7th International Water Resources Management Conference of ICWRS,
18–20 May 2016, Bochum, Germany, IWRM2016-69-2

Regionalization of post-processed ensemble runoff forecasts

J. O. Skøien[1], K. Bogner[2], P. Salamon[1], P. Smith[3], F. Pappenberger[3]

[1]European Commission – Joint Research Centre, Ispra, 21027 (VA), Italy
[2]Swiss Federal Institute WSL - Mountain Hydrology and Mass Movements, Birmensdorf, 8903, Switzerland
[3]European Centre for Medium-Range Weather Forecasts, Reading, RG2 9AX, United Kingdom

Correspondence to: J. O. Skøien (jon.skoien@jrc.ec.europa.eu)

Abstract. For many years, meteorological models have been run with perturbated initial conditions or parameters to produce ensemble forecasts that are used as a proxy of the uncertainty of the forecasts. However, the ensembles are usually both biased (the mean is systematically too high or too low, compared with the observed weather), and has dispersion errors (the ensemble variance indicates a too low or too high confidence in the forecast, compared with the observed weather). The ensembles are therefore commonly post-processed to correct for these shortcomings. Here we look at one of these techniques, referred to as Ensemble Model Output Statistics (EMOS) (Gneiting et al., 2005). Originally, the post-processing parameters were identified as a fixed set of parameters for a region. Later there were methods for regionalizing the post-processed output, but still with regionally constant parameters (Berrocal et al., 2007). In hydrology, Hemri et al. (2013) extended the method to have temporally consistent parameters between time steps for a single location. Engeland and Steinsland (2014) developed a framework which can estimate post-processing parameters which are different in space and time, but still can give a spatially and temporally consistent output. The application of our work is the European Flood Awareness System (http://www.efas.eu), where a distributed model is run with meteorological ensembles as input. We are therefore dealing with a considerably larger data set than Engeland and Steinsland (2014) which is likely to make their method unfeasible in practice. We also want to regionalize the parameters themselves for other locations than the calibration gauges. Lastly, not all forecasts are available for all lead times, as in their approach. We are therefore testing a slightly different approach, where the post-processing parameters are estimated for each calibration station, but with a spatial penalty for deviations from neighbouring stations, depending on the expected semivariance between the calibration catchment and these stations. The estimated post-processed parameters can then be used for regionalization of the postprocessing parameters also for uncalibrated locations using top-kriging in the rtop-package (Skøien et al., 2006, 2014). We will show results from cross-validation of the methodology and although our interest is mainly in identifying exceedance probabilities for certain return levels, we will also show how the rtop package can be used for creating a set of post-processed ensembles through simulations.

1 Introduction

Ensemble modelling has a long history in meteorology, and is also increasingly used in hydrology, mainly using the meteorological ensembles as forcing. By perturbing the initial conditions or parameters of the model, an ensemble of forecasts is produced, assuming that this is a proxy of the uncertainty of the forecast. However, even if the perturbations are sampled from a probability distribution of the conditions or parameters, it is frequent that the resulting ensembles are both biased (the mean is systematically too low or too high) and wrongly dispersed (the ensemble variance indicates a too low or too high confidence in the forecast, compared with the observations afterwards).

It is therefore common to post-process the forecasts. Two commonly methods are frequently used in meteorology: Bayesian Model Averaging (Raftery et al., 2005), which mainly focuses on calibration, and optimization based on the Ensemble Model Output Statistics (Gneiting et al., 2005), referred to as EMOS. Mostly the EMOS-method is calibrated with the use of Continuous Ranked Probability Score (CRPS), which is an indicator which punishes both biases and dispersion errors.

We will here mainly focus on the EMOS-method. In the original contributions in meteorology, it was common to fit a regional set of parameters for the post-processing. From the post-processed distributions for each location, samples were drawn to generate a post-processed ensemble. These were spatially independent, but Berrocal et al. (2007) extended the methodology to use the spatial structure of the errors to generate a spatially structured covariance matrix which can be used to generate spatially consistent samples, based on the Geostatistic output perturbation technique (Gel et al., 2004). Their method is still using the same set of weights for all locations.

