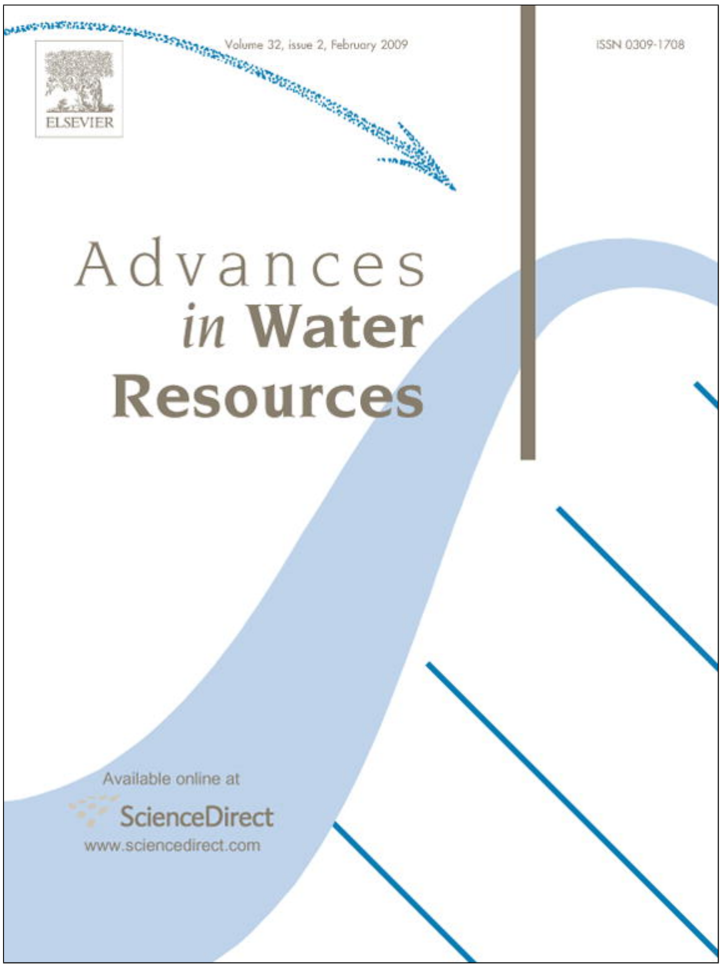
Volume 32, issue 2, February 2009
ISSN 0309-1708
ELSEVIER
Advances
in Water
Resources
Available online at
ScienceDirect
www.sciencedirect.com

Advances in Water Resources 32 (2009) 147–158

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres

Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) II: Ensemble combinations and predictions

Neil R. Viney [a,*], H. Bormann [b], L. Breuer [c], A. Bronstert [d], B.F.W. Croke [e], H. Frede [c], T. Gräff [d], L. Hubrechts [f], J.A. Huisman [g], A.J. Jakeman [e], G.W. Kite [h], J. Lanini [i], G. Leavesley [j], D.P. Lettenmaier [i], G. Lindström [k], J. Seibert [l], M. Sivapalan [m,1], P. Willems [n]

[a] *CSIRO Land and Water, GPO Box 1666, Canberra, ACT 2600, Australia*
[b] *Institute for Biology and Environmental Sciences, Carl von Ossietzsky University, Oldenburg, Germany*
[c] *Institute for Landscape Ecology and Resources Management, Justus-Liebig University, Giessen, Germany*
[d] *Institute for Geoecology, University of Potsdam, Germany*
[e] *Integrated Catchment Assessment and Management Centre, The Australian National University, Canberra, Australia*
[f] *Afdeling Ecologie en Water, Lisec NV, Genk, Belgium*
[g] *ICG-4 Agrosphere, Forschungszentrum Jülich, Germany*
[h] *Hydrologic Solutions, Pantymwyn, United Kingdom*
[i] *Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, USA*
[j] *United States Geological Survey, Denver, CO, USA*
[k] *Swedish Meteorological and Hydrological Institute, Norrköping, Sweden*
[l] *Department of Physical Geography and Quaternary Geology, Stockholm University, Stockholm, Sweden*
[m] *Centre for Water Research, University of Western Australia, Perth, Australia*
[n] *Hydraulics Laboratory, Katholieke Universiteit Leuven, Belgium*

ARTICLE INFO

Article history:
Received 13 July 2007
Received in revised form 20 May 2008
Accepted 22 May 2008
Available online 5 June 2008

Keywords:
Multi-model ensembles
Single-model ensembles
Catchment modelling
Ensemble combination
Uncertainty
Land use change

ABSTRACT

This paper reports on a project to compare predictions from a range of catchment models applied to a mesoscale river basin in central Germany and to assess various ensemble predictions of catchment streamflow. The models encompass a large range in inherent complexity and input requirements. In approximate order of decreasing complexity, they are DHSVM, MIKE-SHE, TOPLATS, WASIM-ETH, SWAT, PRMS, SLURP, HBV, LASCAM and IHACRES. The models are calibrated twice using different sets of input data. The two predictions from each model are then combined by simple averaging to produce a single-model ensemble. The 10 resulting single-model ensembles are combined in various ways to produce multi-model ensemble predictions. Both the single-model ensembles and the multi-model ensembles are shown to give predictions that are generally superior to those of their respective constituent models, both during a 7-year calibration period and a 9-year validation period. This occurs despite a considerable disparity in performance of the individual models. Even the weakest of models is shown to contribute useful information to the ensembles they are part of. The best model combination methods are a trimmed mean (constructed using the central four or six predictions each day) and a weighted mean ensemble (with weights calculated from calibration performance) that places relatively large weights on the better performing models. Conditional ensembles, in which separate model weights are used in different system states (e.g. summer and winter, high and low flows) generally yield little improvement over the weighted mean ensemble. However a conditional ensemble that discriminates between rising and receding flows shows moderate improvement. An analysis of ensemble predictions shows that the best ensembles are not necessarily those containing the best individual models. Conversely, it appears that some models that predict well individually do not necessarily combine well with other models in multi-model ensembles. The reasons behind these observations may relate to the effects of the weighting schemes, non-stationarity of the climate series and possible cross-correlations between models.

1. Introduction

A model is a simplified conceptualization of a complex, possibly chaotic system, which is often characterized by highly variable behaviour in space and time. As such, no model, particularly those

[*] Corresponding author.
E-mail address: neil.viney@csiro.au (N.R. Viney).
[1] Present address: Department of Geography, University of Illinois, Urbana-Champaign, IL, USA.

0309-1708/$ - see front matter Crown Copyright © 2008 Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.advwatres.2008.05.006

associated with natural systems, can ever provide a perfect realization. Indeed, it can even sometimes be difficult to quantify the degree of uncertainty in input data, model structure and model parameterization. Taken together, these uncertainties inevitably lead to considerable uncertainty in model predictions. One of the ways of addressing some of these uncertainty issues is through ensemble modelling. The term 'ensemble modelling' encompasses a large range of approaches to producing predictions of fluxes and properties. A single-model ensemble involves the use of a number of realizations of a single deterministic model. Distinct predictions are obtained for each realization by either perturbing the input data or initial conditions, or by selecting different sets of model parameters. These perturbations may be stochastic or deterministic (e.g. derived from alternative sources). In a multi-model ensemble, several different deterministic models are used. These realizations may or may not use a common input data set.

Ensemble modelling has often been used in the climate and atmospheric sciences, where operational ensembles have been in use for well over a decade. Most studies of the accuracy of multi-model ensemble forecasts in weather prediction report that they tend to outperform individual models [9] and that multi-model ensembles tend to perform better than single-model ensembles [28]. Ensemble modelling has the potential to assist in providing better understanding of the physical processes and in informing the development of better models [7,23].

Until relatively recently, however, ensemble modelling has received little attention in hydrology, where most modelling studies use only one model. There have been several studies comparing predictions from various hydrological models (e.g., [27,19]). In general, these studies have been limited to describing how different models and different modelling approaches can affect prediction accuracy, but usually have not considered the issue of pooling model predictions to arrive at some consensus prediction.

Recently, some new cooperative initiatives such as the Distributed Model Intercomparison Project (DMIP) and the Ensemble Streamflow Prediction (ESP) project in the United States and the international Hydrological Ensemble Prediction Experiment (HEPEX) have begun to explore ensemble modelling in a hydrological setting. These projects are aimed primarily at using single-model ensembles to produce short term streamflow forecasts conditioned on climate forecasts (e.g., [6,11,13]). Other ensemble research has focussed on combining predictions using different parameter sets from a single model structure (e.g., [16]).

The first published study on multi-member hydrologic ensembles appears to be that of Shamseldin et al. [23], in which ensemble predictions are constructed using five rainfall-runoff models applied to 11 catchments. They assessed combinations involving simple averaging, a regression-based scheme and a neural network scheme. While the results are inconsistent from one catchment to another, Shamseldin et al. conclude that the combined models generally produce better predictions than the single models. Georgakakos et al. [10], as part of DMIP, assessed predictions from seven distributed models applied to six catchments in the United States. They found that a simple mean of the five best models in each catchment consistently outperforms the best individual model, but that a weighted mean ensemble, while usually better than the best model, is inferior to the simple mean ensemble. In a further analysis of the DMIP dataset, Ajami et al. [1] assessed the predictions of several types of regression-based ensemble and concluded that they are superior to predictions from an unbiased average ensemble and from individual models. They also found that combination methods involving bias removal are desirable, but only in catchments with stationary flow conditions between the calibration and validation periods.

