Deterministic evaluation of ensemble streamflow predictions in Sweden

Anna Johnell, Göran Lindström* and Jonas Olsson

Swedish Meteorological and Hydrological Institute, SE-601 76, Norrköping, Sweden.
*Corresponding author. E-mail: *Goran.Lindstrom@smhi.se*

Received 1 December 2006; accepted in revised form 21 July 2007

Abstract A system for ensemble streamflow prediction, ESP, has been operational at SMHI since July 2004, based on 50 meteorological ensemble forecasts from ECMWF. Hydrological ensemble forecasts are produced daily for 51 basins in Sweden. All ensemble members, as well as statistics (minimum, 25% quartile, median, 75% quartile and maximum), are stored in a database. This paper presents an evaluation of the first 18 months of ESP median forecasts from this system, and in particular their performance in comparison with today's categorical forecast. The evaluation was made in terms of three statistical measures: bias *B*, root mean square error *RMSE* and absolute peak flow error *PE*. For ESP forecasts the bias ranged between -20% and 80% with a systematic overestimation for Sweden as a whole. A comparison between bias in input precipitation and ESP output, respectively, revealed only a weak relationship, but streamflow overestimation is likely related mainly to model properties. The results from the streamflow forecast comparison showed that the ESP median in deterministic terms performs overall as well as the presently used categorical forecast. Further, ESP has the advantage of providing at least a qualitative measure of the uncertainty in the forecasts, with probability forecasts being the ultimate goal.

Keywords Ensemble; HBV model; prediction; streamflow; Sweden

Nordic Hydrology **Vol 38 No 4–5 pp 441–450** © IWA Publishing 2007

Introduction

The uncertainty in the streamflow response to temperature changes and precipitation amounts is highly dependent on the meteorological and hydrological situation. Typical situations with a high uncertainty include the start of the snow melt period, when a few degrees' temperature difference may have a strong impact on runoff, and periods with intense rainfall. In general, however, operational streamflow forecasting is based on a deterministic meteorological forecast, which produces a single evolution of the streamflow in the forecast period. This is expected to represent the most likely development, but contains no information about the associated uncertainty. An alternative is probabilistic forecasts which produce a possible range of streamflow variation, specified in terms of flow exceedance probabilities (e.g. Krzysztofowicz 2001).

Probabilistic streamflow forecasts can be based on meteorological ensemble forecasts, which are operationally issued at some meteorological services (e.g. Molteni *et al.* 1996). The concept of ensemble forecasting uses the fact that different initial meteorological conditions may be associated with distinctly different evolutions of the weather on a 5–10 d timescale. To take this effect into account, atmospheric models may be run a number of times from slightly perturbed initial conditions, producing an ensemble of possible forecasts (members). By using these forecasts to drive a hydrological model an ensemble of single streamflow forecasts is produced, which may be processed to generate probabilistic forecasts.

doi: 10.2166/nh.2007.022

At the Swedish Meteorological and Hydrological Institute (SMHI), a system for ensemble streamflow prediction (ESP) has been operational since summer 2004. In an operational system errors can easily be adjusted continuously. In the system, 10-d meteorological ensemble forecasts from the European Centre for Medium-range Weather Forecasts (ECMWF) are used to drive the hydrological HBV model. In some recent investigations, hydrological ensemble forecasting based on ECMWF data has been attempted. Much of this work has been performed in connection with the development and application of the European Flood Forecasting System (EFFS; De Roo *et al.* 2003). Ensemble streamflow forecasts generated by the LISFLOOD model have been evaluated mainly for a number of flooding case studies (e.g. Gouweleeuw *et al.* 2005; Pappenberger *et al.* 2005; Werner *et al.* 2005). Pappenberger *et al.* (2005) further combined the ESP uncertainty with the estimated uncertainty associated with the hydrological model. Roulin and Vannitsem (2005) evaluated ESP forecasts in two Belgian catchments. Roulin (2006) further investigated how the ESP forecasts can be combined with a cost–loss model for improved decision-making. These investigations have identified both strengths and weaknesses of the ensemble approach, and in particular it has been concluded that a large amount of historical forecasts is required for proper evaluation (Gouweleeuw *et al.* 2005).