Our application of ensemble forecasting is the European Flood Awareness System (EFAS, http://www.efas.eu), an operational service for flood forecasting in Europe. Different meteorological ensemble forecasts are used for the forecasting. Contrary to the previous applications, we would therefore expect some of the ensembles to be better for some regions, and we do not want a single parameter set for the complete modelling region. Additionally, not all ensembles are available for all lead times, and we would prefer a method which will assure temporal continuity between lead times.

The application of these types of post-processing techniques in hydrology started later. Hemri et al. (2013) developed a method for postprocessing runoff forecasts for individual stations, using the methods of Berrocal et al. (2007) for incorporation of the correlation between lead times. This correlation is likely to be higher for runoff than for meteorological variables. Engeland and Steinsland (2014) presented a method which would fit different weights to different locations and lead times, but still assuming the same number of forecasts for all lead times.

The previous applications in hydrology did not consider forecasts outside the calibration points, similar to what Berrocal et al. (2007) did for meteorological applications. In this paper we will present a methods which will make it easier to make predictions outside calibration points, and also for making simulations of the possible discharge.

2 Data

The analyses in this paper are based on a combination of meteorological forecasts and ensemble forecasts from ECMWF, DWD, COSMO-LEPS and UK Met Office. We use forecasts from a period of almost two years (8 Jan 2012 – 31 Dec 2013). For each day, the forecasts have up to 10 days lead time.

ECMWF: The European Centre for Medium Range Weather produces forecasts for the next 10 days. The forecast from ECMWF is an ensemble with 51 members, in addition to a deterministic forecast

DWD: The German Weather Service produces a deterministic forecast for the next 7 days.

COSMO-Leps: The Cosmo consortium produces an ensemble forecast with 16 members.

UK-MET: The UK Met office produces an ensemble of 24 members.

Each individual forecast has been used as input to the hydrological model LISFLOOD (Van Der Knijff et al., 2010; De Roo et al., 2000), giving an ensemble of runoff values for each forecast day and each lead time. LISFLOOD is a gridded model, which numerically predicts the runoff for each pixel in a 5*5 km grid. The extent of the forecasts and the hydrological model covers most of Europe.

As true values, we are using simulated runoff at 701 stations. The runoff has been simulated from interpolated observed values, using the same model setup of LISFLOOD as for the forecasts. There are some additional stations in the original data set, but these were discarded from the analyses as the runoff appeared to be unreasonably high compared to the estimated basin size, or that some of the forecasts were not available for all lead times and models.

We are using simulated values instead of real observations for comparison, as these will have the same model errors as the forecasts, such as boundary errors and routing errors.

The simulated and forecasted runoff data is divided by catchment area in 1000 km^2. The normalization on area is to work with more area independent values, whereas the use of 1000 km^2 for normalization is to avoid some numerical issues.

3 Method

3.1 Post-processing

The post-processing method we are applying in this paper is based on the Ensemble Model Output Statistics method (Gneiting et al., 2005). Shortly described, the idea is that the mean and variance of a range of forecasts might be biased and wrongly dispersed, so we want to find a weighted mean of the ensemble, whereas the variance can be assumed to fit a regression equation. As we have a combination of deterministic and ensemble forecasts, we will use the deterministic forecasts and the mean of each ensemble forecast for the bias correction, i.e., for a particular station *i* and lead time *l*:

$$Y_{il} = a_{il} + b_{il1}X_{il1} + b_{il2}X_{il2} + \cdots + b_{iln}X_{iln} + e_{il} \quad (1)$$

Where a_{il} is a constant, b_{il1}, …, b_{iln} are weights, e_{il} is an error term averaging to zero, and X_{il1}, ..., X_{iln} are the forecasted variable for this location, deterministic or mean of the ensemble. The deterministic forecasts are from ECMWF and DWD, whereas we use the mean of the ensembles from ECMWF, COSMO and UK Met office, giving n=5 for lead times 1-5. The forecast Y_{il} should then be unbiased, but we would also like to know the variance of e_{il}. This is modelled as a linear function of the variance of all ensembles:

$$\sigma_{il}^2 = Var(e_{il}) = c_{il} + d_{il}S_{il}^2 \quad (2)$$

where S_{il}^2 is the variance of all individual ensemble members, and c_{il} and d_{il} are non-negative coefficients. This gives a Gaussian predictive distribution:

$$N(a_{il} + b_{il1}X_{il1} + b_{il2}X_{il2} + \cdots + b_{iln}X_{iln}, c_{il} + d_{il}S_{il}^2) \quad (3)$$