To date, none of these projects have considered ensemble modelling of the hydrological impacts of land use change. This paper is the second in a series of four describing the Land Use Change on Hydrology by Ensemble Modelling (LUCHEM) project, which involves the application of 10 catchment models to a basin with nested gauges. The first paper [5] describes the background to the project and analyses the individual performances of the 10 models in predicting streamflows for current land use conditions. The third paper [12] describes the results of an application of these 10 models to the prediction of the impacts of land use change and assesses the uncertainty in the land use change predictions. The fourth paper [4] assesses the impact of changes in the resolution of spatial input data on the streamflow predictions of three of the LUCHEM models.

In this paper, we assess the use and effectiveness of a number of model ensembles made up of some or all of the individual models. In comparison with previous studies, this paper includes a larger number of models with a large range of inherent model complexity and assesses a larger range of ensemble combination techniques. The following sections describe the modelling procedure and analyse prediction accuracy both for the individual models and for several types of ensembles including all models.

2. The Dill River catchment

The Dill River in Hesse, Germany, is a tributary of the Lahn River, which ultimately flows westward into the Rhine River. The Dill River at Asslar has a catchment area of 693 km^2. The topography of the catchment is characterized by low mountains and has an altitude range of 155–674 m. The mean annual precipitation of the Dill catchment varies from 700 mm in the south to more than 1100 mm in the higher elevation areas in the north and also exhibits a general west-east gradient. Areas with lower annual precipitation tend to have summer-dominated rainfall patterns, while the wetter parts of the catchment are dominated by winter precipitation patterns. A small proportion (less than 5%) of winter precipitation falls as snow, particularly at higher elevations. Refer to Breuer et al. [5] for further details on the catchment's climate, topography, soil and land cover.

Streamflow in the Dill catchment is generated primarily from interflow processes with relatively little baseflow and surface runoff. Mean annual streamflow for the period 1983–1998 is 438 mm (about 48% of catchment-averaged precipitation). There is a distinct winter peak, with 77% of streamflow occurring in the six months from November to April. There are, however, some significant temporal trends in streamflow patterns during the 19-year period used in this study. The mean annual streamflow recorded in the 1990s is about 20% less than that for the 1980s and this has been accompanied by a significant reduction in runoff coefficient. Most of the reduction in streamflow has occurred during the winter months. There is also some evidence that the winter peak and the period of summer low flows are arriving about one month later during the 1990s than during the 1980s.

3. The models

Ten models with the capability of predicting the impacts of land use change are applied to the Dill catchment. In approximately decreasing order of complexity, they are: DHSVM [26], MIKE-SHE [22], TOPLATS [20], WASIM-ETH [18], SWAT [2], PRMS [15], SLURP [14], HBV [3], LASCAM [24] and IHACRES [8]. In terms of their spatial resolution and the overall number of model parameters, the models represent a broad cross-section of complexity ranging from fully distributed, physically based models with explicit groundwater schemes (DHSVM, MIKE-SHE) to fully lumped, conceptual models (e.g. IHACRES). There are also other more subtle differences among the models, including differences in rainfall interpolation, channel routing and estimation of potential evaporation. Some models use explicit snow accumulation routines, while others treat all precipitation as rainfall.

Each model was prepared, calibrated and operated either by its creator or by a modeller with considerable familiarity in its use. Each modeller was provided with common digital maps of elevation, soil type (discriminated into 149 soil classes and including soil physical characteristics) and land cover. Daily precipitation (16 sites) and weather (2 sites) data were provided, but modellers were free to interpolate and re-process this data by any suitable method. Similarly, the calibration methods and objective functions were different from model to model. All models were calibrated using observed streamflow data for the period 1983–1989 and model predictions were developed for the validation period 1990–1998. A maximum of 3 years of additional weather data (1980–1982) was available for model spin-up. All the models were set up to provide spatially explicit predictions of streamflow at a variety of spatial scales across the catchment. Observed streamflows from three gauged subcatchments were available to assess spatial predictions, but this paper is limited to the analysis of model performance against the observed streamflows at Asslar. A more detailed description of model characteristics, set-up procedures and calibration procedures appears in Ref. [5].

The use of different data interpolation schemes by different models resulted in a considerable range of predictions of basic catchment water balance components like mean annual rainfall, actual evaporation and storage [5]. In an effort to minimise these differences, a second calibration was performed for each model. In the second calibration procedure, modellers were restricted to using common, pre-processed, high-resolution (25 m) fields of vegetation density and daily precipitation (interpolated from the 16 observation sites) and to using other climate data from a single, central observation station. They were invited to recalibrate their models to a similar level of rigour as before and submit a second set of model predictions for both the calibration and validation periods. We refer to this as the homogeneous dataset, while the predictions from the first calibration process are referred to as the original dataset. For the MIKE-SHE model, predictions are available only for the homogeneous dataset. More details on the calibration procedures and an analysis of prediction differences between the two calibration processes are available in Ref. [5].

4. Ensemble construction

Although the second calibration was performed primarily to facilitate better comparison of individual model performances, it also provides an opportunity for ensemble analysis. For each model (except MIKE-SHE), we have two semi-independent realizations of flow predictions, one for the original calibration and one for the calibration using the homogeneous dataset. We may combine these into a single time series of flow predictions for each model by taking the mean of the two predictions. This resulting time series will be referred to as a single-model ensemble (SME). We now use these SMEs (along with the homogeneous predictions of MIKE-SHE) as members of several different types of multi-model ensemble (MME).

Ensemble predictions may be constructed in a number of ways. Perhaps the simplest is the approach we used to construct the SMEs, which is to take the raw mean of the model predictions for each day. Another simple ensemble prediction is to adopt the daily median of all ensemble members. In this study, where 10 models are available, the median is given by the mean of the fifth and sixth highest predictions. Ensembles may also be constructed by including a subset of the 10 models (e.g. the mean of the best three models in calibration). A further elaboration might be to include a mixture of two or more subset ensembles, with the switching between ensembles conditioned upon some attribute of the flow regime or climate. In this paper, we assess the performances of the following ensembles:

- mean of all models,
- median of all models,
- trimmed mean (excluding extreme predictions),
- multiple linear regression of all models,
- principal components regression,
- Bayesian model averaging,
- means of all permutations of models ranging from 1 to 10 members,
- weighted mean of all models (weighted by calibration performance),
- weighted means of all permutations of models,
- conditional ensembles based on season,
- conditional ensembles based on flow stage (i.e. rising or falling),
- conditional ensembles based on flow level (i.e. high and low flows).

In the cases of the regression ensembles, the weighted ensembles and the conditional ensembles, the ensemble members and their relative weightings are determined using predictions from the calibration period only. These weighting are then applied without adjustment during the validation period. For example, for the multi-variable linear regression ensemble, regression coefficients are established during the calibration period and applied unchanged during the validation period.

The multiple linear regression ensemble is constructed by using the 10 SMEs as the independent variables and the observed flow in the calibration period as the dependent variable. Given the high correlation between efficiency and the square of the correlation coefficient for values of both approaching one, it is reasonable to expect that the unconstrained multiple linear regression ensemble represents the optimal linear combination of model predictions during the calibration phase and it should therefore give better efficiencies than the raw mean, the trimmed means and the weighted means – all of which are inferior linear combinations – and any individual model. The regression ensemble must also have zero bias during the calibration period. However, none of these properties will necessarily hold for the validation period. One disadvantage of the regression approaches is that they might include a non-zero intercept. This could result in a nonzero ensemble flow prediction even when all models are predicting zero flow. Regression ensembles might also involve negative coefficients for some models and this too can potentially result in negative flow predictions. To partially overcome these issues we also assess two constrained (zero-intercept) multiple regressions: one involving all 10 models and one including only those models with positive and significantly non-zero coefficients.

Where there is multicollinearity between the independent variables (the individual SMEs) the predictions of a multiple linear regression model may have reduced reliability. Principal components regression (PCR) has been proposed as a method for coping with this collinearity by transforming the predictor variables into a set of orthogonal variables called principal components. These orthogonal variables (or a subset of them) are then suitable for use in ordinary least squares regression. In this study we assess the predictions of PCR ensembles constructed from 1, 2, 3 and 4 principal components.

Recently, a Bayesian model averaging (BMA) method was proposed to optimally combine the predictions of a multi-model ensemble [21,25]. The BMA predictive model can be expressed as

$$p(\Delta \| f_1, \ldots, f_n) = \sum_{i=1}^{n} w_i g_i(\Delta \| f_i),$$

where f is the predictions for each ensemble member, w_i is the weight of the ith model and $g_i(\Delta\|f_i)$ can be interpreted as the probability density function (PDF) of the measurement given the forecast, conditional on the prediction being the best prediction in the ensemble. The weights can be interpreted as the individual model's relative contribution to predictive skill of the ensemble. They are positive and add up to one. In the original approach of Raftery et al. [21], the conditional PDF $g_i(\Delta\|f_i)$ can be approximated with a normal distribution centred on a bias-corrected forecast and having a variance σ^2 as

$$\Delta\|f_i \sim N(a_i + b_i f_i, \sigma^2),$$

where a_i and b_i are bias correction terms that are derived from linear regression between measurement and prediction for the training (calibration) period. Vrugt et al. [25] showed that the use of the intuitively more appropriate gamma distribution did not improve the performance of the BMA method. They did, however, argue that the linear bias correction is too simplistic for streamflow data, and suggested instead to consider a heteroscedastic error with $\sigma^2 = b \cdot f_i$, where b is a slope parameter that relates the forecast to the variance.

