Adv. Sci. Res., 12, 207–218, 2015
www.adv-sci-res.net/12/207/2015/
doi:10.5194/asr-12-207-2015

Advances in Science & Research
Open Access Proceedings

Methodologies to characterize uncertainties in regional reanalyses

M. Borsche[1], A. K. Kaiser-Weiss[1], P. Undén[2], and F. Kaspar[1]

[1]Deutscher Wetterdienst, National Climate Monitoring, Frankfurter Str. 135, 63067 Offenbach, Germany
[2]Swedish Meteorological and Hydrological Institute, Folkborgsvägen 17, 601 76 Norrköping, Sweden

Correspondence to: M. Borsche (michael.borsche@dwd.de)

Received: 25 January 2015 – Revised: 2 October 2015 – Accepted: 16 October 2015 – Published: 27 October 2015

Abstract. When using climate data for various applications, users are confronted with the difficulty to assess the uncertainties of the data. For both in-situ and remote sensing data the issues of representativeness, homogeneity, and coverage have to be considered for the past, and their respective change over time has to be considered for any interpretation of trends. A synthesis of observations can be obtained by employing data assimilation with numerical weather prediction (NWP) models resulting in a meteorological reanalysis. Global reanalyses can be used as boundary conditions for regional reanalyses (RRAs), which run in a limited area (Europe in our case) with higher spatial and temporal resolution, and allow for assimilation of more regionally representative observations. With the spatially highly resolved RRAs, which exhibit smaller scale information, a more realistic representation of extreme events (e.g. of precipitation) compared to global reanalyses is aimed for. In this study, we discuss different methods for quantifying the uncertainty of the RRAs to answer the question to which extent the smaller scale information (or resulting statistics) provided by the RRAs can be relied on. Within the European Union's seventh Framework Programme (EU FP7) project Uncertainties in Ensembles of Regional Re-Analyses (UERRA) ensembles of RRAs (both multi-model and single model ensembles) are produced and their uncertainties are quantified. Here we explore the following methods for characterizing the uncertainties of the RRAs: (A) analyzing the feedback statistics of the assimilation systems, (B) validation against station measurements and (C) grids derived thereof, and (D) against gridded satellite data products. The RRA ensembles (E) provide the opportunity to derive ensemble scores like ensemble spread and other special probabilistic skill scores. Finally, user applications (F) are considered. The various methods are related to user questions they can help to answer.

1 Introduction

Atmospheric reanalyses produce complete and physically consistent data products aiming for a best estimate of the state of the Earth's atmosphere (Dee et al., 2014). In the European Union's seventh Framework Programme (EU FP7) project Uncertainties in Ensembles of Regional Re-Analyses (UERRA), various European meteorological regional reanalyses (RRAs) are developed. Users turn to the spatially highly resolved RRAs for smaller scale information and a more realistic representation of extreme events (e.g. of precipitation) compared to global reanalyses, and would like to derive trends.

The main objectives of the UERRA project are to produce a long-term (several decades) high-resolution climate quality ensemble of European RRAs of Essential Climate Variables (ECVs) and to estimate the associated uncertainties in these RRAs. The RRAs and the uncertainty estimates are to be made publicly available so a large community can benefit from the research. Within the UERRA project and its precursor project European Reanalysis and Observations for Monitoring (EURO4M), gridded data are produced, and data rescue and digitization efforts are undertaken (Brunet et al., 2013). Data rescue efforts are concentrated on filling gaps in data from 1950 onwards. More on the structure, participants, and status of the UERRA project can be found on its website http://www.uerra.eu.

Table 1. Description of the regional reanalysis planned in the UERRA project.

