Seventh Framework Programme

Theme 6 [SPACE]

Project: 607193 UERRA

Full project title:

Uncertainties in **E**nsembles of **R**egional **R**e-**A**nalyses

Deliverable D2.12

Kalman Filter Ensemble DA development

WP no:	2
WP leader:	Met Office
Lead beneficiary for deliverable :	University of Bonn
Name of author/contributors:	Lilo Bach, University of Bonn
Nature:	Report
Dissemination level:	PU
Deliverable month:	21
Submission date: September 28, 2015	Version nr: 1

Project: 607193 - UERRA

Deliverable report D2.12

Feasibility study demonstrating uncertainty estimation capabilities of a COSMO ensemble reanalysis system – PART A

Author

Name	Institution	Address
L. Bach	Meteorological Institute University of Bonn	Auf dem Hügel 20 53121 Bonn Germany

Principal Investigators

Name	Institution	Address
Dr. J. Keller	Hans-Ertel-Centre for Weather Research Meteorological Institute Deutscher Wetterdienst	Frankfurter Straße 135 63067 Offenbach Germany
Prof. Dr. Andreas Hense	Meteorological Institute University of Bonn	Auf dem Hügel 20 53121 Bonn Germany
Dr. C. Ohlwein	Hans-Ertel-Centre for Weather Research Meteorological Institute University of Bonn	Auf dem Hügel 20 53121 Bonn Germany

Abstract

In this report, we present the work on deliverable D2.12. Development steps towards a high-resolution European ensemble reanalysis system are shown placing emphasis on verification of precipitation as **E**ssential **C**limate **V**ariable (ECV). We give an impression of the usability of the reanalysis system for Climate Change Service applications and detect deficiencies that require further development. Both accuracy and uncertainty estimation capabilities of the system are examined.

The work has been conducted as part of Work Package 2 on *Ensemble data assimilation regional reanalysis datasets* committed under the EU-FP7-funded collaborative project entitled *Uncertainties in*

Ensembles of Regional Reanalyses (UERRA: Grant agreement no.: 607193, www.uerra.eu) in cooperation with Dr. C. Schraff, DWD (**D**eutscher **W**etter**d**ienst, FE12).

The report is divided into two parts. In the first part, possible ways for the generation of an ensemble of reanalyses are outlined. The proposed combination of ensemble nudging with a local ensemble transform Kalman filter is not yet feasible. Technically, the implementation process of the data assimilation system can be regarded as finished. However, due to a range of technical and system-relevant issues that are outlined in the section hereafter, large temperature, moisture and precipitation biases occur. Therefore, we do not yet consider it reasonable to use the system for the production of a reanalysis. The effort that has to be made regarding tuning and experimenting with the system has been underestimated. However, ensemble nudging is at a good stage and considered useful and stable. To provide an alternative for the proposed EN-LETKF system in case that it will not reach a state that allows to use it for the generation of a reanalysis data set, the performance of ensemble nudging is examined with a focus on precipitation. The accuracy of the system, i.e. the spatio-temporal coherence of observations and each analysis member is compared to the global reanalysis ERA-INTERIM, a regional HIRLAM reanalysis and the deterministic high-resolution regional reanalysis COSMO-REA6. Moreover, we compare to a dynamical downscaling of ERA-INTERIM using COSMO to demonstrate the value of data assimilation for reanalyses which can be regarded as both weather and climate data sets. Subsequently, we demonstrate the probabilistic capabilities of our ensemble reanalysis systems. Using +6h short-range forecasts from ECMWF-EPS as reference we can show skill for an experiment spanning the month of June in 2011. All in all, the experiments prove that our ensemble nudging system for regional reanalysis has an added value.

1 Possibilities for ensemble reanalysis production

UB's task as part of WP2 in UERRA is to provide a regional ensemble reanalysis system as well as a proof of concept high-resolution data set for Europe. In principle, three data assimilation schemes are available to produce a regional ensemble reanalysis

- ensemble nudging (deliverable D2.11) based on the nudging scheme (Schraff, 1997)
- the local ensemble transform Kalman filter (LETKF) that has been developed for the convective scale (Schraff et. al 2015, submitted).
- a combination of the two that we denote EN-LETKF.

The ensemble nudging component has been developed, tuned and delivered in D2.11. In the last months, more extensive experiments have been performed whose evaluation is demonstrated in the deliverable on hand.

The LETKF for the regional model COSMO is currently under development at DWD and will be in pre-operational mode replacing the nudging scheme in the foreseeable future.

A link of ensemble nudging is considered particularly useful for reanalysis purposes as it combines their positive features yielding low RMSE (LETKF) and a smooth time series with small error spikes (nudging) (Lei et. al, 2012a).

In the next paragraphs, the three different systems are reviewed or introduced and their corresponding advantages and disadvantages are highlighted.

1.1 Ensemble nudging based on deterministic nudging

Note that this paragraph is a review from deliverable D2.11 which has to be included for completeness. Nudging performs a continuous relaxation of the prognostic variables of any numerical weather prediction model towards observations during the forward integration of the model. Additional tendency terms proportional to the observation-model equivalent departures are introduced directly to the prognostic equations. The analysis increments are finally spread to the target grid points within an area of influence. Thereby, a spatial weighting is performed using vertical and horizontal structure functions (Schraff, 1997). The temporal weighting function is designed such that observations are assimilated with maximal weight at the observation time. In contrast to intermittent 3-dimensional data assimilation schemes, asynoptic observations and very high frequent data can be assimilated at appropriate time. Nudging in its current implementation is not dependent on background or observation error covariance matrices. Instead, a static nudging coefficient having units of inverse time determines the strength by which the model state is corrected per model time step. Unlike 4d-Var or the ensemble Kalman filter, nudging in its applied version does not explicitly take into account flow-dependency. However, particularly due to its great performance-cost ratio yielding good analyses at low computational costs without dependence on tangent linear and adjoint models, nudging is used for many applications up to today (Stauffer and Seaman 1990, Stauffer et al. 1991, Seaman et al. 1995, Schraff 1997, Leidner et al. 2001, Otte et al. 2001, Deng et al. 2004, Deng and Stauffer 2006, Schroeder et al. 2006, Dixon et al. 2009, Ballabrera-Poy et al. 2009, Bollmeyer et al. 2015).

Due to the time-continuous manner in which the observations are assimilated, nudging yields smooth, physically consistent time series with little disturbance of the physical balances (e.g. Lei et al. 2012b). This is an advantage over intermittent techniques like the ensemble Kalman filter, where the sudden introduction of large numbers of observations often leads to strong error spikes in the assimilation time window (e.g. Hunt et al. 2004). Nudging is therefore considered an outstanding partner for techniques combining two different data assimilation schemes incorporating their respective advantages. Especially in reanalysis applications at high resolutions, a smoothness of time series should become an increasingly desirable feature for future developments and applications.

Applying ensemble nudging, the different ensemble members are nudged towards probabilistic observations. Following e.g. Houtekamer et al. 1996, a probabilistic observation is given by perturbing the original observation o by means of a random perturbation o' sampled from a normal distribution $o' \sim N(0, \sigma_o)$ with zero mean and a standard deviation given by the observation error σ_o. We assume normally distributed, unbiased, stationary in time as well as spatio-temporally uncorrelated observation errors. The latter is a wide-spread assumption mostly coming along with observation thinning and inflation of the observation error variances (Lahoz et al. 2010).

The perturbation process of observations is implemented into the limited-area model COSMO as part of the nudging scheme. To provide physically sound observations, those exceeding reasonable value ranges are corrected accordingly. E.g., vertical lapse rates becoming super-adiabatic due to perturbation are corrected to prevent an extensive rejection of the probabilistic observations. In principal, observations from all used conventional observing systems including ACAR, AMDAR, TEMPS, PILOT, WIND PROFILER, SYNOP, SHIP and DRIBU undergo the described perturbation process and a suitable quality control thereafter. We have decided to rely on the observation error estimates used by DWD. These have been determined applying the techniques of Hollingsworth and Lönnberg (1986) and Desroziers et al. (2005) to

feedback data from other non-convection resolving NWP systems of similar resolution like COSMO. The latter is of particular importance to guarantee for a reasonable estimation of the representativity component. The DWD observation error estimates have mainly been used for the quality control in the regional NWP system. Recently, their magnitude has been rechecked and partly reconfirmed or updated using feedback output from the new LETKF data assimilation scheme.

1.2 A local ensemble transform Kalman filter for the convective scale

In the COSMO priority project KENDA[1] (**K**ilometer-scale **En**semble **D**ata **A**ssimilation), a local ensemble transform Kalman filter (LETKF) for the convective scale has been implemented under the direction of DWD (C. Schraff, FE12). The implementation follows Hunt et. al, 2007. At the moment, the preparation of a pre-operational mode is running.

A LETKF is an effective version of an ensemble Kalman filter, whereby the analysis is a linear combination of the background ensemble. In the DWD implementation, a square-root filter is applied to the analysis covariance matrix to derive the analysis. In “LETKF” the word “transform” means that the background covariance matrix is transformed into ensemble space in order to perform the analysis in a low-dimensional sub-space leaving the problem to a strongly reduced rank. The word “local” describes that the covariance matrix is localized and that locally independent analyses are performed, i.e. domain and covariance localization are applied. For reduced-rank Kalman filters that are confined to a subspace spanned by the ensemble members, the degrees of freedom are too few to fit the number of observations. Moreover, the low-rank sample covariance matrices contain a great deal of spurious long-range correlations yielding spurious analysis increments. These rank deficiency problems can be reduced if localization is applied. Making use of domain localization, the number of degrees of freedom of the sub-space in which the analysis is constructed is strongly increased as each grid point is updated using a different linear combination of the ensemble perturbations. The fact that only observations in near-distance to the grid point are used assures that the analysis at each grid point is not influenced by distant observations which would be induced by long-range correlations in the background covariance matrix. It is important that near-by grid points see approximately the same observations so that the analysis is balanced and discontinuities leading for example to artificial divergences can be avoided. Covariance inflation additionally suppresses spurious correlations. It is not trivial to choose the right function for covariance localization. This is an area of active research (Flowerdew, 2015, Perianez, 2014). Furthermore, the choice of a good degree of localization is expensive and tuning-intensive as sampling errors, computational efforts and imbalance errors need to be balanced. In the current implementation of KENDA-LETKF, the localization in the vertical is fixed while a simple adaptive localization method is applied in the horizontal.

The LETKF algorithm works as follows. Firstly, a background ensemble is computed. Then, the ensemble perturbations are used to estimate a flow-dependent sample background error covariance matrix. The analysis mean state (best linear unbiased estimate) is computed adding the ensemble mean forecast to a weighted sum of the ensemble perturbations which are nothing else than the deviations of the different ensemble forecasts from the ensemble mean. The weights are determined depending on the deviation between the ensemble members and the observations. Subsequently, the analysis error covariance is estimated which is followed by an identification of the analysis perturbations as a linear combination of the forecast perturbations. The analysis perturbations are determined such that they reflect the analysis error covariance.

