Adv. Geosci., 29, 43–50, 2011
www.adv-geosci.net/29/43/2011/
doi:10.5194/adgeo-29-43-2011

Advances in Geosciences

Multi-model data fusion as a tool for PUB: example in a Swedish mesoscale catchment

J.-F. Exbrayat[1], N. R. Viney[2], J. Seibert[3,4], H.-G. Frede[1], and L. Breuer[1]

[1]Institute for Landscape Ecology and Resources Management, Justus-Liebig-University Giessen, Heinrich-Buff-Ring 26, 35392 Giessen, Germany
[2]CSIRO Land and Water, Canberra, Australia
[3]Department of Geography, University of Zurich, Zurich, Switzerland
[4]Department of Physical Geography and Quaternary Geology, Stockholm University, Stockholm, Sweden

Received: 5 July 2010 – Revised: 11 October 2010 – Accepted: 2 December 2010 – Published: 28 February 2011

Abstract. Post-processing the output of different rainfall-runoff models allows one to pool strengths of each model to produce more reliable predictions. As a new approach in the frame of the "Prediction in Ungauged Basins" initiative, this study investigates the geographical transferability of different parameter sets and data-fusion methods which were applied to 5 different rainfall-runoff models for a low-land catchment in Central Sweden. After usual calibration, we adopted a proxy-basin validation approach between two similar but non-nested sub-catchments in order to simulate ungauged conditions.

Many model combinations outperformed the best single model predictions with improvements of efficiencies from 0.70 for the best single model predictions to 0.77 for the best ensemble predictions. However no "best" data-fusion method could be determined as similar performances were obtained with different merging schemes. In general, poorer model performance, i.e. lower efficiency, was less likely to occur for ensembles which included more individual models.

1 Introduction and scope

Numerous rainfall-runoff models have been developed to describe the water balance and predict runoff at different spatial and temporal scales. However, due to the complexity of natural systems, a lot of the predictive uncertainty is generally linked to the incomplete representation of the different processes involved in modelling flow generation, the so called structural uncertainty (Breuer et al., 2009). Other sources of uncertainty are the initial conditions of the system, measured input and parameter values, regrouped under the general term of stochastic uncertainty.

Depending on the processes they simulate and at which scale the assumptions are made, different models might therefore have different strengths and weaknesses in predicting certain parts of the hydrograph. This has been highlighted in several inter-comparison projects (e.g. Smith et al., 2004; Breuer et al., 2009) The ensemble modelling approach has been proposed to take advantage of these heterogeneities in order to provide more reliable predictions (Viney et al., 2009). Single-models ensembles (SMEs) are obtained from several realisations of the same model structure while exploring the parameter uncertainty. SMEs are for example typical output of the widely used Monte-Carlo based GLUE approach (Beven and Binley, 1992; Beven and Freer, 2001). Multi-model ensembles (MMEs) pool different results obtained from different model structures.

All ensembles can be evaluated in a probabilistic way based on the frequency of prediction of some selected particular events (Renner et al., 2009; Georgakakos et al., 2004). In some other studies the single predictions were combined using different statistical post-processing methods in order to produce "best" forecasts (Shamseldin et al., 1997; Georgakakos et al., 2004; Viney et al., 2009).

In the frame of applying the concept of ensemble modelling to improve the reliability of Predictions in Ungauged Basins (PUB; Sivapalan, 2003) this study evaluates the geographical transferability of parameter sets and combination schemes applied to 5 different rainfall-runoff models: LASCAM (Sivapalan et al., 1996), LASCAM-S (Exbrayat et al., 2010), a self written model based on the snow and soil moisture routines of HBV (Lindström et al., 1997) coupled to the published flow generation equations of INCA (Whitehead et al., 1998) further referred as CHIMP (Combined HBV

Correspondence to: J.-F. Exbrayat
(jean-francois.exbrayat@umwelt.uni-giessen.de)