The main objective of this paper is to present the results from an evaluation of 18 months of operational hydrological ensemble forecasts in 51 catchments in Sweden. This is, to our knowledge, the largest dataset for which medium-range hydrological ensemble forecasts have been evaluated to date. Despite their probabilistic nature, the ESP forecasts are here evaluated in deterministic terms. This is motivated by the fact that, even if the trend is towards probabilistic forecasts, it will likely take a long time before end users will be able to properly handle such forecasts. Thus the importance of categorical forecasts will remain in the foreseeable future and efforts to obtain improved categorical forecasts continue to be important. The ESP median is a potential candidate as a categorical forecast, at least for longer lead times than a few days. A so-called spread-skill analysis, widely used in meteorological ensemble verification (e.g. Scherrer *et al.* 2004) was conducted to evaluate the relationship between ESP spread and forecast error. Besides the streamflow forecasts, the input precipitation forecasts were also evaluated to estimate the errors involved and their relationship with the streamflow errors.

System and database

The ensemble streamflow predictions (ESP) were made within Aegir, a system for automatic real-time hydrological production at SMHI. The system is monitored 24 hours a day. Data are collected from various databases at SMHI and prepared for input to the conceptual, semi-distributed, catchment-scale HBV model (Lindström *et al.* 1997), which is the central part of Aegir. The HBV model is run operationally in 51 indicator catchments and the forecasts are post-processed to generate a range of products that are distributed to other systems and databases. Operational hydrological forecasting is based on a categorical meteorological forecast (here called CAT), which is mainly based on the atmospheric HIRLAM model.

The input meteorological forecasts for ESP calculations are collected from ECMWF. These 10-d forecasts consist of 50 ensemble members and one control forecast. In Aegir, the precipitation and temperature from these forecasts are retrieved and processed for input to the HBV model. In the HBV forecast, autoregressive updating is used to eliminate the model error at the start of the forecast (see, for example, Lundberg 1982).

The output from the ESP calculations by the HBV model consists of 51 streamflow forecasts (Figure 1(a)), which in Aegir are post-processed by calculating five statistical percentiles: minimum (2% probability of non-exceedance), lower quartile (25%), median

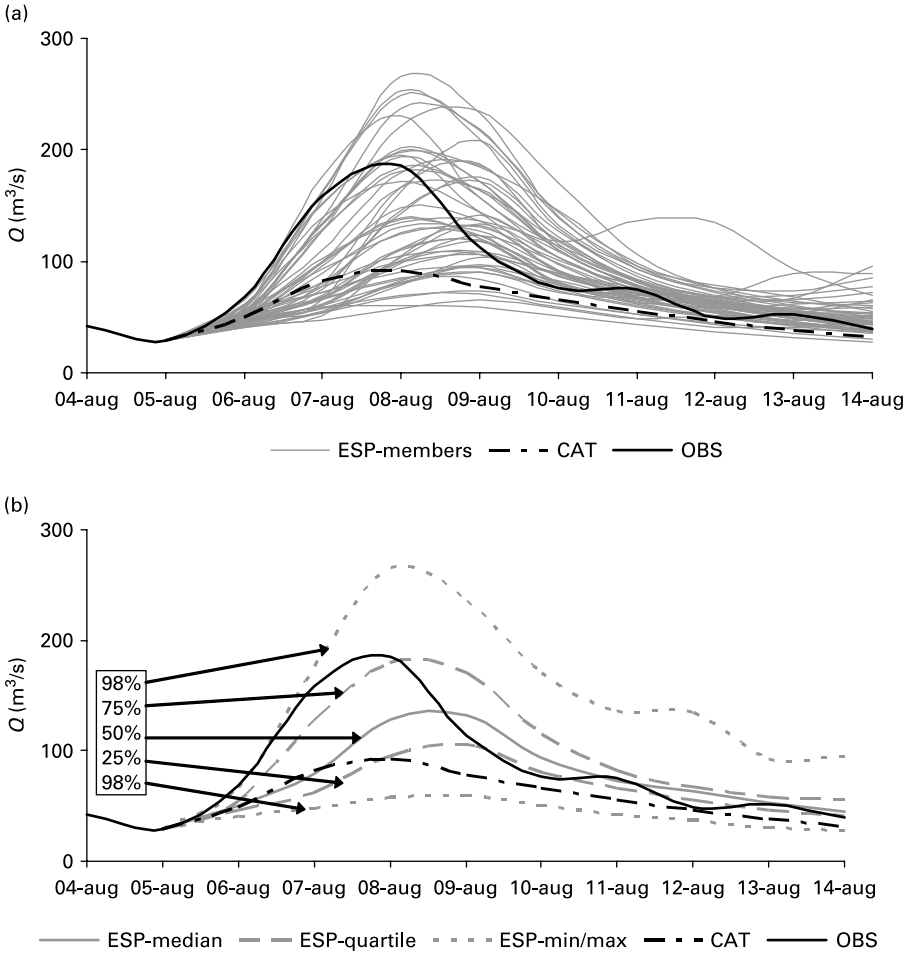

Figure 1 Example of observed discharge (OBS), categorical forecast (CAT) and ESP forecast (ESP) as individual members (a) and converted into statistical percentiles (b)