1 See http://www.cosmo-model.org/content/tasks/priorityProjects/kenda/default.htm

Since the background ensemble usually underestimates spread, different methods need to be applied to inflate the ensemble variance. At DWD, experiments with different methods have been performed (see Schraff et. al 2015, submitted). E.g. multiplicative covariance inflation compares the estimated background and observation error covariance matrices to the real deviations obtainable from observation-background statistics. Thereby, a lack of variance can be determined which can subsequently be multiplicatively inflated. Note that the inflation factor varies in time and space and has upper and lower boundaries. Another method that aims at a boost of ensemble variance is relaxation to prior perturbation (RTPP). Thereby, the analysis perturbations are relaxed towards the background perturbations. A further possibility is relaxation to prior spread (RTPS). Above all, experiments have shown that perturbed lateral boundary conditions are essential to obtain a suitable spread. Particularly experiments with an ICON-ensemble have resulted in fundamental improvements. The system has proven to benefit from a perturbation of the soil moisture in the soil moisture analysis which improves spread and scores in the boundary layer. Application of latent heat nudging to all ensemble members has shown to be very beneficial.

At the moment, the KENDA experiments make exclusively use of conventional observations. It could be shown that the LETKF is already superior to nudging using this set of observations, so that little disagrees with introducing it to operations at this point of time, even though the assimilation of modern observations is not yet ready for operational use. The latter is under extensive development and is expected to lead to further improvement of the convective-scale analysis and forecast quality at DWD. For further details we refer to Schraff et. al 2015, submitted.

1.3 Theoretical advantages of combining ensemble nudging with LETKF

Building on the findings described in the foregoing sections, a combination of ensemble nudging and LETKF can be considered very useful for future developments. On the one hand, ensemble nudging proves to have a comparably good spread. Thus, providing a nudging ensemble to the LETKF may reduce the necessity of covariance inflation, RTPP, RTPS etc. On the other hand, observations from modern observing systems like GPS, satellite data or radar data could be assimilated incorporating the LETKF to ensemble nudging. Application of nudging over long time-windows has the merits that it provides continuously physically-balanced time series and moreover a model trajectory that is close to the one of the true atmosphere over the whole reanalysis time span.

1.3.1 Design of the system and practical problems

The EN-LETKF system makes use of a 4-day integrated full nudging-ensemble that comprises 20 members as initial conditions. Currently, a six-hourly LETKF-cycle is run that is provided with a nudging-ensemble instead of the usual short-range forecast as a background. The observations need to be distributed between the systems. Due to their quasi-continuous availability, we assimilate the observations from AIRREP reports in ensemble nudging. Thus, the ensemble is henceforth only generated through upper-air wind and upper-air temperature perturbations. In the LETKF, observations from SYNOP reports, wind profilers as well as TEMP reports are assimilated. We employ RTPP together with multiplicative covariance inflation. So far, we observe a strong dry-bias in humidity, precipitation as well as a cold-bias in screen-level temperature. This may be due to the following reasons:

(1.) In the KENDA-experiments (see Schraff et. al 2015, submitted) it has been observed that at least 40 members are needed to achieve a meaningful ensemble variance and thus a high-quality analysis. We consider 40 ensemble members a high number for reanalysis purposes and need to

reconsider the performance-cost ratio.

(2.) At the point of time of application for UERRA it had been expected that the integration of assimilation of modern observations at DWD would be at a more developed stage. This would have allowed for assimilating all conventional observations in ensemble nudging and only the modern ones in LETKF, presumably leading to an improvement in analysis quality. To date, we have to divide the set of conventional observations between the two data assimilation systems. We consider it possible that the observation density thus becomes unfeasibly low for the LETKF (which is particularly critical for the adaptive covariance localization) leading to undesirable and so far non-comprehensible results.

(3.) The spread of the nudging ensemble is substantially degraded assimilating only AIRREPs instead of the full set of observations. Thus its original advantage of yielding more spread than the normal background ensemble becomes neglectable. Furthermore, no humidity observations are assimilated in ensemble nudging in the current version which subjects the model fully to the influence of the lateral boundary conditions.

(4.) As mentioned, the KENDA-LETKF has proven to benefit heavily from perturbed lateral boundary conditions. So far, we do not have any global reanalysis ensemble available which could be used to enhance spread. Moreover, it is rather unlikely that a 40-member global ensemble reanalysis will come into existence in the foreseeable future. It is envisaged to prepare a global ICON-ensemble reanalysis. However, this is another great reanalysis effort that must be seen as a future perspective that will offer more possibilities to our regional reanalysis work.

(5.) As described in section 1.2, the LETKF is sensitive to a range of parameters needed for domain and covariance localization. The parameters, such as a localization radius, have been determined for the convective scale (a 2.8 km COSMO version) in extensive tuning experiments at DWD. So far, we have not retuned the parameters for the 12 km scale. It is most likely that the parameters suitable for the convective scale are not usable for the 12 km scale.

It can be summarized that the recent findings at DWD concerning the LETKF, such as an unconditional need of perturbed boundary conditions and a number of at least 40 ensemble members were not foreseeable at the point of time of application for UERRA. At this point, the LETKF was at the stage of implementation and not yet at an experimentation level. Additionally, the combination of LETKF and ensemble nudging seems to be little meaningful as long as the observation stream cannot be expanded using satellite, radar or GPS data. Subdividing the conventional observation stream between LETKF and ensemble nudging we run into problems that do not seem to be trivial at this stage. However, the EN-LETKF seems to have great potential given that its needs are fulfilled so that the idea should not be put aside. Particularly for the convective scale the combination should demand further attention. In case that the ongoing research and experiments with the EN-LETKF do not result in a suitable quality, of the system, ensemble nudging data assimilation system for the production of a 5-year test reanalysis. This is advantageous in many respects, above all, because we could rely on a well-tuned and highly developed implementation of nudging at a great performance-cost ratio and with good uncertainty estimation capabilities. In the next sections, we evaluate the performance of ensemble nudging experiments.

2 Experiments and comparison data sets

2.1 Experimental set-up

Ensemble nudging is implemented in the limited-area NWP model COSMO of the **Co**nsortium for **sm**all-scale **mo**deling. The model is non-hydrostatic and targeted at the representation of meso-alpha and meso-beta processes.

The version of COSMO-EU that is now running at DWD in operational mode for several years, uses a grid spacing of 7 km. For the ensemble reanalysis purpose we adapt it to a grid resolution of 12 km and to the geographical extension of the CORDEX-EUR11 domain (Giorgi, 2009). In the employed version, COSMO has 40 hybrid levels in the vertical. The soil model TERRA makes use of 7 vertical layers going down to approximately 14.5 m depth. The model equations are solved on a rotated latitude-longitude grid that avoids a convergence of the meridians and allows for equidistant grid points. The domain consists of a number of 424 x 412 grid points. The exact domain specification is summarized in Table 1, Figure 1 shows its geographical extension.

Table 1: Domain specification of CORDEX-EUR11.

	CORDEX-EUR11
Rotated north pole coordinates	-162.0°, 39.25°
Lower left corner (rotated coordinates)	-23.375°, -28.375°
Grid spacing	0.11°
Number of grid points (lambda x phi)	424 x 412

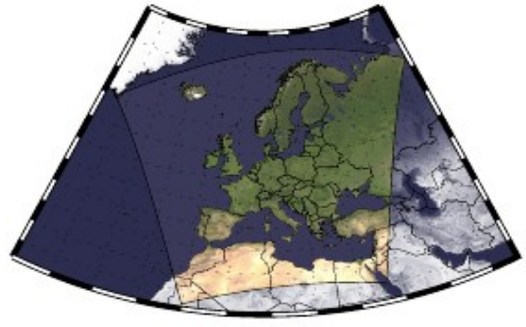


Figure 1: CORDEX-EUR11 domain.

Since there is no suitable global reanalysis ensemble available yet, we make use of the global ECMWF ERA-INTERIM reanalysis as initial and boundary data. To allow for a 3-hourly update of the boundary data we use the analyses at 00 and 12 UTC and reforecasts at +03, +06 and +09 h for consistency reasons. ERA-INTERIM has 0.7° grid resolution (80 km, but used at 0.5º).

Nudging analyses need to be performed in model space. Therefore the range of observations available for assimilation is limited to conventional observations. The observation types and assimilated quantities are summarized in Table 2. Note that the perturbation process of observations is not restricted to a specific variable or observing system, but rather all observations that have status “active” after the quality control are perturbed to generate the nudging ensemble. Additionally to nudging, three offline analysis schemes are applied including analysis of the snow depth, sea surface temperature (SST) analysis and a variational soil moisture analysis (SMA). The snow analysis is performed at 00, 06, 12 and 18 UTC. The SST analysis is performed once a day at 00 UTC. The SMA is applied daily at 00 UTC. For further details we refer to (Bollmeyer, 2015 or Schraff and Hess, 2003). The process cycle is depicted in Figure 2.

We present two case studies comprising a winter month (December 2011, referred to as “winter experiment”) and a summer month (June 2011, referred to as “summer experiment”). Both comprise 20 ensemble members and a control run which is simply a nudging run with original, unperturbed observations. We refer to the ensemble nudging experiments as “C-EN”.

Table 2: Observation variables assimilated from different reports.

	Reports	Assimilated observation variables
Radiosondes	PILOT	Upper-air wind
	TEMP	Upper-air wind, temperature, humidity, screen-level wind, temperature, humidity, geopotential
Aircrafts	AIRREP	Wind, temperature
	AMDAR	Wind, temperature
	ACARS	Wind, temperature
Wind profiler		Upper-air wind
Surface observations	SYNOP	Surface pressure, screen-level wind, 2m humidity
	SHIP	Surface pressure, 10m-wind, 2m-humidity
	DRIBU	Surface pressure, 10m-wind, 2m-humidity

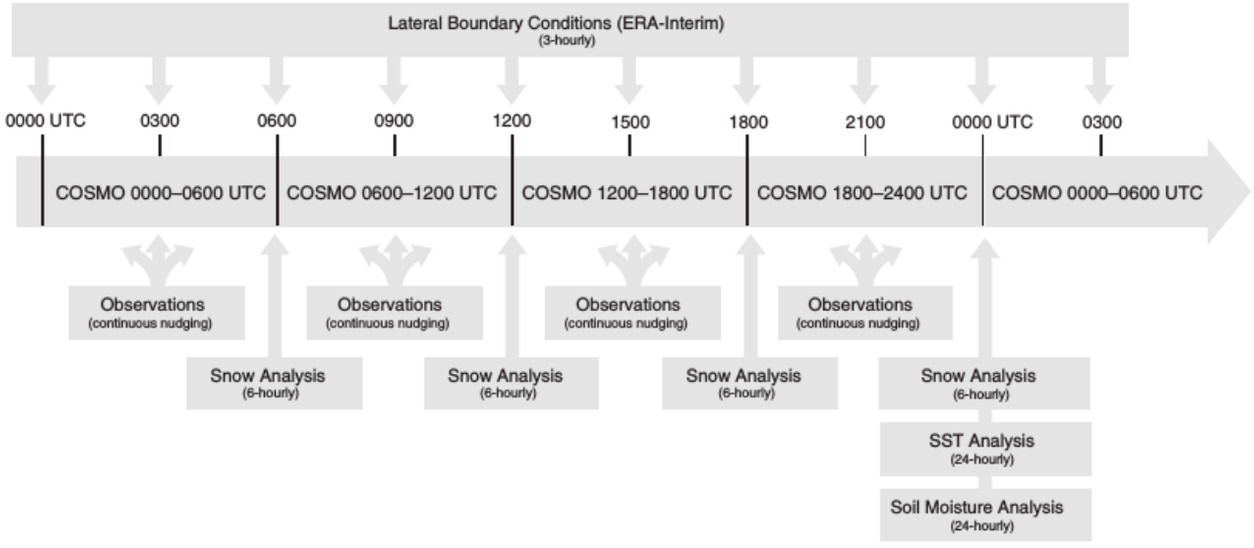


Figure 2: Process cycle of the reanalysis system, adapted from Bollmeyer, 2015, Figure 2.