This can be optimized by minimizing the continuous ranked probability score (CRPS), as described by Gneiting et al. (2005). For each day *d* in the calibration period, the CRPS-error for a certain station *i* and lead time *l* is defined as:

$$crps(F_{ild}, y_{i,l+d}) = \int_{-\infty}^{\infty} [F_{ild}(t) - H(t - y_{i,l+d})]^2 dt \quad (4)$$

Where $H(t - y_{i,l+d})$ is the Heaviside function, which is 0 for $t < y_{i,l+d}$ and 1 for $t \geq y_{i,l+d}$. If F is the CDF of a normal distribution with mean Y_{ild} and variance σ^2, the integral can be replaced with:

$$crps\left[N(Y_{ild},\sigma^2),y_{i,l+d}\right] = \sigma_{ild}\left\{\frac{y_{i,l+d}-Y_{ild}}{\sigma_{ild}}\left[2\Phi\left(\frac{y_{i,l+d}-Y_{ild}}{\sigma_{ild}}\right)-1\right]+2\phi\left(\frac{y_{i,l+d}-Y_{ild}}{\sigma_{ild}}\right)-\frac{1}{\sqrt{\pi}}\right\} \quad (5)$$

where $z=\frac{y_{i,l+d}-Y_{ild}}{\sigma_{ild}}$ is the normalized prediction error and $\Phi(z)$ and $\phi(z)$ represents the CDF and the PDF of a N(0,1) distribution. Equations (4) and (5) can be seen as objective functions when summed over all instances used in the calibration. It is likely that F might violate the normal distribution assumption for runoff variables, however, we will for simplicity use this assumption in this manuscript, and deal with deviations in future work.

3.2 Interpolation of weights

Our region of interest is Europe. We do therefore not expect one set of weights to be sufficient for the whole modelling domain. However, we have the ensembles for all grid cells along a river, and would like to be able to make predictions also for other locations than the calibration locations. The solution is to interpolate the weights along the river network, to have unbiased predictions for each pixel where a prediction is wanted. For this we will use top-kriging (Skøien et al., 2006, 2014). Top kriging is a geostatistical method for interpolation between areas of different spatial support, such as observations along a river network. The method is well explained in the citations above and will only be summarized here for river related applications as follows:

A sample variogram is estimated from the observations for each gauge, as a spatial average if the variable is a spatial aggregate such as runoff and most runoff statistics. The centre of the upstream contributing area is used to compute the distances. Variograms are binned according to the size of each of the catchments, not only distance.

A variogram model is found by jointly fitting regularized variogram values to the binned sample variogram values.

A covariance matrix of expected semivariances between observation catchments and between observation catchments and prediction catchments is found from the variogram model, based on the size and location of the catchments.

Interpolation and cross-validation is performed as in normal kriging, based on the covariance matrices.

The most interesting features of this interpolation method is that it takes into account both network topology (2 locations which are connected on a river network usually gets higher weights than unconnected locations) and spatial proximity. The last feature takes into account both the size and the location of the catchments, not just the distance between the gauges or centres of gravity.

However, the fitting method in Equations (4) and (5) can give poorly correlated weights for neighbouring locations if two or more of the forecasts are highly correlated. For example, if we only had two forecasts and they were equal for a certain location, any combination of weights giving the same sum would give the same error. To force a certain correlation between weights, and variance coefficients between locations (all referred to as parameters below), we use an iterative procedure where we introduce a spatial penalty as a function of the modelled semivariance between two locations and the difference of all the m parameters:

$$S_{pen}=P_c\sum_{i=1}^{n}\frac{1}{\gamma_{0i}}\sum_{j=1}^{m}\frac{2|p_{0j}-p_{ij}|}{|p_{0j}|+|p_{ij}|} \quad (6)$$

Here the parameters of the calibration location is p0j whereas pij is the parameter value for the n locations with the highest correlation to the calibration location. The expected semivariance γ_{0i} between the calibration location and the neigbouring locations is found from a regularized semivariogram for mean runoff. Pc is a penalty coefficient to scale the spatial penalty to the CRPS-error.