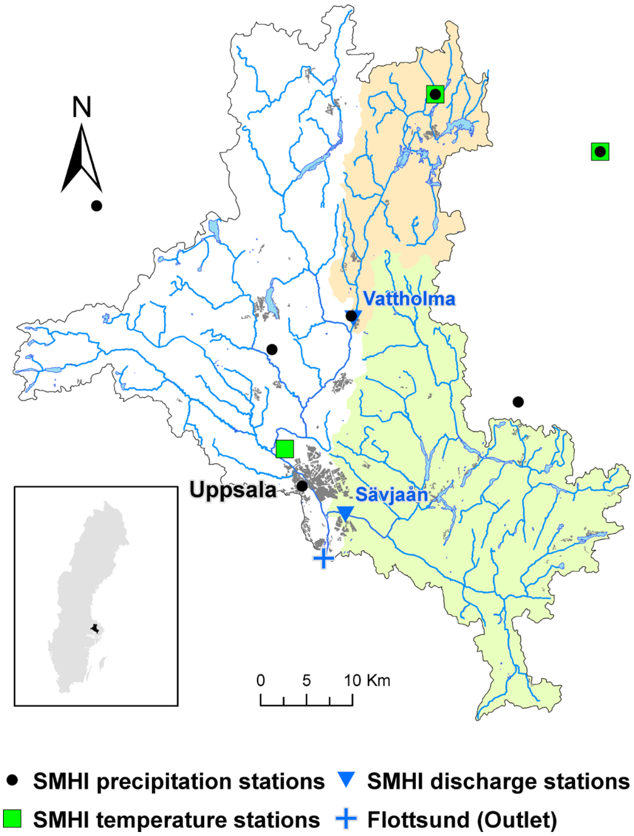

Fig. 1. The River Fyris Catchment.

and INCA Modified in Python, described in Exbrayat et al., 2010), SWAT (Arnold et al., 1998) and HBV-N-D (Lindgren et al., 2007). Different data-fusion methods based on some of those used by Viney et al. (2009) were applied in order to produce large sets of new deterministic MMEs.

This paper is organised as follows. Section 2 presents the catchment and the available data for model application, the models themselves and the different combination methods. In Sect. 3 we present the results for the single models and the newly compiled ensembles in the proxy-basin validation approach. Results are discussed in Sect. 4 and a short summary with conclusions and possible further research directions are presented in Sect. 5.

2 Material and methods

2.1 The river fyris catchment

The study area is located in central Sweden (Fig. 1). The Fyris River catchment has an area of 2000 km^2 and flows into Lake Ekoln which drains into the Baltic Sea. It is a lowland catchment with an elevation ranging between 15 and 115 m a.s.l. Land-use is dominated by mainly coniferous forests (59%) and crop lands (33%). Minor other land cover types are wetlands (4%), urban areas (2%) and lakes (2%) (Lindgren et al., 2007).

Daily records of precipitation (8 gauges) and temperature (3 stations) available from the Swedish Meteorological and Hydrological Institute (SMHI) were used for the chosen 5 years study period (2000 to 2004). Two time series of daily runoff were available over the same period for two non-nested sub-catchments of the Fyris River: Vattholma (281 km^2; light brown in Fig. 1) and Sävja (699 km^2; light green in Fig. 1). These catchments were already studied in another PUB oriented study (Seibert and Beven, 2009). Mean annual runoffs were 219 and 189 mm at Vattholma and Sävja, respectively. As a response to snow melt, high flows usually occur from late autumn to early spring with some thaw-refreezing events leading to high temporal variability during the flood.

2.2 Multi-model members

The five selected models all provide daily runoff predictions. Table 1 gives an overview of different model characteristics and requirements such as the smallest spatial units and input data. There is a good structural variability among the cohort and they may be sorted into an approximate increasing degree of complexity: LASCAM, LASCAM-S, CHIMP, SWAT and HBV-N-D. All these models feature conceptual descriptions of the natural mechanisms involved in flow generation.

LASCAM, LASCAM-S, CHIMP and SWAT were setup in a semi-distributed way. The same sub-catchment delineation, derived from an SRTM digital elevation model, was adopted for each of these 4 models. This spatial disaggregation was obtained with the ArcSWAT extension for ArcGIS (Olivera et al., 2006) which was used for the whole setup of the SWAT model. The sub-catchment delineation divided the Vattholma and Sävja basins into 9 and 28 sub-entities respectively, corresponding to mean sub-catchment areas of 31 and 25 km^2, respectively. While the same parameter sets were applied to each sub-catchment in LASCAM and LASCAM-S, they were independent for each land-use class in CHIMP. SWAT required a disaggregation of each basin into different Hydrological Response Units (HRUs), based on unique combinations of land-use class and soil type. Land-use classes and HRUs in the two latter models were not spatially identified within their sub-catchments and their respective contributions to the flow generation were weighted as a function of their relative areas.

