that the similarity between clusters of historical climate and of projected runoff chance is not significantly higher than chance, selecting models with more accurate historical climates (the performance method) is unlikely to yield models with more accurate projected changes.

3.2.2 Sensitivity to number of models

Both hierarchical and k-means clustering on historical climatology failed the test for sensitivity to the number of models chosen (Fig. 5); both methods were more sensitive than the 95th percentile of the random simulations in at least one watershed, and the k-means method had a higher standard deviation than the average standard deviation of the random simulations in four of the five watersheds. Higher sensitivity from these clustering methods is not surprising because both the hierarchical and k-means methods can add and remove models from a subensemble as the size of the subensemble is increased, whereas the remaining methods do not remove a model after it has been selected and produce ordered results. As a result, the means and other statistical properties of the clustering subensembles tend to be more sensitive to the number of models chosen.

The performance method also failed the test by exceeding the 95th percentile of the random simulations in the Willamette watershed, although the method performed acceptably in other watersheds. In the Willamette, the four highest ranked models project much larger decreases in annual total runoff than most other models, while the tenth-ranked model projects a significant increase in annual runoff. This situation produces a high sensitivity metric and also reveals a weakness of the performance method: whereas the other methods will generally select only one model from a group of models with similar climatology or projected change, the performance method has a tendency to select either all or no models from a group with similar climatology. In this particular case, the first two models chosen by the performance method are the two MIROC-ESM models. These two models project annual total runoff changes of −24.5% and −16.6%, respectively, compared with the complete ensemble mean change of only −4.6%. Of the other models in the top ten, three belong

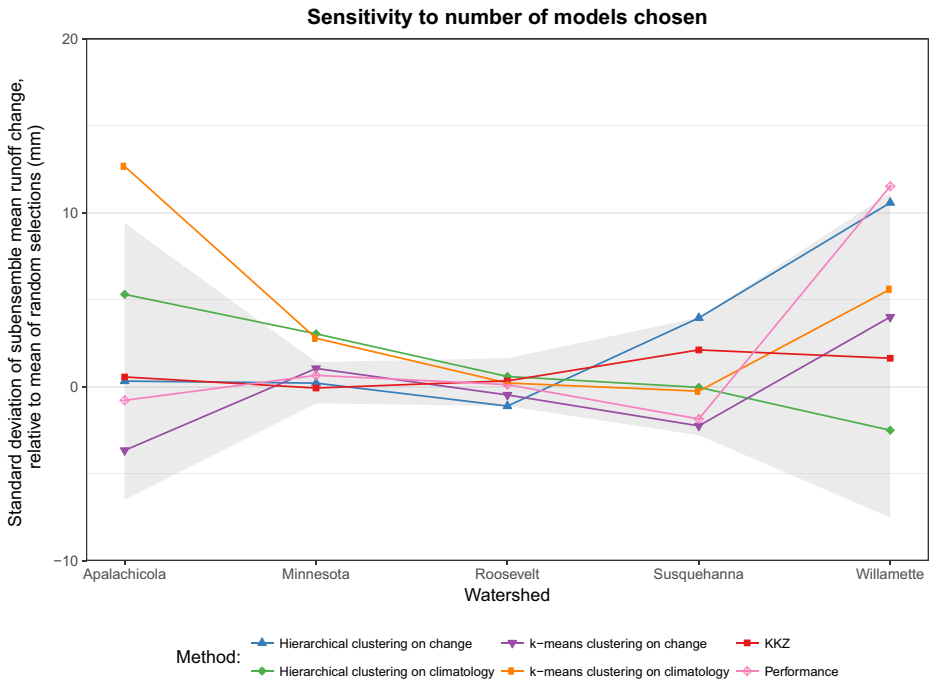

Fig. 5 Evaluation of the sensitivity of the results of the selection methods to the number of models chosen. The y-axis shows the standard deviation of the subensemble mean runoff changes for subensembles with between 5 and 10 models. The gray-shaded region denotes the 5–95 percentile ranges of the random simulations. All values are normalized relative to the mean standard deviation of the random subensembles. Lines are used to connect methods and allow an easier identification of how a method performs across different watersheds

to the CCSM4/CESM family, and two others have some similar components; however, the within-group agreement regarding projected runoff change is not as strong in these models.

