Meteorol Atmos Phys (2014) 123:17–31
DOI 10.1007/s00703-013-0291-3

ORIGINAL PAPER

A sequential ensemble prediction system at convection-permitting scales

Marco Milan · Dirk Schüttemeyer · Theresa Bick · Clemens Simmer

Received: 23 October 2012 / Accepted: 5 October 2013 / Published online: 17 November 2013

Abstract A sequential data assimilation approach (SAM) that incorporates elements of particle filtering with resampling (SIR, Sequential Importance Resampling) is introduced. SAM is applied to the COSMO-DE-EPS, which is an ensemble prediction system for weather forecasting on convection-permitting scales. At the convective scale and beyond, the atmosphere increasingly exhibits non-linear state space evolutions. For an ensemble-based data assimilation system, this requires both an adequate metric that quantifies the distance between the observed atmospheric state and the states simulated by the ensemble members, and a methodology to counteract filter degeneracy, i.e. the collapse of the simulated state space. We, therefore, propose a combination of resampling, which accounts for simulated state space clustering, and nudging. SAM differs from the classical SIR approach mainly in the weighting applied to the ensemble members. By keeping cluster representatives during resampling, the method maintains the potential for non-linear system state development. With three convective case studies, we demonstrate that SAM improves forecast quality compared with the control EPS (EPS without data assimilation) for the first 5–6 h of forecast.

Responsible editor: F. Mesinger.

M. Milan (✉) · C. Simmer
Meteorologisches Institut, University of Bonn, Auf dem Hügel, 20, 53121 Bonn, Germany
e-mail: marco.milan@univie.ac.at

Present Address:
M. Milan
Institut für Meteorologie und Geophysik, University of Vienna, Vienna, Austria

D. Schüttemeyer
European Space Agency, European Space Research and Technology Center (ESA-ESTEC), EOP/SMS, 2201 AG Noordwijk, The Netherlands

T. Bick
Hans-Ertel-Centre for Weather Research, Atmospheric Dynamics and Predictability Branch, Meteorologisches Institut, University Bonn, Bonn, Germany

1 Introduction

In addition to a high-quality model, an accurate image of the initial state of the weather system based on the observations and on a weather forecast model (analysis), is a widely accepted prerequisite for meaningful weather forecasts (Talagrand 1997). As shown in Lorenz (1963a, b), forecast performance additionally depends on flow instabilities, which cause chaotic behaviour and a finite limit of predictability. Predicted states are then extremely sensitive both to model formulations and initial state, and differences among predicted and initial state can amplify during model integration; this behaviour depends on the weather situation and can be exponential (Yoden 2007). Moreover, errors on smaller scales may introduce errors on larger scales, a behaviour, which is known as the inverse error cascade (Leith 1971). Thus the on-going spatial resolution enhancements of weather forecast models, especially for short-range weather forecasting, may also lead to increasing forecast errors because of the additionally simulated small-scale processes; an example is the necessity of new radiation parameterisation schemes in case of high resolution (Müller and Scherer 2005). At the synoptic scale, prediction errors are usually assumed to be Gaussian-distributed. At higher resolutions many processes are simulated directly—instead of being parameterised—and exhibit increasingly non-linear behaviour, e.g. convective events, which often seem to happen in a quasi-random

fashion. The connection between spatial scale and prediction skill is evident in many theoretical and experimental studies (Lorenz 1969; Zepeda-Arce et al. 2000; de Elia et al. 2002; Casati et al. 2004).

The Kalman Filter (KF) provides a variance-minimising solution, but only for quasi-linear system evolutions (Jazwinski 1970). Due to non-linearities, an initially Gaussian probability density function (PDF) will evolve into a non-Gaussian PDF. To handle this behaviour various extensions of the Kalman filter have been formulated; they are, however, not optimal (Kalman 1960; Dee 2005).

The so-called Extended Kalman Filter (EKF) can handle weakly non-linear model behaviour. The error is assumed to evolve according to the tangent linear model, which is derived from the perfect model (van Leeuwen 2003). The assumption of a perfect model also applies to 4-dimensional variational data assimilation (4DVar). Originally, 4DVar schemes assume a Gaussian distribution of observational errors, but recent approaches extend the method to combinations of Gaussian, lognormal observational errors, and to mixed background errors (Fletcher 2010). Thus, both methods—EKF and 4DVar—might fail on the convective scale, where more general PDFs might apply.

Particle Filter (PF) methods (van Leeuwen 2009) take full account of non-linear state developments. The PF, also termed Sequential Monte-Carlo filter (SMC, Doucet et al. 2001), represents the model PDF by a number of randomly selected ensemble members, or particles. The posterior PDF is approximated by $f_{\mathrm{m}}(\psi\ |\mathbf{d})$, where ψ is the model state, and $\mathbf{d}$ are the available observations. SMC (PF) methods appeared in the 1950s (Hammersley and Morton 1954), but were not applied in weather forecasting, probably due to the lack of computing capacity at that time (Doucet et al. 2001). Bird (1978) identified adequate computer technology as a requirement for the use of Monte Carlo methods.

In this work, we introduce a sequential data assimilation approach (SAM), applied to the COSMO-DE-EPS (COSMO, COnsortium for Small scale MOdelling), which is an ensemble prediction system for weather forecasting on convection-permitting scales developed by the German Weather Service (Deutscher Wetterdienst, DWD). SAM combines elements of PF with resampling (SIR, Sequential Importance Resampling), originally termed 'bootstrap filter'.

Following the work of Gordon et al. (1993), and further development by van Leeuwen and Evensen (1996), van Leeuwen (2001, 2003), PF with resampling weights ensembles members according to the probability of available observation, given the state of the ensemble members. Ensemble members with low weights are abandoned, and the original number of members is then restored by adding multiple copies of ensemble members with high weights to construct the posterior PDF (van Leeuwen 2009).

SAM diverges from PF primarily by the weight definition. While PF defines weights by the observation probability given the model state, SAM introduces a metric that approximates the distance between observations and ensemble member state. SAM assumes that the "closeness" of ensemble members states to observations is a strictly monotonic function of the relative importance of ensemble members in the probability density of the observation given the model state; this property assures the correct ranking of the ensemble members.

A well-known problem for PF methods, when applied to high-dimensional systems, is *filter degeneracy*, i.e. a situation when most of the ensemble members have weights close to zero meaning that none or only a small number of the ensemble members have a considerable probability given the observations (Bengtsson et al. 2008; Snyder et al. 2008).

In our approach, we attempt to reduce filter degeneracy via two remedies: first, by clustering prior to filtering and resampling and second, by creating different model evolutions of the multiple copies using a nudging method based on the ratio of observed and modelled precipitation rates.

Ensemble members are clustered according to their mutual similarity; which is expressed by their closeness to observations and quantified by a metric. Moreover, we require that at least one cluster member (even if distances to observations are large) survives filtering. Thus, the ensemble members belonging to the less probable clusters will be discouraged, but the cluster survives at least as a single member. Members with initially low probability given the observations may thus still evolve into a state with higher probability in a subsequent filter time, which might mimic non-linear system state developments (e.g. sudden convection initiation). This approach might also reduce timing errors for convection due to model errors and/or imperfect initial conditions. Only the ensemble members with the highest weights are duplicated (one additional member is created), taking the cluster into account. While the original ensemble member evolves further according to the forward model, its twin is nudged to radar and satellite observation using PIB (Physical Initialisation Bonn, Sect. 3.3).

In Sect. 2, we discuss the PF in more detail. Section 3 describes the applied numerical weather prediction model and the observations used. In Sect. 4 the sequential data assimilation method SAM is explained in more detail. We illustrate, in Sect. 5, the synoptic situations for which we applied the method, and we evaluate its impact in Sect. 6 based on skill scores, reliability curves and other statistical methods. We also compare SAM performance with the EPS without filtering, with the EPS where nudging via PIB is applied to all members, and with a more SIR-like method without resampling (FILTER). Conclusions are drawn in Sect. 7.

2 Particle filter

A PF estimates the posterior PDF of a model state given the observations. An ensemble is considered as a representation of the state PDF by the discrete set of model states represented by the ensemble members. The PF assigns weights to the ensemble members according to their closeness to observations. Generally, weights quantify the probability of a model state (ψ) given the observation ($\mathbf{d}$), $f_{\mathrm{m}}(\psi \mid \mathbf{d})$.