The BMA approach requires the specification of the model weights, the variance and the slope parameter when a heteroscedastic error is assumed. Following Vrugt et al. [25], we estimate these parameters by posing an optimization problem in a maximum likelihood context. A global optimization algorithm is used to maximize the log-likelihood function

$$l(\theta) = \sum_t \log \sum_{i=1}^{n} w_i g_i(\Delta_t\|f_{it}),$$

where θ denotes the parameters to be estimated and the summing is over all observation times t in the training period. After optimization, the deterministic BMA ensemble prediction is simply given by

$$E(\Delta\|f_1,\ldots f_n) = \sum_{i=1}^{n} w_i f_i.$$

In this study, we consider four BMA analyses. The first two consist of the original approach of Raftery et al. [21] with and without bias correction and the other two consider a heteroscedastic error with and without bias correction.

Other ensembles may be constructed by weighting the relative contributions of the models or by using subsets of the 10 models. For example, one of the trimmed mean ensembles is calculated by eliminating the highest and lowest model prediction each day and calculating a mean ensemble prediction from the predicted flows of the remaining eight models. Similar trimmed ensembles may be constructed from the six or four centremost predictions each day. We would expect the predictions of the three trimmed ensembles to lie between those of the mean and median, with the eight-model trimmed mean being nearest to the mean and the four-member trimmed mean being nearest to the median. The mean and median ensembles are equivalent to a 10-model trimmed mean and a two-model trimmed mean, respectively.

In the weighted ensembles, the weightings for each model must be calculated from some statistical property of the model's calibration predictions. A suitable statistic to use is the inverse of the mean square error. This is equivalent to weighting by $1/(1-E)$, where E is the model's calibration efficiency [17]. If we also introduce an optional exponent, k, then the weighting for the ith model is given by

$$W_i(n,k) = \frac{1}{(1-E_i)^k} \Bigg/ \left(\sum_{i=1}^{n} \frac{1}{(1-E_i)^k}\right),$$

where n is the number of models in the ensemble. In this way (provided k is positive), the models with the largest calibration efficiencies make the largest contributions to the weighted ensemble in both calibration and validation periods. The case where $k = 1$ is equivalent to weighting by $1/(1-E)$, while the case where $k = 0$ represents the unweighted mean. The limit as k tends to infinity is equivalent to using the best SME. Thus, the rationale for allowing different values of k is that they allow us to examine ensembles with differeing degrees of relative dominance of the better performing models.

With each unweighted or weighted mean having up to 10 potential members, there is a total of 1023 permutations of MME, ranging from the 10 one-model MMEs (i.e. the 10 SMEs) to the single 10-model MME. For example, there are 252 possible combinations of five-member MMEs. In this paper, we assess all possible permutations.

It is recognized that some models may perform better on some parts of the hydrograph than on others. For example, a model may provide better predictions of summer flows than of winter flows. This is clearly a possibility given that some of the models include snow accumulation and melting routines while others do not. This raises the possibility that better overall predictions might be obtained by using different combinations of models (weighted or not) for different parts of the hydrograph. This approach we refer to as conditional ensembling. Three conditional ensembles are assessed here. As well as assessing the use of separate summer (defined here as May–October) and winter (November–April) MMEs, we also consider separate MMEs for rising and falling flows, and for high and low flows.

For each conditional ensemble it is a requirement that the separate MMEs be constructed using data from the calibration period only. For application to the validation period, we, therefore, need a way of discriminating between the two cases. Furthermore, this discrimination must be based on the SME predictions only, not the observed hydrographs. After some preliminary analysis, it has been found that defining days with rising flows as days on which at least five of the SME predictions are rising is the best indicator of rising flows in the observed record. Days with fewer than five rising SMEs are classified as receding. This includes days with static or zero flows. This scheme has a prediction success rate of 83% during the calibration period. According to this classification criterion, rising flows occur on 28% of days and carry 41% of the total flow. The demarcation for high and low flows is taken as days with the mean MME prediction above and below 1 mm. Mean MME predictions in excess of 1 mm occur on 37% of days and account for 78% of the total flow. Of course, given that most of the flow, and therefore most of the large flow events, occur in winter, it is possible that the flow-dependent conditional ensemble will yield similar results to the seasonal ensemble.

With the emphasis in the LUCHEM project being on predicting the impacts of land use change, we do not consider assessing some of the sequential data assimilation techniques (e.g. the ensemble Kalman filter) that have recently been applied with some success to operational or experimental forecasting ensembles (e.g., [13,25]).

There are many potential metrics of prediction quality. In this study, we restrict ourselves to assessing model performance using the Nash–Sutcliffe efficiency calculated on daily flows, and the total bias. The former gives a measure of how well a model reproduces the fine-scale aspects of the observed time series (albeit biased towards high flows), while the latter assesses the long-term prediction quality.

5. Model predictions

5.1. Predictions of individual models

Breuer et al. [5] report on the calibration and validation performances of the individual models. A brief recap of their results is given here.

A time series of model predictions is shown in Fig. 1 for part of the calibration period. Qualitatively, the models are shown to be providing good predictions of the observed streamflow in terms of timing and magnitude of events. The envelope defined by the range of model predictions encompasses the observed streamflow on 96% of days, with little difference between calibration and validation periods. When this envelope is trimmed to eliminate the largest and smallest prediction, it still includes the observed streamflow on 83% of days. On average, the daily minimum prediction is 52% of the observed flow and the daily maximum prediction is 171% of the observed flow.

Scatter plots for two of the performance statistics (daily Nash–Sutcliffe efficiency and bias) for each of the models are shown in Fig. 2 for the original case. Statistically, the best models are those with efficiencies approaching 1.0 and biases near 0%. For the calibration period (open circles), all but two of the models have negative biases (i.e. they underpredict). However, no model has an absolute bias as high as 10%. The calibration efficiencies range from about 0.6 to 0.9, with the less complex models tending to have higher values.

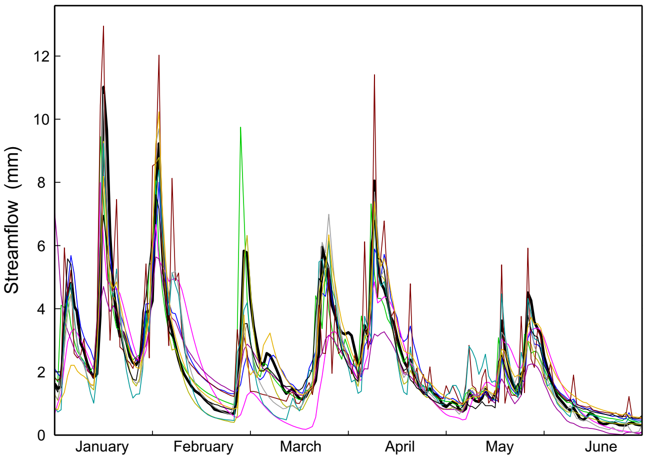

Fig. 1. Time series of observed streamflow (thick black line) in the Dill catchment, 1983, together with the various model predictions for the homogeneous dataset (thin coloured lines). The names of the individual models are immaterial.

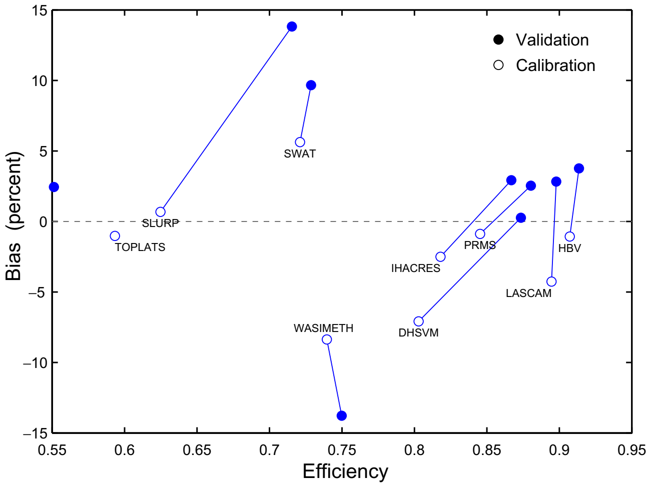

Fig. 2. Bias and efficiency of model predictions for the original dataset for the calibration (open circles) and validation (solid circles) periods. There are no predictions of the MIKE-SHE model available for the original dataset.

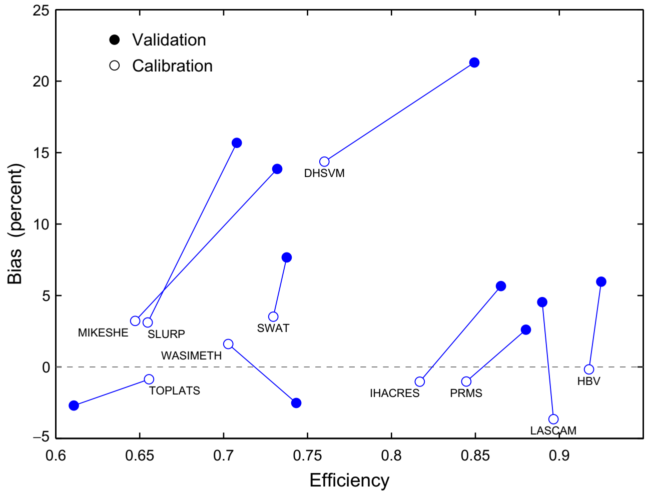

Fig. 3. As for Fig. 2, but for the homogeneous dataset.