2.2 Data sets for comparison

We consult different reanalysis, downscaling and observation data sets for comparison. These comprise ERA-INTERIM global reanalysis data, a regional reanalysis HIRLAM, a dynamical downscaling from ERA-INTERIM using COSMO at 6 km grid spacing, a deterministic nudging reanalysis using COSMO at 6 km grid spacing and rain gauges observations from the German SYNOP network. The specifications of the data sets are summarized in the following tables.

Table 3: Reanalyses used for comparison.

	Institution	Time span	Time steps	Model + DA	Grid spacing	Obs stream	Quantity used	References
ERA-INTERIM	ECMWF	June 2011 Dec 2011	12-hourly analysis window, 00, 12 UTC reanalyses +03,+06,+09 reforecasts	IFS + 4D-Var	0.5 °	Conventional + satellite	3-hourly precip	Dee, 2011
Hirlam	SMHI	June 2011	12-hourly analysis window, 00, 12 UTC reanalyses	Hirlam + 3D-Var	0.2°	Conventional	3-hourly precip	Undén et al. , 2002

			+03,+06,+09 reforecasts					
COSMO-REA6	UB	June 2011	Continuous reanalysis 00,03,06,09,12,15,18,21 UTC	COSMO + Nudging	0.055°	Conventional	3-hourly precip	Bollmeyer, 2015

Table 4: Ensemble prediction system used as a benchmark.

	Institution	**Time span**	**Time steps**	**Model**	**Grid spacing**	**Number of members**	**Quantity used**	**References**
ECMWF-EPS	ECMWF	June 2011 Dec 2011	00UTC + 06 12UTC + 06	IFS	0.25°	50	6-hourly precip	

Table 5: Dynamical downscaling used for comparison.

	Institution	**Time span**	**Time steps**	**Model + DS reanalysis**	**Grid spacing**	**Quantity used**	**References**
COSMO-DOWN6	UB	June 2011	00,03,06,09,12,15,18,21 UTC	COSMO + ERA-INTERIM (3-hourly)	0.055°	3-hourly precip	Bollmeyer, 2015

Finally, Figure 3 shows the spatial distribution of about 1000 rain gauge observations from the DWD SYNOP network which provide the verifying observation used in the evaluation section. We employ both 3-hourly and 6-hourly accumulated precipitation sums. Note that the verification comprises only the German subdomain of the CORDEX-EUR11 area. The observation-grid point assignment is carried out using the method of nearest-neighbour.

Figure 3: Rain gauge stations in the German subdomain providing accumulated precipitation sums as verifying observation to evaluate the performance of ensemble nudging.

3 Evaluation of performance

3.1 Developing a guideline for evaluation

Users of regional reanalyses need information about the quality and usability of the Essential Climate Variables (Bojinski, 2014). It is self-evident that some variables are more valuable for specific applications than others for which the user should rather rely on other climate data sets such as ERA-INTERIM, satellite climatologies like CM-SAF, SYNOP data, dynamical downscalings or gridded observations like EOBS. However, since the establishment of climate centres (such as the Copernicus Climate Change Service), the development of regional reanalysis systems and even more the estimation of their uncertainties are still in their infancy, some pioneering work has to be done to figure out the advantages and disadvantages of the different Essential Climate Variables in the available data sets. This may lead to further generations of regional reanalyses that will focus on the elimination of potential weaknesses of the systems.

As a kind of “first guess”, some regional reanalysis variables are presumed to be superior to their representation in other climate data sets. We refer to that as „added value“. During the initial development process of the reanalysis system it is not possible to check for the value of the whole set of Essential Climate Variables. In a first step, we focus on precipitation. Thereby independence of the verifying variable

is achieved and the necessity of computing short-range reforecasts which merely represent a downstream product can be avoided. Moreover, reforecasts may neither be totally independent of the verifying observation, at least in the case of ensemble nudging. This can be traced back to the form of the temporal nudging weighting functions which assign a non-zero weight to observations with observational time in the near future (this does not hold for precipitation unless latent heat nudging is applied, then the analysis would be dependent on radar observations not on rain gauges). We hypothesize the following indication of added value of precipitation in regional reanalysis (RR):

- RR exhibit spatio-temporal completeness compared to rain gauge observations.
- RR are supposed to have 3-dimensional physical and inter-variable consistency compared to gridded observations.
- RR have higher grid resolution so that processes can be represented non-hydrostatically. Moreover orographic effects, land ocean contrasts and land use effects can be better represented. The observation stream similar to the one used for the production of the lateral boundary conditions (global reanalysis) is assimilated on smaller length scales which yields different, scale-relevant information. *Mesoscale representation of precipitation should lead to better verification than large-scale representation in global reanalyses.*
- Due to data assimilation yielding spatio-temporal accuracy RR are supposed to be usable as both climate and weather data sets whereas dynamical downscaling can only reproduce climatological distributions of variables. *This implies that RR exhibits accuracy whereas dynamical downscalings do not. In particular the latter do not show sharpness from the point of view of forecast verification* (Murphy and Winkler, 1987).

We test the latter two hypotheses and put the first two aside. We go through the following list of reanalysis attributes to test the performance and added value of the system:

- *Accuracy* which describes the spatio-temporal correspondence of model and observations. We investigate accuracy by means of 3-hourly accumulated precipitation and employ the log odds ratio and proportion correct as a measure of performance.
- Attributes of probabilistic reanalysis skill that we observe are
    - *Reliability* which is a statistical indistinguishability of the verifying observations and the ensemble members as well as an agreement of the ensemble probabilities with the corresponding observed frequencies. Climatology is perfectly reliable even though it says nothing about the coherence of observations and model. Reliability can be affected by model bias. We investigate reliability using analysis rank histograms and reliability diagrams as well as the reliability component of the decomposed Brier score.
    - *Spread-skill relationships* that measure the spread's average capability of estimating the analysis error. EPS spread is supposed to resemble the uncertainty underlying the forecast which depends on the “errors of the day”. It indicates predictability, i.e. it is supposed to measure the mean deviation to observations that the forecast will exhibit. In contrast, the spread of an ensemble reanalysis will estimate the uncertainty that evolves

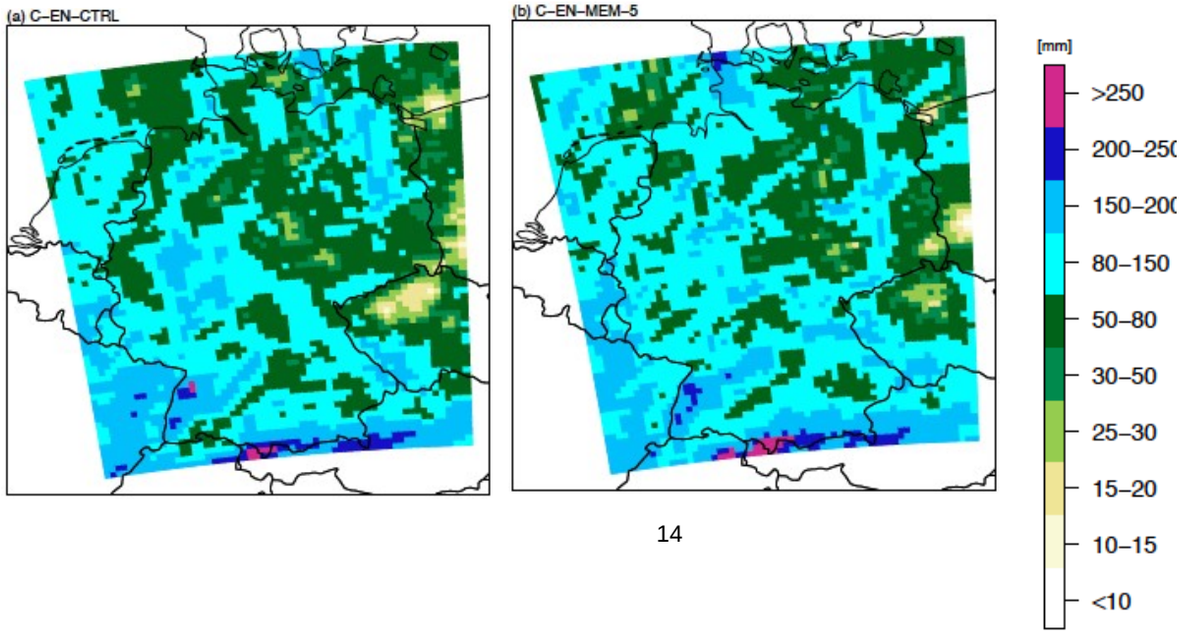
from the component of the NWP system being perturbed as well as random errors that the perturbations project on. E.g. perturbing the lateral boundary conditions may yield a different spread than perturbing the observations. Obviously, only perturbing all uncertain ingredients of the NWP system using profound error estimates and an extensive number of ensemble members would allow for a comprehensive sampling of the true subspace of uncertainty.

- *Resolution* which is the ensemble's ability to assign different ensemble pdfs to different events. It depends strongly on the quality of the model and on the kind of ensemble generation method. Resolution is connected to discrimination which measures if different observed outcomes are correctly distinguished between by the forecasts. Resolution can be investigated by employing the resolution component of Brier score, but also by the Relative Operating Characteristics curve.

For future evaluation we envisage to utilize methods like kernel dressing or fitting suitable parametric distributions to obtain a realistic interpretation of the ensemble probabilities at a limited number of ensemble members. However, to obtain a first impression of the performance of the ensemble nudging system we employ a frequentist interpretation, i.e. the probability for a specific event is estimated by the fraction of ensemble members analysing the event.