For non-linear dynamics, a variance-minimising filter can be derived using Bayes' theorem (van Leeuwen 2003). In Bayesian statistics the unknown model state evolution is represented by the value of a random (multi-dimensional) variable ψ. Using Bayes' theorem, the model state probability density $f_{\mathrm{m}}(\psi)$, is used to find $f_{\mathrm{m}}(\psi \mid \mathbf{d})$, the posterior probability density of ψ given the observations $\mathbf{d}$ (van Leeuwen and Evensen 1996):

$$f_{\mathrm{m}}(\psi|\mathbf{d}) = \frac{f_{\mathrm{d}}(\mathbf{d}|\psi) f_{\mathrm{m}}(\psi)}{f_{\mathrm{d}}(\mathbf{d})} \tag{1}$$

The definition of the probability of the observation $f_{\mathrm{d}}(\mathbf{d})$ is usually assumed to be the marginal probability density of the joint probability density of model states and observations:

$$f_{\mathrm{d}}(\mathbf{d}) = \int f_{\mathrm{d}}(\mathbf{d}, \psi)\, \mathrm{d}\psi = \int f_{\mathrm{d}}(\mathbf{d}|\psi) f_{\mathrm{m}}(\psi)\, \mathrm{d}\psi \tag{2}$$

The variance of an estimate characterises the accuracy of the estimation and the spread of the probability density. The variance-minimising model evolution is equal to the weighted state mean, based on the posterior probability density:

$$\bar{\psi} = \int \psi f_{\mathrm{m}}(\psi|\mathbf{d})\, \mathrm{d}\psi \tag{3}$$

Using discrete probability frequencies, and assuming that all ensemble members have equal a priori probability, we obtain:

$$\bar{\psi} = \frac{\sum_{i=1}^{N} \psi_i f_{\mathrm{d}}(\mathbf{d}|\psi_i)}{\sum_{i=1}^{N} f_{\mathrm{d}}(\mathbf{d}|\psi_i)}, \tag{4}$$

where N is the ensemble size. Thus each ensemble member is weighted by the observation probability given the model state as true (van Leeuwen 2009). This leads to weights given by:

$$w_i = \frac{f_{\mathrm{d}}(\mathbf{d}|\psi_i)}{\sum_{i=1}^{N} f_{\mathrm{d}}(\mathbf{d}|\psi_i)}. \tag{5}$$

The classical PF with resampling (Rubin 1988; Gordon et al. (1993) defines a weight density distribution, and randomly samples from this distribution a sequence of weights, including the ensemble member carrying the weight. Thus the ensemble of drawn ensemble members constitutes the posterior probability density $f_{\mathrm{m}}(\psi \mid \mathbf{d})$. The probability of a weight to be drawn is dependent on its value; larger weights have higher probability to be drawn than lower weights. Due to the discrete ensemble, a particular weight can be drawn several times, and the number of times a weight (and therefore the associated ensemble members) is drawn is equal to the number of identical copies that are made of that ensemble member.

In the new ensemble, all ensemble members have again equal weights. Usually, the resampling restores the total number of particles N. The particle state ψ is then integrated forward in time, by the model f until the next observation time (from observation time $n-1$ to n):

$$\psi^n = f(\psi^{n-1}) + \beta, \tag{6}$$

where β is the stochastic model error, which ensures different evolutions of initially identical copies. Ways of resampling have been developed to avoid filter degeneracy, such as the weight resampling filter from Kim et al. (2003) and the sequential importance resampling and filtering (SIRF) from van Leeuwen (2003), which will be discussed below. Pham (2001) applied jitter to ensemble members with multiple identical copies, which increases ensemble spread and combats filter degeneracy. Jittering adds noises to the ensemble members but does not change the ensemble member weights (van Leeuwen 2009). Still other approaches apply weighting and resampling at different times, e.g. the auxiliary PF (Pitt and Shephard 1999) and the guided SIR (van Leeuwen 2002).

SIR leads to insufficient spread (van Leeuwen 2003), either when system noise provides insufficient state spread or when the ensemble badly approximates the true prior distribution (i.e. when the distance between the best member and the true state is too large). In a high-dimensional state space, the observational state including its PDF covers only a small region of the state space; leading to possibly large distances between ensemble members and the observational state. In order to account for this mismatch during resampling van Leeuwen (2009) uses weights influenced by a so-called proposal density (q):

$$w_i = \frac{1}{A} f_{\mathrm{d}}(\mathbf{d}|\psi_i) \frac{f_{\mathrm{m}}(\psi_i)}{q(\psi_i|\mathbf{d})}, \tag{7}$$

where A is a normalization factor. q is related to a proposed model, and the observation can be included in this density (see Eq. 7). For example, using the PDF from an Ensemble Kalman Filter (EnKF) as proposal density, the particles can be "pushed" to positions in state space where the probability of the particle given the observations is large (van Leeuwen 2010). This solution can be insufficient when the dimension of the state space is too large to be sufficiently covered by the ensemble members.

In our sequential data assimilation method (SAM, see Sect. 4.4), we address nowcasting and short-term weather prediction on convection-permitting scales. Accordingly, forward integration of ensemble member copies, which were created during resampling, must lead to different state space evolutions in time periods much shorter than an hour. EOFs-based (empirical orthogonal functions) or breeding methods (Toth and Kalnay 1993, 1997) could be used but are not advisable for these scales. Breeding perturbations usually result in a negligible spread at short time intervals. Singular vectors, using a tangent linear model, are better approximations of the fastest growing modes, but there is no guarantee that the actual errors will be projected to a significant portion on those modes. The limits of both methods are discussed in Bowler (2006). Approaches following the classical SIR will discourage too many ensemble members by creating multiple copies of only the few best performing ensemble members and lead to low representativeness. SAM inserts only one additional copy of ensemble members with higher weights, and a nudging method (PIB) is applied in order to both add spread and to move the ensemble PDF towards the observations.

3 Data and model

3.1 The COSMO-DE ensemble prediction system (EPS)

We base our study on a pre-operational version of the COSMO-DE Ensemble Prediction System (COSMO-DE EPS, Gebhardt et al. 2011) with 20 members. For our purposes the quality of the original ensemble is important but not a priority; our goal is not to test the quality of a particular given EPS. Rather, our goal is to develop and test a data assimilation method for a high-dimensional system for short-range forecast on convection-permitting scales.

COSMO-DE, which COSMO-DE EPS is based on, is a non-hydrostatic limited area atmospheric weather prediction model, developed by the COnsortium for Small-scale MOdeling (COSMO) led by the German Meteorological Service (Deutscher Wetterdienst, DWD). The ensemble runs in this paper are based on model version 4.11 (see Schulz and Schättler 2005; Baldauf et al. 2011). The model has a spatial resolution of approximately 2.8×2.8 km over an area covering 421 grid cells in the longitudinal direction and 461 grid cells in the latitudinal direction. The atmosphere is vertically resolved into 50 terrain following layers. A Runge–Kutta scheme with a time step of 25 s is used for the numerical integration in time. The scheme is of third order, except for the horizontal advection of water components in the microphysics where a fifth order is used. We switch off the deep convection parameterisation scheme and thus assume that large-scale convection leading to precipitation is sufficiently resolved, while the parameterisation for shallow convection is retained. COSMO-DE has been operational at DWD since April 2007.

The 20 members of the COSMO-DE EPS version applied here have the same initial conditions, but differ by the boundary conditions and the sets of parameterisation during integration. Boundary conditions are taken from a short Range Ensemble Prediction System (SREPS), the so-called AEMet-SREPS (Garcia-Moya et al. 2007) developed by the Spanish weather service (AEMet). AEMet-SREPS is created by driving different regional models, including the COSMO model, each with a spatial resolution of 25 km laterally by the output of four global models: the Integrated Forecast System (IFS, Jakob et al. 1999) from the European Centre for Medium-range Weather Forecast (ECMWF), the Global Model (GME, Majewski et al. 2002) from DWD, the Global Forecast System (GFS, Sela 1980) from the National Centre for Environmental Predictions (NCEP), and the Unified Model (UM, Cullen 1993) from the United Kingdom Meteorological Office (UKMO).

ARPA-SIM in Bologna (Marsigli et al. 2008) nests the COSMO model with a spatial resolution of 10 km into the output of the four COSMO runs of AEMet-SREPS. The 16-member COSMO-SREPS is generated by taking combinations of four AEMet-SREPS runs as initial and boundary conditions with four different settings of the physical parameterisations. The 20 members of COSMO-DE EPS are generated as follows: four members of COSMO-SREPS are selected each with boundary conditions of a different global model. These four members are then integrated using the five different physical parameterisation schemes of COSMO-DE, kept constant over the entire forecast time. See Paulat et al. (2009) for a complete description of COSMO-DE EPS.

3.2 Radar and satellite data

Radar and satellite data are used in the resampling part of SAM. The duplicated ensemble members are nudged via PIB (see Sect. 3.3 and Milan et al. 2008 for more details). PIB requires estimates of surface rainfall and cloud top heights, which are extracted from a quality-controlled rain rate product of DWD (so-called RY) and from products of the Satellite Application Facility on support to NoWCasting and very short-range forecasting (SAFNWC), respectively.