The remaining methods did not exceed the 95th percentile in any of the five watersheds. Clustering on change with k-means performed fairly well and had the lowest sensitivity in two of the five watersheds. Finally, the KKZ method had higher sensitivity than the mean of the random selections in four out of the five watersheds; however, the KKZ results were generally close to the random mean and exhibited less variability across watersheds than the other methods.

4 Discussion

4.1 Method comparison and other considerations

A summary of the results from the model selection method comparison is provided in Table 1. Of all six methods, only the KKZ method met both criteria for suitability. The method converged relatively quickly toward its goal of covering a wide range of spread, and the KKZ results were not significantly more sensitive to the number of models chosen than the random simulations. Although the KKZ method performed well at covering the

Table 1 Summary of the results

Method	Met objective?	Insensitive to n?
KKZ	Yes; covered about 75% of the spread in most watersheds with ≤ 8 models	Yes
Hierarchical and k-means clustering on change	No; did not significantly outperform random in some watersheds, and required larger subensembles than KKZ in others	Yes
Hierarchical and k-means clustering on climatology	No; only modest success at clustering models with similar heritage, and little correlation with projected change	No
Performance	Unlikely; little correlation between historical performance and projected change	No

majority of the spread in the ensembles, covering the *complete* range of spread was difficult in some watersheds, where as many as 22 models were needed to include the full range. It also appears to be difficult to predict how many models are necessary to choose using the KKZ method to cover a certain percent of the spread; for example, only 5 models are necessary to cover 75% of the spread in the Willamette watershed, but 11 are necessary in the Minnesota watershed. However, the varying number of models needed to cover uncertainty may be related to the nature of the model ensemble and the processes driving change and is not necessarily a fault of the KKZ method. Furthermore, the sensitivity of the KKZ subensemble mean runoff change to the number of models chosen exhibited less variability between watersheds than the other selection methods (Fig. 5). Compared with the KKZ method, both the hierarchical and k-means clustering-on-change methods did not perform as well at covering a wide range of model uncertainty.

Both the hierarchical and k-means clustering methods using climatology also failed the evaluation; both methods were strongly sensitive to the number of models chosen in at least one watershed, and both methods did not skillfully identify clusters of projected runoff change using only historical precipitation and potential evapotranspiration data. Grouping the models using the Knutti et al. (2013)-derived clustering (hereafter K13), which is based on the institution that developed each model and on code similarity and is similar to Melsen et al. (2018) and the recommendation of Abramowitz et al. (2019), resulted in higher similarity with clusters of projected runoff change compared with using watershed-averaged PR and PET. One disadvantage of using watershed-averaged values for clustering, which the superior results from the K13 clustering highlight, is that it removes and ignores all spatial covariance both within and outside of the watershed that may be important for identifying models with similar structure and projected changes. A disadvantage of the comparison with the K13 clustering in Fig. 4 a and c is that we treated the K13 clustering as the definitive clustering of model similarity. However, other definitions of model similarity are possible, even when relying solely on information about institution and code similarity. For example, the K13 clustering places CESM1-BGC and NorESM1-M into the same cluster, and although both models share some components, they were also developed and run by two different institutions and it may not necessarily be correct to group these models together. Similarly, the use of the adjusted Rand index in these comparisons ignores how

quantitatively similar different models are and only evaluates whether the results from the models are similar or different enough to place the models into the same or different clusters.