Feature	Met Office	SMHI	HErZ	Météo France
Boundary conditions (forcings)	6 hourly ERA-Interim fields	6 hourly ERA-Interim and/or ERA-40 fields	3 hourly ERA-Interim and/or ERA-20C fields	HARMONIE @ 11 km to > 5.5 km; ALADIN (Horányi et al., 1996) @ 5.5 km
Model and domain	Unified Model, CORDEX EU-11	HARMONIE, Lambert projection, CORDEX EU-11	COSMO, CORDEX EU-11	MESCAN (Soci et al., 2013), Lambert projection
Ensemble members	20	1 (2 for the period 2006 to 2010)	10 to 20	1 (4 to 6 for the period 2006 to 2010)
Deterministic DA method	Hybrid Ensemble-4D-Var	3D-Variational upper-air/OI (optimal interpolation) surface analysis	Nudging	OI surface re-analysis after a static or dynamical downscaling
Ensemble DA method	Ensemble of 4D-Vars	Ensemble of 3D-Vars	LETKF with ensemble nudging	
Time range	1978 to 2013	1961 to 2013	5 years	1961 to 2011
Observation input	various sources of surface, aircraft, upper air, satellite observations, and precipitation	Surface (pressure), SHIP, (SYNOP BUOY, DRIBU); aircraft (AIREP, AMDAR); Upper air (TEMP, PILOT)	Surface (SYNOP (pressure), SHIP, BUOY, DRIBU); aircraft (AIREP, AMDAR); upper air (TEMP, PILOT)	SYNOP (pressure), SHIP, BUOY, 24 h precipitation from rain gauge and T_{min}/T_{max} after pre-processing
Temporal resolution	6 h (analysis), hourly (forecast)	6 h	1 h	6 and 24 h for precipitation
Horizontal resolution	12 km control grid; analysis increments on 24 km; 24 km ensemble	11 km	12 km	5.5 km
Vertical resolution	70 levels from near surface to 80 km	65 levels	40 levels (20 m to 22 km)	only surface

Table 1 summarizes the RRAs which are planned to be produced within the UERRA project at the Met Office (MO), Exeter, UK (developed upon Renshaw, 2013); the Swedish Meteorological and Hydrological Institute (SMHI), Norrköping, Sweden (developed upon Dahlgren et al., 2014 and now Bubnova et al., 1995); DWD's (Deutscher Wetterdienst) Hans-Ertel Centre for Weather Research (HErZ; Simmer et al., 2015) University of Bonn, Germany (Bollmeyer et al., 2015); and Météo France, Toulouse, France (Häggmark et al., 2000; Jansson et al., 2007; Soci et al., 2011, 2013). The RRAs differ in the numerical weather prediction (NWP) model used, the data assimilation system applied, in the boundary conditions applied, and in the observations used for the assimilation. All deterministic RRA lateral boundary conditions are derived from the global ERA-Interim (Dee et al., 2011) and ERA-40 (Uppala et al., 2005)

reanalysis whereas for the ensemble realizations it is planned to take lateral boundary conditions from ERA-20C (Poli et al., 2013) or ERA5, the successor of ERA-Interim, into account.

These three groups (MO, SMHI, HErZ) develop RRAs based on their operational NWP models, which are the Unified Model (Davies et al., 2006), the HARMONIE model (Bubnova et al., 1995; De Troch et al., 2013; Gerard et al., 2009), and the COnsortium for Small-scale MOdelling (COSMO) model (Schättler et al., 2014), respectively. They plan to produce a deterministic run and a set of ensembles of reanalyses with up to 20 members each. The data assimilation methods employed by MO are a hybrid 4DVar (Clayton et al., 2013) for the deterministic run and an ensemble of 4D-Vars (Rawlins et al., 2007) for the ensemble runs. SMHI has implemented a 3D-Var (Berre, 2000; Fischer et al., 2005) for the deterministic run and an ensemble of (2) 3D-Vars (using the ALADIN and ALARO-0 physics versions as described in De Troch et al., 2013) for the ensemble runs. HErZ uses nudging and is developing a Local Ensemble Transform Kalman Filter (LETKF) (Hunt et al., 2007; Harnisch and Keil, 2015) with Ensemble Nudging for the deterministic and ensemble runs, respectively. Ensembles are created either by perturbed initial conditions (MO), disturbed model physics (SMHI), disturbed observations (HErZ), or a combination of these. Météo-France performs an additional statistical and dynamical downscaling of the RRAs produced at SMHI followed by high resolution reanalysis for the surface only in order to drive inter alias hydrological physical models. The data sets are planned to span several decades, and a few years for the more experimental set-ups. SMHI produces in addition a cloudiness reanalysis (based on Häggmark et al., 2000) using a consistent satellite data set.