RY is based on the lowest elevation scans of the 16 German operational C-Band Doppler radars (example in Fig. 1). The product is derived from scans with elevation angles between 0.5° and 1.8°, which are combined to

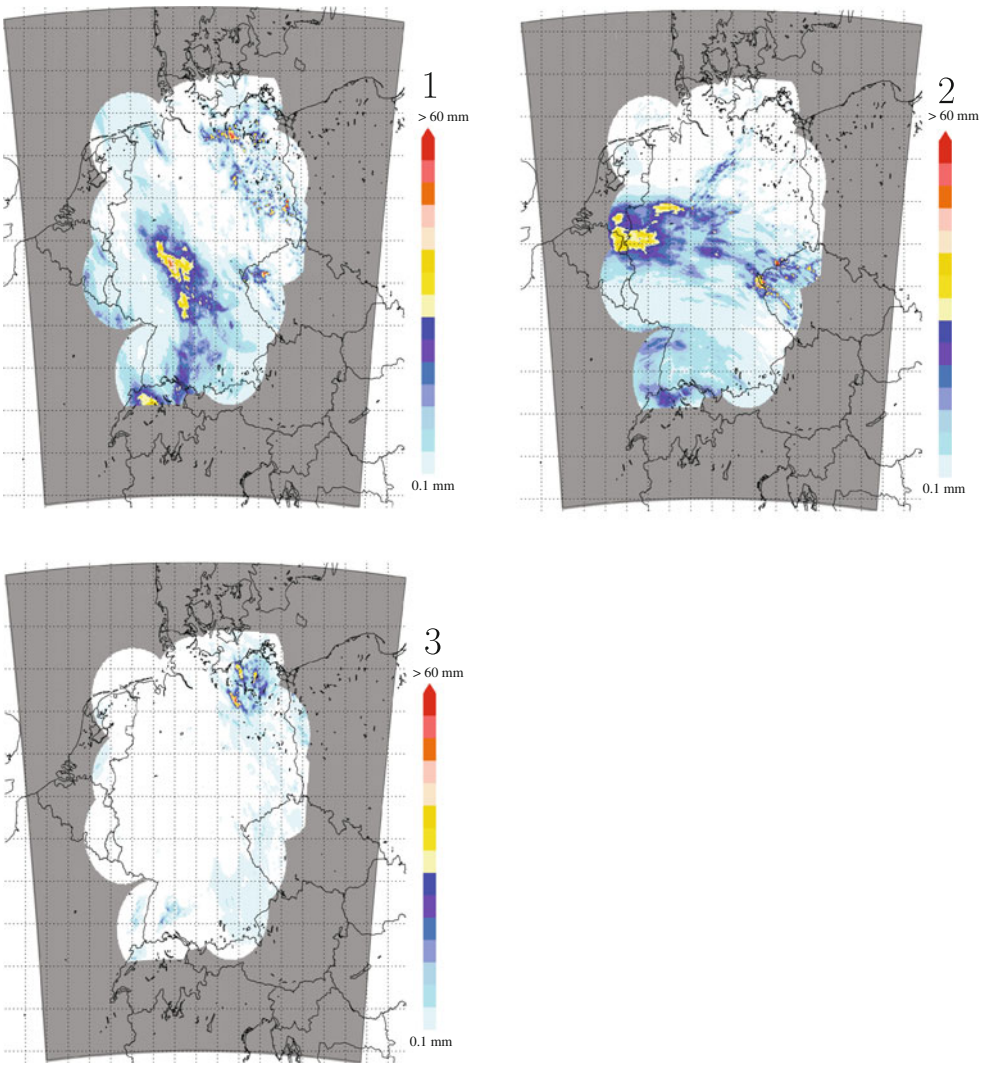

Fig. 1 RADAR precipitation sum (mm) during the analysed period (00–16 UTC). **1** *CASE 1* 8 August 2007, **2** *CASE 2* 9 August 2007, **3** *CASE 3* 12 August 2007

minimise blocking by orography. The maximum range for each radar is 128 km. For more information, see the description of the DWD weather radar network on http://www.dwd.de. Typical problems with radar data, such as anomalous propagation and attenuation, are filtered out by DWD. Compositing is achieved by selecting the radar observations closest to the ground. The original product has a temporal resolution of 5 min and a spatial resolution of 1 km.

Cloud top heights are derived from the SAFNWC products (http://nwcsaf.inm.es/) for cloud top temperature and height (CTTH) and from the cloud type product (CT). SAFNWC products are obtained from DWD with temporal resolution of 15 min. For further information see SAFNWC (2004).

3.3 PIB

PIB was originally derived by Haase et al. (2000) and developed into its current form by Milan et al. (2008) and Milan (2010). PIB nudges vertical wind profiles (w in m/s), specific water vapour (q_v in kg/kg), cloud water content (q_c in kg/kg) and cloud ice content (q_i in kg/kg) to the observations. Only a short description is given here; for a comprehensive explanation, see the cited articles.

First a radar-based surface precipitation field is estimated for every model time step. Nudging is initiated when model precipitation and radar-observed precipitation differ by more than 20 % at a grid point, which approximates the uncertainty of radar-based precipitation estimates. For every grid point with estimated precipitation rate above 0.1 mm/h for which the upper condition applies, a simple single-column cloud and precipitation model is used to adjust the simulated cloud base and top height, and the profiles of vertical wind and humidity. At grid points with estimated precipitation rates below 0.1 mm/h, PIB reduces the water vapour content, the cloud water content and the cloud ice content based on satellite information.

Based on an identical twin experiment, Milan et al. (2008) showed PIB's ability to maintain the main features of model storm evolutions both during the assimilation window and during the free forecast. Aside from precipitation evolution, the tests also investigated CAPE, cloud top, cloud bottom and mass flux convergence near the cloud base. Real data experiments (Milan 2010) covered 1-month simulations with initialisation every 8 h (at 00, 08, 16 UTC). PIB forecasts were compared with LHN forecasts (Latent Heat Nudging, Stephan et al. 2008) and with a forecast without radar data assimilation (CONTROL). While CONTROL had the tendency to underestimate precipitation, especially for stronger rain rates, PIB and LHN succeeded in reducing this error for convective situations and showed similar skills.

4 Resampling and filtering

4.1 Metric

In our approach, we borrow from PF the idea of weighting particles according to their closeness to the observations. In contrast to PF, where the weights are drawn from the posterior PDF, the metric used in SAM is based on two Objective Skill Scores (OSS, for a description of various skill scores see Jolliffe and Stephenson 2004), which quantify the difference between model-generated precipitation ψ and radar-estimated precipitation $\mathbf{d}$ interpolated to the model space ($H^{-1}(\mathbf{d})$). The inverse of the radar and/or satellite observation operators H are, however, never actually calculated. The applied OSS are variants of the Frequency Bias (FBI) and the Equitable Threat Score (ETS). The modified frequency bias ($\mathrm{FBI}_{\mathrm{mod}}$) introduced by Weusthoff et al. (2011),

$$\mathrm{FBI}_{\mathrm{mod}} = \begin{cases} 1 - \frac{1}{\mathrm{FBI}}: & \mathrm{FBI} > 1 \\ 1 - \mathrm{FBI}: & \text{else} \end{cases} \tag{8}$$

ranges between 0 and 1, where 0 is the perfect score. The ETS has a range between −1/3 and 1 (with perfect score 1), but we choose a modified ETS ($\mathrm{ETS}_{\mathrm{mod}}$) to have a score with the same range as $\mathrm{FBI}_{\mathrm{mod}}$:

$$\mathrm{ETS}_{\mathrm{mod}} = (1 - \mathrm{ETS}) \cdot \frac{3}{4} \tag{9}$$

Our metric M combines both scores via

$$M(\psi, H^{-1}(\mathbf{d})) = (\mathrm{ETS}^2_{\mathrm{mod}} + \mathrm{FBI}^2_{\mathrm{mod}})^{\frac{1}{2}} \tag{10}$$

For a perfect model forecast M is 0.

4.2 Clustering

Prior to filtering, we group the ensemble members using hierarchical clustering analysis (Wilks 2006) based on the metric M. Given N ensemble members, the method starts with N clusters each containing one ensemble member. The two closest clusters (given the metric) merge into one new cluster and reduce the number of clusters from N to $N - 1$. The procedure is iterated until all ensemble members are in a cluster, creating a series of cluster sets with increasing levels of aggregation.

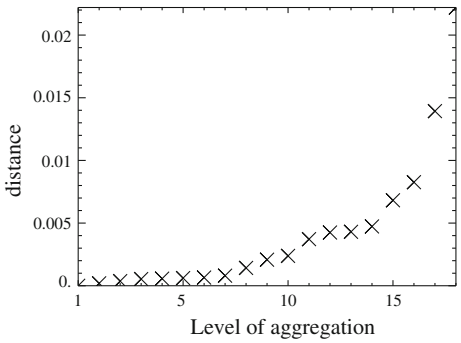

Fig. 2 Example of average distances between clusters as a function of the level of aggregation, with the "knee" at level 15

The distance between two arbitrary clusters, $C1$ and $C2$, is determined following Sneath and Sokal (1973), as the average distance computed from all possible ensemble members pairs (d_c)

$$d_c = \frac{1}{n_1 \cdot n_2} \sum_{\psi_i \in C1} \sum_{\psi_j \in C2} d(\psi_i, \psi_j), \tag{11}$$

where n_1 and n_2 are the number of ensemble members within the individual clusters. $d(\psi_i, \psi_j)$ is the length of the vector between ψ_i and ψ_j, in the two-dimensional space defined from $\mathrm{ETS}_{\mathrm{mod}}$ and $\mathrm{FBI}_{\mathrm{mod}}$. Since clustering is based on scores computed over the whole model domain, quite different ensemble members in terms of precipitation distribution can reside in the same cluster, this is a consequence in the definition of the "closeness" of the member to the observation.

The distances between the clusters increase with successive merging, and a suitable level of aggregation must be defined for the final cluster selection. One must either fix the number of desired clusters, or equivalently fix a threshold for the minimum distance above the clusters are separated. We have opted for the second choice in order to give the clustering the freedom to change the number of clusters. To this goal we examine the mean cluster distance as a function of the level of aggregation and select the strongest increase per level of aggregation (also named *knee*) of this strictly monotonic function as the appropriate level of aggregation (see, e.g. Fig. 2 where the *knee* is analysed at aggregation level 15).

Salvador and Chan (2004) introduced the objective L-method to determine the knee: the graph of N points is approximated by a pair of straight lines. Given one point on

the graph representing aggregation level a, with $2 \leq a \leq N-1$, the left line is the linear interpolation of all points on the left of a, including a, while the right line is the linear interpolation of all points on the right of a, including a. For each of the resulting $N-2$ pairs of lines, one computes the quality of fit the line pairs to the points via the RMSD and select the a with the minimum RMSD as the optimal level of aggregation. If the minimum RMSD is, however, larger than the one computed for the line of the linear interpolation taken all points, the highest level of aggregation (one cluster) is selected.