The HBV-N-D model is a fully distributed adaptation of the concepts of the semi-distributed HBV model (Lindström et al., 1997) based on the D8 single flow-direction algorithm (O'Callaghan and Mark, 1984). The HBV-N-D model application used in this study is based on the same setup used by Lindgren et al. (2007) and features 250 m × 250 m grid cells which are associated with a specific land-use type. The running-time of this model was the limiting factor of this study and explains the choice of a relatively short 5 years

Table 1. Main model characteristics.

Model	Smallest spatial unit	Climate forcings
LASCAM	Sub-catchment	Daily rainfall and annual PET
LASCAM-S	Sub-catchment	Daily rainfall, mean temperature and annual PET
CHIMP	Land-Use class	Daily rainfall, temperature and PET
SWAT	HRU	Daily rainfall, minimal and maximal temperature*
HBV-N-D	Grid cell	Daily rainfall and mean temperature, monthly PET

HRU: Hydrological Response Unit; PET: Potential Evapotranspiration *Climate forcing for PET calculation is dependent on the PET method selected, in this case we used the Hargreaves method

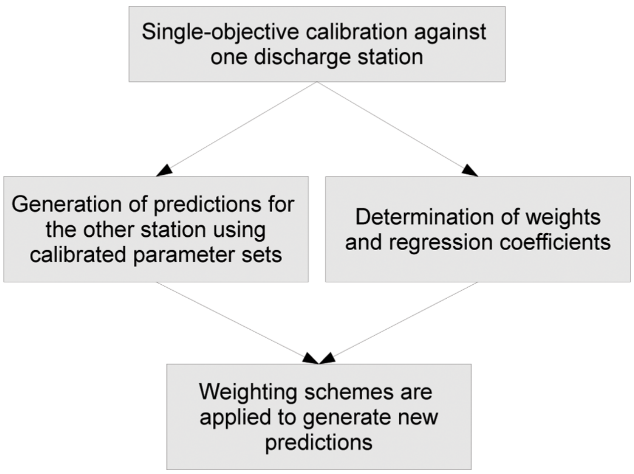

Fig. 2. Ensemble construction methodology.

evaluation period as well as discrepancies in the calibration procedure (Sect. 2.3).

Daily potential evapotranspiration was computed with the temperature-based Hargreaves method (Hargreaves and Samani, 1985) and aggregated to the required time period for each model (Table 1). Snowmelt and snowpack processes were simulated for each calculation unit based on the empirical degree-day approach in all models except LASCAM which does not include any snow routine. We therefore developed the LASCAM-S model by implementing a similar method based on the equations published by Lindström et al. (1997) applied at the sub-catchment scale adopted in LASCAM.

2.3 Ensemble construction and assessment

A summary of the methodological approach used in this study is presented in Fig. 2. This flow chart gives an overview of the different steps we followed in order to create our new model-fusion based forecasts. Each model was calibrated once against each daily discharge record. The calibration was realised in a single-objective way using the Shuffled Complex Evolution optimisation algorithm (Duan et al., 1992) for all the models except for the time-consuming HBV-N-D for which the Parameter Estimator PEST (PEST; Doherty, 2004) was chosen. The optimisation criterion OF was defined as the average of the Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970) calculated for predicted discharge values directly (Eq. 1) and the efficiency obtained with logarithmic values (Eq. 2).

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N} (O_i - S_i)^2}{\sum_{i=i}^{N} \left(O_i - \overline{O}\right)^2} \quad (1)$$

$$\mathrm{lnNSE} = 1 - \frac{\sum_{i=1}^{N} (\ln O_i - \ln S_i)^2}{\sum_{i=1}^{N} \left(\ln O_i - \ln \overline{O}\right)^2} \quad (2)$$

In Equations (1) and (2), O_i and S_i are observed and simulated discharges at time step i while $\bar{O}$ is the mean observed runoff over the N considered time steps. NSE is more sensitive to higher values while lnNSE is also sensitive to lower ones (Krause et al., 2005). The criterion OF therefore gives a better account of the global quality of the prediction. It ranges between $-\infty$ and 1, corresponding to the poorest fit and a perfect match between observations and predictions, respectively.