An estimation of uncertainty is required by users on different temporal and spatial scales for enabling and enhancing applications of RRA products. Section 2 outlines the methods, relying on assimilation feedback, comparison against station observations, gridded station observations, gridded satellite remote sensing based data products, and ensembles. We also discuss deriving uncertainty estimation from user applications. The advantages and disadvantages of the various methods are discussed and they are related to various scientific and user questions. The meteorological parameters for which each method is feasible depend mainly on practical considerations. In Sect. 3, the various skill scores are summarized, the suitability of which depends on the meteorological parameters and the question sought to be answered. In Sect. 4, a summary and recommendations for best practises are given.

2 Methods to estimate uncertainties in RRAs

Table 2 summarizes six methods which have been discussed within the UERRA project and correspond to those discussed in the EU FP7 project Coordinating Earth Observation Data Validation for Re-analysis for Climate Services (CORE-CLIMAX), see Kaiser-Weiss et al. (2015). Each of the methods is discussed in this section, benefits and drawbacks pointed out, and the principle aim pursued with each method condensed into a scientific and a user question. Furthermore, the discussion below addresses both random and systematic uncertainties. A systematic uncertainty (which can depend on time as well as location or meteorological situation) is referred to as bias. Random uncertainties are characterized with the root mean square (RMS) against some data set considered as truth, in case of absence of error estimates of the latter.

2.1 Method A: feedback statistics

Here, we distinguish between virtually independent observations (not assimilated yet) and dependent observations (observations passed through the assimilation system prior to the production of the reanalysis field). Strictly speaking, any observation system biased relative to the model will yield the analysis biased, and in this sense, measurements at later points in time are not strictly independent (as they are subject to this bias, too). Strictly independent data are hard to come by, especially for longer time periods, because they would naturally be selected for assimilation. Integrated parameters, such as precipitation, are strictly independent for most reanalysis systems. Further, minimum and maximum temperatures are not assimilated (though would not be strictly independent in case of biases). In order to maintain a reasonable large data base, we have to loosen our requirements on independence such that we use virtually independent data as if they were strictly independent.

We regard the feedback as the output of the assimilation system in observation space. The feedback comprises the assimilated observations (bias corrected where applicable) (o), the background or "free forecast" (the short range forecast used in the data assimilation) ($\mathbf{H}_b$), and the analysis ($\mathbf{H}_a$). The background and the analysis are brought into observation space with the matrix of the linearized observation operator $\mathbf{H}$, so a direct comparison between observations and model parameters can be performed in an optimal way. We check the feedback for trends and seasonal dependency (which are not desirable).

Usually, feedback statistics are a standard output of the data assimilation system and are frequently used by the producers for quality control. Note that, when comparing different systems, the bias correction might differ. For one system, an approximately Gaussian distribution is expected for both $o - \mathbf{H}_b$ and $o - \mathbf{H}_a$, where the absolute value of $o - \mathbf{H}_b$ is the bias between observations and modelled observations, i.e. even if the difference is zero, both terms could be biased versus the (unknown) truth.

Comparison between the different RRAs has to be done with observations assimilated by all of the UERRA reanaly-

Table 2. Methods and data sources suitable to derive uncertainty estimates for regional reanalyses.