4.3 Choice of members to resample and filter

SAM determines the members to remove and to double, based on the average metric $\bar{M}_i$ for each cluster i and on the number of ensemble members m_i contained in a cluster (N the size of the EPS). We assume that a cluster with more ensemble members is more probable than a cluster with fewer members. This is accomplished by selecting equal weights for ensemble members within a cluster via

$$W_n(i) = \frac{m_i}{N} \tag{12}$$

and a cluster weight, which is related to the metric via

$$W'_{\mathrm{M}}(i) = \frac{\min(\bar{M})}{\bar{M}_i} \tag{13}$$

With the normalization

$$W_{\mathrm{M}}(i) = \frac{W'_{\mathrm{M}}(i)}{\sum_i W'_{\mathrm{M}}(i)} \tag{14}$$

both weights are now combined into one weight by

$$Wr(i) = \frac{W_n(i) + 5W_{\mathrm{M}}(i)}{\sum_i (W_n(i) + 5W_{\mathrm{M}}(i))} \tag{15}$$

The constant 5 is heuristically set to give more importance to the metric M. Other metrics can be chosen, but we do not expect considerable sensitivity to its choice. These sensitivity studies are beyond the scope of the current work.

For filtering we take the normalized inverse:

$$Wf(i) = \frac{W'f(i)}{\sum_i (W'f(i))}, \tag{16}$$

where

$$W'f(i) = \frac{1}{Wr(i)} \tag{17}$$

With each filter step, we remove five members and create five identical copies of other five members. The number of members to filter and to resample is fixed to maintain the ensemble characteristics and to avoid too few remaining members (i.e. loss of representativeness of the EPS). From the ith cluster we remove n_{f} members:

$$n_{\mathrm{f}}(i) = INT(5 \cdot Wf(i)) \tag{18}$$

Following this rule, the total number of ensemble members removed (filtered, N_{f}) can be larger or lower than five. In the case N_{f} is larger/lower than five, we remove less/more ensemble members of the best/worst cluster. In each case, the worst cluster must survive with at least one ensemble member. In the same way, but using Wr, we choose the five members to be resampled.

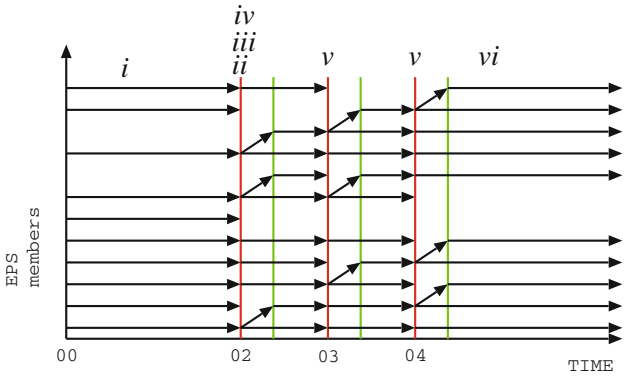

Fig. 3 Schematics of the SAM chain including the step levels from Sect. 4.4

4.4 SAM

SAM filters and resamples a given ensemble based on precipitation field differences between ensemble members and observations. The structure is sketched in Fig. 3; the list below refers to the figure.

(i) The ensemble is started at 00 UTC and integrated until 02 UTC.
(ii) The distances between the members and the observations are computed based on our metric and clustered as described in Sect. 4.2.
(iii) Five ensemble members are discarded (filtered) and five duplication of other better performing members are created (for details see Sect. 4.3).
(iv) While the remaining ensemble members are integrated in time for one hour, the additional members created by duplication are integrated in time while subject to nudging via PIB during the first 15 min.
(v) Steps (ii) until (iv) are repeated twice.
(vi) The data assimilation interval is followed by 11 h of free run (until 16 UTC, in fact the final non-PIB ensemble members have 12 h of free run while the final PIB-particles have 11 h and 45 min).

This setup is chosen because of the PIB'S ability to nudge the model state in the direction of the observed state using very short assimilation windows (Milan et al. 2008; Milan 2010). We compare SAM to the original EPS (ORIGINAL), with a pure filtered EPS without resampling (FILTER) and to PIB-EPS (an ensemble where the timing is identical to the one in SAM, but all ensemble members are subject to PIB).

5 Case studies

We tested SAM for three convective cases in August 2007; Fig. 1 shows for each case the radar-derived fields of the precipitation sum fields integrated from 00 to 16 UTC.

The first case (8th August) is characterized by three small low-pressure systems over western Poland, between Greece and Ukraine, and over France, respectively, embedded in an anticyclonic zone over central Europe. According to the radar observations, a stratiform precipitation field moves across south-western Germany, while an area with convective activity resides in north-eastern Germany. The rain gauge network of DWD indicates a maximum hourly rain rate in Baden-Württemberg of 10 mm during the first 12 h of the day.

In the second case (9th August), central Europe is governed by a single low-pressure system. A frontal zone separates eastern Europe with subtropical warm air from western Europe with colder humid air. Strong rain affects western and southern Germany as well as Switzerland during the day. The accumulated daily precipitation locally exceeds 40 mm. During the afternoon, a convective line approaches the Berlin area.

The third case (12th August) is dominated by subtropical wet/maritime air transported to the midlatitudes due to a high-pressure region over central Europe and a through over western Germany. Under these conditions, which are favourable for convection initiation, multicells with strong precipitation rates develop in a very limited region in north-eastern Germany during the morning.

6 Results

In the following, we compare the quality of the predicted precipitation fields for the three cases by comparison with the radar observations.

The performance of four ensembles is discussed: ORIGINAL, PIB, SAM, and FILTER. ORIGINAL denotes the COSMO-DE EPS as described in Sect. 3.1, with no filtering, resampling or nudging applied (16-h free forecast starting at 0 UTC). In PIB, all members of COSMO-DE EPS were subject to data assimilation using the PIB approach described in Sect. 3.3. SAM is described in Sect. 4. FILTER only applies the filter part of SAM without the duplication part: accordingly, each filter event reduces the ensemble size by removing the least probable ensemble members, taking clustering into account. In this way—similar to the standard PF—the importance of the remaining members in the ensemble increases.

We quantify the quality of the ensemble forecasts (except FILTER) by comparing *resolution, reliability and sharpness* (Jolliffe and Stephenson 2004). Such a detailed analysis cannot be applied to FILTER due to its low final ensemble size.

Resolution is the forecast system's ability to distinguish between different observed frequency distributions. Reliability quantifies the capacity of the EPS to forecast unbiased estimates of the observed frequencies associated with different forecast probability values, or the average agreement between forecasted and observed states. If we choose a specific event from the observations, e.g. precipitation sums above 1 mm, the variable x_0 attains the value 1 if the event happens, and 0 otherwise. The PDF of the forecast ($q(x)$) is estimated from the ensemble, i.e. the probability distribution of the event happening; $q(x)$ and x_0 are then compared in every grid point. An EPS achieves perfect reliability, when the probability that the event occurs in the observations ($x_0 = 1$) given the PDF of the forecast, is equal to the PDF of the forecast:

$$f(q) = p(x_0 = 1|q(x)) = q(x). \tag{19}$$

Sharpness is the forecast's tendency to divert from the climatology; thus it is a measure of the forecast's variability. Sharpness is not a verification measure because of its independence from observations. Large frequencies of both zeros and ones indicate a high degree of sharpness. The low ensemble size of FILTER does not allow for the estimation of reliability.

6.1 ETS, FBI and relative entropy

The hourly Equitable Threat Score (ETS) and the frequency bias (FBI) are computed for the forecast period using a threshold of 0.1 mm/h to distinguish between yes/no precipitation (Wilks 2006) to quantify forecast quality. In the result figures (Figs. 4, 5, 6), the different global models at the lateral boundaries conditions for the ensemble members are indicated by different colours. Both scores must be judged in conjunction, since a larger FBI also tends to have a higher ETS (Mesinger 2008). Thus, a higher ETS represents a "better score" only if the frequency bias is the same. Moreover, the ETS is influenced by the spatial structure of the characteristics within the regions, this is a general problem in COSMO (observed in CLM, the climate version of COSMO,

by Bachner et al. 2008; Wang et al. 2013). Note also that the first one or two forecast hours (depending on the synoptic situation) should be interpreted with care, because a dynamic model needs spin-up time to produce realistic precipitation.

The forecasted probability of the occurrence of the event p_j, which can be estimated from the ensemble, is compared with the observation x_j, and given the value 1 or 0 depending on whether the event (in our case precipitation above 0.1 mm) occurred or not. Then the Brier score (B) can be computed,

$$B = \frac{1}{n}\sum_{j=1}^{n}(p_j - x_j)^2, \tag{20}$$

where n is the number of compared grid points. A perfect EPS has a Brier score equal to 0 since $p_j = x_j$ for all j. In order to compare a given forecast with a reference forecast, the Brier Skill Score (BSS) is appropriate,

$$\mathrm{BSS} = 1 - \frac{B}{B_{\mathrm{ref}}}, \tag{21}$$

where B_{ref} being the Brier score of the reference forecast. BSS is equal to 1 for a perfect deterministic forecast system and is 0 or negative for a forecast system that performs similarly or poorer than the reference forecast. The integral of the Brier scores computed for all possible thresholds x constitutes the Continuous Ranked Probability Score (CRPS):

$$\mathrm{CRPS} = E\left(\int_{-\infty}^{\infty}\left[F(x) - H(x - x_0)\right]^2 \mathrm{d}x\right) \tag{22}$$

$$H(x - x_0) = \begin{cases} 1: & x \geq x_0 \\ 0: & x < x_0 \end{cases}, \tag{23}$$

where $F(x) = p(X \leq x)$ is the cumulative density function (CDF) of the EPS system and where H is the Heaviside function. The PDF of the EPS is computed using ensemble kernel dressing (Bröcker and Smith 2008) based on gamma functions. In practice, the CRPS is computed discretely. For a comprehensive description see Hersbach (2000).