(50%), upper quartile (75%) and maximum (98%) (Figure 1(b)). These values are stored in a database which has been used for deterministic evaluation of the ESP forecasts.

The evaluated streamflow data base comprises historical ESP and CAT forecasts as well as observations during an 18-month period (July 2004–January 2006) in all 51 catchments. The catchments are distributed all over Sweden with sizes ranging from 8 km^2 to 6110 km^2 and average streamflow from 0.05 m^3/s to 124 m^3/s. The amount of missing observations during the period was approximately 9%. Six of the catchments were excluded from the evaluation due to either too few available observations in the evaluation period or excessive model errors.

Evaluation methods

Prior to the evaluation of the streamflow forecasts (Q), an evaluation of the input precipitation forecasts (P) was performed. In this evaluation, the values used as input to the HBV model were compared with daily observations. These observations were in the form of catchment averages, calculated from meteorological stations inside and in the vicinity of each evaluated catchment. The values were adjusted in the HBV model, to compensate for

altitude differences between stations as well as different elevation zones within the catchment.

In terms of ensemble forecasts a natural candidate for deterministic evaluation is the ensemble median forecast, which is the one used here. Systematic errors in the meteorological and hydrological forecasts can, however, result in streamflow forecasts that are systematically too high or too low. Thus, another percentile could, on average, be more similar to the observed values than the median.

Forecast performance was evaluated in terms of two commonly used statistical measures, bias and root mean square error (e.g. Maidment 1993). The bias B is used to investigate the forecast error in terms of mean streamflow, and is calculated as

$$B = \overline{Q_f} - \overline{Q_o} = \frac{1}{n}\sum_{i=1}^{n} Q_f(i) - \frac{1}{n}\sum_{i=1}^{n} Q_o(i) \tag{1}$$

where $Q_f(i)$ is forecasted and $Q_o(i)$ is observed streamflow during the n-day evaluation period. Also absolute bias $B_{\text{abs}} = \left|\overline{Q_f} - \overline{Q_o}\right|$ was used. The root mean square error $RMSE$ is expressed as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} [Q_f(i) - Q_o(i)]^2} \tag{2}$$

and is thus a measure of the typical daily error. As the streamflow level varies substantially between different catchments, both B and $RMSE$ were divided by the average observed streamflow $\overline{Q_o}$ and thus they are relative measures (%).

Evaluation in terms of B and $RMSE$ was performed for two different sets of data. The first comprised all days during the evaluation period, i.e. all streamflow levels. The second set comprised only days when $Q_o(i)$ at the time of the forecast was larger than $\overline{Q_o}$: thus the number of days included is about half the total number of days in the evaluation period. The second evaluation focuses on streamflow above the mean level, which is more relevant for flood forecasting than streamflow below the mean level. In a third type of evaluation we focused only on the highest streamflow peaks. This was done by dividing the evaluation period into three six-month periods, and for each period identifying the maximum value of $Q_o(i)$, $Q_o^{\max}$. A relative peak error PE was defined as

$$PE = \frac{Q_f(i) - Q_o^{\max}}{Q_o^{\max}} \tag{3}$$

where $Q_f(i)$ is the forecasted streamflow (i.e. the ESP median and CAT forecasts made 1–9 d earlier) on the day of the observed maximum value. Also the absolute peak error PE_{abs} was evaluated.

In addition to the evaluation of the ESP median, a so-called spread-skill analysis was performed to evaluate the prospect of using the spread in the ESP forecast as an indicator of the expected forecast error. As a measure of the ESP spread on a certain day, the range between the upper quartile and the lower quartile, IQR, was used (Figure 1(b)). For each catchment and forecast lead time, five classes were defined representing "very small", "small", "medium", "large" and "very large" spread. The classes were based on a ranked list of the IQR values obtained during the evaluation period, with 20% of the forecasts falling into each class. Thus class "very small" comprised percentiles 0–20 in the IQR-distribution, class "small" percentiles 20–40, etc. The corresponding forecast skill was defined by the mean absolute error MAE of the ESP median, i.e.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|Q_f(i) - Q_o(i)\right| \quad (4)$$

where the averaging was made over the m days within each spread class. Also MAE was divided by $\overline{Q_o}$ to become a relative error.

