Ensemble modelling of the hydrological impacts of land use change

[1]Viney, N.R., [2]B.F.W. Croke, [3]L. Breuer, [4]H. Bormann, [5]A. Bronstert, [3]H. Frede, [5]T. Gräff, [6]L. Hubrechts, [3]J.A. Huisman, [2]A.J. Jakeman, [7]G.W. Kite, [8]J. Lanini, [9]G. Leavesley, [8]D.P. Lettenmaier, [10]G. Lindström, [11]J. Seibert, [12]M. Sivapalan and [13]P. Willems

[1]CSIRO Land and Water, GPO Box 1666, Canberra, ACT 2600, Australia. E-Mail: neil.viney@csiro.au, [2]ICAM, SRES, The Australian National University, Canberra, ACT, Australia, [3]Justus-Liebig University, Giessen, Germany, [4]Carl von Ossietzsky University Oldenburg, Germany, [5]University of Potsdam, Germany, [6]Lisec NV, Genk, Belgium, [7]Hydrologic Solutions, Pantymwyn, United Kingdom, [8]University of Washington, USA, [9]United States Geological Survey, Denver, USA, [10]Swedish Meteorological and Hydrological Institute, Norrköping, Sweden, [11]Stockholm University, Sweden, [12]University of Illinois, Urbana-Champaign, USA, [13]Katholieke Universiteit Leuven, Belgium

Keywords: *Ensemble modelling, hydrology, land use change*

EXTENDED ABSTRACT

Ensemble modelling, whereby predictions from several models are pooled in an attempt to improve prediction accuracy, has often been used in the climate and atmospheric sciences, but until recently, has received little attention in hydrology. One of the key aims of the ensemble approach is to reduce uncertainty in the modelled predictions. 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 produce ensemble predictions of the effects of several projected land use changes. 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.

Overall, the simpler models tend to perform better in both calibration and validation, but while all models tend to show improved performance during the less-extreme validation period, this improvement is greatest for some of the more complex models. Despite the disparity in model performance, three ensemble predictions made up of various combinations of the 10 model predictions outperform all of the individual models. In calibration, the ensemble based on a multi-variable regression of all models provides the best predictions, but its prediction accuracy declines to a greater extent than all of the models in terms of both its bias and Nash-Sutcliffe efficency when used in the validation period. In the validation period, the best predictions are provided by an ensemble consisting of the daily median model predictions. The predictions of this median ensemble also improve more between calibration and validation than any of the other models, thus indicating its robustness.

The calibrated models are applied to three land use change scenarios. In the scenarios, the projected patterns of land use are based on assumed average field sizes of 0.5 ha, 1.5 ha and 5.0 ha, respectively. These contrast with a current average field size of about 0.7 ha. There is broad agreement among the models on the expected hydrological change (Figure 1). This, coupled with the validation success of the mean and median ensembles, suggests that we can predict with some confidence the direction and magnitude of streamflow changes associated with the three scenarios.

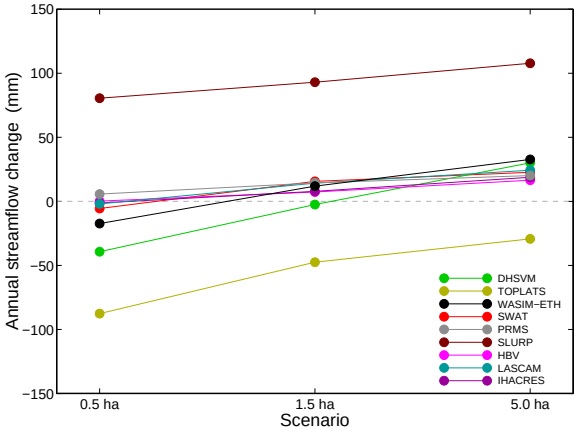

Figure 1. Changes in predicted mean annual streamflow for the three land use change scenarios relative to the current land use. Each line represents a different model. No predictions are available for the MIKE-SHE model for this data set.