Method	Data source	Parameter	Details	Scientific questions	User questions
A: feedback statistics	Radiosonde soundings	Temperature, wind speed, and relative humidity	Focus on lower troposphere; bias and RMSE of time series; store in ODB format	How stable are the regional reanalyses (RRAs) with respect to multi-annual trends on a spatial scale of roughly 100 km?	How well represented are trends and climatologies of wind speed relevant for wind energy?
B: station observations	B1: (independent) mast station data; B2: (dependent, i.e. assimilated) station data	B1: wind speed B2: T_{min}, T_{max}, and number of days of threshold exceedance of temperature and precipitation	There are many more suitable observations available for B2 than for B1.	At which time scales can we find which correlations between reanalysis fields and station observations?	On which time scales of variability and for which parameters can we use the RRAs similar to the use of station measurements?
C: gridded station observations	Gridded data products for the Nordic region and the UK; E-OBS, APGD	Precipitation; T_{min} and T_{max}	To consider whether a part of underlying station observations was assimilated into the reanalysis.	What differences do we get with different products when determining the effective spatial and temporal scales of the RRAs?	Which scales of the RRAs (temporal, spatial) can be interpreted?
D: gridded satellite data products	Satellite data products of CM-SAF and CCI	Global radiation; cloud liquid water path; total cloud cover; precipitation; snow water equivalent		How well do the RRAs compare to the satellite observations – or exceed their quality?	Does the RRA or the satellite provide the better data product for the user applications?
E: Ensemble based comparison	Data with uncertainty estimates;	Precipitation; T_{min}, T_{max}, T_{mean};	Ensemble based uncertainty estimates will be performed on (1) data with uncertainty estimates.	Does the ensemble provide a useful spatially and temporally resolved uncertainty estimate?	Which uncertainty characteristics can be interpreted from the ensembles, for which user relevant parameters?
	Products as in methods A through D	Parameters as in A through D	(2) the basis of methods A through D		
F: User related models		T_{mean}; T_{max} and T_{min} pseudo analysis; wind speed; precipitation;	SURFEX by Météo France and HYPE by SMHI		Is the result of a user model forced by RRAs significantly better than with the original forcing?

sis systems, which do not differ much in their bias correction. Especially suitable for this purpose are radiosondes (temperature, wind speed and direction, and relative humidity). The core scientific interest in applying feedback statistics for validating the RRAs is how stable the RRAs are with respect to multi-annual trends on different spatial scales for these parameters. For instance, users of wind energy applications want to know how well represented the wind speed at heights relevant to wind energy is, especially with respect to trends and frequency distributions. Feedback statistics can of course be applied to all sorts of assimilated observations in order to perform the afore mentioned task. However, assimilated radiosonde data best fit the purpose of this method due to the fact that they are used throughout the different RRA systems as anchor-data (i.e. no bias corrections applied). In contrast, aircraft data are handled specifically for each assimilation system (due to data thinning, data selection, and individual pre-processing) which renders inter-comparison of that data source much more difficult.

When, firstly, comparing the mean or RMS of $o - \mathbf{H}_b$ between different reanalysis systems, it is crucial that comparable forecast lengths are selected. If this is valid $o - \mathbf{H}_b$ is a good measure to start with as a means of comparing against virtually independent observations. A smaller $o - \mathbf{H}_b$, i.e. a closer match, can be caused by a number of reasons: (1) the background is closer to the truth – desired from the user per-

spective, (2) the bias removal prior to the assimilation was successful – desired for the assimilation, (3) the background and the observations have the same bias against the truth – unlikely when a number of observing systems yield similar results, (4) the model is biased due to the assimilation of biased observations in the previous step – also unlikely when a number of observing systems yield similar results. Secondly, in a successful assimilation, the average RMS of $o - \mathbf{H}_a$ should be smaller than the average RMS $o - H_\mathrm{b}$ for each observing system. Note that the mean or RMS of $o - \mathbf{H}_a$ is harder to compare between different reanalysis systems.

Desroziers et al. (2005) provide a methodology of the diagnosis in observation space and show that the variance of $o - \mathbf{H}_b$ is the sum of the variance of background error in observation space $(\widetilde{\sigma b})^2$ and the variance of observation error $(\widetilde{\sigma o})^2$, compare their Eq. (1):

$$\frac{1}{N}\left(\sum_{i=1}^{N}\left(o_i - \mathbf{H}_{b_i}\right)\left(o_i - \mathbf{H}_{b_i}\right)\right) = \left(\widetilde{\sigma b}\right)^2 + (\widetilde{\sigma o})^2.$$

Further, they show that the background error covariance in observation space can be related to the product of $\mathbf{H}_a - \mathbf{H}_b$ and $o - \mathbf{H}_b$, see their Eq. (2), and the observation error covariance can be related to the product of $o - \mathbf{H}_a$ and $o - \mathbf{H}_b$, see their Eq. (3). Finally, if the error covariances are correctly specified in the analyses, the product of $\mathbf{H}_a - \mathbf{H}_b$ and $o - \mathbf{H}_a$ should relate to the analysis-error covariance (see their Eq. 4).