The PDF of accumulated precipitation over the 11 h of free forecast (from 05 UTC until 16 UTC, Figs. 4, 5, 6 right panels) from all EPS systems is used to test the ability to forecast the precipitation distribution based on the relative entropy of the model probability compared with the radar probability as distance measure. The relative entropy [Kullback–Leibler divergence (Kullback 1968)], is defined as

$$\mathrm{rel.entropy} = \sum_{i}^{n}\left(Z_{\mathrm{RAD}} \log \frac{Z_{\mathrm{RAD}}}{Z_{\mathrm{MOD}}}\right), \tag{24}$$

where n is the number of bins into which precipitations sums are classified, Z_{RAD} the frequency of values in a given bin from radar observed precipitation and Z_{MOD} the frequency of values in a given bin from model precipitation. A perfect simulation will have a relative entropy equal to 0. The relative entropy is computed for all ensemble members. Its mean is taken as the minimum variance state and the standard deviation as uncertainty.

CASE 1 ORIGINAL shows good ETS (at 08 UTC values around 30 %, Fig. 4 left panel, a1) and good FBI scores (values close to one over the forecast period after 2 h of spin up, Fig. 4 middle panel, a1). During most of the free forecast, all ensemble members show similar quality; only towards the end of the free forecast ensemble members driven by the GME global model lose quality (green lines). SAM clearly enhances the ETS over the forecast time without a substantial negative influence on FBI. Due to the initially good quality of the ensemble members driven by GME, SAM resamples some of these members, which reduces the ensemble ETS and increases FBI towards the end of the free forecast. Due to our clustering-based scheme, also ensemble members driven by other global models are resampled and/or maintained, which ameliorates the ensemble quality. The filter part of SAM manages to choose some of the members with better scores even at the end of the forecast (see ETS and FBI scores in Fig. 4 left and middle panels, c1) and achieves good persistence.

PIB results in even higher ETS, but this improvement is accompanied by a notable negative effect on the FBI. The influence of the driving global models is less pronounced (at least for the ETS), as all ensemble members are nudged to the observations, leading to similar quality among all members.

BSS and CRPS are based on hourly precipitation, computed over the free forecast, and then averaged. The CRPS for all ensembles (excluding FILTER) are very similar, while the BSS for precipitation above 0.1 mm/h is positive and the highest for SAM (Table 1).

The PDFs of precipitation accumulation (Fig. 4 right panel, a1, c1 and d1) obtained from ORIGINAL, SAM and PIB are similar, while the relative entropies are the highest for SAM and PIB. The distribution of the precipitation sum during the free run does not depend on the filter; thus the relative entropy does not vary significantly between the ensemble systems (Fig. 4 right panel, b1). Since the variance of the relative entropy also relates to the "dispersion" of the ensemble members compared with the observations, some of the original dispersion is lost by PIB as expected.

CASE 2 The control run (ORIGINAL) initially shows good quality in terms of ETS, after 02 UTC (Fig. 5 left panel, a2); but ETS decreases from around 40 to 10 % in the ensuing 14 h. The ensemble members driven by UM have—most of the time—the best ETS. The FBI (Fig. 5 middle panel, a2) is close to 1 between 02 and 06 UTC followed by a slight underestimation of the precipitation area; in this time period there are no significant differences between the members belonging to different driving

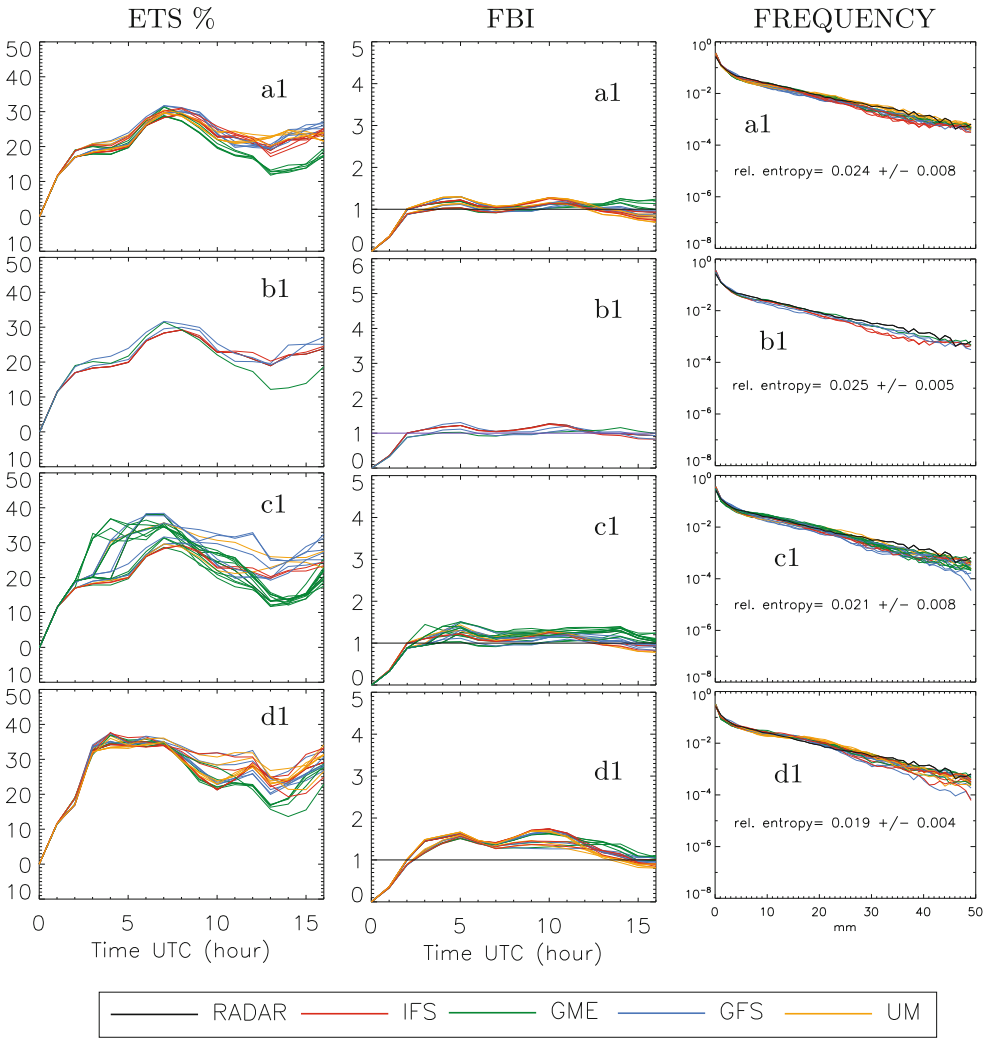

Fig. 4 *CASE 1* ETS (*left panel*), FBI (*middle panel*) for the hourly precipitation, threshold 0.1 mm/h; Distribution of precipitation sum (mm, *right panel*) during the free forecast run (from 05 to 16 UTC). The values of the relative entropy including uncertainty are added to the plots. **a** ORIGINAL, **b** FILTER, **c** SAM, **d** PIB

Table 1 Brier Skill Score (BSS) with 0.1 mm and Continuous Ranked Probability Score (CRPS) for all case studies

	BSS SAM 0.1 mm	BSS PIB 0.1 mm	CRPS ORIG.	CRPS SAM	CRPS PIB
CASE 1	0.09	0.03	1.016	1.020	1.016
CASE 2	0.12	0.16	0.997	0.988	1.065
CASE 3	0.13	0.001	0.639	0.639	0.639

models. SAM and, even more so, PIB exhibit a higher ETS (up to 10 %) compared with ORIGINAL for 5–6 h of the forecast time. SAM leads to a somewhat better FBI during the period, i.e. the members are better distributed around 1. The better performing ensemble members driven by UM are kept and other members still profit from nudging towards the observations. The filter part of SAM (see results for FILTER) manages to choose those members, which have good scores until 12 UTC. Notably, many of the GFS-driven members are removed. GME drives some of the surviving members, but they have low ETS scores at the end of the forecast. Thus, for the last four forecast hours, ETS scores for SAM fall below the ones for PIB. All members driven by UM have similar performance during the filtering, but clustering and duplication using PIB prevents narrowing of the ensemble to UM-driven members. SAM duplicates only one GFS-driven member and no UM-driven members. PIB leads to a clear overestimation of the precipitation probability between 02 UTC and 07 UTC. The influence of the driving models is again less pronounced for PIB than for ORIGINAL and SAM. SAM slightly improves both BSS and CRPS compared with CONTROL, while PIB has a slightly worse CRPS (Table 1). The ETS and FBI scores for PIB are very similar, while SAM maintains some of the original spread.