Note that the verification scores and functions used in this deliverable can be found in Jollife and Stephenson, 2012 and citations of the corresponding original publications therein.

3.2 A first look on precipitation fields using monthly precipitation climatology



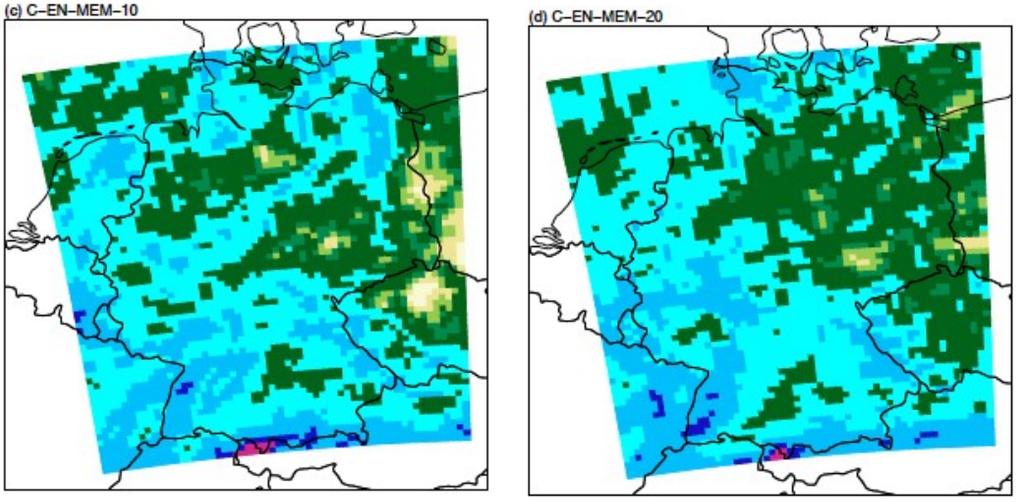


Figure 4: Monthly precipitation climatology for June 2011. a) Control run of ensemble nudging. b) to d) Arbitrarily chosen ensemble members.

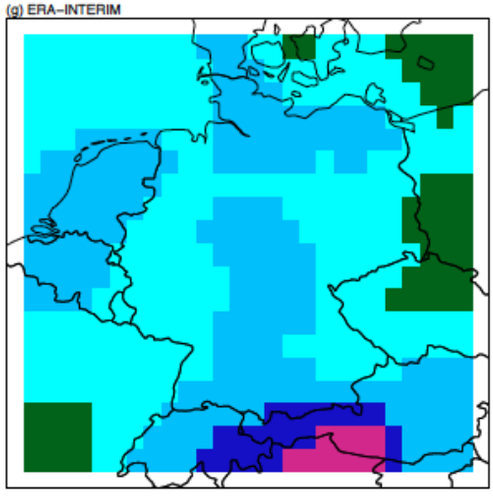


Figure 5: Monthly precipitation climatology for ERA-INTERIM, June 2011.

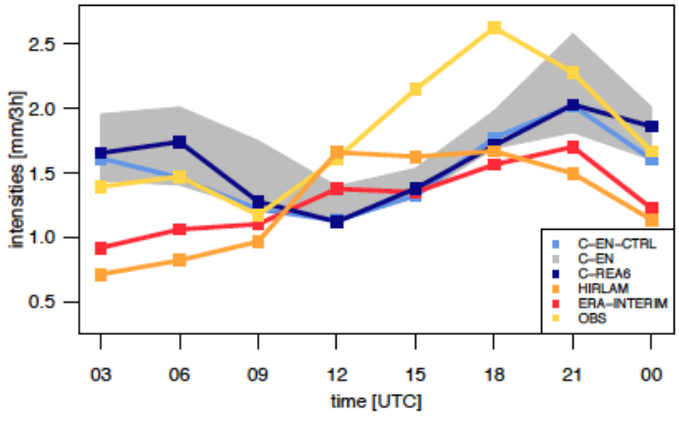


Figure 6: Diurnal cycle of precipitation for reanalyses. and rain gauges.

Figures 4 and 5 show monthly integrated precipitation for the nudging control run and three ensemble members as well as for ERA-INTERIM. One of the main applications of reanalysis data is climate monitoring which among other things is interested in the temporal evolution of monthly climatologies or anomalies from the long-time mean[2]. Visual observation of these climatologies already reveals important features of the ensemble nudging reanalysis. On the one hand, mesoscale variability catches the eye when comparing to ERA-INTERIM. On the other hand, the nudging ensemble members exhibit enough similarity to exclude randomness that might be induced by the observation perturbations, but at the same time a degree of variability that indicates uncertainty in the details of the precipitation patterns.

Figure 6 displays the diurnal cycle of precipitation (mm/3h) for ERA-INTERIM, Hirlam, the ensemble nudging control run (C-EN-CTRL), ensemble nudging (C-EN), COSMO-REA6 (C-REA6) and rain gauge

2 See https://www.wmo.int/pages/themes/climate/climate_monitoring.php

observations (OBS). For more details concerning the data sets see section 2.2 of this deliverable. It is a well-known problem that COSMO places the precipitation maximum several hours too late (see e.g. Bollmeyer, 2015). This can be observed for all reanalyses. However, the upper limit of the nudging ensemble is capable of reaching the maximum value, even though shifted in time. Thus, the observation perturbations have a positive impact on the representation of the diurnal cycle of precipitation. Secondly, the diurnal cycle gets more pronounced with increased resolution. Surprisingly, COSMO-REA6 does not have a significantly better diurnal cycle than C-EN-CTRL which has only half of the grid spacing.

3.3 Measuring performance of three-hourly integrated precipitation

To obtain an impression of the performance of precipitation in the reanalysis experiments, we make use of a contingency table for binary events. As verifying observation we use about 1000 rain gauge measurements (see Figure 3). We are both interested in the agreement of the climatological distribution of precipitation events in reanalysis and observations using the corresponding marginal distributions and in their spatio-temporal coherence represented by their joint distributions. To investigate the first aspect we employ the frequency bias given by

$$FB = \frac{a+b}{a+c} \quad FB \in [0, \infty].$$

It compares the number of „yes events“ in the reanalysis (hits „a“ and false alarms „b“) to the number of true „yes events“ (hits „a“ and misses „c“). The frequency bias disregards the spatio-temporal coherence of observations and model. However, this is considered by the log odds ratio which is obtained by

$$LOR = \log\left(\frac{ad}{bc}\right) \quad LOR \in [-\infty, \infty].$$

Therein, „d“ are the correct negatives. The log odds ratio gives great weight to the correct negatives (as these usually represent the main part of the events) so that it tends to assign better scores to rarer events. It measures the ratio of the odds of making a hit to the odds of making a false alarm. The proportion correct score compares the correctly captured events to all events

$$PC = \frac{a+d}{a+b+c+d} \quad PC \in [0,1].$$

It is very useful, however, for rare extreme events the proportion of possible hits becomes so low that it is not longer meaningful. Therefore we introduce a weighted proportion correct, which is weighted by means of the probability of detection. Thereby it allows to distinguish between accurate and non-accurate systems:

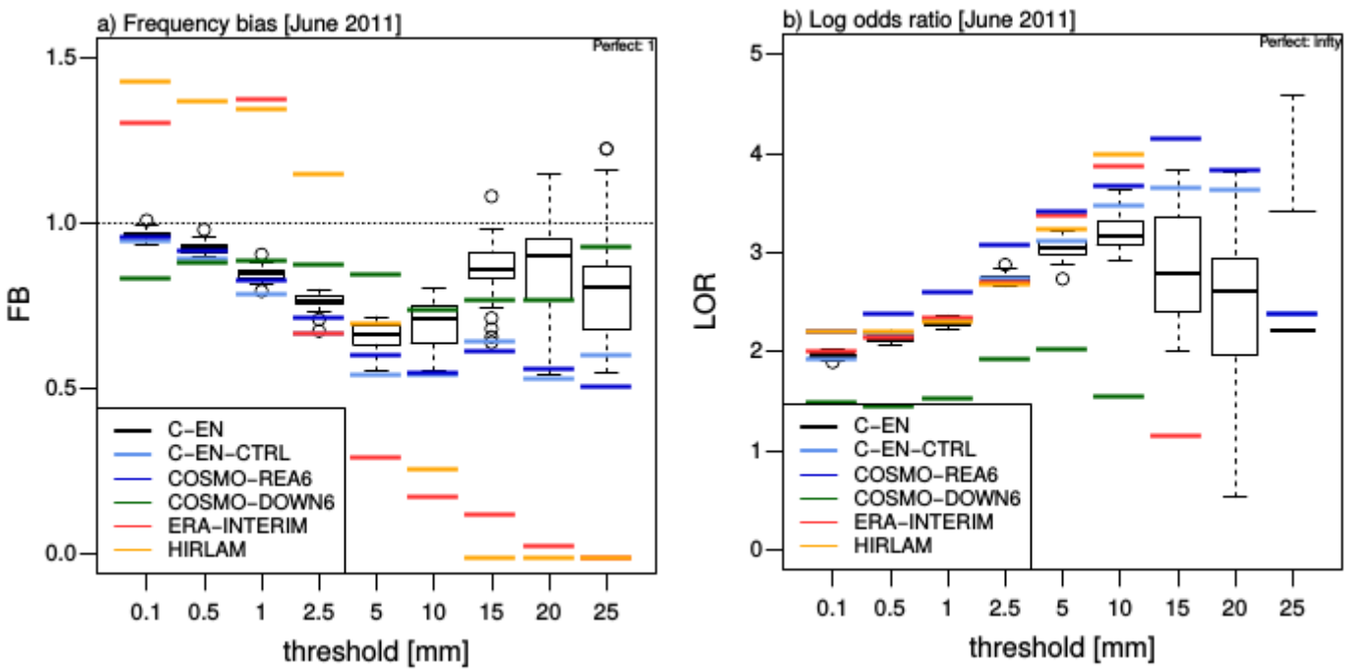
$$wPC = \frac{a+d}{a+b+c+d} * \frac{a}{a+c} \quad wPC \in [0,1].$$

In the following, we focus on an evaluation of the summer experiment, since to date we do not have all reanalyses available for comparison for the winter experiment. The frequency bias displayed in Figure 7a shows that the nudging ensemble underestimates the “yes events” at the lower thresholds, particularly at 2.5 and 5 mm. The obs perturbations have a positive impact as they increase the number of yes events in the ensemble members compared to the control run (C-EN-CTRL) which is presented in a light blue.