Results and discussion

When averaged over all catchments during the entire evaluation period, the observed precipitation is 2.34 mm/d. The corresponding value for the forecasts is 2.35 mm/d. There is a slight variation for different days in the forecast (between 2.28 and 2.40) but no apparent trend. In total there is thus only a very small bias in the precipitation forecasts, but for an individual catchment the bias may be substantial. This is illustrated in Figure 2, showing the relationship between observed and forecasted precipitation in all catchments. In some catchments the difference exceeds 1 mm/d (i.e. ~400 mm/yr), which corresponds to a relative bias B of more than 50%. Figure 3(a) shows B for all catchments. In general, although difficult to see clearly on the map, B is larger in small catchments than in large catchments. This is expected as the precipitation from the coarse ECMWF grid should be less representative the smaller the catchment.

Further analysis of the ensemble forecast precipitation showed that the daily standard deviation was underestimated by ~15% in the ensemble forecasts, indicating that high precipitation intensities are underestimated. Another consequence of the coarse grid is that the frequency of dry days ($P = 0$) is underestimated in the ensemble forecasts by ~12%. As the total precipitation was well described in the ensemble forecasts, this means that the intensity during wet days ($P > 0$) was underestimated by ~15%.

Also the streamflow evaluation was performed by comparison with daily observations in each catchment during the evaluation period. Figure 3(b) shows the relative bias B for streamflow in all catchments, averaged over all days in the forecast. Whereas the P bias was generally within the range -50% $< B <$ 50%, with approximately 50% of the catchments having a positive bias, for Q the range is -20% $< B <$ 80% with 75% of the catchments having a positive bias. As for the P forecasts, B is generally larger in small catchments. Despite this

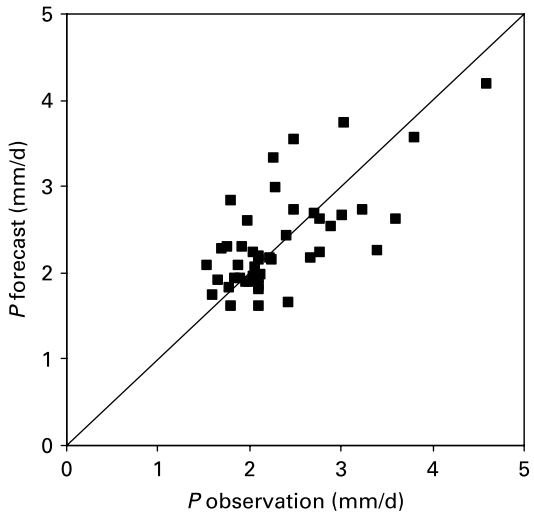

Figure 2 Relationship between observed and ECMWF forecasted mean precipitation in the 45 catchments