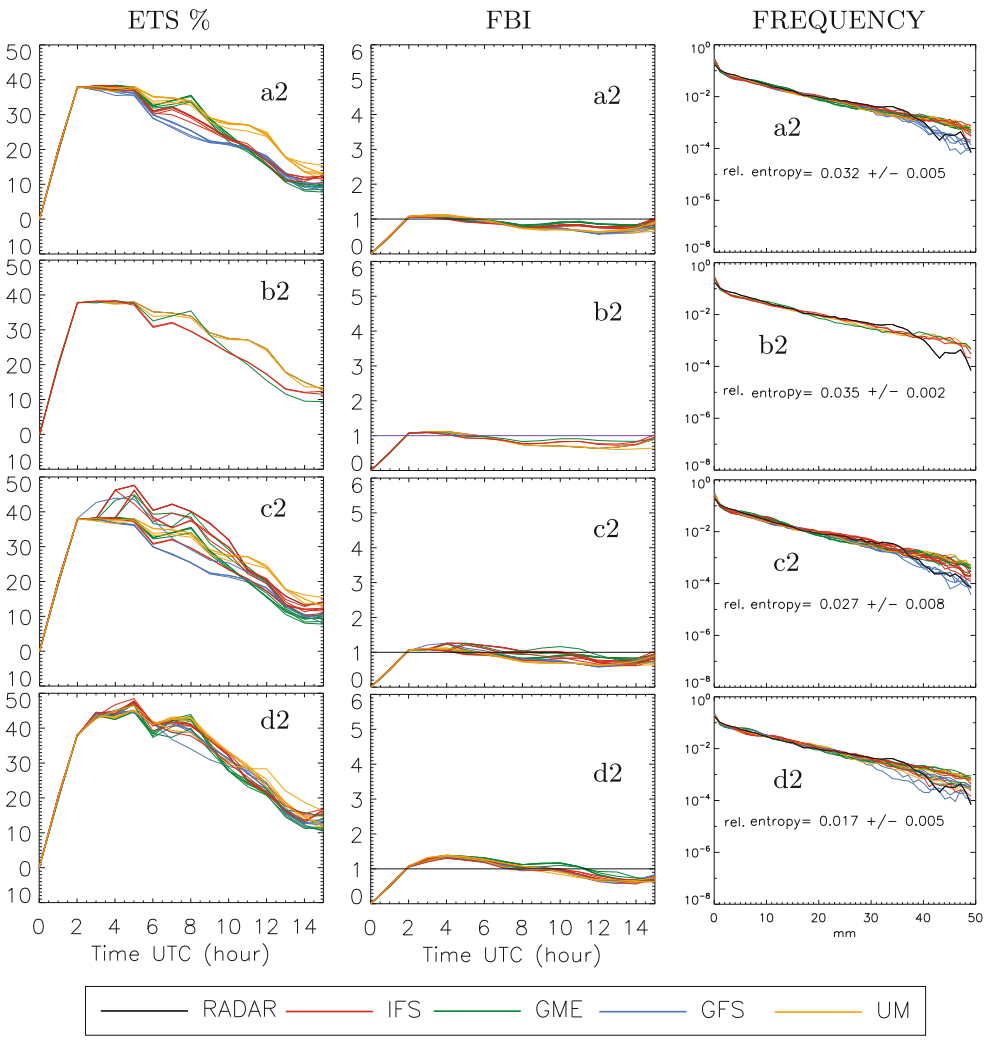

Fig. 5 *CASE 2* ETS (*left panel*), FBI (*middle panel*) for the hourly precipitation, threshold 0.1 mm/h; Distribution of precipitation sum (mm, *right panel*) during the free forecast run (from 05 to 16 UTC). The values of the relative entropy including uncertainty are added to the plots. **a** ORIGINAL, **b** FILTER, **c** SAM, **d** PIB

The PDFs of precipitation accumulation (Fig. 5 right panel) are again similar, but SAM and even more PIB clearly lead to better relative entropies. In addition, the ensemble PDFs are better distributed around the radar PDF also for higher thresholds. Only GFS-driven ensemble members manage to approximate the distribution of precipitation accumulation above 40 mm well. This behaviour is maintained in SAM and PIB. FILTER removes all members driven by GFS, as a consequence of the metric. The metric is based only on yes/no precipitation and ignores precipitation amounts. In this case, the variance of PIB is lower than that in ORIGINAL and SAM.

CASE 3 Precipitation forecasts are difficult for this case, because of the relatively small region where the convective precipitation occurs. Small errors in positioning and/or timing lead to high false alarms and missed events, which dominate ETS and FBI (Wilks 2006; Hamill 1999).

The ETS for ORIGINAL (Fig. 6 left panel, a3) is around 0 %, and thus comparable in quality with a random forecast. The forecast overestimates precipitation between 02 and 10 UTC (FBI $>$ 1 in Fig. 6 middle panel, a3), followed by underestimation after 10 UTC. Between 04 UTC and 10 UTC the FBI for the UM-driven members is the highest. SAM and PIB again result in higher ETS over the entire free run, with the best score for PIB (Fig. 6 left panel, c3–d3), but the FBI suggests a considerable overestimation particularly for the PIB variant (Fig. 6 middle panel, c3–d3). SAM keeps no UM-driven ensemble members, and no GME-driven members are resampled. We presume that the general poor quality of all ensemble members between 02 UTC and 04 UTC leads to an unsuitable cluster attribution, followed by suboptimal filtering and resampling of the members. In other words, we believe that the closeness

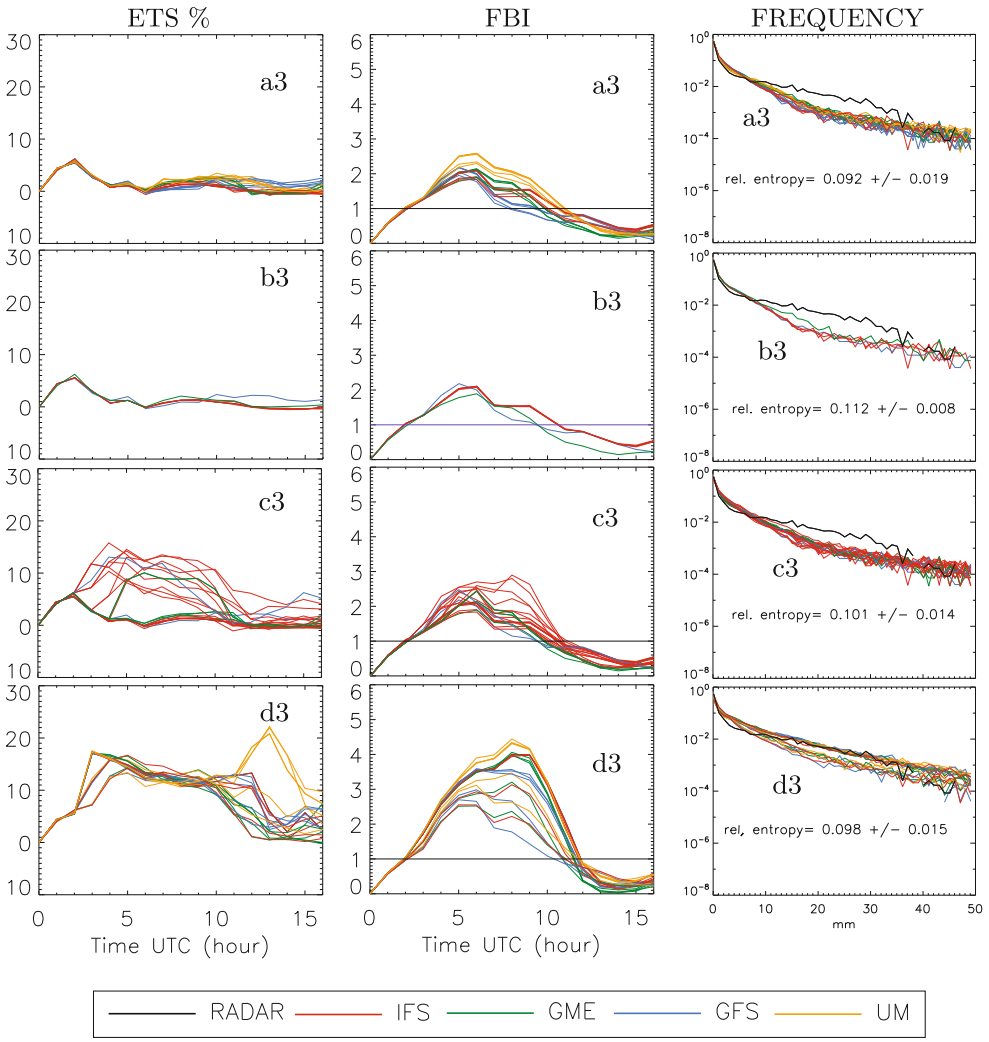

Fig. 6 *CASE 3* ETS (*left panel*), FBI (*middle panel*) for the hourly precipitation, threshold 0.1 mm/h; Distribution of precipitation sum (mm, *right panel*) during the free forecast run (from 05 to 16 UTC). The values of the relative entropy including uncertainty are added to the plots. **a** ORIGINAL, **b** FILTER, **c** SAM, **d** PIB

of all members to each other in the beginning led to a random filter and resampling behaviour.

The BSS clearly favours SAM (Table 1), but not PIB, while the CRPS is almost identical for all ensemble systems. Given the poor quality of the original EPS for all members, FILTER does not improve the ensemble.