It is not currently possible to determine whether the projections from the subensembles created by the performance method met the objective of providing more accurate future projections than the other subensembles. In some cases, cross-validation could be used to evaluate the performance method by comparing the means or trends during a time period used for selection with the means or trends during a second out-of-sample time period (e.g., Terando et al. 2012). In this study, cross-validation would be difficult to apply for a number of reasons, including the application of bias correction to the climate model data, the weak signal of greenhouse gas forcing in the historical simulation, and the long time periods necessary to avoid the influence of internal variability. However, the weak relationship between historical climate and projected change, which is consistent with other studies (Whetton et al. 2007; Knutti et al. 2010), provides evidence that the performance objective is unlikely to be met.

4.2 Differences between watersheds

Although the behavior of most methods was roughly consistent across the different watersheds, there were some notable differences between regions. For example, although the KKZ method was able to obtain about 75% spread coverage with between 5 and 7 models in 3 of the watersheds, 11 models were needed to reach 75% coverage in the Minnesota watershed (Fig. 3). Compared with the Apalachicola and Willamette watersheds, where high spread coverage is reached with only a few models, model-projected runoff change in the Minnesota watershed is less correlated with projected precipitation and potential evapotranspiration changes (Figs. S2 to S6 in the Supporting Information), so any method based on relationships between PET and PR change and runoff change is likely to be less successful. Correlations within variables and between seasons are also generally lower in the Minnesota watershed compared to the Willamette and especially the Apalachicola watershed, so even if runoff was perfectly predictable from these variables, more models would be required to cover the spread in all four seasons. On the other hand, correlations between the variables and seasons in the Susquehanna watershed are similarly weak, yet the KKZ method needs fewer models to cover most of the spread in the Susquehanna watershed than it does in the Minnesota watershed.

4.3 Common limitations

By selecting individual model simulations, all of the methods examined in this study are based on an implicit assumption that the future climate will lie within or along the bounds of the full-model ensemble (Terando et al. 2012). However, the future climate may lie outside of the bounds of the ensemble because the model point estimates in the ensemble neglect uncertainties that may be substantial, especially at the regional scale considered here (Sriver et al. 2015). The number of models in the ensemble (29) may also be too small to reliably include values from the tails of the actual probability distribution. As a result, all methods may underestimate the plausible range of future change.

A second limitation is that none of the methods in this study assigns any probability to the future climate state; however, probabilistic information about future climate conditions is important for decision-making (Krzysztofowicz 2001; Ramos et al. 2013). Various methods, such as hierarchical models (Najafi and Moradkhani 2013; Kerkhoff et al. 2015) and Bayesian model averaging (Bastola et al. 2011), have been employed to produce

projections that are probabilistic and are not limited to the range of climates contained within the climate model ensemble. However, unlike the model selection methods studied here, these methods do not solve the problem of computational expense unless they are applied to the results after model selection.

One limitation of this study is that we have focused solely on selecting from the range of uncertainty represented by multiple GCMs and have ignored other sources of uncertainty. For example, Melsen et al. (2018) found that although GCM uncertainty dominates over hydrological model uncertainty in projections of runoff change for the USA, hydrological model uncertainty is larger in some regions of the USA. From the perspective of a study seeking to apply model selection to cover GCM uncertainty, some a priori knowledge of dominant sources of uncertainty in the study region may be useful for determining how many GCMs to select and whether to also include multiple downscaling methods, hydrological models, parameter choices, etc., in the experimental design.

A second possible limitation of this study is that we have examined only mean conditions for four seasons for each watershed; however, extreme conditions, such as flood events and prolonged droughts, may also be important to consider in some applications. Adapting all of the selection methods used in this study to focus on metrics relevant to changes in extremes would be straightforward, and, provided that there is a reasonable correlation between the climate model inputs used in the selection methods and the hydrological extremes of interest, we expect that the results of this study will also apply. Our finding of the superiority of the KKZ method is in agreement with studies that have used the KKZ and k-means clustering-on-change methods to cover uncertainty in a number of extreme indices (Cannon 2015a; Wang et al. 2018).