1 INTRODUCTION

A model is a simplified conceptualisation of a complex, possibly chaotic system, which is often characterised by highly variable behaviour in space and time. As such, no model, particularly those associated with natural systems, can ever provide a perfect realisation. Indeed, it can even sometimes be difficult to quantify the degree of uncertainty in input data, model structure and model parameterisation. 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 realisations of a single deterministic model. Distinct predictions are obtained for each realisation 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 realisations 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 (Georgakakos *et al.*, 2004) and that multi-model ensembles tend to perform better than single-model ensembles (Ziehmann, 2000).

Ensemble modelling has, however, 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., Ye *et al.*, 1997, Perrin *et al.*, 2001). 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 DMIP and ESP in the United States and the international HEPEX project have begun to explore ensemble modelling in a hydrological setting. These projects are aimed primarily at producing short term streamflow forecasts conditioned on climate forecasts. In what appears to be the only published study on multi-member hydrologic ensembles, Georgakakos *et al.* (2004), as part of DMIP, assessed predictions from seven distributed models applied to six catchments. They found that a simple mean of the five best models in each catchment consistently outperformed the best individual model, but that a weighted mean ensemble, while usually better than the best model, was inferior to the simple mean ensemble.

To date, none of these projects have considered ensemble modelling of the hydrological impacts of land use change. This paper describes the application of ten catchment models to a basin with nested gauges.

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 characterised by low mountains and has an altitude range of 155–674 m. The mean annual precipitation of the Dill catchment varies fom 700 mm in the south to more than 1100 mm in the higher elevation areas in the north and 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 of winter precipitation falls as snow, particularly at higher elevations.

Just over half of the catchment is forested (with nearly even proportions of deciduous and coniferous species), while 21 % is pasture, 9 % is fallow, 6 % is cropped (winter rape, winter barley, oats) and the remaining 10 % is either urban or water. However, the pattern of land use across the catchment is highly fragmented, with an average field size of less than 1 ha and few individual land use patches covering more than 1 km^2. Social, political and economic pressures are slowly transforming land uses away from cropping and leading to increasing pasture and forest cover. This transformation is also accompanied by increasing patch sizes.

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. Three subcatchments of the Dill catchment are gauged. These

are the Aar catchment (area 133 km^2; mean annual streamflow 328 mm), the Dietzhölze catchment (area 80 km^2; mean annual streamflow 551 mm) and the Obere-Dill catchment (area 62 km^2; mean annual streamflow 483 mm).

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 (Wigmosta *et al.*, 1994), MIKE-SHE (Refsgaard and Storm, 1995), TOPLATS (Peters-Lidard *et al.*, 1997), WASIM-ETH (Niehoff *et al.*, 2002), SWAT (Arnold *et al.*, 1998), PRMS (Leavesley and Stannard, 1995), SLURP (Kite, 1995), HBV (Bergström, 1995), LASCAM (Sivapalan *et al.*, 1996) and IHACRES (Jakeman *et al.*, 1990). 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 three years of additional weather data (1980–1982) was available for model spin-up.

4 MODEL PREDICTIONS

4.1 Predictions of individual models

A time series of model predictions is shown in Figure 2 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

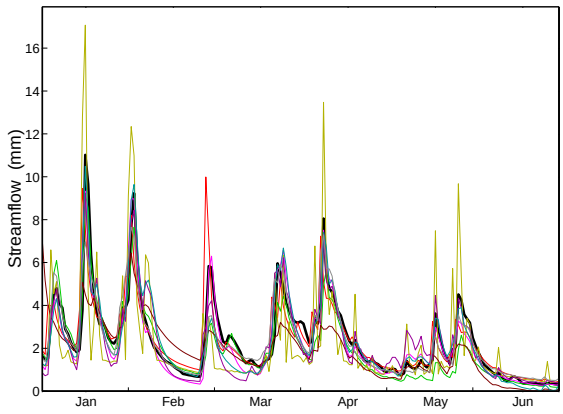

Figure 2. Time series of observed streamflow (thick black line) in the Dill catchment, 1983, together with the various model predictions (thin coloured lines). Refer to Figure 1 for legend.