Ensemble nudging is significantly closer to a perfect FB of 1 than ERA-INTERIM (red) and Hirlam (orange) which produce too many events of small precipitation amounts and too few events at the higher thresholds. The deterministic nudging reanalysis (COSMO-REA-6) (darkblue) at a grid resolution of 6 km is comparable to C-EN at the smaller thresholds. From 10 mm upwards the observation perturbations are very beneficial regarding frequency bias as the whole ensemble outperforms COSMO-REA6 and C-EN-CTRL. The outliers of the ensemble even overestimate the extremer events. A clear advantage of C-EN compared to Hirlam and ERA-INTERIM can be observed. The COSMO dynamical downscaling from ERA-INTERIM to 6 km (green) has a frequency bias competitive or even superior to to the one of C-EN which proves its usability for climatological studies that focus on frequencies.

Regarding the log odds ratio shown in Figure 7b, the downscaling is by far worse. Here, the added value of both regional and global reanalysis which employ data assimilation to keep the model trajectory as close as possible to the true trajectory of the atmosphere becomes obvious. At the lower thresholds, Hirlam, ERA-INTERIM and ensemble nudging are roughly comparable, whereas at 5 and 10 mm the first two are better.The coarser resolved reanalyses tend to produce smaller numbers of events at the higher thresholds (see frequency bias) so that they have less hits, but also more correct negatives which have the main weight in the score. The control run lying in the upper quantiles of the nudging ensemble indicates that the obs perturbations lead to more precipitation by destabilizing vertical profiles, however, not necessarily in the right places. This leads to significantly lower numbers of correct negatives at the high thresholds which in turn yields a lower log odds ratio. COSMO-REA6 and C-EN-CTRL benefit from their higher resolution which allows for a representation of extremer precipitation amounts. However, the observation perturbations degrade the accuracy of extremer events (compare the ensemble to the control run) as they place the precipitation to the wrong locations yielding lower amounts of correct negatives that lead to a lower log odds ratio.

The proportion correct in Figure 7c proves the superiority of COSMO-REA6 and C-EN over Hirlam and ERA-INTERIM at the lower thresholds. Here, it shows that COSMO-REA6 does have more accuracy than the other reanalyses (see Figure 7d). At 2.5 and 5 mm all reanalyses are fairly comparable. For the higher thresholds, the proportion correct is weighted by the probability of detection which reveals significantly more accuracy of the higher resolved COSMO reanalyses. At 15 mm the wPC tends to zero for Hirlam and ERA-INTERIM. At 20 and 25 mm decision thresholds the only reanalysis that maintains a degree of accuracy is COSMO-REA6.



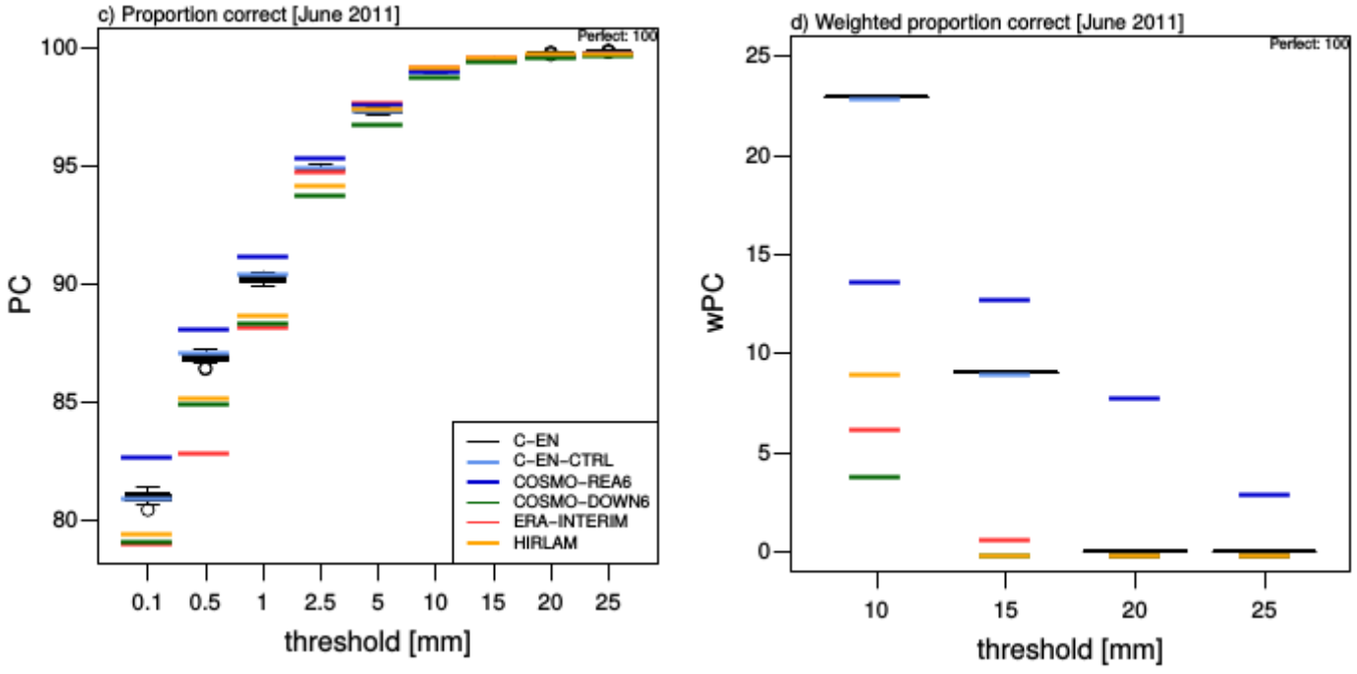


Figure 7: a) Frequency bias, b) log odds ratio, c) proportion correct and d) weighted proportion correct of the summer ensemble nudging experiment in comparison to ERA-INTERIM (red), Hirlam (orange), COSMO-REA6 (dark blue), C-EN-CTRL (light blue) and COSMO-Downscaling (green).

Conclusions:

- The obs perturbations have a positive impact on the frequency bias.
- The obs perturbations tend to degrade accuracy at the extreme thresholds, therefore the uncertainty estimation at the high thresholds cannot be expected to be meaningful at a small ensemble size.
- C-EN has added value compared to ERA-INTERIM and Hirlam. This expresses in

- a significantly better frequency bias yielding value for climatological studies.
- more accuracy which is particularly proven by the proportion correct and its weighted version.
    - particularly the representation of extreme events is significantly better than in the coarser resolved reanalyses.

- C-EN has added value compared to the dynamical downscaling of ERA-INTERIM using COSMO at 6 km grid spacing
    - a superior accuracy of C-EN is shown both by the log odds ratio and by the proportion correct.

Thus all hypotheses regarding the added value of the representation of precipitation in regional reanalyses (see section 3.1) could be proven for the summer experiment with the ensemble nudging regional reanalysis system. In the next sections we examine the probabilistic and uncertainty estimation capabilities of the system.

3.4 Capability of uncertainty estimation and reliability

The analysis rank histograms for 3-hourly accumulated precipitation shown in Figure 8 provide a measure of reliability. Reliable ensemble systems consist of ensemble members that are statistically indistinguishable and sample the true distribution of possible outcomes. In analysis rank histograms, deviations from the true pdf (or the observational pdf that may be contaminated by observational errors) can express as a bias, i.e. systematic error in the expectation value of the pdf (if it is a normal distribution). Another problem is under-dispersiveness meaning that the sampled pdf is too sharp. The opposite problem is over-dispersiveness.

To avoid artefacts due to a limited sample size we exclude the all-zero events from the data sets (both observation and all ensemble members indicate zero precipitation, usually the observation would obtain a random rank drawn from a uniform distribution).

The analysis rank histogram for the summer experiment displayed in the upper picture of Figure 8 shows both a weak low- and a high-bias. The high-bias can be traced back to events that are not captured by any of the ensemble members. This is presumably a matter of extreme events. The low-bias arises from events for which the whole ensemble over-estimates the precipitation. The experiment has a spread-skill ratio of 0.79. Here, spread is measured in terms of standard deviation and the model-obs deviation in terms of RMSE.

The winter experiment shows optically more pronounced under-dispersiveness and achieves a weaker spread-skill ratio (0.67). This may indicate that the uncertainty estimation works slightly better for summer conditions. However, a larger sample size would allow to divide the data into subsets to obtain further insight into biases in dependence of thresholds and locations.

The reliability diagrams shown in Figure 9 are computed based on 6-hourly accumulated precipitation from +6h ECMWF-EPS forecasts started at 12 and 00 UTC and based on C-EN reanalyses at 06 and 18 UTC. The observed relative frequencies are conditioned on the ensemble probabilities. The error bars are so-

called consistency bars which are estimated based on consistency sampling following Bröcker and Smith, 2007. The method was developed to cope with the fact that not even a perfectly reliable NWP system would yield an exactly diagonal diagram due to limited sample size. In one resampling cycle, the whole set (N) of reanalysis probabilities is sampled into a new order. A corresponding set of binary observations is generated drawing an independent uniformly distributed random variable of sample size N which is assigned 1 where it is smaller than the resampled reanalysis probability and 0 elsewhere (index per index). By definition, the resampled reanalysis set is reliable for the new binary observations. In our case, this cycle is repeated 5000 times. We plot bars that extend from the 5% to the 95% quantiles. Thus, where the reliability curve falls within the consistency bars, it is reliable.