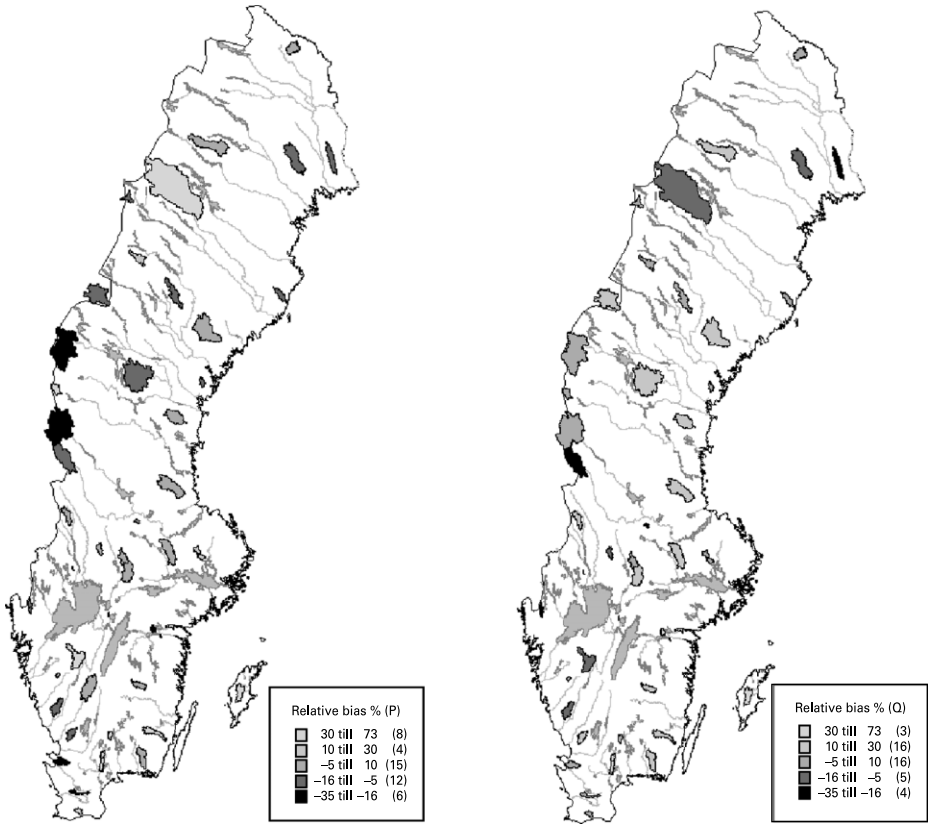

Figure 3 Relative bias in ECMWF precipitation forecasts (a) and ESP forecasts (b) in the 45 catchments

qualitative similarity, the relationship between the P and Q bias for individual catchments is very weak, which is shown in Figure 4. The streamflow bias is thus not primarily caused by bias in the precipitation input but rather by the implementation of the HBV model.

Moving over to the comparison between the ensemble streamflow forecast (ESP) and the categorical forecast (CAT), the result in terms of B_{abs} averaged over all catchments is shown in Figure 5(a). For both the ESP median and the CAT forecast, B_{abs} increases with increasing length of the forecast, from ~5% on day 1 up to ~20% on day 9. The ESP median has a slightly lower absolute bias than the CAT forecast for days 1 and 2 in the forecast. For day 3 the result is identical, and for days 4–9 the CAT forecast has a slightly lower bias. On average, B_{abs} for ESP median is 0.8% higher than the CAT forecast. The difference is small, however. The result in terms of B is shown in Figure 5(b). As discussed in connection with Figure 3(b), in total B is positive, i.e. streamflow is systematically overestimated in the forecasts. Further, this overestimation increases with the forecast lead time, from ~3% on day 1 to almost 15% on day 9. For days 5–7 the overestimation is smaller for CAT, whereas the ESP median performs better for the other days. On average, B for the ESP median is 0.3% lower than the CAT forecast.

The reason for the systematic overestimation (Figure 5(b)) has not been completely clarified and will be further investigated, but it is likely related mainly to limitations in the HBV model calibration. It is difficult to reproduce the full streamflow variability, but often peak flows are somewhat underestimated and low or base flows somewhat overestimated. In the measure B, the latter effect will likely dominate. The systematic increase of B with forecast length is probably related to the autoregressive updating. This procedure eliminates

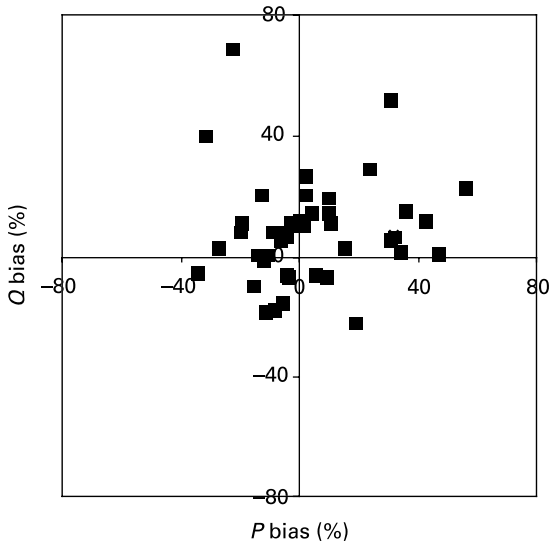

Figure 4 Relationship between bias in ECMWF precipitation forecasts and ESP forecasts in the 45 catchments

bias at the time of the forecast, but as the effect of the updating gradually wears off with time in the forecast, B increases.