When the predictions in the validation period are assessed (solid circles in Fig. 2), the relative positions of the models remain largely unchanged. However, nine of the 10 models have increased (i.e. more positive) biases, to the extent that they are all now overpredicting. All but one of the models also have increased efficiencies in the validation period.

For some models the differences in calibration statistics between the original and homogeneous datasets are quite small, with the latter typically yielding slightly better efficiencies and slightly more positive bias. However, for other models, especially DHSVM and WASIM-ETH, efficiencies decline and bias increases substantially. When the homogenized calibrations are used in the validation period (Fig. 3), once again, most models have increased biases and, except for LASCAM and TOPLATS, also have increased efficiencies. The trajectories of movement between calibration and validation are similar to those in Fig. 2.

5.2. Single-model ensembles

The bias and efficiency of the SMEs are compared to those of the individual calibrations in Fig. 4 and Table 1. By definition, the bias of a SME constructed from the mean of its members is equal to the

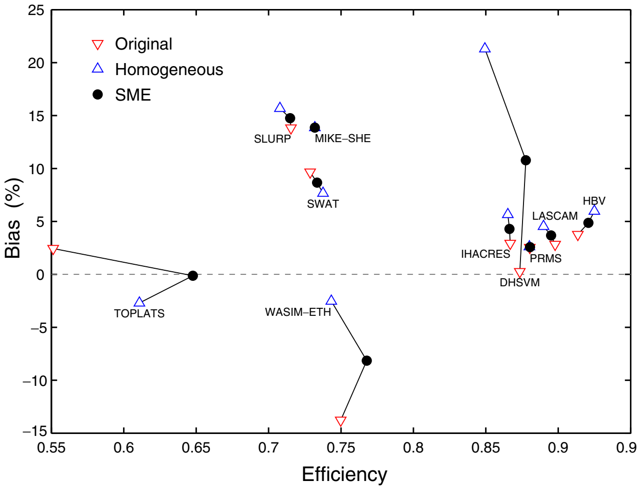

Fig. 4. Bias and efficiency of single-model ensembles for the validation period, compared with the corresponding statistics for the original and homogeneous datasets.

Table 1
Calibration and validation efficiencies for the original and homogeneous data sets and the SMEs

Model	Calibration			Validation		
	Orig.	Homog.	SME	Orig.	Homog.	SME
DHSVM	0.803	0.760	0.803	0.873	0.849	0.878
MIKE-SHE	–	0.647	0.647	–	0.732	0.732
TOPLATS	0.593	0.656	0.670	0.551	0.611	0.648
WASIM	0.740	0.703	0.742	0.750	0.743	0.768
SWAT	0.721	0.730	0.726	0.729	0.738	0.733
PRMS	0.845	0.845	0.845	0.880	0.880	0.880
SLURP	0.625	0.655	0.644	0.715	0.708	0.715
HBV	0.907	0.918	0.913	0.914	0.925	0.921
LASCAM	0.895	0.897	0.897	0.898	0.890	0.895
IHACRES	0.818	0.817	0.818	0.867	0.865	0.866

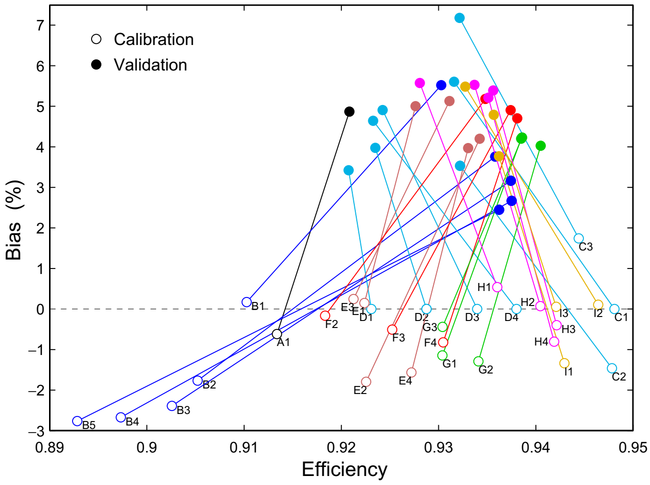

Fig. 5. Bias and efficiency of selected multi-model ensembles for the calibration (open circles) and validation (closed circles) periods, compared with the corresponding statistics for the best single-model ensemble. Labels for each ensemble are defined in Table 2.

mean of the biases of the two model realizations. However, in both calibration and validation periods, the efficiency of the SME is always greater than the mean efficiency of the two ensemble members. In fact, in three of the nine calibration cases and four of the nine validation cases, the ensemble efficiency exceeds those of both ensemble members. These include both time periods for TOPLATS and WASIM-ETH.

5.3. Multi-model ensembles using simple averaging

In this section, we use the predictions of the SMEs calculated above (along with the homogeneous predictions of MIKE-SHE) as members of several different types of simple multi-model ensembles that are based on raw averages.

The prediction capabilities of selected simple MMEs (labelled with the letter B) are shown in Fig. 5 and Table 2. The mean ensemble has the best efficiency in the calibration period and the median has the worst. However, in the validation periods, the median has the largest increase in efficiency of the simple ensembles and the smallest increase in bias. All five ensembles improve in efficiency in the validation period. The mean always has a larger (i.e. more positive) bias than the median for both calibration and validation periods.

Interestingly, although the raw mean is a simple ensemble and includes contributions from some models that have quite modest performance statistics, its validation efficiency is still superior to that of the best SME. Similarly, the trimmed mean and median ensembles, despite having poorer calibration statistics than the best SME are all clearly better than the best SME in the validation period.

Table 2
Bias and efficiency of selected MMEs for the calibration and validation periods

MME	Calibration		Validation	
	Bias (%)	Efficiency	Bias (%)	Efficiency
A1. Best SME	−0.6	0.913	4.9	0.921
B1. Mean	0.2	0.910	5.5	0.930
B2. Trimmed Mean (8 models)	−1.8	0.905	3.8	0.936
B3. Trimmed Mean (6 models)	−2.4	0.903	3.2	0.937
B4. Trimmed Mean (4 models)	−2.7	0.897	2.7	0.938
B5. Median	−2.8	0.893	2.4	0.936
C1. Linear regression (unconstrained)	0.0	0.948	5.6	0.932
C2. Linear regression (constrained)	−1.5	0.948	3.5	0.932
C3. Linear reg. (constrained, coeff>0)	1.7	0.944	7.2	0.932
D1. PCR (1 component)	0.0	0.923	3.4	0.921
D2. PCR (2 components)	0.0	0.929	4.0	0.923
D3. PCR (3 components)	0.0	0.934	4.9	0.924
D4. PCR (4 components)	0.0	0.938	4.6	0.923
E1. BMA (homoscedastic)	0.1	0.922	5.0	0.928
E2. BMA (homoscedastic, unbiased)	−1.8	0.923	4.2	0.934
E3. BMA (heteroscedastic)	0.2	0.921	5.1	0.931
E4. BMA (heteroscedastic, unbiased)	−1.6	0.927	4.0	0.933
F2. Weighted 10-model ($k = 0.5$)	−0.2	0.918	5.2	0.935
F3. Weighted 10-model ($k = 1.0$)	−0.5	0.925	4.9	0.937
F4. Weighted 10-model ($k = 1.5$)	−0.8	0.930	4.7	0.938
G1. Seasonal 10-model ($k = 1.5$)	−1.1	0.930	4.2	0.938
G2. Rising/falling 10-model ($k = 1.5$)	−1.3	0.934	4.0	0.941
G3. High/low 10-model ($k = 1.5$)	−0.4	0.930	4.2	0.939
H1. Best unweighted calibration	0.5	0.936	5.6	0.928
H2. Best weighted calib. ($k = 0.5$)	0.1	0.940	5.5	0.934
H3. Best weighted calib. ($k = 1.0$)	−0.4	0.942	5.4	0.936
H4. Best weighted calib. ($k = 1.5$)	−0.8	0.942	5.2	0.935
I1. Best seasonal ($k = 1.5$)	−1.3	0.943	3.8	0.936
I2. Best rising/falling ($k = 1.5$)	0.1	0.946	5.5	0.933
I3. Best high/low flow ($k = 1.5$)	0.1	0.942	4.8	0.936

5.4. Regression-based ensembles

The multiple linear regression ensemble (C1 in Fig. 5 and Table 2) has the best calibration efficiency of any model tested in this study, but unlike the simple averaging ensembles, its performance degrades significantly between the calibration and validation periods (in both bias and efficiency). Its validation efficiency is inferior to those of the median and trimmed mean ensembles, and its validation bias is larger than any other ensemble in Fig. 5 and Table 2.

In the multiple linear regression ensemble, there is considerable multicollinearity between the independent variables. Cross-correlation coefficients of up to 0.96 are observed between some pairs of SMEs. Furthermore, three of the SMEs have negative coefficients in the linear regression ensemble, while two others have negative coefficients that are not significantly different to zero. When the intercept is constrained to zero (C2 in Fig. 5 and Table 2) efficiencies are identical to the unconstrained case, but biases are more negative, for both calibration and validation. When we further limit the regression by eliminating SMEs with negative or non-significant coefficients (C3 in Fig. 5 and Table 2) the efficiency in validation is depressed slightly, but in validation it is unchanged. However, biases are significantly more positive than any other ensemble in Fig. 5 and Table 2.