Before Desroziers et al. (2005), Hollingsworth and Lönnberg (1986) and Lönnberg and Hollingsworth (1986) applied a curve fitting of the covariances as a function of distance to separated background errors from observation error based on the assumption that forecasts errors are horizontally homogeneous and observation errors are horizontally uncorrelated, where it is understood that the observation error includes the representativeness error (also known as sampling error) as well as instrumental errors. This means a separation of background bias and observation bias by inspection of the spatial structure of mean background departures mean$(o - \mathbf{H}_b)$ for, e.g. each season or time of the day. It is well known that the NWP models used in the data assimilation exhibit biases that have both diurnal and seasonal variations. Thus, inspecting horizontal plots of biases at station locations (or possibly from a gridding interpolation procedure) will show if there is a significant background model bias. This is the case if there are spatially consistent mean departures that are sizeable compared with the RMS of the departures. In addition, this can be found by looking at the mean of the analysis increments mean$(a - b)$ in grid point space.

On the other hand, if there are large variations of the mean background departures from station to station without horizontal consistency one may conclude that the bias is in the observation itself, or in some cases, due to poor representativeness of the NPW model background at the specific location. It may be due to occasional large departures between model orography and the one of the station or very different surface properties. The latter is expected to be the case only for a small portion of the stations and manual inspections of the data can reveal if it is the case. Either way, stations with large biases should be excluded for evaluations of random errors and used with special care when studying trends.

Multi-annual trends in the RRAs are determined by the boundary conditions (lateral boundary from the global reanalysis as well as the lower boundary, especially by the soil moisture and sea surface temperature), and the assimilated observations, where any trend in the observation system bias (relative to the model) would influence trends in the reanalysis. The latter might be expected to happen in regions where the observation coverage changed significantly over time.

The advantages of applying method (A) is that it provides a relative measure and can detect discontinuities and breaks as well as slowly increasing or decreasing systematic errors of the analyses (which all would influence trends). These results are dependent on the weight the observations are given in the different data assimilation systems, thus they need to be inter-compared between the different RRAs with care. A drawback of this method is that inter-comparison between the feedback statistics of different reanalysis systems is in principle difficult mainly because the handling of the observations may differ from one centre to another. For some of the parameters of interest (like, e.g. precipitation), the bias and RMSE are not sufficient scores to perform a thorough statistical analysis, thus more than these standard parameters should be evaluated based on the feedback. Note this method is limited in application to parameters which are assimilated.

2.2 Method B: station observations

Method (B) describes the comparison against point measurements from station observations. Interpreting single grid cells from the regional reanalysis and taking it as a proxy for a single point poses several questions from a theoretical point of view, such as that of the representativeness of the NWP model for a given observation location and observing method. However, the benefit of the method is that from a practical point of view this is often the easiest approach. Users of the RRAs are very much interested to answer the question how well the reanalyses compare against (their own) independent observation time series. This boils down to the question at what time scales correlations can be found between RRA fields and station observations.

In contrast to the previously described feedback methodology, this method allows to compare parameters which are not assimilated. For an independent validation and uncertainty estimate the reference data are required not to be assimilated into the RRAs. This leads to the scientific question on which time scales of variability and with which parameters the RRAs can be used to compare with station measurements.

The major drawback of the method is that a point measurement cannot be expected to closely match the nearest reanalysis grid point even if the measurement happens to be located exactly at the centre of the model grid point. This is due to the fact that the point measurement is representative for a limited area around the measurement, whereas the model grid point represents a much larger area, corresponding to the inherent spatial resolution, which can be expected to be larger than the nominal resolution, i.e. span several model grid cells. Not only the difference in the spatial but also the temporal representativeness needs to be considered, because the point measurement takes place more or less instantaneous, whereas the model value represents an average over a longer time period. Hakuba et al. (2014) (and references therein) provide a study on the representativeness of ground-based point measurements of surface solar radiation compared to gridded satellite data products detailing the points mentioned above.

And most importantly, the vertical representation usually is different between the model and the point measurement because the model levels are seldom exactly at the height of the measurement as, e.g. 2 m temperature or 10 m wind speed. This is complicated by the fact that the model topography is smoothed and thus different to the real one and there are limits to which extent small scale variability can be modelled or parameterized.