Then, by comparing predictions obtained with the calibrated parameter sets with the observed data of the corresponding calibration station, we determined the different weights to be applied to our predictions according to the following methods (Table 2). Most of these methods were already tested in a previous split-sample application case (Viney et al., 2009). The weighted-mean approach (WM) used the value of the OF criterion as a weighting coefficient, giving more weight to the better performing members of the fusion procedure. Weight values were in that case independent of the number of members to be merged together. On the other hand, un-constrained and constrained multiple linear regression coefficients (UR and CR schemes) were obtained by using the observations as dependent variables and each possible combinations of 2 to 5 model realisations as independent variables. We therefore obtained different sets of coefficients depending on the combination itself.

The final step of the ensemble generation was to use the calibrated parameter sets to create new single predictions at the alternate subcatchment in the frame of the proxy-basin

Table 2. Overview of the applied merging schemes for ensemble generation.

Merging scheme	Description	Abbr.
Mean	Daily mean of the predictions	ME
Weighted mean	Daily weighted mean of the predictions with weights set as the value of the criterion OF	WM
Median	Daily median value of the prediction	MD
Un-constrained multiple linear regression	Observations are used as dependent variables while predictions are used as independent ones and are assigned different weighting coefficients	UR
Constrained multiple linear regression	Same as above with an interception constrained through the origin	CR

Table 3. Single run results for the different parameter sets. OF = objective function according to Eqs. (1) and (2); PB% = percent bias according to Eq. (3).

Model	Vattholma				Sävja			
	Calib		Proxy		Calib		Proxy	
	OF	PB%	OF	PB%	OF	PB%	OF	PB%
LASCAM	0.64	−7.9	−0.08	56.6	0.67	−13.2	−9.53	−74.0
LASCAM−S	**0.84**	−0.6	0.01	49.8	0.79	−7.9	−2.76	−81.2
CHIMP	0.79	−6.2	0.46	**−17.2**	0.75	−9.4	0.56	**8.3**
SWAT	0.81	−1.8	**0.70**	−20.2	0.78	8.1	0.49	30.7
HBV−N−D	0.83	**−0.2**	0.58	−19.3	**0.84**	**−1.5**	**0.62**	18.8

approach. Simple daily mean (ME) or daily median (MD) were used as data-fusion methods to combine the new model realisations along the aforementioned weights and regression coefficients fitted at the other station. All the methods were applied to every possible combination of 2 to 5 models leading to the creation of 130 new deterministic predictions for each station. Eventually negative regression coefficients could lead to the occurrence of unrealistic negative predictions leading the corresponding MMEs to be disqualified. The quality of the ensemble predictions was finally evaluated by computing the aforementioned OF criterion and the percent bias (PB%, which should be close to zero), calculated by

$$\text{PB\%} = \frac{\sum_{i=1}^{N}(S_i - O_i)}{\sum_{i=1}^{N} O_i}. \quad (3)$$

Notations correspond to those used in Eqs. (1) and (2). We finally compared the multi-model ensembles with their members and the directly calibrated single runs based on these goodness-of-fit descriptors.

3 Results

Calibration and validation results of the single models at the two discharge stations have been summarised in Table 3. The label Calib was used when the corresponding station was the calibration one, while the label Proxy was used when the station was used as validation one (i.e. with geographically transferred parameter sets inherited from the other subcatchment calibration). Calib runs always significantly outperformed the Proxy ones. The LASCAM model gave the worst results in both calibration and proxy-basin application. The upgraded LASCAM-S yielded the best calibration results at Vattholma and second best at Sävja but its predictive quality was really lowered in both proxy-basin approaches. The three other models (i.e. CHIMP, SWAT and HBV-N-D) also showed good calibration results for each sub-catchment. In the proxy-basin context they always showed significantly better performance than LASCAM and LASCAM-S.

The prediction quality of the different model-combination schemes which were applied was illustrated with scatter plots of the criteria values for each station in Figs. 3a and b for Vattholma and Sävja, respectively. A number of different model combinations outperformed the best single model (Proxy runs in Figs. 3a and b) and while regression and median MMEs were the best at Vattholma, mean and weighted-mean methods also performed well at Sävja. However, no MMEs outperformed the four best Calib models in either subcatchment.