Principal components regression can be used to overcome some of the problems with linear regression methods. PCR ensembles are constructed for $p = 1, \ldots, 4$, where p is the number of principal components used. The resulting ensemble statistics are presented in Fig. 5 and Table 2 (ensembles D1–D4). There is an increase in calibration efficiency as more components are added, but little improvement in validation efficiencies. As was the case for the linear regression MME, the PCR ensembles all have efficiencies that decline in the validation period. In both calibration and validation, the PCR efficiencies are less than those of the linear regression MME.

5.5. Bayesian model averaging ensembles

Results of the BMA analyses are presented in Fig. 5 and Table 2 (ensembles E1–E4). There is little difference between the performances of the two error models. However, the use of the linear bias correction slightly improves efficiency in both error models and also leads to lower (more negative) prediction biases. Unlike the linear regression and PCR methods, the BMA ensembles yield efficiencies that are greater in the validation period than in the calibration period.

5.6. Unweighted and weighted selective ensembles

Fig. 6 shows the ranges of calibration and validation efficiencies for the various permutations of unweighted mean MMEs and provides further demonstration of the effectiveness of ensembling. For both calibration and validation, there is a clear trend towards increased median efficiencies and decreased ranges of efficiency as the number of models in a MME increases. In both cases, the maximum efficiency occurs for $n = 5$, but this occurs for different combinations of models. All MMEs with three or more members ($n \geqslant 3$) have efficiencies greater than the median of the respective SMEs ($n = 1$). In calibration, 49% of ensembles with $n > 1$ have efficiencies that exceed the highest efficiency of their constituent SMEs, while in validation, this number increases to 75%. For all values of n, the validation efficiencies exceed those of calibration.

Next, we investigate which combinations of models are likely to produce the greatest increases in prediction efficiency. For each of the permutations with $n > 1$, we can calculate the variance of efficiencies of the constituent SMEs. We then divide the population of MMEs into three groups: those with the lowest third of variances (i.e. those MMEs whose members have the most similar efficiencies), those with the highest third of variances (i.e. those whose members have widely differing efficiencies) and those with the middle third. In the calibration period, 85% of the ensembles with the lowest variances have ensemble efficiencies that exceed the maximum efficiency of the constituent SMEs (not shown). In contrast, only 21% of ensembles with the highest variances improve on the best SME.

Weighted ensembles ($k > 0$) generally provide better prediction efficiencies in calibration and validation than the unweighted ensembles ($k = 0$). Analysis using a number of values of k indicates that a value of 1.5 provides the largest prediction efficiencies. As k increases beyond 1.5 efficiencies begin to decline. Fig. 7 compares validation efficiencies for three values of k. Within each ensemble class (i.e. each value of n) there is a general trend of increasing efficiencies and decreasing ranges as k increases from 0 to 1.5.

Adopting $k = 1.5$ as the optimal weighting parameter, we compare calibration and validation efficiencies for the 1023 individual MMEs in Fig. 8a. All but one of the MMEs with $n \geqslant 2$ have calibration and validation efficiencies that exceed the five worst SMEs. Indeed, there are relatively few MMEs with validation efficiencies poorer than the worst nine SMEs. There also appears to be a clear correlation ($r^2 = 0.75$) between calibration and validation efficiencies in Fig. 8a, with the models with large calibration efficiencies also tending to have large validation efficiencies. However, closer inspection of the upper right corner of Fig. 8a suggests that this observation does not hold throughout the range of efficiencies. In Fig. 8b, it emerges that none of the ensembles with the 78 best calibration efficiencies is among those with the best 78 validation efficiencies.

The 10-model MME is only just in the upper quartile of calibration efficiencies (252nd of 1023), but is 80th best in validation. However, only 26 MMEs are better than the 10-model MME in both calibration and validation. In contrast, the best of the SMEs (HBV) ranks only 586th in calibration and 780th in validation. In other words, more than three-quarters of the MMEs have better validation efficiencies than the best individual model. These include all 175 models with seven or more members. The effectiveness of combining predictions from several models is also highlighted by the observation that only seven of 1013 ensembles with $n \geqslant 2$ have validation efficiencies that are less than any of their constituent SMEs.

There appear to be subtle synergies between various models in the MMEs. The best calibrated model (HBV) appears in all of the top 423 calibration MMEs; no other model appears in more than 256 of these, although all are in at least 196. However, the representation of SMEs in the best validation models is more revealing. DHSVM, which has only the fourth best validation efficiency of the

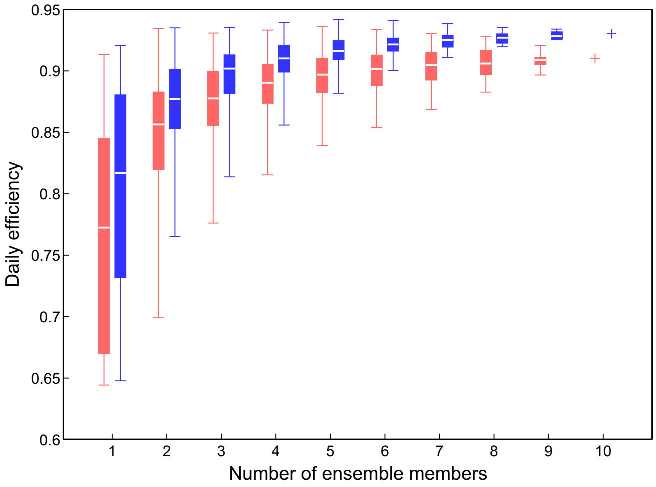

Fig. 6. Boxplots of the unweighted mean ensembles with differing numbers of members. For each value of n, the calibration period is shown on the left in red and the validation period on the right in blue. The data marks in the boxes are minimum, first, second and third quartiles, and maximum.

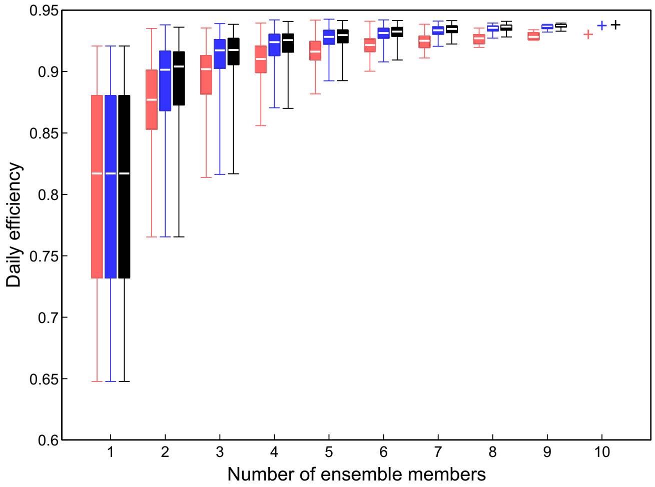

Fig. 7. Comparison of validation boxplots for different values of the weighting exponent, k. In each triplet of boxes the left (red) box is for $k = 0$, the centre (blue) box for $k = 1$ and the right (black) box for $k = 1.5$.

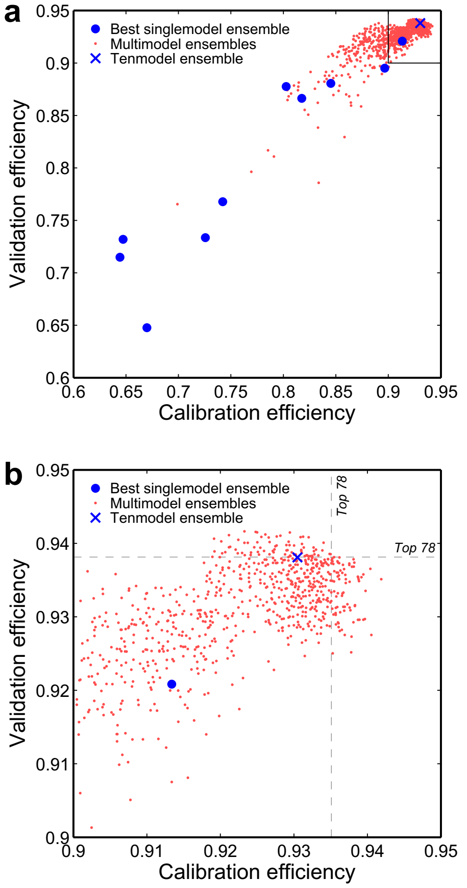

Fig. 8. Calibration and validation efficiency of (a) all permutations of weighted ($k = 1.5$) mean ensembles and (b) the models in the upper right part of (a).

10 SMEs (Table 1) and one of the largest biases, is a member of all of the best 134 validation MMEs. This is more than HBV, the best validation SME, a constituent of 133 of the best 134. On the other hand, LASCAM, which has the second best SME in the validation period and is close to the median in bias, is unrepresented in any of the best 49 MMEs. All other SMEs are represented at least 20 times. The addition of LASCAM to the top 96 validation models that do not already include it, universally results in a decrease in validation efficiency. Of the 10 nine-member ensembles, the one that doesn't include LASCAM has the highest validation efficiency. In contrast, the addition of either DHSVM or HBV to an ensemble that doesn't already include them always results in an increase in validation efficiency. In the case of WASIM-ETH, this is true for all but one ensemble.