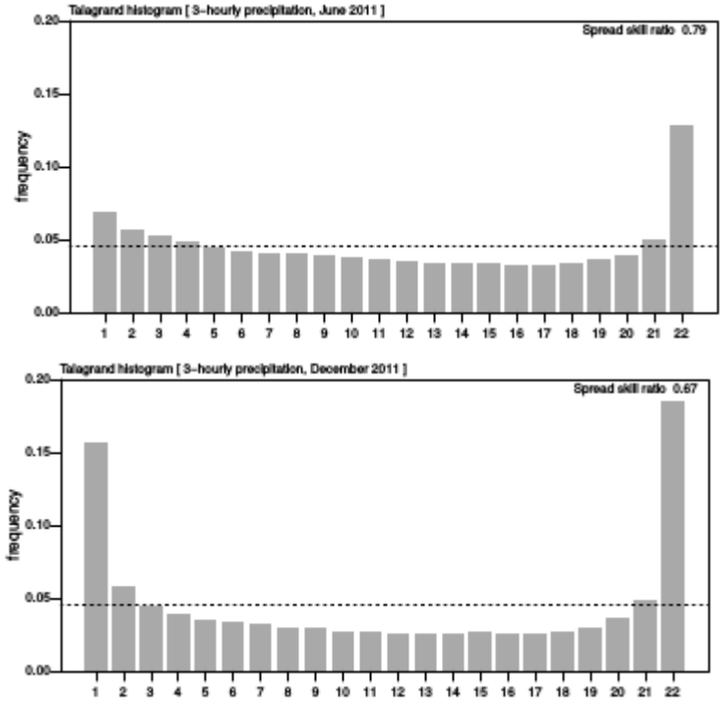
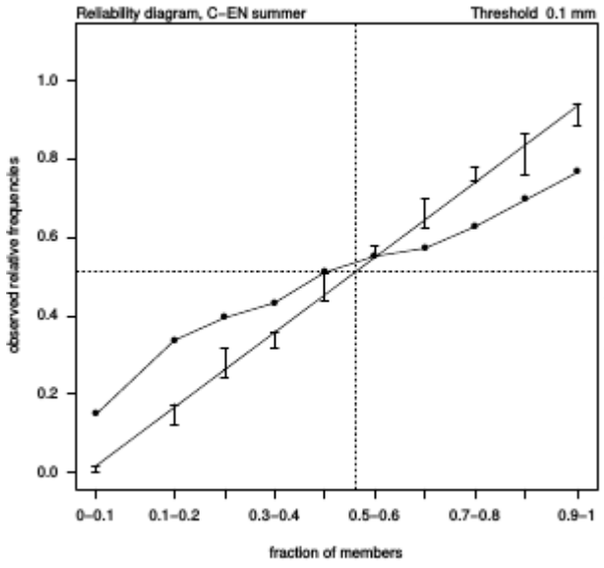
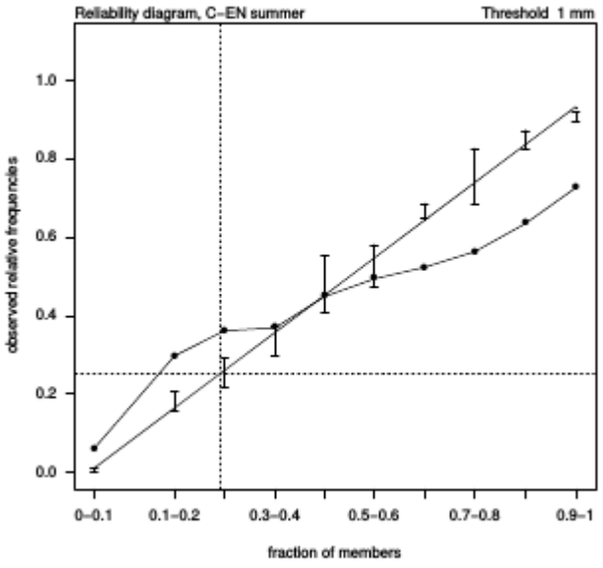
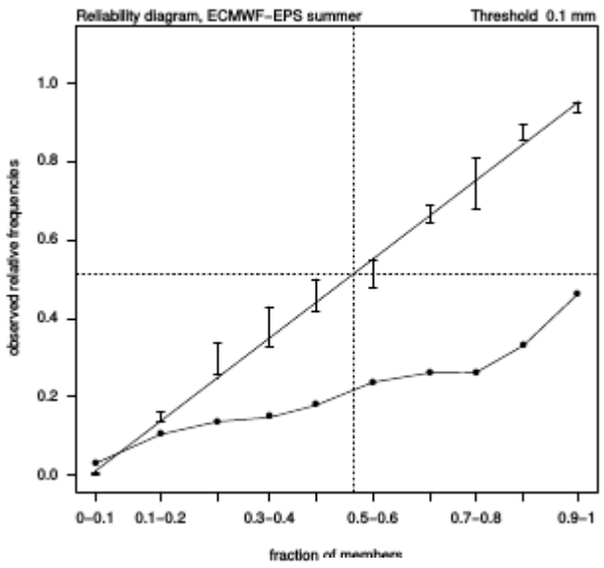


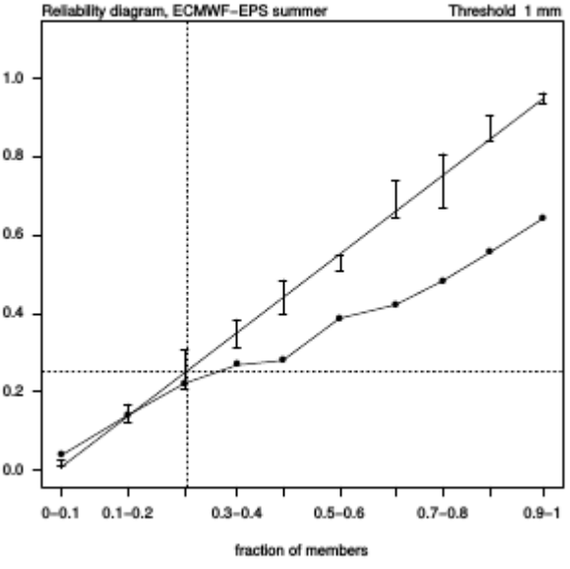
Figure 8: Analysis rank histograms computed from 3-hourly accumulated precipitation using rain gauge observations in the German subdomain.

At 0.1 mm decision threshold, a strong over-forecasting can be observed for the ECMWF-EPS for all probabilities. At 1 mm the reliability improves, but the nudging ensemble is still closer to the diagonal. Similar holds for the higher thresholds (not shown here). The consistency bars show that neither of the systems is perfectly reliable.


Reliability diagram, C−EN summer
Threshold 0.1 mm
observed relative frequencies
0.0
0.2
0.4
0.6
0.8
1.0
0−0.1
0.1−0.2
0.3−0.4
0.5−0.6
0.7−0.8
0.9−1
fraction of members


Reliability diagram, C−EN summer
Threshold 1 mm
observed relative frequencies
0.0
0.2
0.4
0.6
0.8
1.0
0−0.1
0.1−0.2
0.3−0.4
0.5−0.6
0.7−0.8
0.9−1
fraction of members


Reliability diagram, ECMWF−EPS summer
Threshold 0.1 mm
observed relative frequencies
0.0
0.2
0.4
0.6
0.8
1.0
0−0.1
0.1−0.2
0.3−0.4
0.5−0.6
0.7−0.8
0.9−1


Reliability diagram, ECMWF−EPS summer
Threshold 1 mm
0.0
0.2
0.4
0.6
0.8
1.0
0−0.1
0.1−0.2
0.3−0.4
0.5−0.6
0.7−0.8
0.9−1
fraction of members

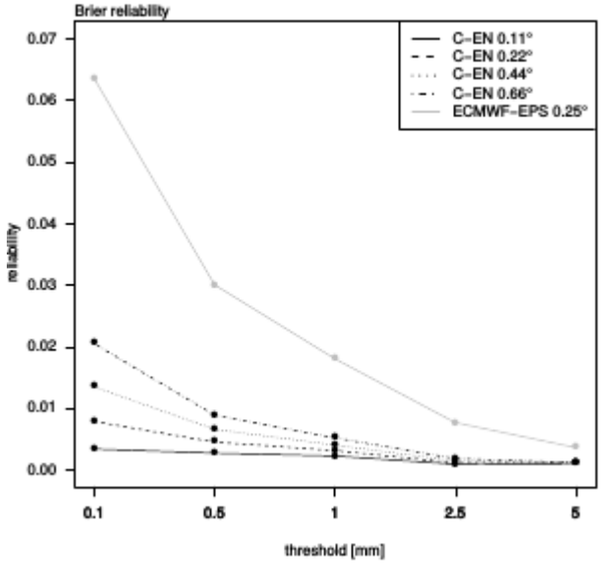

Brier reliability
C−EN 0.11°
C−EN 0.22°
C−EN 0.44°
C−EN 0.66°
ECMWF−EPS 0.25°
reliability
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.1
0.5
1
2.5
5
threshold [mm]

Figure 9 Reliability diagrams, C-EN (top) and ECMWF-EPS (bottom) for 0.1 mm (left) and 1 mm (right), Brier reliability.

The superiority of C-EN with respect to reliability is confirmed by the reliability component of the Brier score (for a definition, see Hersbach 2000) displayed on the bottom of Figure 9 for the summer experiment. It is much closer to zero for C-EN and even stays there after spatial averaging to increasingly coarse grid spacings (dashed lines). Towards higher thresholds the Brier reliability of the ECMWF-EPS increases. For the winter experiment we obtain similar results (not shown here).

It can be concluded that the nudging ensemble is more reliable than the ECMWF-EPS for the chosen time span and subdomain considering 6-hourly integrated precipitation sums. The analysis rank histograms for 3-hourly accumulated precipitation indicate that the ensemble has weak biases at both low and high precipitation amounts. It tends to over-forecast smaller precipitation amounts and under-forecast extreme precipitation. This is rather a problem of grid resolution and convection schemes, i.e. a matter of model biases, than of ensemble generation.

Conclusion:

- Here, the C-EN is more reliable than the ECMWF-EPS short-range forecasts.
- The uncertainty estimation capabilities concerning precipitation estimated by the spread-skill ratio are okay, but improvable.

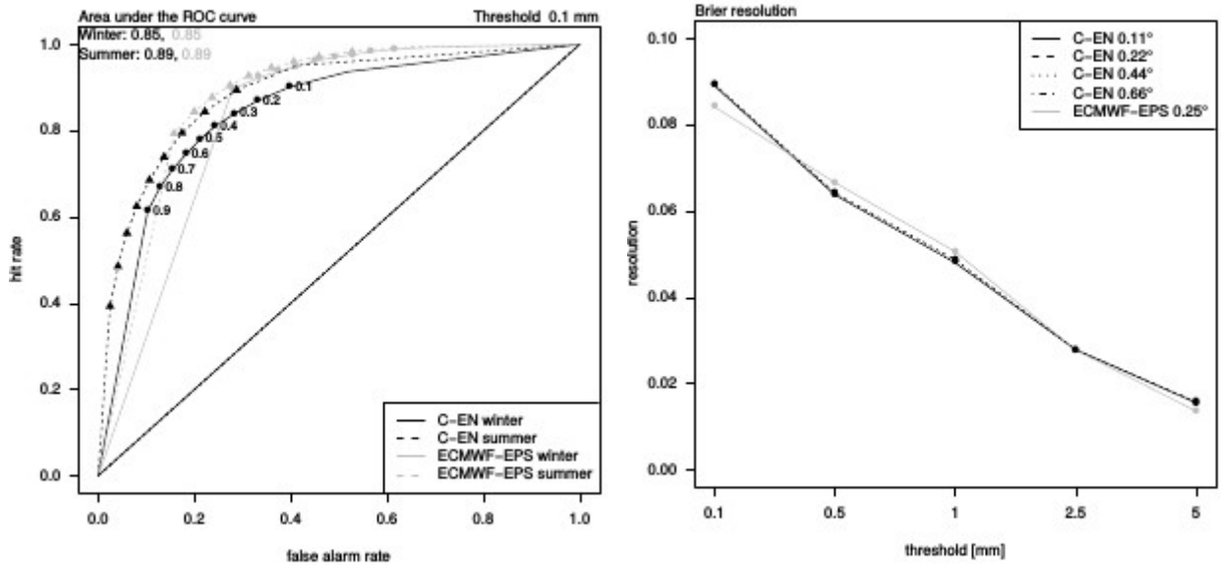


Figure 10: Roc curves for the ECMWF-EPS (grey) and C-EN (black) for winter (continuous lines + points) and summer (dashed lines + triangles) experiments at a threshold of 0.1 mm. Brier resolution component for 5 different thresholds for ECMWF-EPS and C-EN with the latter averaged to coarser resolutions shown by the dashed and pointed lines (summer experiment).