In Figs. 4a and b the distribution of the criteria values obtained was represented with boxplots as a function of the number of ensemble members and merging scheme for Vattholma and Sävja, respectively. As already shown in Figs. 3a and b, the highest OF values obtained with MMEs

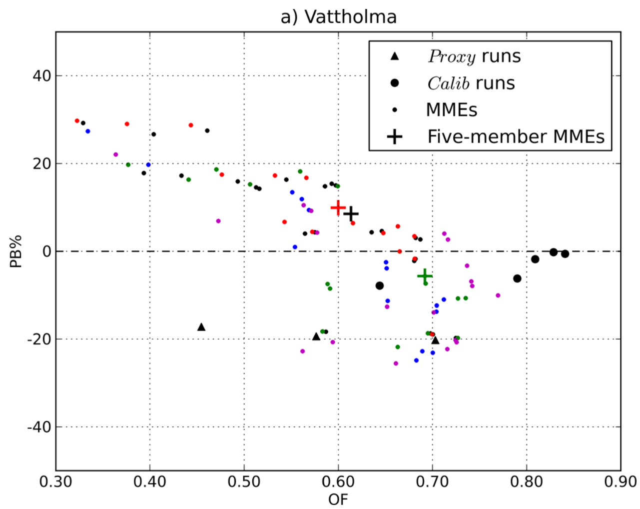

Fig. 3a. Criteria value for the different multi-model ensemble (MME) predictions at Vattholma. Two poor Proxy runs with low OF values are not shown and colours correspond to the different data-fusion methods (red: ME; black: WM; green: MD; blue: UR; magenta: CR; see Table 2 for abbreviations).

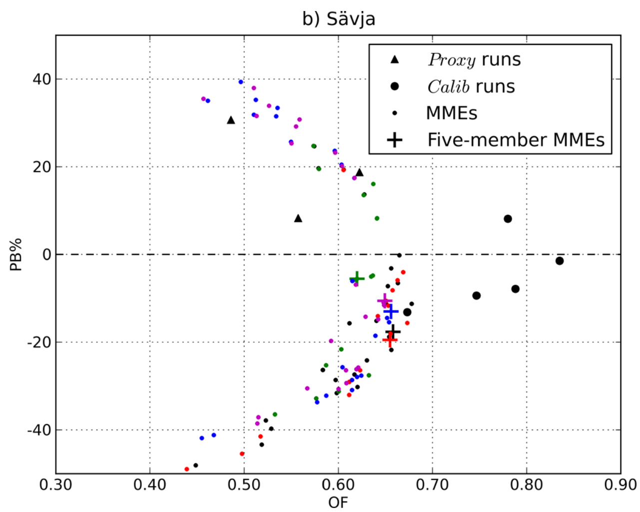

Fig. 3b. Criteria value for the different multi-model ensemble (MME) predictions at Sävja. Two poor Proxy runs with low OF values are not shown and colours correspond to the different data-fusion methods (red: ME; black: WM; green: MD; blue: UR; magenta: CR; see Table 2 for abbreviations).

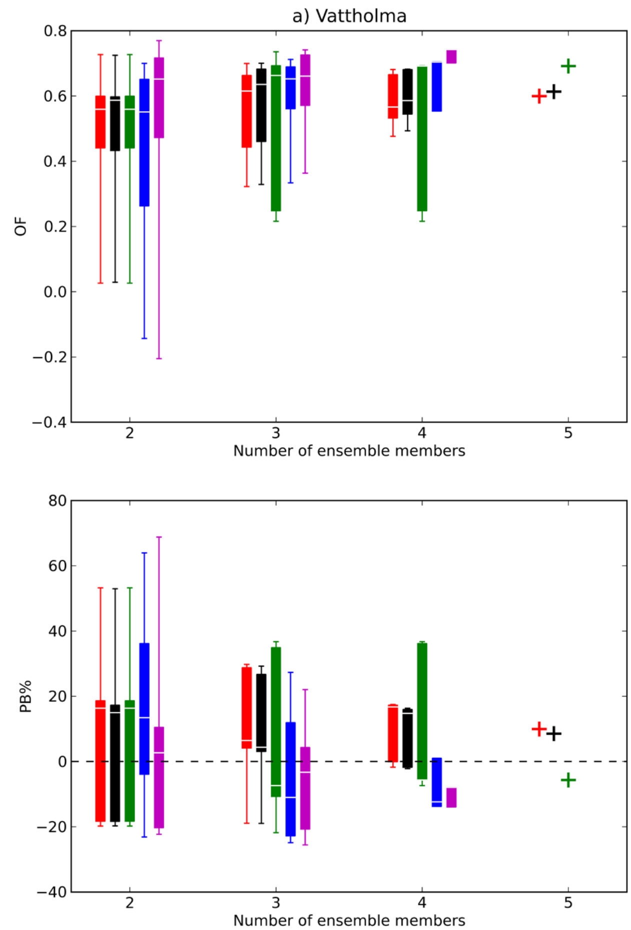

Fig. 4a. Criteria values distribution depending on applied scheme (red: ME; black: WM; green: MD; blue: UR; magenta: CR; see Table 2 for abbreviations) and number of ensemble members at Vattholma. Missing crosses indicate unrealistic negative prediction obtained with UR or CR schemes. OF = objective function according to Eqs. (1) and (2); PB% = percent bias according to Eq. (3).