When only streamflow levels above the mean $\overline{Q_o}$ are taken into account the result in terms of B_{abs} becomes similar to the picture in Figure 5(a). However, for B the situation is quite different from Figure 5(b), but only a minor bias was found which did not vary systematically with forecast length. For the ESP median, B averaged over all days in the forecast is -2.9%, and for CAT the value is only 0.4%. Thus the ESP median systematically slightly underestimates streamflows above the mean flow, whereas CAT in total forecasts these flows well. This confirms that the total overestimation (Figure 5(b)) is related to periods with low flow.

In terms of *RMSE* the ESP median performed slightly better than the deterministic CAT forecast, both when evaluated for all streamflow levels (Figure 6) and when evaluated for streamflow above $\overline{Q_o}$ (not shown but very similar to Figure 6). Judging from the *RMSE* analysis,

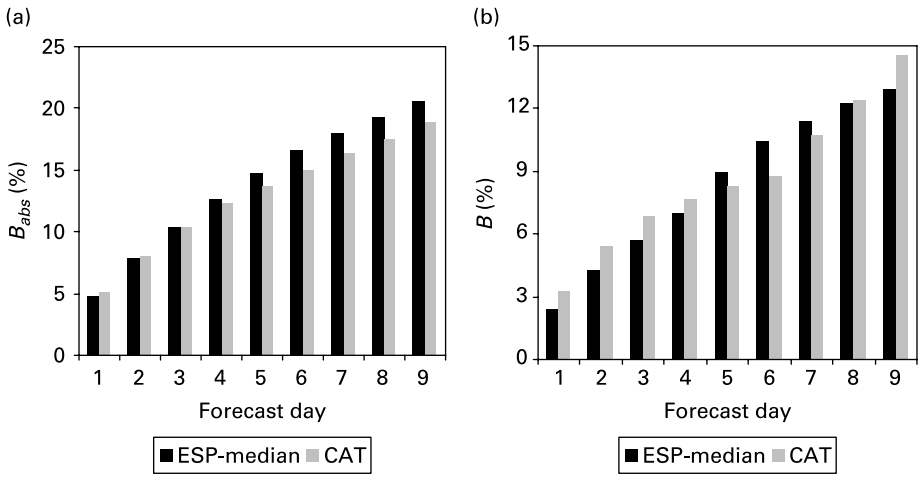

Figure 5 B_{abs} (a) and B (b) of the ESP median and CAT forecasts, averaged over all 45 catchments, as a function of forecast day

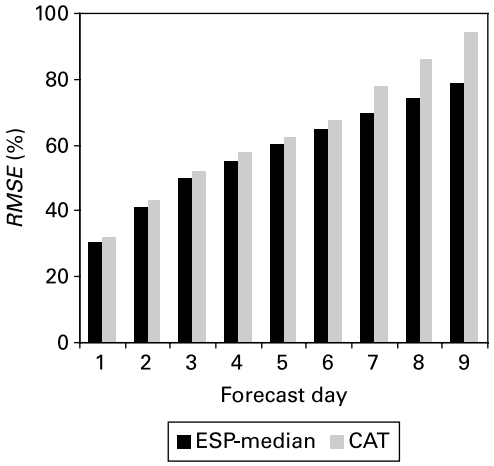

Figure 6 *RMSE* of the ESP median and CAT forecasts, averaged over all 45 catchments, as a function of forecast day

the greatest gain from using the ESP median instead of the traditional deterministic forecast occurs for longer forecast lead times. On average, the ESP median *RMSE* is 5.0% lower than CAT, varying from 1.5% on day 1 to 16% on day 9. It may be remarked that the *RMSE* values seem high, varying from ~30% on day 1 to 80–90% and even more for the longest lead times. However, the *RMSE* by construction gives a high penalty for single large errors, such as a severely underestimated peak flow. Further, in some small catchments $\overline{Q_o}$ is very low and thus errors expressed as a percentage may become very high even if the actual error (in m^3/s) is small.

Figure 7(a) shows the results from the analysis of the highest streamflow levels in the data set in terms of PE_{abs}. The result has been averaged over the three six-month periods used. On the x axis, D-1 denotes the day before Q_o^{max} was recorded and D-9 nine days before. Concerning the difference between EPS median and the CAT forecast there is no clear pattern. In general the ESP median error is equal to or slightly lower than the error in the CAT forecast, but on average the difference is 0%. Figure 7(b) shows the result in terms of *PE*, which shows that the total peak error is almost entirely associated with underestimation. For day 1 the underestimation is identical for the ESP median and the CAT forecast, but for