3.5 Resolution

The ROC curve displayed in Figure 10 shows the resolution of the ensemble nudging system in comparison to the ECMWF-EPS based on 6-hourly precipitation sums accumulated to 06 and 18 UTC. The ROC curve is a signal detection curve for binary data whereby the hit rate is displayed versus the false

alarm rate over probabilistic decision thresholds (0 to 1 by 0.1). In a perfect ensemble system the curve would run from (0,0) to (0,1) to (1,1), i.e. low decision thresholds correspond to high hit rates and high false alarm rates whereas higher decision thresholds should come along with high hit rates and low false alarm rates. That is the more confident the ensemble is about the occurrence of an event the better the hit rate and the lower the false alarm rate should be. The closer the curve is to the diagonal the less the ensemble system can discriminate between events and the less resolution it has. The ECMWF-EPS is displayed in grey and C-EN in black.

The summer experiments are illustrated as dashed lines with triangles and the winter experiments as continuous lines with points. Both systems are able to discriminate between different events. However, for 0.1 mm ensemble nudging is shifted to pairs of lower false alarm rates and lower hit rates whereas the ECMWF-EPS is shifted to pairs of higher hit rates coming along with higher false alarm rates. Reason for that is presumably the previously discussed problem of coarser resolved models that produce higher numbers of small-amount precipitation events yielding higher hit rates together with higher false alarm rates.

The area under the ROC curve is equivalent at 0.1 mm for both systems and both experiments. The resolution component of the Brier score confirms that the resolution of ECMWF-EPS and C-EN is comparable for all regarded thresholds.

Conclusion:

Here, the resolution of C-EN and ECMWF-EPS short-range forecasts is comparable.

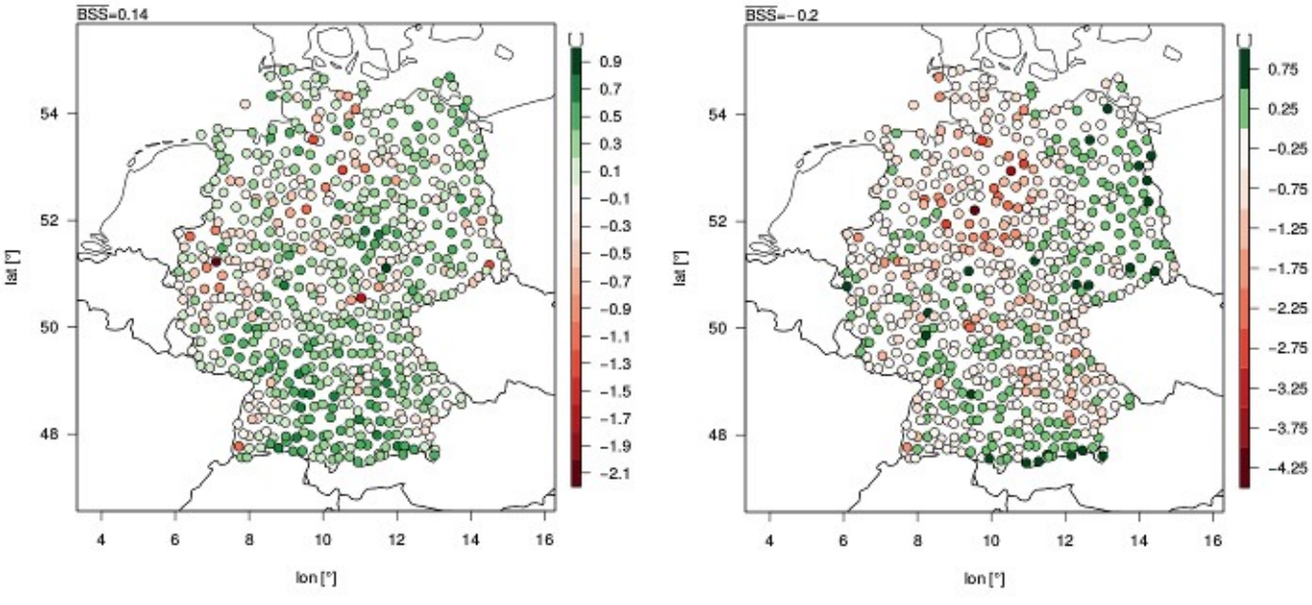


Figure 11: Spatial distribution of the brier skill score the summer (left) and the winter experiments (right).

3.6 Skill of C-EN using ECMWF-EPS as reference

Skill describes the quality of a NWP system compared to a benchmark. We assess the probabilistic skill of the ensemble nudging system using the ECMWF-EPS as a reference. As metrics we employ the Brier skill score (BSS) as well as the continuous ranked probability skill score (CRPSS). The CRPS is nothing else than the Brier score integrated over an infinite number of decision thresholds so that it considers the full range of precipitation amounts. Both skill scores are defined as

$$SS = 1 - \frac{S}{S_{ref}} \quad ,$$

where S can be replaced by any of the score metrics and S_{ref} is the reference system's score. The Brier score is given by

$$BS = \frac{1}{N}\sum_{t=1}^{N}(p_t - o_t)^{2.}$$

Therein, p_t is the ensemble probability while o_t is a binary observation for a decision threshold. N is the sample size that may include different locations and time steps. The CRPS is given by

$$CRPS = \int_{-\infty}^{\infty}(P(x) - P_a(x))^2 dx \quad \text{with}$$

$$P(x) = \int_{-\infty}^{x}\rho(y)\,dy \quad \text{and}$$

$$P_a(x) = H(x - x_a) \quad .$$

It is the difference between the predicted (in our case analysed) and occurred cumulative distribution functions P(x) and $P_a(x)$. H is the Heaviside function. Note that both Brier score and CRPS can be decomposed into a reliability, resolution and uncertainty part (Hersbach, 2000 for the CRPS and Murphy, 1973). The skill scores are computed for 6-hourly accumulated precipitation at 06 and 18 UTC. +6 ECMWF-EPS forecasts of precipitation are used as a benchmark. The Brier skill score for 1 mm threshold displayed in Figure 11 shows a local improvement over the ECMWF-EPS of up to 90% for the summer experiment and one of up to 75% for the winter experiment. For summer, the mean Brier skill score is 0.14 and thus shows an average superiority of ensemble nudging while in winter it is -0.2 proving superiority of the ECMWF-EPS. However, note that the sample size for each station is only 60 (data for 06 and 18 UTC, 30 days) so that these results have to be handled with care. The superiority of ensemble nudging in summer and of the ECMWF-EPS in winter is confirmed through Table 5 which summarizes the Brier score and Brier skill score for different decision thresholds. This basic conclusion is not affected by the uncertainties underlying the computation. Note that here, 1000 bootstrap samples comprising 90% of the data have been drawn to compute the Brier scores and their uncertainty (standard deviation) while for Figure 11 the sample has been divided according to locations and the scores have been computed for each location separately.´

The CRPSS shown in Figure 12 indicates a local improvement of up to 75%. On average, it is 0.14 which indicates a superiority of ensemble nudging for the experimental month even if all possible thresholds are

considered as done by the CRPS. However, in winter the C-EN is outperformed by the ECMWF-EPS as the mean CRPSS is -0.19. Locally, C-EN is up to 175% worse than the ECMWF-EPS. However, just as outlined for the Brier skill score, these results are only preliminary due to a limited sample size and will revisited as soon as a longer time span will be available for the experiments.

Table 5: Brier score for a range of decision thresholds for ECMWF-EPS and C-EN 6-hourly precipitation sums for June and December 2011. Estimated from 1000 bootstrap samples drawing 90% of the data, the uncertainty is about 10^{-3},, rounded to the third decimal place.

Threshold [mm]	C-EN June	ECMWF-EPS June	BSS June	C-EN December	ECMWF-EPS December	BSS December
0.1	0.12	0.19	0.58 ∓ 0.01	0.165	0.191	0.13∓0.01
0.5	0.01	0.12	0.44 ∓ 0.01	0.137	0.130	-0.05∓0.01
1	0.08	0.1	0.37 ∓ 0.01	0.118	0.105	-0.12∓0.01
2.5	0.05	0.06	0.33 ∓ 0.01	0.0742	0.062	-0.18∓0.01
5	0.34	0.04	0.34 ∓ 0.01	0.039	0.032	-0.24∓0.02
10	0.02	0.03	0.35 ∓ 0.02	0.025	0.019	-0.3∓0.03
15	0.02	0.03	0.29 ∓ 0.02	0.014	0.01	-0.41∓0.05
20	0.01	0.01	0.1 ∓ 0.02	0.007	0.004	-0.53∓0.07

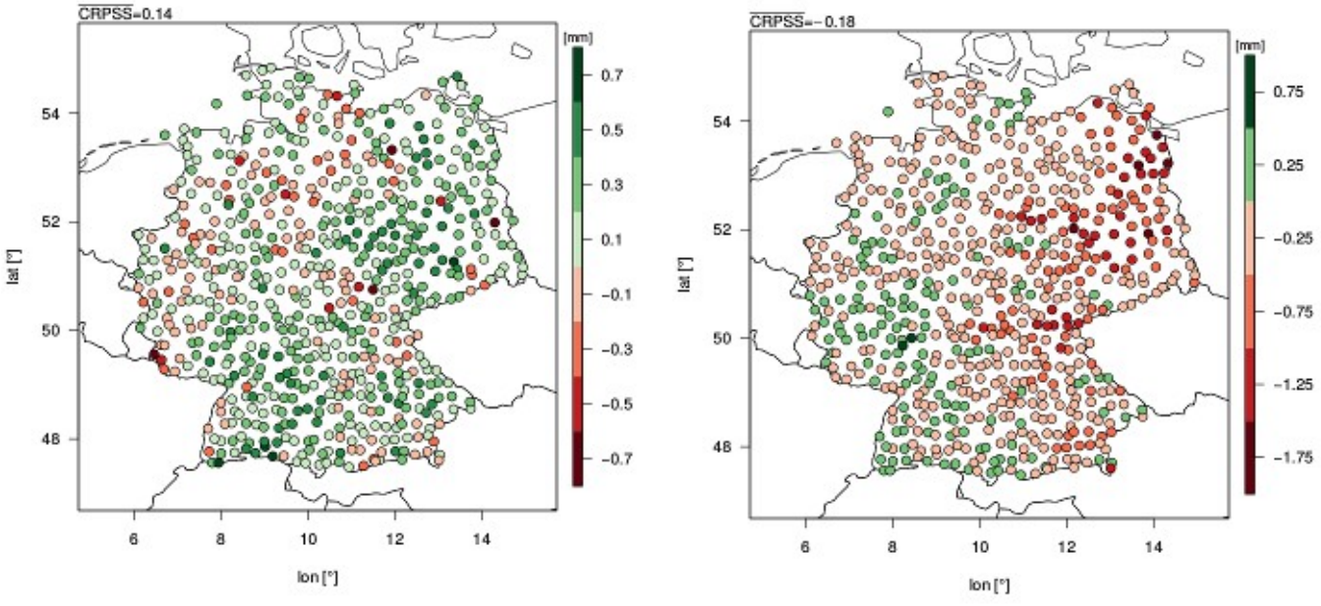


Figure 12: Spatial distribution of the CRPS skill score for the summer (left) and winter (right) experiments.