Of course, the reality is that for any given value of *k*, there are really only two of the 1023 possible permutations that could justifiably be chosen for blind prediction on the basis of their calibration performance alone. These are the 10-member MME, which incorporates all available SMEs, and the permutation with the best calibration performance. As is evident in Fig. 8b, there are many candidate permutations with better validation efficiencies than these, but we have no way of objectively choosing them based on their calibration performance.

The calibration and validation measures for some of the choosable unweighted and weighted MMEs appear in Table 2 and Fig. 5, where they are labelled with the letters F and H. The unweighted 10-model MME is the same as the mean MME. For the 10-model MMEs, the sequence of model weightings from B1 through F2 and F3 to F4 clearly shows a trend of increasing efficiencies in both periods as the weighting factor, *k* increases. There is also a decrease in validation bias. For all values of *k*, the permutation with the best calibration efficiency is always a five-member ensemble comprising SWAT, DHSVM, HBV, TOPLATS and LASCAM. Whilst these MMEs have some of the highest calibration efficiencies in Fig. 5, their performance degrades in the validation period. In validation, these five-member ensembles always have slightly lower efficiencies and slightly larger biases than the corresponding 10-member MMEs.

5.7. Conditional ensembles

5.7.1. Ensembles for different seasons

HBV has the best calibration efficiencies in winter; LASCAM in summer. As well as LASCAM, IHACRES, DHSVM and SWAT all have higher calibration efficiencies in summer than in winter. For SWAT, MIKE-SHE and TOPLATS, the absolute differences between summer and winter calibration efficiencies exceed 0.1. These seasonal differences mean that combined seasonal MMEs made up of the best calibration models in summer and winter are likely to contain different constituents. They also mean that a weighted ensemble with the same members in summer and winter is likely to feature different effective weightings for the various constituent SMEs for each season.

In the calibration period, there is a general tendency towards slight underprediction in summer and slight overprediction in winter. In the validation period, streamflows in both seasons are overpredicted, more so in summer.

Analysis of seasonal models with various values of *k* again shows that $k = 1.5$ gives the best predictions in both periods. However, in comparison with the non-seasonal selective ensembles, the use of seasonal switching results in only minor improvements in efficiencies in both periods and small reductions in bias (more negative in calibration; less positive in validation). For example, comparing the 10-member seasonal ensemble (G1) with F4 in Fig. 5 shows little change in efficiencies. In calibration, the best combination of seasonal models is an ensemble consisting of six SMEs for summer and five SMEs for winter. The resulting MME is denoted as I1 in Fig. 5, but shows only modest improvement in overall efficiency over the corresponding non-seasonal MME (H4).

5.7.2. Ensembles for rising and receding flows

Among the SMEs in the calibration period, LASCAM has the best efficiency for rising flows and HBV has the best efficiency for receding flows. Only LASCAM and MIKE-SHE have better calibration efficiencies for rising flows than for receding flows. There are large differences (more than 0.1) between rising and receding efficiencies in the calibration period for SWAT, MIKE-SHE, PRMS and WASIM-ETH.

In calibration, there is a general tendency to slightly underpredict rising flows and slightly overpredict recessions. In validation, the tendency is towards more substantial overprediction of rising flows and a moderate overprediction of recession flows.

With a weighting exponent of $k = 1.5$, the 10-member MME based on rising and falling flow levels (model G2 in Fig. 5) has significantly better efficiencies for both calibration and validation than the corresponding non-conditional model (F4). Of all the ensembles tested in this study that could reasonably be chosen for blind prediction on the basis of their calibration performances alone, this model has the best validation efficiency. Indeed, its validation efficiency would place it in the top 15 models in Fig. 8b.

The best calibration ensemble based on rising and falling flows consists of four SMEs for the rising limb and five SMEs for the falling limb, and is denoted I2 in Table 2 and Fig. 5. Its calibration efficiency is very high and it has little calibration bias, but in the validation period its performance is poorer than the corresponding non-conditional MME (H4).

5.7.3. Ensembles for high and low flows

In calibration, there is a general tendency to slightly underpredict high flows and moderately overpredict low flows. In validation, both are moderately overpredicted. All the SMEs have substantially poorer calibration efficiencies for low flows than for high flows, with four having negative low flow efficiencies. HBV has the highest efficiencies for both high and low flows.

The 10-member ensemble (G3 in Table 2 and Fig. 5) has similar calibration and validation efficiencies to the corresponding non-conditional ensemble (F4), but is slightly less biased. The best ensemble in calibration (I3) has five members for high flows and eight for low flows. It too, shows little improvement over the performance of the corresponding non-conditional model (in this case, H4).

The results shown for the flow-dependent ensembles use a mean MME prediction threshold of 1 mm to discriminate between high and low flow days. Tests using a variety of thresholds between 0.5 and 7.0 mm (which are exceeded on 61% and 1.5% of days, respectively) show that choice of threshold has a negligible impact on prediction efficiency in the validation period.

5.7.4. Model weights

Weights for the ensembles that seek to find a time-independent combination of SMEs are listed in Table 3. In all cases weights ar determined solely from predictions in the calibration period. In terms of calibration efficiency, the best ensembles are the multiple linear regression ensembles, especially C1 and C2 (Table 2). They are characterized by large weights for the best-performed models, athough these are inflated slightly by the negative weights for some models. Other well-performed calibration ensembles are those for the best weighted calibrations (ensembles H2–H4), where the ensembles are dominated by just two models (HBV and LASCAM). In validation, the best ensembles with constant weighting are F3 and F4, which contain non-negative contributions from all 10 SMEs. However, they too are dominated by HBV and LASCAM with only one other SME (PRMS) having weights that are consistently greater than average.

In contrast with these, the ensembles that place least weight on the best-performed models (e.g. B1, D1, D2 and F2) have considerably poorer efficiencies in calibration or validation or both. The most extreme weight for any model is for one of the BMAs (E4), where HBV provides 61 % of the ensemble with no other model contributing more than 16%. While this yields efficiencies that are greater than A1 (for which the effective HBV weighting is 100%), its efficiencies are less than those of many of the ensembles that weight the models slightly more equitably (e.g. F3, F4).

Finally, it is interesting to note that the five models with non-negative linear regression coefficients (C1–C3) are the same as the five models with the best calibration efficiencies (H1–H4). Those five do not include some models that contribute strongly to some of the other ensembles. In particular, PRMS has above average weights in all ensembles except the linear regressions, yet is not included in H1–H4.

6. Discussion

Prediction uncertainty arises from three sources: data uncertainty, model structural uncertainty and parameter uncertainty. Ensemble modelling can be used to reduce any of these uncertainties. In this study we do not explore parameter uncertainty (this is best done using a Monte Carlo approach in a single member ensemble). A multi-model ensemble approach generally helps reduce prediction uncertainty by sampling models with a range of structural uncertainties. Different models have different strengths and weaknesses. Some models will predict better than others in different parts of the hydrograph (e.g. baseflow or peak flows, summer or winter). In an ensemble, the deficiencies in one model may be masked by the strengths in others or even by a compensating weakness in another model. In the original calibrations in this study, each model uses the input data in different ways to construct precipitation, potential evaporation and vegetation density fields. In this way, the ensembles based on the original calibration encompass a wide range of input data and associated uncertainty. The use of the homogeneous data set is an attempt to isolate differences in model structural uncertainty by providing consistent input data for each model.

Thus, ensemble modelling provides an estimate of the most probable state of the system. In certain circumstances, particularly for single-model ensembles, it can also provide an estimate of the range of possible outcomes. For multi-model ensembles, this may be unreliable as it is dependent on the prediction accuracy of the ensemble members. Nonetheless, the observation here that 96% of observed daily flows fall within the envelope defined by the daily range of predictions, suggests that this envelope might be an approximate representation of the 95% confidence interval.

The calibration statistics (Table 1, Figs. 2 and 3) indicate that the semi-distributed conceptual models tend to provide the best fits to the calibration period. This is possibly related to the generally lar-

Table 3
Model weights for selected MMEs

MME	DH	MS	TO	WA	SW	PR	SL	HB	LA	IH
B1	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100
C1	0.215	−0.004	0.118	−0.044	0.156	−0.006	−0.148	0.485	0.312	−0.119
C2	0.221	−0.002	0.124	−0.040	0.157	−0.012	−0.129	0.482	0.281	−0.098
C3	0.143		0.136		0.100			0.461	0.179	
D1	0.108	0.101	0.123	0.116	0.111	0.113	0.081	0.121	0.121	0.101
D2	0.163	−0.011	0.109	0.014	0.128	0.171	0.072	0.159	0.133	0.153
D3	0.156	0.019	0.157	−0.056	0.082	0.182	0.125	0.370	0.063	−0.021
D4	0.139	−0.139	0.235	−0.053	0.055	0.161	0.094	0.371	0.075	−0.013
E1	0.060	0.005	0.002	0.033	0.209	0.192	0.000	0.331	0.108	0.058
E2	0.021	0.005	0.014	0.062	0.065	0.219	0.000	0.468	0.147	0.000
E3	0.081	0.005	0.000	0.020	0.160	0.227	0.000	0.386	0.025	0.096
E4	0.000	0.005	0.055	0.066	0.049	0.155	0.000	0.609	0.061	0.000
F2	0.100	0.074	0.077	0.087	0.084	0.112	0.074	0.150	0.137	0.104
F3	0.093	0.052	0.056	0.071	0.067	0.119	0.052	0.212	0.178	0.101
F4	0.082	0.034	0.038	0.055	0.050	0.118	0.034	0.282	0.216	0.092
H1	0.200		0.200		0.200			0.200	0.200	
H2	0.181		0.140		0.154			0.274	0.251	
H3	0.154		0.092		0.111			0.350	0.293	
H4	0.123		0.057		0.075			0.422	0.323	

ger numbers of optimizable parameters in this type of model as compared to the distributed models, which tend to have many parameters that must be prescribed a priori, but few optimisable parameters. The use of manual calibration for many of the distributed models may also compromise their calibration efficiencies. However, in the validation period, the prediction efficiencies of some of the distributed models tend to increase more than those of the semi-distributed models (Table 1, Figs. 2 and 3). A notable exception here is that the most lumped model also increases efficiency quite significantly between calibration and validation.