It is possible to correct for some of the above mentioned deficiencies. For interpolating the model value to the observation height, either the vertical observation operator can be used – as done in the intrinsic handling in method (A) – or some simplified interpolation of the model levels to the height above ground where the observation took place can be applied. Still, the problem remains that the model topography and the real one differ and some local topographic effects are not modelled.

If the assimilated, dependent observations together with the observation operators are used, this method (B) is identical to method (A). Then the difference is more a technical one, in terms of separate data handling of the observations and RRA fields. However, in practice, the full observation operators of the RRA cannot readily be applied off-line. Hence, the advantage of method (B) is that the same observations can be used for all RRA systems, irrespectively how they were used (or not used) in each of the systems.

Most of the representative station observations are assimilated into the RRAs, leaving only few high quality independent measurements for the uncertainty estimation. Therefore, the reference data are divided into two groups, namely (B1) independent observations, mainly wind speed from tall mast stations, and (B2) dependent observations, which were chosen to include T_{min}, T_{max}, and threshold values of temperature as well as precipitation. Refer to Sect. 3 for a discussion on scores and skill scores to use for this method and parameters.

2.3 Method C: gridded station observations

Method (C) describes validation against gridded data fields which are spatially interpolated station observations. This is, similar to method (B), a handy comparison for users of traditional data sources, who consider switching to or also including reanalysis data for their specific applications. Several data products exist which cover the European continent or a sub-region thereof. Mainly two data products, which will be extended and improved within the UERRA project, will be used for the validation performed within this project. The E-OBS data set (Haylock et al., 2008; van den Besselaar et al., 2011) is created by statistical interpolation of European land station observations and consists of daily values of temperature (minimum, mean, and maximum), precipitation, and sea level pressure. The projection of the data is provided either on a regular grid or a rotated pole grid in 0.25° or 0.5° and 0.22° or 0.44° horizontal resolution, respectively. The E-OBS product is continuously updated and new versions of the product are released frequently. The second gridded dataset used for the validation is the Alpine Precipitation Grid Dataset (APGD), see Isotta et al. (2013) for details. The APGD is a high-resolution statistically interpolated precipitation data set, based on roughly 5500 rain gauge measurements, and covers the Alpine region, a sub-European land area.

Advantages with this method of validation are that the gridded data products have already been very carefully prepared, covering desirable parameters such as temperature and precipitation, and are ready to use. In combination with skill scores (see Sect. 3 for details) such as the equitable threat score (ETS) (Gandin and Murphy, 1992) or fractional skill score (FSS) (Roberts and Lean, 2008), uncertainty estimates concerning spatial and temporal scales as well as trend analyses can be performed. The main scientific interest to pursue with this method is to determine the effective spatial and temporal resolution of the RRAs and answer the user question which scales of the RRAs can be interpreted.

Caution needs to be exercised, as with method B, that the underlying station observations are independent of the RRAs. Furthermore, it is hard to answer how representative the gridded data products are in comparison with the reanalyses at the interpolated points. The smoothed topography plays a significant role because the parameters of interest (i.e. precipitation) are temporally and spatially highly variable surface parameters. We have to make sure that we do not automatically arrive at the conclusion that the best reanalysis is the one with the most similar topography, or the most similar correlation length scales, compared to the topography of the gridded data product.

Another principle concern with gridded data products is that the number of station data contributing to each grid cell varies so that the grid cells come with a regional varying quality. One way to reduce this effect is to demand a thresh-

old value of a minimum number of stations contributing to a single grid cell.

The analysis could comprise a simple comparison (bias, RMSE, and frequency distribution) as well as more sophisticated skill scores as the ETS, FSS, and thresholds or the likes as applicable in order to estimate the effective spatial and temporal scale of the product. In addition, the temporal change (trends) of the above mentioned statistical properties needs to be analysed.

Suitable for this method of comparison are fields of precipitation, T_{min}, and T_{max}, because there is large user interest in these variables and products covering whole of Europe exist. Furthermore, the spatial aggregation remedies the high local differences which make the point comparisons in (B) so difficult. Uncertainty information of the gridded data product would be required, to be provided either as a range or an ensemble of grids, to use for evaluation purposes. Refer to Isotta et al. (2015) for a successful example application of this method for precipitation in the Alpine area.