The calibration is done station by station. In the first iteration, no spatial penalty is added, as many neighbouring stations have not been computed yet. In the second iteration, Pc is set equal to 1. At the end of the second iteration, this coefficient is recomputed as two times the ratio between the CRPS-error and the spatial penalty. This Pc is used for the third iteration. In the calibration, the most recent parameter values are always used, i.e., if a neighbor has already been updated, this value is used instead of the one from the previous iteration. The locations are visited in a random order for each iteration.

3.3 Simulations of runoff

A common usage of post-processing in meteorology is to create simulations of the variable of interest. This can also be done with the post-processing we are presenting here, based on the calibrated parameters and the semivariogram above. The simulation method is based on the Sequential Gaussian Simulation method (Deutsch and Journel, 1998), combined with Kriging with uncertain data (KUD) (de Marsily, 1986; Merz and Blöschl, 2005).

We start with the weighted mean and uncertainties for each calibration location.

In a random order, we visit all calibration locations and prediction locations, and do the following step for each of them:

1. For a new location, we predict the mean and the kriging variance, using the weighted mean for the calibration locations, and previously simulated locations as observations. For the KUD prediction, we use the weighted ensemble variance for the calibration locations.
2. Sample a value from the predictive distribution (traditionally assumed to be Gaussian) with the prediction as mean and the kriging variance as variance. Add this to the set of observed/simulated values. This simulated value will in the subsequent simulations have an uncertainty of zero in the KUD prediction.
3. Replace the weighted mean with a simulated value if the simulation concurs with a calibration location.

Simulation of runoff values implies some numerical challenges. First of all, when we have multiple points along a river segment, the contributing area of these points might be almost similar in many cases. This can create highly correlated neighbours, which can again create singular or close to singular covariance matrices. We have found some methods to automatically remove some of these neighbours, still there might be some numerical challenges in solving the kriging equations. There are therefore some cases where numerical issues can give a negative kriging variance, which is first of all physically impossible, second, makes it impossible to draw a value above. We are still in the process of finding robust solutions for these cases, in the meantime there will be some points where we are not able to make simulations.

A second issue is that runoff values are typically above zero. Using random sampling from a Gaussian distribution can give negative observations. We are therefore instead assuming a long-normal distribution in this case, log-transforming the predictive mean and variance with σ_l^2 as the logtransformed variance: $\sigma_l^2 = \log(\frac{\sigma^2}{\mu} + 1)$ before sampling. This ensures positive runoff values, but should be seen as an approximation to the correct solution and is likely to be slightly biased (Clark, 1998). We will further investigate better approaches for this case.

4. Results

4.1 Fitting of EMOS-parameters and effect of spatial penalty

The initial fitting of the parameters are done without the spatial penalty. We can therefore easily see how much the use of the penalty increases the CRPS-error and how we reduce the spatial errors as in Figure 1. The CRPS-error increases marginally for all catchments, but the largest increase is less than 25 percent and 75% of the catchments increase less than 5%. The spatial penalty reduces considerably with the iterations, 55% of the catchments see the spatial error reduced to less than 1/2. The last panel shows that the CRPS-error is dominating the total error for most catchments, and is considerably larger for a large group of them, whereas there are some (around 30%) where the spatial error dominates.

We can also notice that there is quite a large range in the errors, both CRPS and spatial error in approximate range from 1-1000. We have not analyzed the reason for this, although it is likely not related to area. First, the runoff has been divided by catchment area, second, we have plotted (not shown) both errors and ratios between errors against catchment area, without finding any strong relationships.