As noted by Cannon (2015a), a potential limitation of the KKZ and clustering-on-change methods is that the model selections are sensitive to the presence of outliers. For example, in the Roosevelt watershed, a comparison of the models selected by the KKZ and performance methods (not shown) indicates that the majority of the spread in the ensemble is produced by a few of the worst performing models. In this case, the KKZ and clustering-on-change methods that are designed to obtain high spread coverage with small ensembles do so by prioritizing selection of all of the outlying, poorly performing ensemble members. If the KKZ or clustering methods were to be used for a study in this region, it would be worthwhile to examine whether the outlying models have errors so severe that future climate changes cannot be properly represented. With strong evidence that future climate changes cannot be properly simulated, the models could be removed from the ensemble prior to application of the selection algorithm. Some studies have included outlier screening and removal steps in their model selection methods. However, we are hesitant to do so for three reasons. First, the screening and removal steps introduce subjective, human judgment into procedures that are otherwise objective and automated. Second, poor historical performance may often be irrelevant given the typically weak correlation between historical climate and future projections in climate models. Third, erroneous discarding of outlying models is likely to result in a substantial underestimate of risk and uncertainty.

5 Conclusion

Of the six selection methods tested, the KKZ method was the only method that was capable of both meeting its objective with relatively small subensemble sizes and producing results that were not strongly sensitive to the number of models selected. However, regardless of the selection method used, the results from top–down style assessments should be carefully

interpreted. Projections from small model subensembles are likely to contain a large degree of uncertainty as a result of the choice of selection method and the number of models chosen, yet the results from these small subensembles are also likely to not cover the range of extreme changes contained in the full ensemble. It may also be difficult to predict exactly how many models are necessary to meet certain criteria, and subensembles with good properties in one region are not guaranteed to maintain those properties in other regions. On the other hand, results from large subensembles or complete ensembles may be affected by the inclusion of structurally and statistically similar models. At the very least, studies selecting data from a small number of models for use in downscaling and impacts assessment should assess the properties of the selected models in context of the models that were not selected, and studies using large or complete ensembles should consider the impact of model dependence on the results.

Acknowledgments We thank the anonymous reviewers and editor for comments that improved this manuscript.We also thank the World Climate Research Programme's Working Group on Coupled Modelling, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison, the Global Organization for Earth System Science Portals, and the climate modeling groups (Table S1) for their roles in CMIP, as well as the providers of the downscaled hydrology projections.

Funding information Funding was provided by the National Science Foundation (CBET-1360286), Pennsylvania Sea Grant (NA10OAR4170063), and National Oceanic and Atmospheric Administration's National Centers for Coastal Ocean Science (NA16NOS4780207 to Virginia Institute of Marine Science).

References

Abramowitz G, Bishop CH (2015) Climate model dependence and the ensemble dependence transformation of CMIP projections. J Clim 28(6):2332–2348

Abramowitz G, Herger N, Gutmann E, Hammerling D, Knutti R, Leduc M, Lorenz R, Pincus R, Schmidt GA (2019) ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing. Earth Syst Dyn 10(1):91–105

Al Aamery N, Fox JF, Snyder M (2016) Evaluation of climate modeling factors impacting the variance of streamflow. J Hydrol 542:125–142

Arnell NW, Gosling SN (2013) The impacts of climate change on river flow regimes at the global scale. J Hydrol 486:351–364

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

Bishop CH, Abramowitz G (2013) Climate model dependence and the replicate Earth paradigm. Clim Dyn 41:885–900

Brekke LD, Dettinger MD, Maurer EP, Anderson M (2008) Significance of model credibility in estimating climate projection distributions for regional hydroclimatological risk assessments. Clim Chang 89:371–394

Brekke L, Thrasher BL, Maurer EP, Pruitt T (2013) Downscaled CMIP3 and CMIP5 climate projections: release of downscaled CMIP5 climate projections, comparison with preceding information, and summary of user needs. Technical report, U.S. Department of the Interior, Bureau of Reclamation Technical Services Center, Denver

Brekke L, Wood A, Pruitt T (2014) Downscaled CMIP3 and CMIP5 climate and hydrology projections: release of hydrology projections, comparison with preceding information, and summary of user needs. Technical report, U.S. Department of the Interior, Bureau of Reclamation Technical Services Center, Denver