2.4 Method D: gridded satellite data products

Method (D) concerns the validation against satellite based observations. Long-term climate quality satellite data are produced by the Satellite Application Facility on Climate Monitoring (CM SAF) and the European Space Agency's (ESA) climate change initiative (CCI) (Hollmann et al., 2013). The characteristics of satellite data products depend on the observing system, i.e. whether it was produced by geostationary or low Earth orbiting satellites (GEO, LEO) and on the part of the spectrum the observation was taken, i.e. in the optical or microwave. Data products of GEO satellites exhibit a high temporal (up to 15 min) and spatial (about 5 km) resolution but provide only a limited area (non-global) coverage. LEO satellites have the capacity to provide global products but provide data only in swaths and can only provide few samples of a specific location per day (depending on the latitude). Products based on observations taken in the optical and near infra-red spectrum (e.g. global radiation) provide higher spatial resolution (up to a few kilometres) than products based on microwave observations (e.g. precipitation) that have a much coarser resolution of about 25 km. For studies with focus on Europe, such products can be derived from the observations of the SEVIRI-instrument (Spinning Enhanced Visible and InfraRed Imager) which is on-board of EUMETSAT's METEOSAT-satellites (from 2004 onwards only). UERRA will produce a 20 year cloud cover reanalysis (based on Häggmark et al., 2000) from METEOSAT and NOAA AVHRR satellite data.

Potentially useful parameters for RRA uncertainty characterization include global radiation, cloud liquid water path, and snow water equivalent. The CM SAF CLoud property dAtAset using SEVIRI (CLAAS) provides, amongst other, daily means of cloud properties and solar radiation (Stengel et al., 2014) and is freely available through Stengel et al. (2013). Snow water equivalent (SWE) data (Takala et al., 2011) is freely provided by the European Space Agency's (ESA) GlobSnow and GlobSnow-2 projects (Luojus et al., 2010, http://www.globsnow.info). Precipitation and total cloud cover are available as satellite products but need to be used with care. The precipitation products have a coarser horizontal resolution of $0.25° \times 0.25°$, do not all cover whole of Europe, and underestimate precipitation in all seasons (Kidd et al., 2012). The total cloud cover satellite product is difficult to compare against model output because of different definitions of total cloud cover between model and satellite instrument.

In order to characterize uncertainties, skill scores in addition to the usually applied bias and RMSE are needed to determine the effective temporal and spatial resolution of the RRAs and satellite data products. Climatologies of frequency distributions, numbers of threshold value exceedance, and any variation over time has to be captured. Another possibility is to apply the fractional skill score (refer to Sect. 3) to the RRAs and the satellite data sets by comparing against a high resolved (station based) reference and investigate different temporal and spatial scales. This result would be of value for users of RRAs in order to judge which data product is likely to best suited for their applications. The analysis can be complicated by the fact that the RRAs may feature a finer effective resolution than the satellite data products, thus only scales corresponding to the satellite data (temporal and/or spatial) can be compared with this method.

2.5 Method E: ensemble based comparison

Method (E) describes validation performed on ensembles of regional reanalyses which are developed in the UERRA project. For instance, MO plans a set of about 20 ensemble members for their RRA employing a 4D-Var data assimilation system. The benefit of an ensemble-based reanalysis is that it inherently provides estimates of the analysis error, as noted by Whitaker et al. (2009). Additionally, the spatio-temporal evolution of the background error is estimated from the ensemble-spread. The ensemble spread describes possible variability and uncertainty within the reanalysis. Whitaker and Loughe (1998) have examined the relationship between the ensemble spread and ensemble mean skill and found that the spread can be used as a measure of skill for the ensemble mean if the spread is very large or very small compared to its climatological value. Many more ensemble based verification methods have been developed and are listed in Sect. 3. The drawbacks of an ensemble-based reanalysis is that the Ensemble Prediction System (EPS) costs considerable additional computing time and therefore always results into a trade-off between resolution and the number of ensemble members calculated. Additionally, it cannot be guaranteed that the ensemble members cover the full physical space of possible realizations.