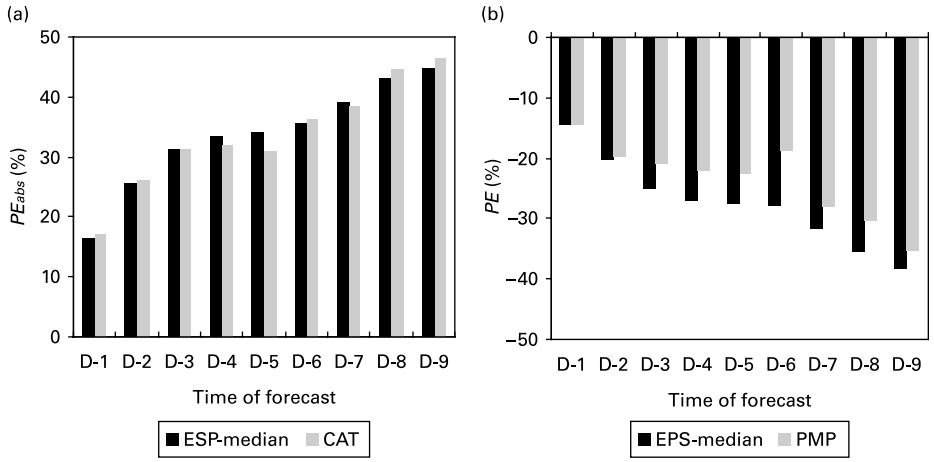

Figure 7 PE_{abs} (a) and *PE* (b) of the ESP median and CAT forecasts of Q_o^{max} made 1–9 d before, averaged over all 45 catchments

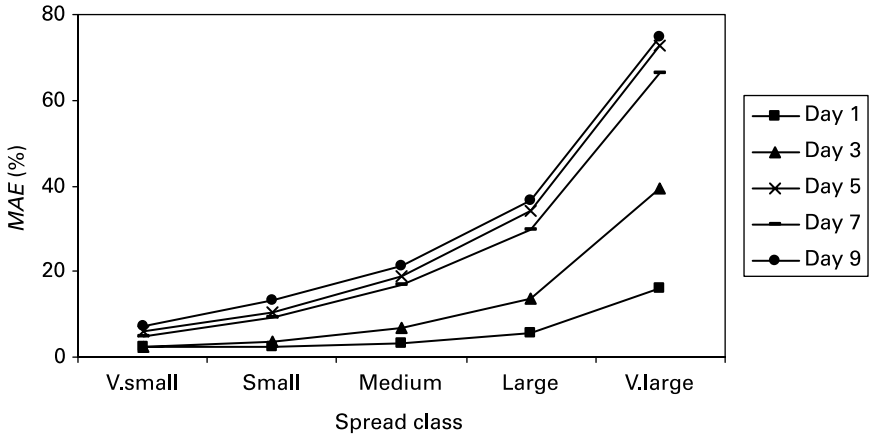

Figure 8 Relationship between ESP spread class and the subsequent forecast error in terms of *MAE* of the ESP median, averaged over all 45 catchments

longer lead times the underestimation is more pronounced in the ESP median, on average by 4%. This is likely related to the smoothed nature of the EPS median and the consequent inability of fully describing extreme events.

It may be remarked that the peak flow underestimation by both types of forecasts is partly related to the chosen strategy of performing the evaluation for days with high observed streamflow. Because of model uncertainty it is common for simulated peak flows to differ somewhat from the observed ones, both with respect to timing and magnitude. A simulated peak flow may occur one or a few days after or before the observed peak, and even if the timing is correct some underestimation is common. Figure 7(b) thus does not imply that all peak flows are consistently underestimated by 15–40%, but the picture is also affected by timing errors.

The result from the spread-skill analysis is shown in Figure 8. It is apparent that a very small ESP spread is indicative of a comparatively small error also in the forecasted streamflow. As the ESP spread increases, so does the associated forecast error. It may be noted that the difference in *MAE* between spread classes "very small", "small" and "medium" is rather small, and it is not until the spread reaches the classes "large" and "very large" that the forecast error substantially increases. It may further be noted that *MAE* for forecast days 5, 7 and 9 is overall rather similar, indicating that the ESP median forecast error exhibits only a limited increase after day 5. This tendency confirms the interpretation made in connection with Figure 6, i.e. that the ESP forecast has its main strength in the latter part of the 9-d forecast period. Overall, the results from the spread-skill analysis confirm the ability of the ESP forecast spread to function as an indicator of the forecast uncertainty.