Cannon AJ (2015a) Selecting GCM scenarios that span the range of changes in a multimodel ensemble: application to CMIP5 climate extremes indices. J Clim 28(3):1260–1267

Cannon AJ, Sobie SR, Murdock TQ (2015b) Bias correction of GCM precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes? J Clim 28(17):6938–6959

Casajus N, Périé C, Logan T, Lambert M-C, de Blois S, Berteaux D (2016) An objective approach to select climate scenarios when projecting species distribution under climate change. PLoS ONE:11
Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature. Geosci Model Dev 7(3):1247–1250
Chen J, Brissette FP, Lucas-Picher P (2016) Transferability of optimally-selected climate models in the quantification of climate change impacts on hydrology. Clim Dyn 47:3359–3372
Chen J, Brissette FP, Lucas-Picher P, Caya D (2017) Impacts of weighting climate models for hydro-meteorological climate change studies. J Hydrol 549:534–546
Frontier S (1976) ÉTude de la décroissance des valeurs propres dans une analyse en composantes principales: Comparaison avec le modèle du bâton brisé. J Exp Mar Biol Ecol 25:67–75
Gibson JR, Najjar RG (2000) The response of Chesapeake Bay salinity to climate-induced changes in streamflow. Limnol Oceanogr 45(8):1764–1772
Hagemann S, Chen C, Clark DB, Folwell S, Gosling SN, Haddeland I, Hanasaki N, Heinke J, Ludwig F, Voß F, Wiltshire AJ (2013) Climate change impact on available water resources obtained using multiple global climate and hydrology models. Earth Syst Dyn 4:129–144
Hartigan JA, Wong MA (1979) Algorithm AS 136: a k-means clustering algorithm. J R Stat Soc Ser C (Appl Stat) 28(1):100–108
Held IM, Soden BJ (2006) Robust responses of the hydrological cycle to global warming. J Clim 19(21):5686–5699
Herger N, Abramowitz G, Knutti R, Angélil O, Lehmann K, Sanderson BM (2017) Selecting a climate model subset to optimise key ensemble properties. Earth Syst Dyn Discuss 2017:1–24
Hirabayashi Y, Mahendran R, Koirala S, Konoshima L, Yamazaki D, Watanabe S, Kim H, Kanae S (2013) Global flood risk under climate change. Nat Clim Chang 3:816–821
Holman IP, Allen DM, Cuthbert MO, Goderniaux P (2012) Towards best practice for assessing the impacts of climate change on groundwater. Hydrogeol J 20:1–4
Houle D, Bouffard A, Duchesne L, Logan T, Harvey R (2012) Projections of future soil temperature and water content for three southern quebec forested sites. J Clim 25:7690–7701
Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218
Jackson DA (1993) Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecol 74(8):2204–2214
Johnson TE, Butcher JB, Parker A, Weaver CP (2012) Investigating the sensitivity of U.S. streamflow and water quality to climate change: U.S. EPA Global Change Research Program's 20 Watersheds Project. J Water Resour Plan Manag 138(5):453–464
Justić D, Rabalais NN, Turner RE (2005) Coupling between climate variability and coastal eutrophication: evidence and outlook for the northern Gulf of Mexico. J Sea Res 54(1):25–35
Katsavounidis I, Kuo C, Zhang Z (1994) A new initialization technique for generalized Lloyd iteration. IEEE Signal Process Lett 1(10):144–146
Kaufmann L, Rousseeuw PJ (1990) Finding groups in data: an introduction to cluster analysis. Wiley, New York
Kerkhoff C, Künsch HR, Schär C (2015) A Bayesian hierarchical model for heterogeneous RCM–GCM multimodel ensembles. J Clim 28(15):6249–6266
Knutti R, Furrer R, Tebaldi C, Cermak J, Meehl GA (2010) Challenges in combining projections from multiple climate models. J Clim 23:2739–2758
Knutti R, Masson D, Gettelman A (2013) Climate model genealogy: generation CMIP5 and how we got there. Geophys Res Lett 40:1194–1199
Krzysztofowicz R (2001) The case for probabilistic forecasting in hydrology. J Hydrol 249(1-4):2–9
Leduc M, Laprise R, De elía R, Šeparović L (2016) Is institutional democracy a good proxy for model independence? J Clim 29:8301–8316
Li H, Sheffield J, Wood EF (2010) Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching. J Geophys Res 115:D10101
Liang X, Lettenmaier DP, Wood EF, Burges SJ (1994) A simple hydrologically based model of land-surface water and energy fluxes for general-circulation models. J Geophys Res Atmosph 99(D7):14415–14428
Liang X, Wood EF, Lettenmaier DP (1996) Surface soil moisture parameterization of the VIC-2l model: evaluation and modification. Glob Planet Chang 13:195–206
Maraun D (2013) Bias correction, quantile mapping, and downscaling: revisiting the inflation issue. J Clim 26(6):2137–2143
Maraun D, Shepherd TG, Widmann M, Zappa G, Walton D, Gutiérrez JM, Hagemann S, Richter I, Soares PMM, Hall A, Mearns LO (2017) Towards process-informed bias correction of climate change simulations. Nat Clim Chang 7(11):764–773