were 0.77 for Vattholma and 0.68 for Sävja. These corresponded to two different model fusion methods: constrained regression with two members at Vattholma and weighted-mean of three models at Sävja. The best MMEs achieved improvements of +0.07 and 0.06 in comparison to the corresponding best Proxy run (Table 3). However, as illustrated in Figs. 3a and b, different model fusion methods provided predictions almost as good, especially at Sävja, with either more or less members (Figs. 4a and b). Comparatively, PB% values close to 0 were achieved several times in both cases (Figs. 3a and b) and more frequently with simple mean and weighted-average methods. The five-member MMEs were never the best predictors but the corresponding median always had low biases and even outperformed the other five-member MMEs for OF at Vattholma. However, in this latter case, the two regression-based new predictions were disqualified since negative flow values were predicted.

4 Discussion

As expected the implementation of the snow module into LASCAM significantly improved the prediction quality of the calibrated models at each station for both OF and PB%. In the two calibration/proxy-basin cases, no single model could be pointed out as the global best performer

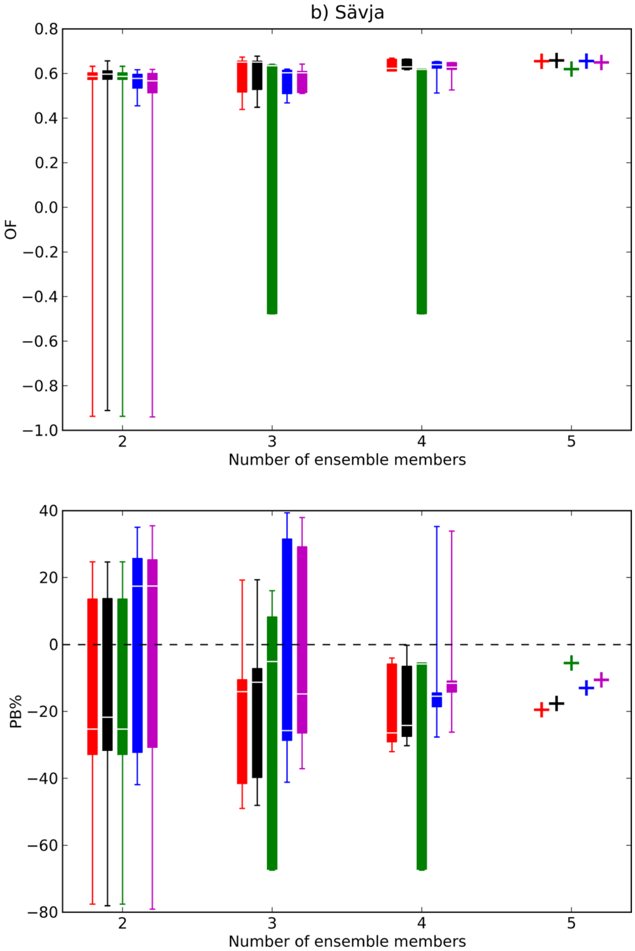

Fig. 4b. Criteria values distribution depending on applied scheme (red: ME; black: WM; green: MD; blue: UR; magenta: CR; see Table 2 for abbreviations) and number of ensemble members at Sävja. Missing crosses indicate unrealistic negative prediction obtained with UR or CR schemes. OF = objective function according to Eqs. (1) and (2); PB% = percent bias according to Eq. (3).

for each considered station and criterion. Very different predictions (according to the metrics) have been obtained even though the two studied catchments were very similar. The more distributed CHIMP, SWAT and HBV still obtained better results with transferred parameter values than LASCAM and LASCAM-S. The overall heterogeneity of model predictions was considered as a good starting point for the ensemble generation following data-fusion methods (Shamseldin et al., 1997).