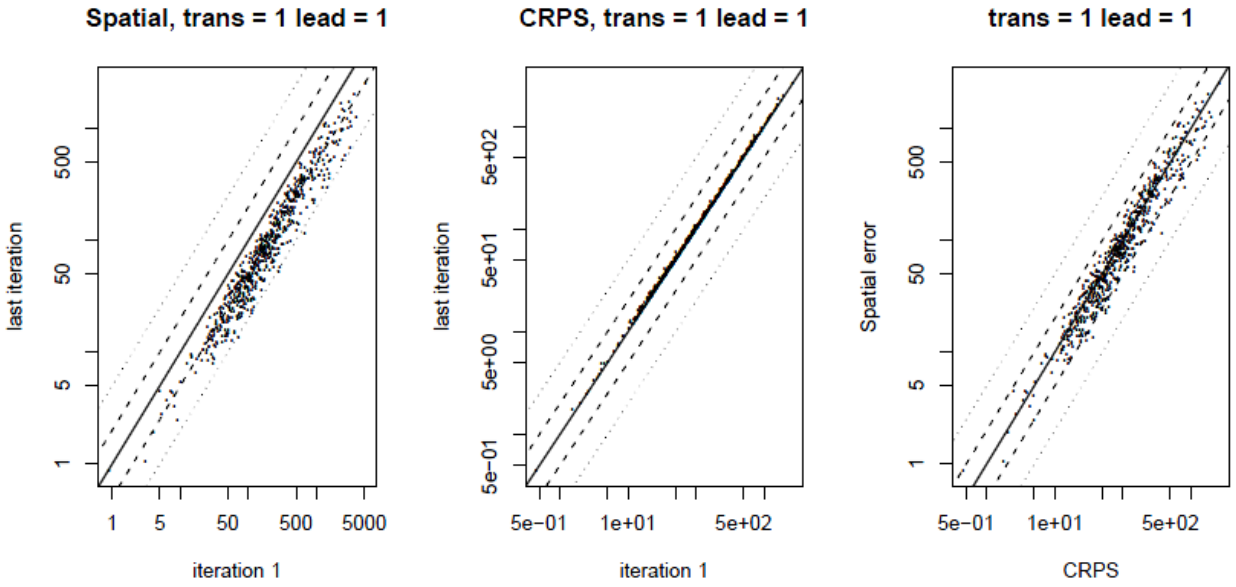

Figure 1. Left and central panel: Development of CRPS-errors and spatial errors from first to last iteration for lead time of one day. Right: Comparison of CRPS and spatial error after last iteration. Solid line represents 1:1, whereas dashed lines represent 2:1 and 1:2 and stapled lines 5:1 and 1:5.

4.2 Interpolation of EMOS-parameters

Table 1 gives an overview of r^2 of the cross-validated EMOS-parameters from Equations (1) and (2). The variable names in the equations are given in brackets for the column names. We can see that there is a good correspondence between the fitted parameters and the interpolated parameters. Some of the cells are given a color code, where the one red cell has $r^2 < 0.6$, there is one orange cell with $r^2 < 0.7$, three yellow cells with $r^2 < 0.8$ and 8 green cells with $r^2 < 0.9$. The remaining 59 parameters have $r^2 > 0.9$. This means that the top-kriging method can well be used for interpolating EMOS-parameters between different locations on the stream network, at least when the parameters have been fitted using a spatial penalty, as we have done in this manuscript. We have not yet examined in detail the reasons for the poorer results for a few locations and lead times.

Lead	Var1 (c_{il})	Var2 (d_{il})	Intercept (a_{il})	DWD (b_{il1})	ECMWF (b_{il2})	ECMWF mean (b_{il3})	COSMO mean (b_{il4})	UK mean (b_{il5})
1	0.98	0.74	0.90	0.92	0.96	0.93	0.98	0.90
2	0.98	0.88	0.85	0.92	0.95	0.93	0.95	0.93
3	0.98	0.89	0.90	0.92	0.87	0.93	0.95	0.95
4	0.98	0.95	0.94	0.92	0.92	0.94	0.94	0.97
5	0.98	0.94	0.92	0.91	0.88	0.95	0.94	0.97
6	0.91	0.57	0.88	0.76	0.65	0.79	NA	0.80
7	0.96	0.97	0.96	0.94	0.91	0.94	NA	0.97
8	0.97	0.97	0.85	NA	0.94	0.94	NA	0.97
9	0.96	0.97	0.97	NA	0.96	0.95	NA	0.97
10	0.97	0.96	0.97	NA	0.94	0.94	NA	0.97

Table 1. Cross-validation r^2-values for EMOS-parameters for different lead times. Background colors of cells refer to the cell-value, where the colors red, orange, yellow and green refer to values below 0.6, 0.7, 0.8 and 0.9, respectively.

4.3 Simulation of runoff fields

With the fitted parameters, we have an estimate of the predictive mean and uncertainty for each of the calibration locations. From these, we can simulate the specific runoff for each pixel along the river network. Figures (2) and (3) show the results from 4 simulations for the first and the 10th forecast day, respectively, based on forecasts from February 17th for a region on the German-Polish border. The forecast indicates a flood event to the end of this period (10th forecast day), so the predictions and the simulations are considerably higher for Figure 3 than for Figure 2. The dots show the predictive mean for the calibration locations based on the fitted EMOS-parameters, whereas the pixels show the simulated values based on the variogram and the predictive uncertainty from the calibration locations. We can see that the simulations are relatively close to the predicted mean for locations close to the calibration locations, whereas the deviations between simulations can be considerably larger in the smaller tributaries far from the calibration locations.