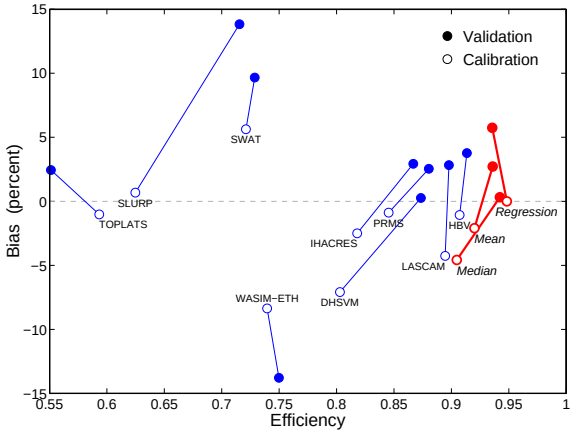

Figure 3. Bias and efficiency of model predictions (thin blue lines) and ensemble predictions (thick red lines) for the original dataset for the calibration (open circles) and validation (solid circles) periods. There are no predictions of the MIKE-SHE model available for the original dataset.

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.

Scatter plots for two of the performance statistics (daily Nash-Sutcliffe efficiency and bias) for each of the models are shown in Figure 3. 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 (that is, 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.

When the predictions in the validation period are as-

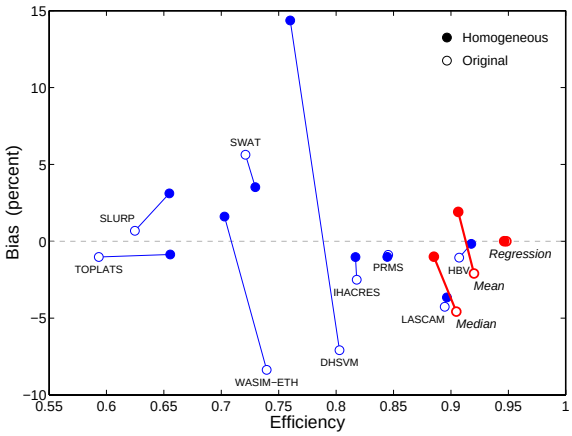

Figure 4. Bias and efficiency of model predictions (thin blue lines) and ensemble predictions (thick red lines) for the calibration period for the original (open circles) and homogeneous (solid circles) datasets.

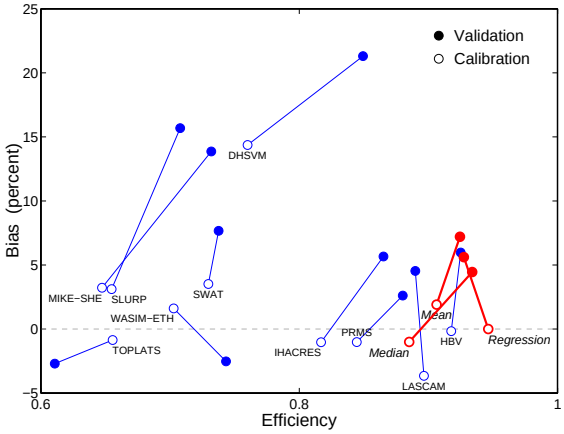

Figure 5. As for Figure 3, but for the homogeneous dataset.

sessed (solid circles in Figure 3), the relative positions of the models remain largely unchanged. However, all models (except WASIM-ETH) have increased biases, to the extent that they are all now overpredicting. All models (except TOPLATS) also have increased efficiencies in the validation period. These increases are particularly noticeable for IHACRES, PRMS and DHSVM, but less so for the two models with the largest calibration efficiencies (HBV and LASCAM).

In an attempt to highlight structural differences among the models, as opposed to differences in input preprocessing, a second calibration was performed for each model. This involved using common fields (at the 25 m scale) of catchment rainfall, potential evaporation and vegetation density. The recalibrations were intended to be performed with the same level of rigour as the original calibrations and are hereafter referred to as the homogeneous calibrations.

The recalibration statistics (solid circles) are shown in Figure 4 alongside those for the original calibration. For some models the differences are quite small and typically yield slightly better efficiencies and slightly more positive bias, but for others, especially DHSVM and WASIM-ETH, efficiencies decline and bias increases substantially.