Conclusion:

- Employing the Brier skill score and the CRPS skill score can be shown for the test period of June 2011 for ensemble nudging compared to the ECMWF-EPS. Findings from the foregoing sections suggest that this is due to a better reliability of ensemble nudging.
- The winter experiment for C-EN shows slightly inferior average skill scores. This is probably due to a resolution worse than the one of the ECMWF-EPS. The physical processes that the observation perturbations project on have to be subjected to further investigation.

4. Summary and conclusion

The objective of the UB work leading into the deliverable on hand was to appraise the feasibility of a new ensemble reanalysis system. Originally, the idea was to develop a combination of a new ensemble nudging technique and a LETKF newly implemented at DWD. This EN-LETKF system seems to be very promising for the application to regional reanalysis as it allows to produce temporally smooth and physically balanced time series (close to the trajectory of the true atmosphere) and a high analysis quality incorporating modern observations. Of course, this would also be feasible using for example 4D-Var, however, such a data assimilation system is not available for the regional scale in Germany. To date, the conditions that would allow for a successful application of the EN-LETKF are not fulfilled. This is a matter of the necessary ensemble size of at least 40 members, a matter of tuning to the system from 2.8 km to 12 km, a matter of a lack of perturbed boundary conditions and above all a matter of a too small observation stream with a lack of modern observations that forces a division of the conventional observations between ensemble nudging and LETKF leading to an underexposed observation density. So far, it is not clear which of these points leads to the observed problems with the system. Presumably, it is a combination of all unfulfilled conditions that is possibly aggravated by technical problems or errors.

However, at this point of time, we can show that the ensemble nudging data assimilation system is beneficial for the purpose of computing a regional ensemble reanalysis. Certainly, nudging is about to be disestablished as obsolete method for operational NWP, however, for regional reanalysis it proves to be very useful. This is not only due to low computational costs (e.g. compared to 4D-Var) which allows for a considerable number of ensemble members, but also due to a good accuracy of the analysed fields and probabilistic and uncertainty estimation capabilities that we have shown in the course of this deliverable.

We have hypothesized that high-resolution regional reanalyses have an added value regarding precipitation compared to global and coarser resolved reanalyses as well as dynamical downscalings. Indeed, we could show that our sample of data based on an experiment with ensemble nudging for June 2011 agrees with theses hypotheses. Ensemble nudging outperforms the Hirlam and ERA-INTERIM reanalyses both regarding frequency bias for a number of thresholds (making the data set meaningful for climatological investigations based on frequency) as well as regarding the proportion correct score which shows that it has higher accuracy both at the lower and the extreme decision thresholds. Just as would be expected, the dynamical downscaling, which has twice the resolution of the ensemble nudging runs is competitive concerning the frequency bias, but is poor in terms of accuracy.

The probabilistic capabilities of the system are promising. For the chosen time period ensemble nudging

yields a good reliability and resolution, whereby the latter is a bit worse than the one of the ECMWF-EPS in winter. This is reflected by inferior Brier and CRPS scores in winter. The average skill scores are slightly positive in summer and slightly negative in winter. This shows that the probabilistic capabilities of ensemble nudging are on average comparable to the ones of the ECMWF-EPS which is currently the gold-standard ensemble. Anyway, it should be kept in mind that ensemble nudging is a reanalysis ensemble whilst the ECMWF-EPS is used for NWP. The uncertainty estimation capabilities measured by means of the spread-skill ratio appear to be promising. They are expected to improve further if more uncertain ingredients of the system like the lateral boundary conditions or model physics are perturbed.

References

Ballabrera-Poy, J., Kalnay, E., and Yang, S. (2009). Data assimilation in a system with two scales – combining two initialization techniques. *Tellus*, 61A: 539-549.

Bojinski, S., Verstraete, M., Peterson, T.C., Richter, C., Simmons, A. and Zemp, M. (2014). The concept of essential climate variables in support of climate research, applications and policy. BAMS. 1431-1443.

Bollmeyer, C., Keller, J., Ohlwein, C., Wahl, S., Crewell, S., Friederichs, P., Hense, A., Keune, J., Kneifel, S., Pscheidt, I., Redl, S., and Steinke, S. (2015). Towards a high-resolution reanalysis for the European CORDEX domain. *Quaterly Journal of the Royal Meteorological Society*, 141:1-15.

Bröcker, J. and Smith, L.A. (2007). Increasing the reliability of reliability diagrams. Weather and forecasting. 22:651—661.

Dee, D.P., Uppala, S.M., Simmons, Berrisford, P., Poli, P. Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechthold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, Bormann, N., Delsol, C., Dragani, R., Fuentes, M. Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Holm, E.V., Isaksen, L., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.-J., Park, B.K., Peubey, C., de Rosnay, P., Tavolato, C., Thepaut, J.-N. And Vitart, F. (2011). The ERA-INTERIM reanalysis: configuration and performance of the data assimilation system. Part A. *Quaterly Journal of the Royal Meteorological Society.* 137:553-597.

Deng, A., Seaman, N., Hunter, G., and Stauffer, D. (2004). Evaluation of interregional transport using the MM5-SCIPUFF system. *Journal of Applied Meteorology*. 43:1864-1886.

Deng, A. And Stauffer, D. (2006). On improving 4-km mesoscale model simulations. *Journal of Applied Meteorology*, 45:361-381.

Desroziers, G., Berre, L., Chapnik, B., and Poli, P. (2005). Diagnostics of observation, background and analysis-error statistics in observation space. *Quaterly Journal of the Royal Meteorological Society*, 131:3385-3396.

Dixon, M., Li, Z., Lean, H., Roberts, N., and Ballard, S. (2009). Impact of data assimilation on forecasting convection over the Unitied Kingdom using a high-resolution version of the Met Office unified model. *Monthly Weather Review.* 137:1562-1585.

Flowerdew, J. (2015). Towards a theory of optimal localization. *Tellus A*, 67: 1-18.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems). *Weather and forecasting*. 15:559-570.

Hollingsworth, A. An Lönnberg, P. (1986). The statistical structure of short-range forecast errors as determined from radiosonde data. Part 1: The wind field. *Tellus*, 38A:111-136.

Houtekamer, P., Lefaivre, L., Derome, J., Ritchie, H., and Mitchell, H. (1996). A system simulation approach to ensemble prediction. *Monthly Weather Review*, 124:1225-1242.

Hunt, B.R., Kalnay, E., Kostelich, E., Ott, E., Patil, D., and Co-Authors. (2004). Four-dimensional ensemble Kalman filtering. *Tellus*, 56A:273-277.

Hunt, B.R., Kostelich, E.J. and Szunyogh, I. (2007). Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*. 230:112-126.

Jolliffe, I.T. and Stephenson, D.B. (2012). Forecast verification – a practioner's guide in atmospheric science. Wiley-Blackwell. 2Nd Edition.

Lahoz, W., Khattatov, B., and Ménard, R., editors. (2010). *Data Assimilation. Making Sense of Observations.* Springer.

Lei, L., Stauffer, D., and Deng, A. (2012b). A hybrid nudging-ensemble Kalman filter approach to data assimilation. Part II: Application in a shallow-water model. *Tellus*, 64.

Lei, L., Stauffer, D., Haupt, S.E., and Young, G. (2012a). A hybrid nugding-ensemble Kalman filter approach to data assimilation. Part I: Application in the Lorenz system. *Tellus A*, 64.

Leidner, S., Stauffer, D., and Seaman, N. (2001). Improving short-term numerical weather prediction in the California coastal zone by dynamic initialization in the marine boundary layer. *Monthly Weather Review*, 129: 275-294.

Otte, T., Seaman, N., and Stauffer, D. (2001). A heuristic study on the importance of anisotropic error distribution in data assimilation. *Journal of Applied Meteorology*. 47:1853-1867.

Perianez, A., Reich, H. and Potthast, R. (2014). Optimal localization for Ensemble Kalman Filter systems. Journal of the Meteorological Society of Japan. 92:585-597.

Schraff, C. (1997). Mesoscale data assimilation and prediction of low stratus in the Alpine region. *Meteorology and Atmospheric Physics,* 64:21-50.

Schraff, C. and Hess, R. (2012). A description of the nonhydrostatic regional COSMO-model. Part III: Data assimilation. Technical report, Deutscher Wetterdienst, Offenbach, Germany.

Schraff, C., Reich, H., Rhodin, A., Schomburg, A., Stephan, K., Perianez, A. and Potthast, R.(submitted 2015) Kilometer-scale ensemble data assimilation for the COSMO model (KENDA). *Quaterly Journal of the Royal Meteorological Society.* 00:1-20.

Schroeder, A., Stauffer, N., Seaman, N., Deng, N., Gibbs, A., and Co-Authors (2006). An automated high-resolution, rapidly relocatable meteorological nowcasting and prediction system. *Monthly Weather Review*, 134:1237-1265.

Seaman, N., Stauffer, D., and Lario-Gribbs, A. (1995). A multi-scale four-dimensional data assimilation

system aoolied in the San Joaquin Valley during SARMAP. Part I: Model design and basic performance characteristics. *Journal of Applied Meteorology*. 34:1739-1761.

Stauffer, D. and Seaman, N. (1990). Use of four-dimensional data assimilation in a limited-area mesoscale model. Part I: Experiments with synoptic data. *Monthly Weather Review*, 118:1250-1277.

Stauffer, D., Seaman, N., and Binkowski, F. (1991). Use of four-dimensional data assimilation in a limited-area mesoscale model. Part II: Effects of data assimilation within the planetary boundary layer. *Monthly Weather Review.* 119:734-754.

Undén, Per, L. Rontu, H. Järvinen, P. Lynch, J. Calvo, G. Cats, J. Cuxart, K. Eerola, C. Fortelius, J. Antonio Garcia-Moya, C. Jones, G. Lenderlink, A. McDonald, R. McGrath, B. Navascues, N. Woetman Nielsen, V. Ødegaard, E. Rodriguez, M. Rummukainen, R. Rõõm, K. Sattler, B. Hansen Sass, H. Savijäarvi, B. Wichers Schreur, R. Sigg, H. The, A. Tijm, 2002: HIRLAM-5 Scientific Documentation, HIRLAM-5 Project, c/o Per Undén SMHI, S-601 76 Norrköping, SWEDEN