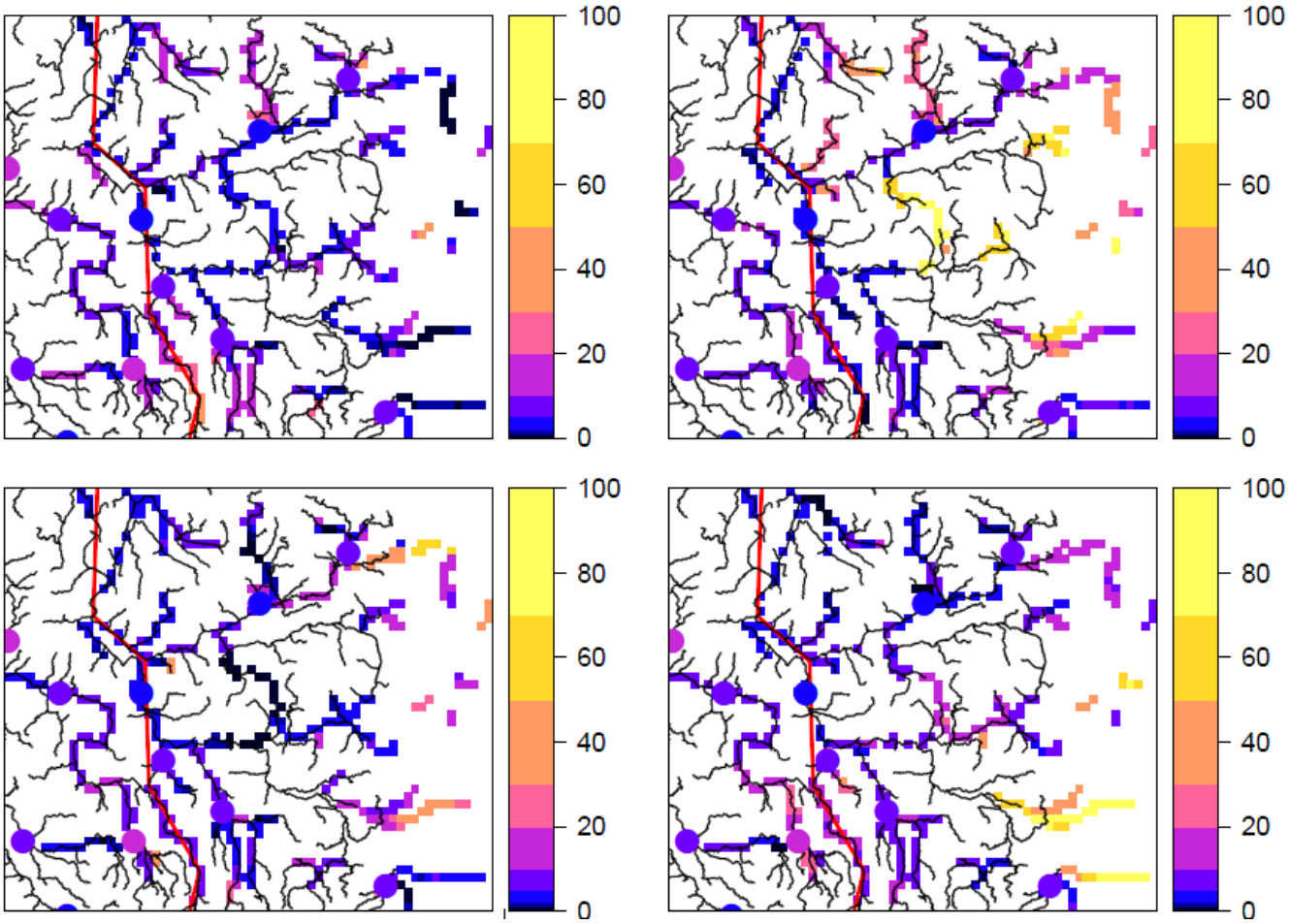

Figure 2. Predicted (dots) and simulated specific runoff (pixels) for one day lead time for a region on both sides of the German/Polish border (red line).