When the homogenised calibrations are used in the validation period (Figure 5), 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 Figure 3.

The patterns of prediction statistics are similar for the other three flow gauges (not shown). In all three cases, the models, on average, tend to underpredict in the calibration period and overpredict in the validation period. The increases in bias are greatest for the Obere-Dill catchment. When averaged over all four catchments and over both datasets (Table 1), HBV, TOPLATS and PRMS are the least biased in calibration and TOPLATS and PRMS are the least biased in validation. These latter two, together with WASIM-ETH, show the smallest changes in bias between calibration and validation.

Table 1. Model statistics averaged over all four subcatchments and for both datatsets. The unit of MSE change is percent. Note that the statistics for MIKE-SHE are for the homogenised data set only.

	Bias (%)		Efficiency		MSE
Model	Cal.	Val.	Cal.	Val.	Change
DHSVM	5.0	14.7	0.73	0.83	−33.2
MIKE-SHE	8.9	24.5	0.57	0.57	−1.8
TOPLATS	−2.9	1.0	0.64	0.62	4.9
WASIM	−4.8	−7.6	0.65	0.72	−17.6
SWAT	5.1	11.9	0.70	0.72	−5.5
PRMS	−3.1	2.8	0.82	0.87	−25.0
SLURP	5.1	21.2	0.60	0.68	−18.9
HBV	1.4	10.0	0.89	0.89	−5.5
LASCAM	−3.3	7.6	0.87	0.88	−4.3
IHACRES	−4.5	4.1	0.78	0.83	−23.6
Mean	0.1	8.1	0.88	0.91	−22.7
Median	−3.6	4.7	0.85	0.91	−40.7
Regression	0.0	7.0	0.93	0.90	38.2

The ranking of model efficiencies is also similar to Dill, with HBV, LASCAM and PRMS having the best efficiencies for both calibration and validation, and for both datasets (Table 1). The models also tend to show increased efficiencies (decreased mean square errors) in the validation period for all three catchments. When averaged over the four catchments and over both datasets, the models with the greatest proportional decrease in mean square error (MSE) are DHSVM, PRMS and IHACRES. The models with the least decreases are TOPLATS (which actually increases MSE) and MIKE-SHE, while LASCAM, SWAT and HBV all have small decreases.

4.2 Ensemble predictions

Ensemble predictions may be constructed in a number of ways. Perhaps the simplest 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. A third method that is also adopted here is to construct a multi-variable linear regression during the calibration period and to apply that regression during the validation period. One disadvantage of this approach is that it might include a non-zero intercept. It might also involve negative coefficients for some models and can potentially result in negative flow predictions. 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 regression ensemble represents the optimal linear combination of model predictions during the calibration phase and it should therefore give better efficiencies than both the raw mean, which is an inferior linear combination, 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.

The prediction capabilities of the three ensembles are shown in Figures 3 and 5. For both datasets the regression ensemble has the best efficiency in the calibration period and the median has the worst. However, in the validation periods, the median is the best ensemble in terms of both efficiency and bias. These patterns are similar for the other three catchments (not shown). The performance of the regression model degrades universally between calibration and validation periods (in both bias and efficiency), while the mean and median always improve in efficiency. The mean always has a larger (i.e., more positive) bias than the median for both calibration and validation periods. When averaged over all four catchments and both datasets (Table 1), the regression ensemble increases MSE between calibration and validation by 38.2 %, while the mean and median ensembles decrease MSE by 22.7 % and 40.7 %, respectively.

Interestingly, although the raw mean is a simple ensemble and includes contributions from some models that have quite modest performance statistics, it generally gives validation predictions that have better or at least similar efficiencies to the best of the individual models. HBV and LASCAM sometimes have calibration efficiencies that are better than the mean and median ensembles. However, there is only one instance—the homogeneous dataset in Obere-Dill catchment (where HBV has the best efficiency)—in which any individual model has a validation efficiency that exceeds those of any of the ensembles. This is also the only case in which all ten models have positive bias in the validation period.