Conclusions

The evaluation of precipitation in the ECMWF ensemble forecasts showed that there is a negligible volume bias for Sweden as a whole, but for individual catchments the bias may reach ±50% of the local mean precipitation. As expected, the bias was generally higher in small catchments than in large. This tendency was also found in the streamflow evaluation, but in this case the bias was generally positive, i.e. streamflow was systematically overestimated in the forecasts. The relationship between precipitation bias and streamflow bias was very weak, but streamflow bias was found to originate mainly from the limited ability of the model to fully reproduce the streamflow variability, in particular during low flow periods. This indicates that, for reducing the bias in the streamflow forecasts,

improvements in the implementation and calibration of the hydrological model are more important than bias correction in the precipitation forecasts.

Concerning the comparison between the ESP median forecast and today's categorical forecast, the difference in performance is overall small. As expressed in the accuracy measures used here, the ESP median forecast is as accurate as the categorical forecast. Some results, in particular from the *RMSE* analysis, indicated an improved performance for long lead times (6–9 d) by the ESP median, which is encouraging. On the other hand, due to its smoothed nature, the ESP median underestimated peak flows to a somewhat higher degree than did the categorical forecast.

The main gain from ensemble forecasts is, however, their ability to provide a range of possible streamflow levels, rather than a single value. The results from the spread-skill analysis showed a clear relationship between the size of this range and the associated forecast error. Thus, even such a simple measure such as the ESP spread may provide an added value to the forecasts, in the form of an estimation of its expected error. The ESP system may further be developed to provide proper probabilistic forecasts, explicitly specifying the probability of reaching different streamflow levels. Such a probabilistic development and evaluation is currently ongoing and the results will be presented elsewhere.

Acknowledgements

The financial support from Elforsk, Räddningsverket (the Swedish Rescue Services Agency) and the SMHI is gratefully acknowledged. We thank two anonymous reviewers for their constructive criticism of the original manuscript.

References

De Roo, A., Gouweleeuw, B., Thielen, J., Bartholmes, J., Bongioannini-Cerlini, P., Todini, E., Bates, P., Horritt, M., Hunter, N., Beven, K., Pappenberger, F., Heise, E., Rivin, G., Hils, M., Hollingsworth, A., Holst, B., Kwadijk, J., Reggiani, P., van Dijk, M., Sattler, K. and Sprokkereef, E. (2003). Development of a European flood forecasting system. *Int. J. River Basin Mngmnt.*, **1**, 49–59.

Gouweleeuw, B.T., Thielen, J., Franchello, G., De Roo, A.P.J. and Buizza, R. (2005). Flood forecasting using medium-range probabilistic weather prediction. *Hydrol. Earth Syst. Sci.*, **9**, 365–380.

Krzysztofowicz, R. (2001). The case for probabilistic forecasting in hydrology. *J. Hydrol.*, **249**, 2–9.

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

Lundberg, A. (1982). Combination of a conceptual model and an autoregressive error model for improving short time forecasting. *Nordic Hydrol.*, **13**, 233–246.

Maidment, D.R. (Ed.) (1993). *Handbook of Hydrology*. McGraw-Hill, New York.

Molteni, F., Buizza, R., Palmer, T.N. and Petroliagis, T. (1996). The E.C.M.W.F. ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.*, **122**, 73–119.

Pappenberger, F., Beven, K.J., Hunter, N.M., Bates, P.D., Gouweleeuw, B.T., Thielen, J. and De Roo, A.P.J. (2005). Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions with the European Flood Forecasting System (EFFS). *Hydrol. Earth Syst. Sci.*, **9**, 381–393.

Roulin, E. (2006). Skill and relative economic value of medium-range hydrological ensemble predictions. *Hydrol. Earth Syst. Sci. Disc.*, **3**, 1369–1406.

Roulin, E. and Vannitsem, S. (2005). Skill of medium-range hydrological ensemble predictions. *J. Hydrometeorol.*, **6**, 729–744.

Scherrer, S.C., Appenzeller, C., Eckert, P. and Cattani, D. (2004). Analysis of the spread-skill relations using the ECMWF ensemble prediction system over Europe. *Weather and Forecasting*, **19**, 552–565.

Werner, M., Reggiani, P., De Roo, A., Bates, P. and Sprokkereef, E. (2005). Flood forecasting and warning at the river basin and at the European scale. *Natural Hazards*, **36**, 25–42.