The PDFs of precipitation accumulation (Fig. 6 right panel) for ORIGINAL and SAM are quite similar, and the respective relative entropies are within their uncertainty limits. All ensemble members overestimate precipitation frequencies below 5 mm over 12 h, while precipitation amounts between 5 and 35 mm are underestimated. In the PIB variant, some members better approximate the radar PDF, but the relative entropy is not improved above its uncertainty limits.

6.2 Reliability and sharpness

The value of $f(q)$ (Eq. 19) quantifies the reliability of an EPS system. $f(q)$ is estimated by counting the relative frequency of the observed event for cases for which the event (in this case an hourly rain rate above 0.1 mm) was forecasted to occur with the probability q. An instructive way to visualize reliability quality is to plot $f(q)$ as a function of the probability that the event occurs in the EPS ($q(x)$) (*reliability diagram*, Wilks 2006). In such a diagram, a perfect reliable system lies in principle on the diagonal (from Eq. 19), but sampling effects might cause deviations. The amount of sampling-induced variability can be visualized by plotting reliability diagrams for the same forecast system using a randomly chosen ensemble member in place of the observations (Wilks 2006).

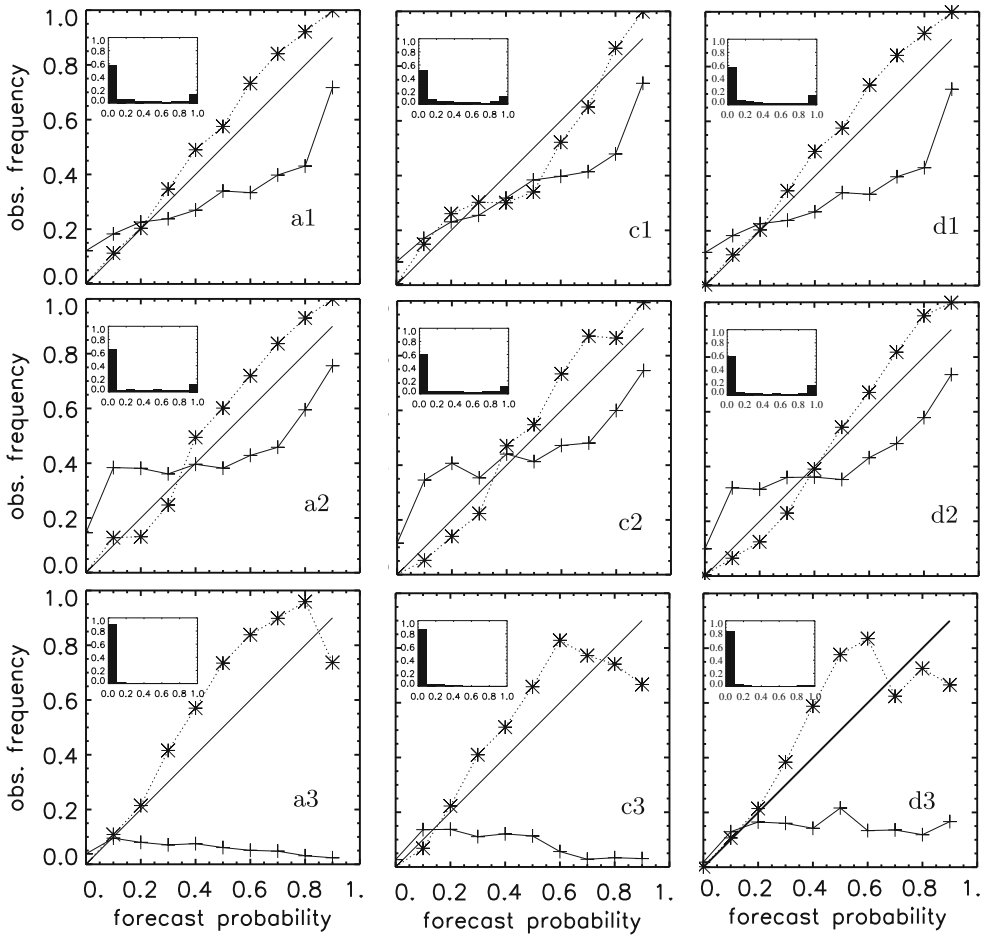

Fig. 7 Reliability plot for precipitation during the free forecast run (from 05 to 16 UTC) and the three cases (1 to 3). *Full line* reliability diagram for the ensemble forecast. *Dashed line*: reliability diagram for a perfect ensemble forecast where "observation" is defined as one of the ensemble members. *Upper left box* sharpness graph. **a** ORIGINAL, **c** SAM, **d** PIB. The rows depict the three cases. E.g. **a1** output for ORIGINAL in CASE 1

For CASE 1, the reliability diagrams (Fig. 7, first row) show forecasted probabilities close to observed frequencies for all EPS (except FILTER). From the diagrams, we can suppose that ORIGINAL underforecasts events that are associated with smaller forecast probabilities and overforecasts events associated with larger forecast probabilities. Overall, SAM performs somewhat better, especially for the lower probability values, while PIB is very similar to ORIGINAL. All EPS systems have high sharpness (see the upper left corners in subfigures of Fig. 7).

The reliability diagrams for CASE 2 (Fig. 7, second row), also indicate forecasted probabilities close to observed frequencies. Similarly to CASE 1, ORIGINAL underforecasts events associated with smaller forecast probabilities, while events associated with larger forecast probabilities are overforecasted. In this case, both SAM and PIB lead to no improvements. All systems have high sharpness.

For CASE 3 (Fig. 7, third row), the very low forecast skill (Sect. 6.1) leads to reliability curves close to the horizontal (Atger 2004). Thus all EPS have a very low reliability, i.e. they overpredict almost all observation frequencies. Both SAM and PIB perform better than ORIGINAL, especially for the lower probability values, where PIB behaves best. All systems again have a high sharpness.

7 Conclusions

We have developed and applied SAM, a new ensemble-based data assimilation approach, that employs elements of the particle filter, to an ensemble prediction system at the convection-permitting scale for nowcasting and short-term forecasts. Using the classical particle filter, the weights in high-dimensional systems tend to collapse, and very large ensembles are required to avoid collapse (Snyder et al. 2008). SAM is designed for a relatively small ensemble of 20 members, based on the COSMO-DE EPS. We have attempted to reduce filter degeneracy and to keep the distribution of the ensemble close to the observations using an approach based on clustering combined with a nudging method (PIB) that uses radar and satellite observations.

In order to evaluate the effect of the assimilation method excluding any clustering and filtering, we have compared

SAM performance with that of the original EPS and with the original EPS with all members nudged towards observations using PIB. We have also quantified the pure effect of cluster-based filtering (without resampling) to check the persistence of the quality for the chosen members. Different skill scores and performance indices have been computed for objective comparisons.

Filtering/resampling based on clustering maintains representativeness of the EPS. Ensemble members differ from each other by their driving global models (lateral boundary conditions) and physical parameterisations. In order to better accommodate chances for non-linear developments in model evolutions, SAM does not remove all ensemble members with initially low performance (given the metric). Members driven by the same global model tend to behave similarly, but clustering reduces the probability that all members with the same boundary conditions are removed, which leads to a better representativeness of the EPS.

Generally, the SAM and PIB variants of the EPS enhance the quality skill scores. ETS is the highest for both variants over the complete forecast time for all three case studies with PIB outperforming SAM. However, PIB often overestimates rain events leading to a higher FBI; in this regard, SAM outperforms PIB. This behaviour must be pointed out when ETS skills are compared, because overestimating precipitation can already lead to a higher ETS (Mesinger 2008). The EPS variants FILTER, SAM and PIB change the distribution of the precipitation sum only marginally during the free run. For all three cases, the relative entropy is only slightly changed. In two of the three cases the precipitation PDF of PIB ensemble members are very similar and reduce ensemble spread.

Regarding resampling/duplication, the selection of the ensemble members is decisive for SAM and FILTER. Our choice of clustering and metrics also might influence the results, but a sensitivity study using other clustering methods, metrics or combination of metrics (e.g. Weusthoff et al. 2011) exceeds the scope of this paper and is suggested as a follow-up study. The clustering approach is promising due to its potential to catch up possible non-linear evolutions of the dynamic system and since it does not reside on Gaussian approximations.

From our results, PIB alone would nudge all ensemble members to radar observations, which is accompanied by a loss of ensemble spread and thus a reduced ability to account for uncertainties in chaotic dynamical systems. In SAM, the radar data assimilation, together with filtering/resampling improves the forecast quality without side effects on the ensemble spread of the preexisting EPS.

Further studies are needed to test the sensitivity of SAM to filter degeneracy. This could be done by repeating steps (ii) until (iv) (Sect. 4.4 and Fig. 3) in a continuous assimilation cycle over a longer period, e.g. more than 1 day.

Acknowledgments Funding of this work by the Deutsche Forschungsgemeinschaft (DFG) under Grants SI606/7-3 and SI606/13-1 is gratefully acknowledged. We thank the Deutscher Wetterdienst (DWD) for providing the COSMO-DE model and the precipitation data, and the Satellite Application Facility on support to NoWCasting and very short-range forecasting (SAFNWC) for providing the satellite-derived cloud information. We also acknowledge funding of the contributions by Mrs. Theresa Bick by the Hans-Ertel-Centre for Weather Research, Atmospheric Dynamics and Predictability Branch. The authors would like to thank also the reviewers, Christoph Schraff and the other two anonymous, for their valuable comments and suggestions to improve the quality of the paper.