4.3 Predictions of the impacts land use change

Three scenarios of land use change are considered and are based on land use simulations involving different field sizes. The three scenarios are based on predicted land uses associated with average field sizes of 0.5 ha, 1.5 ha and 5.0 ha. In general, increasing field sizes are associated with decreasing areas of forested land and increasing areas of cropland.

The impacts of land use change are assessed by running each of the models calibrated for current land uses with the changed land use scenarios, using weather input from the period 1983–1998. The resulting annual streamflow predictions for the Dill catchment appear in Figure 1. The slopes of the lines in Figure 1 are quite similar, indicating that all the models are in broad agreement about the the relative increases in predicted streamflow as field size increases. However, some models have predictions that are substantially offset either above (SLURP) or below (TOPLATS) the main cohort. For the models in Figure 1, the mean changes in mean annual streamflow, relative to the baseline case, for each of the three scenarios are -7 mm, 13 mm and 27 mm, respectively. For the third scenario, this represents an increase of about 6 %. In percentage terms, the increases tend to be greater in summer (11 % for the third scenario) than winter (5 %), but the patterns of change among models and among scenarios for summer and winter (not shown) are similar to those for the entire year.

There is substantial spatial variability within the catchment in the predicted streamflow changes (not shown). For the three internal subcatchments, predicted increases in streamflow are greatest in the two wettest subcatchments. For Dietzhölze, the mean predicted increase for the third scenario is 93 mm, (17 % of mean annual flow) and in Obere-Dill it is 86 mm (18 % of mean flow). In contrast, the mean increase in Aar subcatchment is just 10 mm, or 3 %.

5 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 used 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.

Ensemble modelling thus provides an estimate of the most probable state of the system. In certain circumstances, particulary 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, Figure 4) indicate that the semi-distributed conceptual models (especially HBV and LASCAM) tend to provide the best fits to the calibration period. This is possibly related to the generally larger numbers of optimisable 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 (most notably DHSVM) tend to increase more than those of the semi-distributed models (Table 1, Figures 3 and 5). A notable exception here is that the most lumped model, IHACRES, also increases efficiency quite significantly between calibration and validation.

All models except WASIM-ETH show increased bias in the validation period, a period that is characterised by reduced 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.

Figure 4 indicates 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 homogenised data set, which included nearest-neighbour interpolation of precipitation. These two distributed models both used inverse-distance interpolation in the original data set. On the other hand, LASCAM, which also used 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.

All three simple ensembles consistently outperform all models in terms of model efficiency. This happens despite the modest prediction statistics of some of the models. This finding is in agreement with experiences in the atmospheric sciences and also with the outcomes of Georgakakos *et al.* (2004).

Among the three ensembles, the regression-based ensemble is consistently best in calibration, but its performance degrades 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 reduced in the validation period, there is potential for those negative contributions to the ensemble to behave in unfavourable ways. The median ensemble, although having the weakest predictions of the three in calibration, consistently has the best validation statistics. It is not clear why it should outperform the mean in validation. It is interesting to note that each model contributes directly to the Dill catchment median ensemble at least 4 % of the time and that no model contributes more than 19 % of the time.

Many other, more sophisticated ensembles can be readily envisaged. An obvious example would be a weighted mean ensemble with weights dependent on calibration statistics (e.g., efficiency), so that the stronger models have a greater impact on the ensemble. Such an ensemble was not tested here, but we might speculate that while it would almost certainly have poorer calibration statistics than the regression ensemble, it might provide better validation statistics than either the regression or mean ensembles. Other weighted ensembles might involve selecting and weighting ensemble members differently for different characteristics of the hydrograph (for example, summer and winter, high flows and low flows, rising limbs and falling limbs, presence or absence of snow). Testing of such ensembles remains for a more detailed study than this one.

In predicting the impacts of land use change, there is strong agreement between the models on the relative streamflow changes associated with each scenario (Figure 1). Despite this, there is not uniform agreement on how far these flows should deviate from the baseline (i.e., current) case. However, the two main outliers are the models with the lowest efficiencies.