Masson D, Knutti R (2011) Climate model genealogy. Geophys Res Lett:38
Maurer EP, Wood AW, Adam JC, Lettenmaier DP (2002) A long-term hydrologically based dataset of land surface fluxes and states for conterminous United States. J Clim 15:3237–3251
Maurer EP, Pierce DW (2014) Bias correction can modify climate model simulated precipitation changes without adverse effect on the ensemble mean. Hydrol Earth Syst Sci 18(3):915–925
McSweeney CF, Jones RG, Lee RW, Rowell DP (2015) Selecting CMIP5 GCMs for downscaling over multiple regions. Clim Dyn 44:3237–3260
Melsen LA, Addor N, Mizukami N, Newman AJ, Torfs PJJF, Clark MP, Uijlenhoet R, Teuling AJ (2018) Mapping (dis)agreement in hydrologic projections. Hydrol Earth Syst Sci 22(3):1775–1791
Mendlik T, Gobiet A (2016) Selecting climate simulations for impact studies based on multivariate patterns of climate change. Clim Chang 135:381–393
Milligan GW, Cooper MC (1986) A study of the comparability of external criteria for hierarchical cluster analysis. Multivar Behav Res 21:441–485
Milly P, Wetherald RT, Dunne KA, Delworth TL (2002) Increasing risk of great floods in a changing climate. Nature 415:514–517
Najafi MR, Moradkhani H (2013) A hierarchical Bayesian approach for the analysis of climate change impact on runoff extremes. Hydrol Process 28(26):6292–6308
Nijssen B, Lettenmaier DP, Liang X, Wetzel SW, Wood EF (1997) Streamflow simulation for continental-scale river basins. Water Resour Res 33(4):711–724
Nijssen B, O'Donnell GM, Hamlet AF, Lettenmaier DP (2001) Hydrologic sensitivity of global rivers to climate change. Clim Chang 50:143–175
Ott I, Duethmann D, Liebert J, Berg P, Feldmann H, Ihringer J, Kunstmann H, Merz B, Schaedler G, Wagner S (2013) High-resolution climate change impact analysis on medium-sized river catchments in Germany: an ensemble assessment. J Hydrometeorol 14:1175–1193
Pierce DW, Cayan DR, Maurer EP, Abatzoglou JT, Hegewisch KC (2015) Improved bias correction techniques for hydrological simulations of climate change. J Hydrometeorol 16(6):2421–2442
Rabalais NN, Turner RE, Diaz RJ, Justić D (2009) Global change and eutrophication of coastal waters. ICES J Mar Sci 66:1528–1537
Ramos MH, van Andel SJ, Pappenberger F (2013) Do probabilistic forecasts lead to better decisions? Hydrol Earth Syst Sci 17(6):2219–2232
Riahi K, Rao S, Krey V, Cho C, Chirkov V, Fischer G, Kindermann G, Nakicenovic N, Rafaj P (2011) RCP 8.5-a scenario of comparatively high greenhouse gas emissions. Clim Chang 109:33–57
Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65
Sanderson BM, Knutti R, Caldwell P (2015) A representative democracy to reduce interdependency in a multimodel ensemble. J Clim 28:5171–5194
Schewe J, Heinke J, Gerten D, Haddeland I, Arnell NW, Clark DB, Dankers R, Eisner S, Fekete BM, Colón-González FJ, Gosling SN, Kim H, Liu X, Masaki Y, Portmann FT, Satoh Y, Stacke T, Tang Q, Wada Y, Wisser D, Albrecht T, Frieler K, Piontek F, Warszawski L, Kabat P (2014) Multimodel assessment of water scarcity under climate change. Proc Natl Acad Sci 111(9):3245–3250
Sriver RL, Forest CE, Keller K (2015) Effects of initial conditions uncertainty on regional climate variability: an analysis using a low-resolution CESM ensemble. Geophys Res Lett 42(13):5468–5476
Steinschneider S, McCrary R, Mearns LO, Brown C (2015) The effects of climate model similarity on probabilistic climate projections and the implications for local, risk-based adaptation planning. Geophys Res Lett 42:5014–5022
Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Philosophical Transactions of the Royal Society A: Mathematical. Phys Eng Sci 365:2053–2075
Teng J, Vaze J, Chiew FHS, Wang B, Perraud J-M (2012) Estimating the relative uncertainties sourced from GCMs and hydrological models in modeling climate change impact on runoff. J Hydrometeorol 13:122–139
Terando A, Keller K, Easterling WE (2012) Probabilistic projections of agro-climate indices in North America. Journal of Geophysical Research Atmospheres:117
Vetter T, Reinhardt J, Flörke M, Griensven A, Hattermann F, Huang S, Koch H, Pechlivanidis IG, Plötner S, Seidou O, Su B, Vervoort RW, Krysanova V (2017) Evaluation of sources of uncertainty in projected hydrological changes under climate change in 12 large-scale river basins. Clim Chang:141
Vicuna S, Maurer EP, Joyce B, Dracup JA, Purkey D (2007) The sensitivity of California water resources to climate change scenarios. J Amer Water Resour Assoc (JAWRA) 43(2):482–498
Wang H-M, Chen J, Cannon AJ, Xu C-Y, Chen H (2018) Transferability of climate simulation uncertainty to hydrological climate change impacts. Hydrol Earth Syst Sci 22:3739–3759