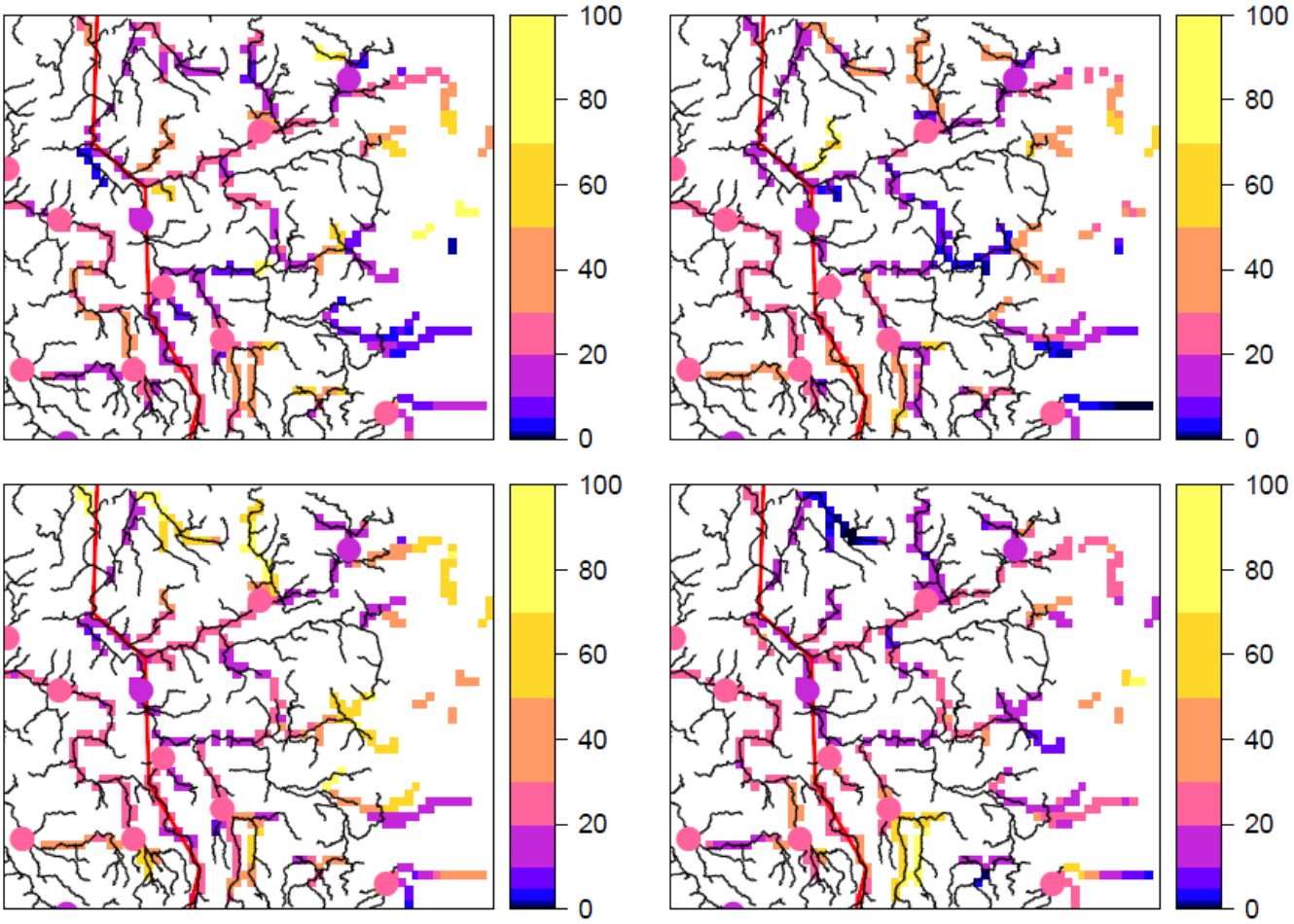

Figure 3. Predicted (dots) and simulated specific runoff (pixels) for 10 days lead time for a region on both sides of the German/Polish border (red line).

5. Conclusions

We have used the EMOS-method for post-processing of runoff predictions from an ensemble forecasting method. The results indicate that it is well possible to use top-kriging for interpolating the EMOS-parameters along the river network as long as the parameters have been fitted with a method which forces some degree of spatial continuity between the parameters.