The various gauged subcatchments of the Dill include a range of land uses, with some land use types more

prevalent in some parts of the catchment than in others. Thus we would expect those models that demonstrate superior predictions across all subcatchments to provide the most reliable predictions of the impacts of land use change. This of course assumes that all predictions are made without recalibration for each subcatchment. It is encouraging therefore, that apart from the two outliers in Figure 1, the predictions of the remaining models do not vary greatly. This, coupled with the validation success of the mean and median ensembles, suggest that we can conclude, with strong—but still unquantified—confidence that the streamflow changes associated with the three land use scenarios are approximately equal to the (virtually identical) mean and median of the model predictions: −7 mm, 13 mm and 27 mm, respectively.

6 CONCLUSIONS

Ten models have been applied to the Dill catchment to predict streamflow at four sites. 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.

The study has confirmed the potential for multi-model ensembles to provide hydrological predictions whose accuracy exceeds those of individual models. Of the three simple ensembles tested, the median ensemble, which includes the daily median model prediction is shown to be superior to the mean and regression based ensembles during the validation period.

The study has also demonstrated the advantages of a multi-model approach to predict the impacts of land use change. Although the predicted streamflow changes in this study are quite small, there is strong agreement among the models on the direction and magnitude of change for each scenario.

7 ACKNOWLEDGEMENTS

This study has been supported by the German Science Foundation within the scope of the Collaborative Research Centre (SFB) 299.

8 REFERENCES

Arnold, J.G., R. Srinivasan, R.S. Muttiah and J.R. Williams (1998), Large area hydrologic modeling and assessment. Part I: Model development, *Journal of the American Water Resources Association*, 34, 73–88.

Bergström, S. (1995), The HBV Model, In: V.P. Singh (ed.), *Computer models of watershed hydrology*, Water Resources Publications, Highland Ranch, Colorado, USA, 443–476.

Georgakakos, K.P., D. Seo, H. Gupta, J. Schaake and M.B. Butts (2004), Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, *Journal of Hydrology*, 298, 222–241.

Jakeman, A.J., I.G. Littlewood and P.G. Whitehead (1990), Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments, *Journal of Hydrology*, 117, 275–300.

Kite, G.W. (1995), The SLURP model, In: V.P. Singh (ed.), *Computer models of watershed hydrology*, Water Resources Publications, Highland Ranch, Colorado, USA, 521–562.

Leavesley, G.H. and L.G. Stannard (1995), The precipitation runoff modeling system — PRMS. In: V.P. Singh (ed.). *Computer models of watershed hydrology*, Water Resources Publications, Highland Ranch, Colorado, USA, 281–310.

Niehoff, D., U. Fritsch and A. Bronstert (2002), 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, *Journal of Hydrology*, 267, 80–93.

Perrin, C., C. Michel and V. Andréassian (2001), Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *Journal of Hydrology*, 242, 275–301.

Peters-Lidard, C.D., M.S. Zion and E.F. Wood (1997), A soil-vegetation-atmosphere transfer scheme for modeling spatially variable water and energy balance processes, *Journal of Geophysical Research*, 102, 4303–4324.

Refsgaard, J.C. and B. Storm (1995), MIKE SHE. In: V.P. Singh (ed.), *Computer Models of Watershed Hydrology*, Water Resources Publications, Highland Ranch, Colorado, USA, 809–846.

Sivapalan, M., J.K. Ruprecht and N.R. Viney (1996), Water and salt balance modelling to predict the effects of land-use changes in forested catchments. 1. Small catchment water balance model, *Hydrological Processes*, 10, 393–411.

Wigmosta, M.S., L.W. Vail and D.P. Lettenmaier (1994), A distributed hydrology-vegetation model for complex terrain. *Water Resources Research*, 30, 1665–1679.

Ye, W., B.C. Bates, A.J. Jakeman, N.R. Viney and M. Sivapalan, (1997), Performance of conceptual rainfall-runoff models in low-yielding catchments, *Water Resources Research*, 33, 153–166.

Ziehmann (2000), Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models, *Tellus*, 52A, 280–299.