All models except one show increased bias in the validation period, a period that is characterized by reduced rainfall and runoff coefficients. This perhaps highlights the potential problems associated with applying models in situations that are even only slightly different to the periods of calibration. It also has implications for the use of rainfall-runoff models in predicting the impacts of climate change. This is usually done by calibrating a model to current conditions then simulating using a modified or synthetic future climate input. If, as evidenced by the results of this study, there is a widespread tendency for models to overpredict in climates with reduced rainfall, then hydrologists may be routinely understating the hydrological impacts of changes towards a drier climate.

Figs. 2 and 3 and Table 1 indicate considerable variability in model calibration response between the two input data sets. Most prominent is the substantial increase in bias by DHSVM and WASIM-ETH for the homogenized data set, which includes nearest-neighbour interpolation of precipitation. These two distributed models both use inverse-distance interpolation in the original data set. On the other hand, LASCAM, which also uses inverse-distance interpolation originally, shows little change in bias. This is presumably due to LASCAM's semi-distributed lumping of precipitation input and its greater calibration flexibility. Nonetheless, the experiences of DHSVM and WASIM-ETH, together with TOPLATS, which shows increased efficiency, highlight the importance of uncertainties in model input for distributed models.

In validation, even the simplest of MMEs (the mean, trimmed mean and median ensembles) consistently outperform all the SMEs in terms of model efficiency. This happens despite the modest prediction statistics of some of their constituent models. This finding is in agreement with experiences in the atmospheric sciences and also with the outcomes of Georgakakos et al. [10]. The median ensemble and the closely related trimmed mean ensembles, although having the weakest predictions of the simple ensembles in calibration, consistently have the best validation statistics. That these ensembles outperform the mean in validation indicates the merits of trimming each day's extreme predictions. It is interesting to note that each model contributes directly to the median ensemble at least 9% of the time and that no model contributes more than 30% of the time. There is negligible difference in the various contributions between the calibration and validation periods. It is also interesting that the most frequent contributors are not the two models with the best prediction statistics (HBV and LASCAM), but PRMS (30%), DHSVM (26%), SWAT (25%) and TOPLATS (20%). Typically, these are the models with the mid-ranking overall biases across the combined calibration and validation periods (Figs. 2 and 3). The simple, unweighted ensembles have one advantage over the weighted ensembles in that they do not depend on an assessment of calibration performance and can therefore be applied in ungauged catchments.

The linear regression ensembles (especially those involving all 10 SMEs) are best in calibration, but their performances degrade noticeably in validation. This degradation is possibly associated with differences in model cross-correlations between the calibration and validation periods. When two models have highly correlated predictions there is greater scope for one of them to have a negative regression coefficient. If that correlation is altered in the validation period, there is potential for those negative contributions to the ensemble to behave in unfavourable ways. Removing the regression intercept has negligible impact on performance of the linear regression. Removal of the variables with negative coefficients reduces the calibration efficiencies slightly and increases prediction bias, but has little effect on validation efficiency.

The principal component regression ensembles remove the undesirable aspects of high cross-correlations in the independent variables. However, their validation predictions, although less biased, have lower efficiencies than the multiple linear regression ensembles. In fact, it appears that the PCR ensembles provide only marginal improvement over the predictions of the best SME.

Ajami et al. [1] advocate the inclusion of a bias removal step in regression-based ensembles. However, in the Dill case study, the mean of the model predictions differs from the mean of the observations in the calibration period by less than 0.2%. This means that any bias correction will have negligible impact on either the ensemble predictions or their performance statistics. The general observation from this study that regression-based combination techniques do not provide better validation predictions than a simple mean ensemble is broadly consistent with the conclusions of Shamseldin et al. [23] and Doblas-Reyes et al. [9]. Doblas-Reyes et al. [9] attributed this to the lack of robustness of the regression coefficients. In contrast, Ajami et al. [1] found that the use of regression-based ensembles does confer greater prediction quality, especially if biases are removed and stationarity is preserved between the calibration and validation periods. Their regression-based ensembles give relatively poorer predictions on the one catchment they tested with a significant decrease in mean streamflow between the calibration and validation periods. A change of similar magnitude prevails in the Dill case study and may perhaps explain the poorer performance of the multiple linear regression and PCR ensembles reported here.

Despite this, the bias correction employed in the BMA analysis showed modest improvements in efficiency and prediction bias during the validation period. However, there is little difference between the performances of the BMAs with different error models. It should be noted that the true strength of the BMA approach is more obvious for probabilistic ensemble analysis. This aspect will be explored in future work.

According to Figs. 6 and 7, the largest efficiencies in both calibration and validation are achieved for a combination of five models, regardless of weighting exponent. This is consistent with the conclusions of Georgakakos et al. [10] and Ajami et al. [1], who found that four to five ensemble members give optimal results. Georgakakos et al. [10] note that the optimal number of models is partly dependent on the nature of the ensemble members. Where the ensemble members are all of reasonable quality, the ensemble predictions are likely to show significant improvement over the predictions of the individual models. In contrast, where the quality of the ensemble members is variable, the ensemble is less likely to show significant improvement over the best of the member models and the use of weighted averaging is likely yield better predictions than an unweighted mean. This observation is borne out in this study, where first, there is a clear inverse relationship between constituent variance and the degree by which an ensemble improves upon its best constituent, and second, the proportion of all ensembles that are better than each constituent increases as the weighting exponent is increased from $k = 0$ to $k = 1.5$.

Table 2 indicates that in this study a weighting exponent of $k = 1.5$ provides better calibration and validation ensembles than the other exponents tested. A larger value of k places greater weight on the better performing models, and less weight on the weaker models. In contrast, a value of $k = 0$ weights each model uniformly. The results presented here clearly show that there are advantages in giving greater weight to the better models. Despite this, it is also evident from the way the minimum efficiencies in Figs. 6 and 7 increase steadily with increasing values of n that even the worst performing SMEs can add value to an ensemble prediction. This is further demonstrated by the presence of each SME in at least 9% of days in the median MME and in almost half of the best 423 unweighted calibration ensembles. In other words, every model brings to the ensembles some useful information that may not be adequately modelled by other SMEs.

Despite the fact that there can be substantial differences in the ranking of the SMEs for the various different states of the conditional ensembles, the use of conditional switching brings only slight improvements over the corresponding non-conditional ensembles. Of the three conditional ensembles tested, only the one that discriminates between rising and falling limbs yields appreciable gains in calibration and validation efficiencies. Ajami et al. [1] extended the seasonal conditional model to generate separate ensembles for each calendar month, but found little improvement in overall prediction quality. They attribute this lack of improvement to the likelihood that stationarity assumptions are more easily violated when multi-model techniques are applied monthly. Despite there being considerable variations in individual model performance in different months for the Dill dataset (not shown), it is likely that stationarity issues will also compromise a monthly switching ensemble here. As a consequence, this option has not been explored in this study. It is likely that better conditional ensembles could be achieved if the individual models had been calibrated separately for each conditional state. This approach has yielded good results with SMEs by combining differently parameterized model realizations in a probablistic framework [16]. The cost of this approach however is a substantially increased overhead in model calibration (and a large increase in the number of parameters required to describe the catchment) and for this reason it has not been adopted in this study.

In all the conditional and non-conditional selective ensemble approaches tested, the ensemble (or ensemble pair) that is identified as the best in the calibration period invariably degrades in performance in the validation period. In contrast, the corresponding 10-member ensembles invariably improve and their validation performance is better than that of the best calibration ensembles. The deterioration of the best calibration ensembles is probably due to the non-stationarity of the hydrological conditions in the Dill catchment.

Three aspects of the weighted ensemble predictions are unexpected and require closer scrutiny. First, the optimal five-member combinations found in this study do not exactly correspond to the five best SMEs in either period. In calibration, the best combination includes the seventh and eighth best SMEs, but not the third and fourth. In validation, it includes the sixth, but not the second. This implies that simply constructing an ensemble by choosing the best five SMEs is not necessarily the best option.

Second, that none of the best 78 calibration ensembles is among the best 78 in validation is unexpected. Even if the efficiencies were distributed randomly with no relationship between calibration and validation efficiencies, we would expect about six MMEs to be common to the top 78 in each period. The probability that there are none (given a random distribution) is just 0.0016. That there is an obvious correlation between calibration and validation efficiencies in Fig. 8a makes these results even more extraordinary.

Third, the second best validation SME (LASCAM) does not figure in any of the best 49 validation MMEs, yet the fourth best validation SME (DHSVM: a SME with a relatively large bias) appears in all of the best 134 validation ensembles. Furthermore, when LASCAM is added to the best validation MMEs that do not include it, the resulting ensemble efficiencies decline. This is not the behaviour in combination one would expect from a model that, in isolation, remains the second best in the validation period.