Different models combinations gave a large range of predictions and good improvements were realised by some of them. As plotted in Figs. 4a and b the best MMEs considering the criterion OF were different between the two stations: constrained regression schemes were usually more efficient at Vattholma (magenta dots in Fig. 4a) while highest improvements were realised with median and weighted-mean schemes at Sävja (green and black dots in Fig. 4b). This was achieved in both cases regardless of the number of merged members. This statement is consistent with Abrahart and See (2002) who showed that the most efficient data-fusion methods depended on the particular application case. However, as Georgakakos et al. (2004) demonstrated, the simple mean of five model predictions consistently outperformed the best single model prediction in several catchments, it was only true at Sävja in our study.

A general trend was that using more members in the selected compilation methods constrained the distribution of the OF and PB% values (Figs. 4a and b). This was similar to the results obtained by Viney et al. (2009) for simple mean and weighted combinations in both calibration and validation periods of a split-sample approach. More precisely, lower values of OF were less likely to occur with more members except with the "median" scheme. On the other hand this latter method applied to the five model predictions gave the closest value to 0 for PB% among the five-member MMEs while keeping OF values close to the best achieved by any other combination method. This could be explained because CHIMP was the most frequent model to participate in this MME while also providing the least biased of the Proxy predictions.

The simplest averaging schemes (i.e. mean and weighted-mean) showed similar results in terms of criteria values distribution (MD and WM in Figs. 4a and b). There was a slight shift of boxplots, and therefore the distribution, towards zero bias and higher efficiencies for "weighted-mean" schemes, thus illustrating the effect of the weighting process. This occurred even if the poorly fitted LASCAM-S Proxy runs were given heavy weights in response to good calibration results (Table 3).

Viney et al. (2009) also showed that applying multiple-linear regression coefficients gave the best results in terms of NSE for calibration periods. But these ensembles were outperformed by some other combinations in a split-sample calibration context. The proxy-basin validation scheme adopted in our study showed that the MMEs based on constrained regressions were the best performers at Vattholma but not at Sävja. Still, they obtained OF values close to the best ones (Figs. 3b and 4b) in this latter case. There was no real advantage in using these more complicated regression merging schemes as it resulted in unrealistic negative runoff predictions 14 and 2 times for Vattholma and Sävja, respectively.

Some multi-model predictions gave PB% values close to zero (Figs. 3a and b) while the best Proxy run for this criterion had a bias of 8.3% (CHIMP at Sävja, Table 3). This could be attributed to an inter-model balance as LASCAM and LASCAM-S usually had opposed biases in comparison to the other Proxy model realisations (Table 3). It therefore moved the merged predictions towards bias values closer to 0. Moreover, even with the poor efficiencies illustrated by low OF values for LASCAM and LASCAM-S (Table 3), these MMEs could also achieve good results for this latter criterion at Vattholma. At Sävja they even obtained the best OF (Fig. 4b) while Table 3 showed that in that case, the LASCAM and LASCAM-S Proxy runs were very biased with negative OF.

These results represented another illustration of the advantage of combining strengths of different predictions which is the surrounding philosophy in model combination (Shamseldin et al., 1997). Such MMEs were able to provide at least good estimates of the global water balance over the study period even though some of them were partly based on the worst single models (which indeed provided the final prediction with interesting information).

5 Conclusions

Several interesting results could be either deduced or confirmed in regards of previously published studies but this time in the frame of a PUB application. First, no optimal combination schemes could be identified, even though the two catchments investigated were similar, which did not increase the transferability of the different methods. Still, the study showed that the more members that were merged together, the lower the risk of getting bad predictions. In the case of a PUB, this offers a minimum guarantee that the newly compiled predictions would be closer to reality even while including very bad single predictors. For example, using the simple daily median value of the five single model predictions provided good results with a low bias and could be identified as a good all-round compromise.

However, even if good results were obtained with some of the data fusion methods, none of them could outperform the calibration process as a result of a poor transferability of single parameter values. A probable limitation of this study was therefore to consider only one realisation per model in a deterministic way. According to the equifinality theory (Beven and Freer, 2001), different parameter combinations are able to give evenly good predictions. Due to small and unquantifiable heterogeneities between catchments, a common optimal parameter set is not likely to exist even while considering two basins in the same hydro-climatic context. Therefore, an idea for next PUB predictions would be to study the transferability of optimised (i.e. constrained) parameter ranges of the predictions after analyses of numerous realisations of the same model (i.e. SMEs) and to introduce probabilistic rather than deterministic predictions.