Weigel AP, Knutti R, Liniger MA, Appenzeller C (2010) Risks of model weighting in multimodel climate projections. J Clim 23:4175–4191
Weiland FCS, van Beek LPH, Weerts AH, Bierkens MFP (2012) Extracting information from an ensemble of GCMs to reliably assess future global runoff change. J Hydrol 412-413:66–75
Whetton P, Macadam I, Bathols J, O'Grady J (2007) Assessment of the use of current climate patterns to evaluate regional enhanced greenhouse response patterns of climate models. Geophys Res Lett:34
Wilby RL, Dessai S (2010) Robust adaptation to climate change. Weather 65(7):180–185
Wilcke RAI, Bärring L (2016) Selecting regional climate scenarios for impact modelling studies. Environ Modell Softw 78:191–201
Willmott CJ, Matsuura K (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 30:79–82
Willmott CJ, Matsuura K, Robeson SM (2009) Ambiguities inherent in sums-of-squares-based error statistics. Atmos Environ 43(3):749–752
Wood AW, Leung LR, Sridhar V, Lettenmaier DP (2004) Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. Clim Chang 62:189–216
Zubler EM, Fischer AM, Fröb F, Liniger MA (2016) Climate change signals of CMIP5 general circulation models over the Alps—impact of model selection. International Journal of Climatology:3088–3104

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.