References

Atger F (2004) Relative impact of model quality and ensemble deficiencies on the performance based probabilistic forecasts evaluated through the Brier score. Nonlinear Process Geophys 11:399–409

Bachner S, Kapala A, Simmer C (2008) Evaluation of daily precipitation characteristics in the CLM and their sensitivity to parameterizations. Meteorol Z 17:407–420. doi:10.1127/0941-2948/2008/0300.

Baldauf M, Seifert A, Förstner J, Majewski D, Raschendorfer M, Reinhardt T (2011) Operational convective-scale numerical weather prediction with the COSMO model: description and sensitivities. Mon Weather Rev 139:3887–3905

Bengtsson T, Bickel P, Li B (2008) Curse-of-dimensionality revisited: collapse of the particle filter in very in very large systems. In: Probability and statistics: Essays in honor of David A. Freedman, vol 2. Institute of Mathematical Statistics, USA, pp 316–334

Bird G (1978) Monte Carlo simulation of gas flow. Annu Rev Fluid Mech 10:11–31

Bowler N (2006) Comparison of error breeding, singular vectors, random perturbations, and ensemble Kalman filter perturbation strategies on a simple model. Tellus 58A:538–548

Bröcker J, Smith L (2008) From ensemble forecasts to predictive distribution functions. Tellus A 4:663–678

Casati B, Stephenson D, Ross G (2004) A new intensity-scale approach for a verification of spatial precipitation forecasts. Meteorol Appl 11:141–154

Cullen M (1993) The unified forecast/climate model. Meteorol Mag 122:81–94

Dee D (2005) Bias and data assimilation. Q J R Meteorol Soc 131:3323–3343

de Elia R, Laprise R, Denis B (2002) Forecasting skill limits of nested, limited-area models: a perfect-model approach. Mon Weather Rev 130:2006–2023

Doucet A, de Freitas N, Gordon N (2001) Sequential Monte Carlo methods in practice, 1st edn. Springer, Berlin

Fletcher SJ (2010) Mixed Gaussian-lognormal four-dimensional data assimilation. Tellus 62A:266–287

Garcia-Moya J-A, Callado A, Escriba P, Santos C, Santos-Munoz D, Simarro J (2011) Predictability of short-range forecasting: a multimodel approach. Tellus A 63(3):550–563. doi:10.1111/j.1600-0870.2010.00506.x

Gebhardt C, Theis S, Paulat M, Bouallègue ZB (2011) Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. Atmos Res 100:168–177

Gordon N, Salmond S, Smith AFM (1993) Novel approach to nonlinear/non-Gaussian bayesian state estimation. IEE Proc 140:107–113

Haase G, Crewell S, Simmer C, Wergen W (2000) Assimilation of radar data in mesoscale models: physical initialization and latent heat nudging. Phys Chem Earth 25:1237–1242

Hamill TM (1999) Hypothesis tests for evaluating numerical precipitation forecasts. Weather Forecast 14:155–167

Hammersley JM, Morton KW (1954) Poor man's Monte Carlo. J R Stat Soc B 16:23–38

Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast 15:559–570

Jakob C, Andersson E, Beljaars A, Buizza R, Fisher M, Gerard E, Ghelli A, Janssen P, Kelly G, McNally P, Miller M, Simmons A, Teixeira J, Viterbo P (1999) The IFS cycle CY21R4 made operational in October 1999. Technical Report, ECMWF Newsletter, f87

Jazwinski AH (1970) Stochastic process and filtering theory. Academic Press, New York

Jolliffe I, Stephenson DB (2004) Forecast verification, Wiley, New York, Chap Glossary

Kalman R (1960) A new approach to linear filtering and prediction problems. Trans ASME J Basic Eng 82:35–45

Kim S, Eyink L, Restrepo J, Alexander F, Johnson G (2003) Ensemble filtering for nonlinear dynamics. Mon Weather Rev 131:2586–2594

Kullback S (1968) Information theory and statistics, 2nd edn. Dover Publications Inc., New York

Leith CE (1971) Atmospheric predictability and two-dimensional turbulence. J Atmos Sci 28:145–161

Lorenz E (1963) Deterministic non-periodic flow. J Atmos Sci 20:130–141

Lorenz E (1963) The predictability of hydrodynamic flow. Trans NY Acad Sci Ser II 25:409–432

Lorenz E (1969) The predictability of a flow which possesses many scales of motion. Tellus 21:289–307

Majewski D, Liermann D, Prohl P, Ritter B, Buchhold M, Hanisch T, Paul G, Wergen W, Baumgardner J (2002) The operational global icosahedral–hexagonal gridpoint model GME: description and high-resolution tests. Mon Weather Rev 130:319–338

Marsigli C, Montani A, Paccagnella T (2008) The COSMO-SREPS ensemble for short-range: system analysis and verification on the MAP D-PHASE DOP. In: Proceedings of the joint map d-phase scientific meeting-cost 731 mid-term seminar

Mesinger F (2008) Bias adjusted precipitation threat scores. Adv Geosci 16:137–143

Milan M (2010) Physical initialisation of precipitation in a mesoscale numerical weather forecast model. PhD thesis, University of Bonn. https://hss.ulb.uni-bonn.de/2010/2041/2041.htm

Milan M, Venema V, Schüttemeyer D, Simmer C (2008) Assimilation of radar and satellite data in mesoscale models: a physical initialization scheme. Meteorol Z 17:887–902

Müller MD, Scherer D (2005) A grid- and subgrid-scale radiation parameterization of topographic effects for mesoscale weather forecast models. Mon Weather Rev 133:1431–1442

Paulat M, Theis S, Gebhardt C, Ben Bouallegue Z, Buchhold M, Ohl R (2009) COSMO-DE-EPS-construction, diagnoses and verification of a limited-area ensemble prediction system on the convective scale. In: 9th EMS annual meeting, 9th European conference on applications of meteorology (ECAM) abstracts, held 28 Sept–2 Oct 2009, Toulouse, France, vol 1. id. EMS2009-379, p 379. http://meetings.copernicus.org/ems2009/

Pham D (2001) Stochastic methods for sequential data assimilation in strongly nonlinear systems. Mon Weather Rev 129:1194–1207

Pitt M, Shephard N (1999) Filtering via simulation: auxiliary particle filters. J Am Stat Assoc 94:590–599

Rubin DB (1988) Using the SIR algorithm to simulate posterior distributions. Oxford University Press, Oxford

SAFNWC (2004) User manual for the PGE01-02-03 of the SAFNWC/MSG: scientific part. http://nwcsaf.inm.es

Salvador S, Chan P (2004) Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In: Proceedings of the 16th IEE international conference on tools with AI, pp 576–584

Schulz J, Schättler U (2005) Kurze Beschreibung des Lokal-Modells LME und seiner Datenbanken auf dem Datenserver des DWD. German Weather Service (DWD), Research Department, P.O. 100465, D-63004 Offenbach

Sela J (1980) Spectral modeling at the National Meteorological Center. Mon Weather Rev 108:1279–1292

Sneath J, Sokal M (1973) Unweighted pair-group method using arithmetic averages. In: Francisco S (ed) Numerical taxonomy, Freeman, San Francisco, pp 230–234

Snyder C, Bengtsson T, Bickel P, Anderson J (2008) Obstacles to high-dimensional particle filtering. Mon Weather Rev 136:4629–4640

Stephan K, Klink S, Schraff C (2008) Assimilation of radar-derived rain rates into the convective-scale model COSMO-DE at DWD. Q J R Meteorol Soc 134:1315–1326

Talagrand O (1997) Assimilation of observations, an introduction. J Meteorol Soc Jpn Spec Issue 75:191–209

Toth Z, Kalnay E (1993) Ensemble forecasting at NMC: the generation of perturbations. Bull Am Meteorol Soc 74:2317–2330

Toth Z, Kalnay E (1997) Ensemble forecasting at NCEP and the breeding method. Mon Weather Rev 125:3297–3319

van Leeuwen P (2001) An ensemble smoother with error estimates. Mon Weather Rev 129:709–728

van Leeuwen P (2002) Ensemble Kalman filters: sequential importance resampling and beyond. In: Proceedings of the ECMWF workshop on the role of the upper ocean in medium and extended range forecasting, Reading, United Kingdom. ECMWF

van Leeuwen P (2003) A variance-minimizing filter for large-scale applications. J Meteorol Soc Jpn Spec Issue 131:2071–2084

van Leeuwen P (2009) Review particle filtering in geophysical systems. Mon Weather Rev 137:4089–4114

van Leeuwen P (2010) Nonlinear data assimilation in geosciences: an extremely efficient particle filter. Q J R Meteorol Soc 136:1991–1999

van Leeuwen P, Evensen G (1996) Data assimilation and inverse methods in terms of a probabilistic formulation. Mon Weather Rev 124:2898–2913

Wang D, Menz Ch, Simon T, Simmer C, Ohlwein C (2013) Regional dynamical downscaling with CCLM over East Asia. Meteorol Atmos Phys 121(1–2):39–53. doi:10.1007/s00703-013-0250-z

Weusthoff T, Leuenberger D, Keil C, Craig G (2011) Best member selection for convective-scale ensembles. Meteorol Z 2:153–164

Wilks D (2006) Statistical methods in the atmospheric sciences. Elsevier, Amsterdam

Yoden S (2007) Atmospheric predictability. J Meteorol Soc Jpn 85:77–102

Zepeda-Arce J, Foufoula-Georgiou E, Droegemeier K (2000) Space-time rainfall organization and its role in validating quantitative precipitation forecasts. J Geophys Res 105:10129–10146