Acknowledgements. The meteorological and hydrological data was obtained from the Swedish Meteorological and Hydrological Institute (SMHI).

Edited by: A. Weerts
Reviewed by: two anonymous referees

References

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

Arnold, J. G., Srinivasan, R., Muttiah, R. S., and Williams, J. R.: Large area hydrologic modeling and assessment part I: model development, J. Am. Water Resour. As., 34, 73–89, 1998.

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrol. Process., 6, 279–298, 1992.

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, J. Hydrol., 249, 11–29, 2001.

Breuer, L., Huisman, J., Willems, P., Bormann, H., Bronstert, A., Croke, B. F. W., Frede, H.-G., Gräff, T., Hubrechts, L., Jakeman, A., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D., Lindström, G., Seibert, J., Sivapalan, M., and Viney, N. R.: Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM) I: Model intercomparison with current land use, Adv. Water Resour., 32, 129–146, 2009.

Doherty, J.: PEST-Model Independent Parameter Estimation User Manual: 5th edn., Watermark Numerical Computing, Brisbane, Australia, 2004.

Duan, Q., Sorooshian, S., and Gupta, V.: Effective and Efficient Global Optimization for Conceptual Rainfall-Runoff Models, Water Resour. Res., 28, 1015–1031, 1992.

Exbrayat, J.-F., Viney, N. R., Seibert, J., Wrede, S., Frede, H.-G., and Breuer, L.: Ensemble modelling of nitrogen fluxes: data fusion for a Swedish meso-scale catchment, Hydrol. Earth Syst. Sci., 14, 2383–2397, doi:10.5194/hess-14-2383-2010, 2010.

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

Hargreaves, G. and Samani, S.: Reference crop evapotranspiration from temperature, Appl. Eng. Agric., 1, 96–99, 1985.

Krause, P., Boyle, D. P., and Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci., 5, 89–97, 2005, http://www.adv-geosci.net/5/89/2005/.

Lindgren, G., Wrede, S., Seibert, J., and Wallin, M.: Nitrogen source apportionment modeling and the effect of land-use class related runoff contributions, Nord. Hydrol., 38, 317–331, 2007.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, J. Hydrol., 201, 272–288, 1997.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290, 1970.

O'Callaghan, J. F. and Mark, D. M.: The extraction of drainage networks from digital elevation data, Comput. Vision Graph., 28, 323–344, 1984.

Olivera, F., Valenzuela, M., Srinivasan, R., Choi, J., Cho, H., Koka, S., and Agrawal, A.: ArcGIS-SWAT: A geodata model and GIS interface for SWAT, J. Am. Water Resour. As., 42, 295–309, 2006.

Renner, M., Werner, M., Rademacher, S., and Sprokkereef, E.:

Verification of ensemble flow forecasts for the River Rhine, J. Hydrol., 376, 463–475, 2009.

Seibert, J. and Beven, K. J.: Gauging the ungauged basin: how many discharge measurements are needed?, Hydrol. Earth Syst. Sci., 13, 883–892, doi:10.5194/hess-13-883-2009, 2009.

Shamseldin, A. Y., O'Connor, K. M., and Liang, G. C.: Methods for combining the outputs of different rainfall-runoff models, J. Hydrol., 197, 203–229, 1997.

Sivapalan, M.: Prediction in ungauged basins: a grand challenge for theoretical hydrology, Hydrol. Process., 17, 3163–3170, 2003.

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

Smith, M. B., Seo, D., Koren, V. I., Reed, S. M., Zhang, Z., Duan, Q., Moreda, F., and Cong, S.: The distributed model intercomparison project (DMIP): motivation and experiment design, J. Hydrol., 298, 4–26, 2004.

Viney, N. R., Bormann, H., Breuer, L., Bronstert, A., Croke, B. F. W., Frede, H.-G., Gräff, T., Hubrechts, L., Huisman, J. A., Jakeman, A. J., Kite, G. W., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M., and Willems, P.: Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) II: Ensemble combinations and predictions, Adv. Water Resour., 32(2), 147–158, 2009.

Whitehead, P., Wilson, E., and Butterfield, D.: A semi-distributed integrated nitrogen model for multiple source assessment in catchments (INCA): Part I – model structure and process equations, Sci. Total Environ., 547–558, 1998.