We have also shown that it is possible to use top-kriging for simulation of runoff at uncalibrated locations, using the variogram and post-processed predictive distributions at the calibration locations. Using these simulations is a different approach than interpolating the EMOS-parameters to create uncorrelated predictive distributions for each locations along the river network. Such simulations have, as far as we know, not yet been used in forecasting, and the possible usages still need further analyses. One important aspect is that the uncertainty of the smaller tributaries will not only be based on meteorological uncertainty, as for ensemble modelling with a hydrological model with a single parameter setup, it will also include the modelling uncertainty.

We notice that there are still a few issues which have to be further improved in the analyses presented here. First of all, much of the theory is developed for variables with a normal distribution. However, runoff usually does not follow a normal distribution. We will in the near future analyse the possibilities for using transformations to be able to work with more

normalized variables. We did use a lognormal transform for the simulations. However, the way it was done is not well founded in geostatistical theory, and will need further improvements. Some of the simulated values in the tributaries are extremely large, which can well describe the statistical uncertainty, but maybe not so much the meteorological uncertainty. Further comparisons of the simulations here and the pixel results from a distributed ensemble model will be necessary.

Berrocal, V. J., Raftery, A. E. and Gneiting, T.: Combining Spatial Statistical and Ensemble Information in Probabilistic Weather Forecasts, Mon. Weather Rev., 135(4), 1386–1402, doi:10.1175/MWR3341.1, 2007.

Clark, I.: Geostatistical estimation and the lognormal distribution. Geocongress. Pretoria, RSA., [online] Available from: Available at: http://uk.geocities.com/drisobelclark/resume/papers/Geocongress1998.zip (verified 01. Nov. 2006), 1998.

Deutsch, C. V. and Journel, A. G.: GSLIB: geostatistical software library and user's guide. Second edition, Oxford University Press; Applied Geostatistics Series. [online] Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-0031916947&partnerID=tZOtx3y1, 1998.

Engeland, K. and Steinsland, I.: Probabilistic postprocessing models for flow forecasts for a system of catchments and several lead times, Water Resour. Res., 50(1), 182–197, doi:10.1002/2012WR012757, 2014.

Gel, Y., Raftery, A. E. and Gneiting, T.: Calibrated Probabilistic Mesoscale Weather Field Forecasting, J. Am. Stat. Assoc., 99(467), 575–583, doi:10.1198/016214504000000872, 2004.

Gneiting, T., Raftery, A. E., Westveld, A. H. and Goldman, T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, Mon. Weather Rev., 133(5), 1098–1118, doi:10.1175/MWR2904.1, 2005.

Hemri, S., Fundel, F. and Zappa, M.: Simultaneous calibration of ensemble river flow predictions over an entire range of lead times, Water Resour. Res., 49(10), 6744–6755, doi:10.1002/wrcr.20542, 2013.

Van Der Knijff, J. M., Younis, J. and De Roo, A. P. J.: LISFLOOD: a GIS- based distributed model for river basin scale water balance and flood simulation, Int. J. Geogr. Inf. Sci., 24(2), 189–212, doi:10.1080/13658810802549154, 2010.

De Marsily, G.: Quantitative hydrogeology, Academic Press Inc., London., 1986.

Merz, R. and Blöschl, G.: Flood frequency regionalisation - Spatial proximity vs. catchment attributes, J. Hydrol., 302(1-4), 283–306, 2005.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M.: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, Mon. Weather Rev., 133(5), 1155–1174, doi:10.1175/MWR2906.1, 2005.

De Roo, A. P. J., Wesseling, C. G. and Van Deursen, W. P. A.: Physically based river basin modelling within a GIS: The LISFLOOD model, in Hydrological Processes, vol. 14, pp. 1981–1992, John Wiley & Sons Ltd. [online] Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-0034254644&partnerID=tZOtx3y1, 2000.

Skøien, J. O., Blöschl, G., Laaha, G., Pebesma, E., Parajka, J. and Viglione, A.: rtop: An R package for interpolation of data with a variable spatial support, with an example from river networks, Comput. Geosci., 67, 180–190, doi:10.1016/j.cageo.2014.02.009, 2014.

Skøien, J. O., Merz, R. and Blöschl, G.: Top-kriging - geostatistics on stream networks, Hydrol. earth Syst. Sci., 10, 277–287, 2006.