It is likely that these three anomalous observations have common causes. One cause is undoubtedly associated with the combination weights for $k = 1.5$. As one of the best SMEs in the calibration period, LASCAM has a large weighting. It contributes 22% of the mass to the 10-member ensemble (compared to 28% and 12% for the best and third best SMEs). However, it is one of only two SMEs whose validation efficiency is less than its calibration efficiency (Table 1). Although the decrease is quite small, the fact that most of the other models have increased efficiencies means that LASCAM is effectively over-weighted in the validation period. Given its performance in validation, a more realistic validation weighting would be 17%. In contrast, DHSVM, which has a big increase in efficiency between calibration and validation, is effectively under-weighted in the validation period. The impact of weighting, however, does not entirely explain these anomalies, since all three are also present in the unweighted ensembles with $k = 0$ (although to a lesser extent for the second and third anomalies). Part of the explanation may lie in the nonstationarity between the calibration and validation periods. However, it is not clear why, in the absence of weighting, this would affect some SMEs more than others, given that (except for WASIM-ETH) their biases increase by approximately the same amount. A third possible reason could relate to cross-correlations between the SMEs, with different models interacting differently in combination with others. Of particular note is the observation that the cross-correlations between each pair of SMEs are universally greater in the validation period than in the calibration period. These synergies could be tested further by assessment of ensemble performance on the gauged subcatchments of the Dill River or by application of the models to different basins with different climatic characteristics. These will be the subjects of future studies.

This study has identified the six-member trimmed mean as being one of the best combination methods for prediction of streamflows in the validation period. This ensemble, together with the unweighted mean ensemble, is used in the companion paper by Huisman et al. [12] to predict the impacts of land use change in the Dill catchment. Both are shown to produce consistent trends in the response of streamflow to land use change, and their predictions are of a similar magnitude and direction to those of an alternative probabilistic method. The consistency and coherency of these trends lead to increased confidence in the scenario predictions.

7. Conclusions

Ten models have been applied to the Dill catchment to predict streamflow. The general model performance is satisfactory during both calibration and validation periods. The semi-distributed models tend to perform best during both periods, but do not improve their fits during the less-demanding validation period as much as some of the distributed models that do not require as much calibration.

Single-model ensembles are constructed for nine of these models using two separate input data sets and parameter sets. They give prediction efficiencies that always exceed the mean efficiencies of the individual realizations, and in some cases, exceed the best efficiency of the individual realizations.

These single-model ensembles are then used to construct multi-model ensembles using a variety of combination techniques. The study has confirmed the potential for multi-model ensembles to provide hydrological predictions whose accuracy exceeds those of individual models. Of the simple averaging ensembles tested, the trimmed mean ensembles, which includes the daily model predictions of the central four or six models are superior to the mean and regression-based ensembles during the validation period. The regression-based combination techniques (multiple linear regression and principal components regression) give good predictions in the calibration period, but their performances degrade considerably in the validation period.

Of the weighted ensembles, the best 10-member validation predictions are obtained from an ensemble that ascribes relatively large weights to the best performing calibration models. The results of this study also confirm that even the weakest of the individual models brings useful information to any ensemble it is included in and will usually improve the predictions. The use of conditional switching between summer and winter ensembles and between high and low flows yield little improvement in overall prediction quality. However, a conditional ensemble that discriminates between rising and receding flows shows moderate improvement.

For each of the multi-model combination techniques, the best performing ensemble – usually one containing about five members – is not neccessarily the one that contains the best five individual models. Furthermore, with considerable non-stationarity in climate prevalent during the study period, the best ensembles in the calibration period are typically not the best in the validation period. Some models that predict well individually in the validation period do not combine well with other models in the multi-model ensembles, while other models with more modest predictions appear to revel in combination. The reasons for these anomalies appear to relate to the weighting schemes, the non-stationarity of the climate series and cross-correlations between models, but further work is required to pinpoint the exact reasons behind these observations.

This study has used a larger number of models, and with a larger range in complexity, than have been reported in other hydrological ensemble studies. However, one limitation is that it has been applied to only one catchment. It would be quite instructive to explore whether the results obtained for the Dill catchment apply generally for the same models in other catchments with different hydrological regimes.

Acknowledgements

This study has been supported by the German Science Foundation within the scope of the Collaborative Research Centre (SFB) 299. The authors thank Lu Zhang and Santosh Aryal for their comments on the manuscript.

References

[1] Ajami NK, Duan Q, Gao X, Sorooshian S. Multimodel combination techniques for analysis of hydrological simulations: application to Distributed Model Intercomparison Project results. J Hydrometeorol 2006;7:755–68.

[2] Arnold JG, Srinivasan R, Muttiah RS, Williams JR. Large area hydrologic modeling and assessment. Part I: Model development. J Am Water Resour Assoc 1998;34:73–88.

[3] Bergström S. The HBV Model. In: Singh VP, editor. Computer models of watershed hydrology. Highland Ranch, Colorado, CO, USA: Water Resources Publications; 1995. p. 443–76.

[4] Bormann H, Breuer L, Gräff T, Huisman JA, Croke B. Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) IV: Model sensitivity on data aggregation and spatial (re-)distribution. Adv Water Res 2008; this issue.

[5] Breuer L, Huisman JA, Willems P, Bormann H, Bronstert A, Croke BFW, et al. Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) I: Model intercomparison of current land use. Adv Water Resour 2008; this issue.

[6] Carpenter TM, Georgakakos KP. Intercomparison of lumped versus distributed hydrologic modelling ensemble simulations on operational forecast scales. J Hydrol 2006;329:174–85.

[7] Clemen RT. Combining forecasts: a review and annotated bibliography. Int J Forecast 1989;5:559–83.

[8] Croke BFW, Jakeman AJ. A catchment moisture deficit module for the IHACRES rainfall-runoff model. Environ Mod Software 2004;19:1–5.

[9] Doblas-Reyes FJ, Hagedorn R, Palmer TN. The rationale behind the success of multi-model ensembles in seasonal forecasting – II. Calibration and combination. Tellus 2005;57A:234–52.

[10] Georgakakos KP, Seo D, Gupta H, Schaake J, Butts MB. Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. J Hydrol 2004;298:222–41.

[11] Gourley JJ, Vieux BE. A method for identifying sources of model uncertainty in rainfall-runoff simulations. J Hydrol 2006;327:68–80.

[12] Huisman JA, Breuer L, Bormann H, Bronstert A, Croke BFW, Frede H, et al. Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) III: Scenario analysis. Adv Water Resour 2008; this issue.

[13] Kim Y, Jeong D, Ko IH. Combining rainfall-runoff model outputs for improving ensemble streamflow prediction. J Hydrol Eng 2006;11:578–88.

[14] Kite GW. The SLURP model. In: Singh VP, editor. Computer models of watershed hydrology. Highland Ranch, Colorado, CO, USA: Water Resources Publications; 1995. p. 521–62.

[15] Leavesley GH, Stannard LG. The precipitation runoff modeling system – PRMS. In: Singh VP, editor. Computer models of watershed hydrology. Highland Ranch, Colorado, CO, USA: Water Resources Publications; 1995. p. 281–310.

[16] Marshall L, Nott D, Sharma A. Towards dynamic catchment modelling: a Bayesian hierarchical mixtures of experts framework. Hydrol Proc 2007;21:847–61.

[17] Nash JE, Sutcliffe JV. River flow forecasting through conceptual models. I. A discussion of principles. J Hydrol 1970;10:282–90.

[18] Niehoff D, Fritsch U, Bronstert A. Land-use impacts on storm-runoff generation: scenarios of land-use change and simulation of hydrological response in a meso-scale catchment in SW-Germany. J Hydrol 2002;267:80–93.

[19] Perrin C, Michel C, Andréassian V. Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. J Hydrol 2001;242:275–301.

[20] Peters-Lidard CD, Zion MS, Wood EF. A soil-vegetation-atmosphere transfer scheme for modeling spatially variable water and energy balance processes. J Geophys Res 1997;102:4303–24.

[21] Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. Using Bayesian model averaging to calibrate forecast ensembles. Mon Wea Rev 2005;133:1155–74.

[22] Refsgaard JC, Storm B. MIKE SHE. In: Singh VP, editor. Computer models of watershed hydrology. Highland Ranch, Colorado, CO, USA: Water Resources Publications; 1995. p. 809–46.

[23] Shamseldin AY, O'Connor KM, Liang GC. Methods for combining the outputs of different rainfall-runoff models. J Hydrol 1997;197:203–29.

[24] Sivapalan M, Ruprecht JK, Viney NR. Water and salt balance modelling to predict the effects of land-use changes in forested catchments. 1. Small catchment water balance model. Hydrol Proc 1996;10:393–411.

[25] Vrugt JA, Robinson BA. Treatment of uncertainty using ensemble methods: comparison of sequential data assimilation and Bayesian model averaging. Water Resour Res 2007;43:W01411.

[26] Wigmosta MS, Vail LW, Lettenmaier DP. A distributed hydrology-vegetation model for complex terrain. Water Resour Res 1994;30:1665–79.

[27] Ye W, Bates BC, Viney NR, Sivapalan M, Jakeman AJ. Performance of conceptual rainfall-runoff models in low-yielding catchments. Water Resour Res 1997;33:153–66.

[28] Ziehmann. Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. Tellus 2000;52A:280–99.