Due to the fact that the ensembles of the RRAs are calculated on a lower resolution (about twice as coarse), one interesting scientific question is to what extent the ensembles provide a better estimate on spatial and temporal uncertainty than the deterministic reanalysis. Specifically, a user of the RRAs would be interested in which uncertainty characteristics can be interpreted from the ensembles for user relevant parameters.

Simple metrics such as the ensemble mean and spread can be used to characterize the inherent variability of each RRA and the possible temporal and spatial variation thereof. If an ensemble has been tested to be reliable (refer Sect. 3), it may be used to produce probability density functions (PDFs) or cumulative distribution functions (CDFs) of a given parameter. This allows users to retrieve, for this parameter, information about the probability of exceedance of a given threshold. In addition, probabilistic scores can be applied against station observations, gridded station data, and satellite data products (methods B, C, and D) where in principle all the above considerations hold true.

When performing verification of deterministic or ensemble based (re)analyses or forecasts, their skill is assessed against observations which are assumed to be exact. There is an inconsistency when accepting that the model may be uncertain while the observations are assumed to be certain which can lead to misguided verification results. Therefore, in recent years, the effect of observation errors on the verification of ensemble prediction systems in particular was analysed and new scores and methods were developed. Saetra et al. (2004) investigated the effects of observation errors on the statistics for ensemble spread and reliability. They show that rank histograms are highly sensitive to the inclusion of observation errors, whereas reliability diagrams are less sensitive. Candille and Talagrand (2008) introduce the "observational probability" method by defining observation uncertainty as a normal distribution which guarantees that the uncertainty is variable in both mean and spread. Santos and Ghelli (2012) extend the previous work and developed the "observed observational probability" method which includes uncertainty in the verification process to variables that are non-Gaussian distributed, in particular precipitation.

2.6 Method F: user models

Users regarding reanalysis as an additional data source simply use the reanalysis data fields and draw conclusions based on whether their application improved compared to their traditional forcing data. Though this is a valid approach for a certain case, it does not give scientifically sound uncertainty estimates. We discuss this method here because it is expected to be a popular way.

As one example for user models driven by RRA fields, the surface and soil model SURFEX (Masson et al., 2013) from Météo-France is considered in the UERRA project. The SURFEX model computes soil variables such as surface and deep soil temperature, soil moisture, and snow characteristics. The SURFEX drainage and run-off forces the hydrological model TRIP (Oki and Sud, 1998) to compute river discharges.

For validation, the discharge of catchments is compared to RRA precipitation over the same catchments. Though a modelled discharge closer to the observed ones is highly desirable for the application, conclusions from the validation results are not straightforward. Poor validation would not necessarily mean the RRA of the UERRA project was poor. It may be that the user model was tuned to good performance with the traditional input data. Likewise, good validation results may reflect the model tuning rather than the quality of the RRA.

3 Verification skill

For the various methods described above, verification scores and skill scores can be applied in order to provide an estimate of uncertainty for the RRAs. Which score can be applied also depends on the parameter but mainly on the question which is sought to be answered. Here, a short summary is given of suitable scores and skill scores which are intended to be applied within the various methods and to the appropriate parameters.

In Sect. 2.2, the described method is based on point (station) reference measurements. Frequency distributions of climatologically relevant threshold values can be used as verification measures such as how often, e.g. a temperature, precipitation, or wind speed threshold was exceeded (or fell below). In addition, there are a number of different skill scores which are suitable to quantify the uncertainty estimation and which are applicable to this method. To start with, continuous metrics such as the bias, RMSE, and correlation are good measures to get a first impression of the data product in relation to its reference. For instance, for temperature or wind speed these measures can easily be applied. However, for precipitation more advanced but commonly used scores and skill scores need to be applied based on contingency tables. These include but are not limited to the probability of (false) detection POD (POFD), false alarm ratio, critical success index, the equitable threat score, and Heidke skill score, (for a comprehensive discussion of statistical methods refer to Wilks, 2011).

When the reference data source is not a single station observation but a gridded data product, as introduced in Sects. 2.3 and 2.4, then the uncertainty estimation can be performed in two ways. It is possible to either verify pointwise by applying the above mentioned classical metrics or use new spatial-based verification approaches. These new verification approaches include fuzzy verification methods, e.g. the fractional skill score (FSS) (Roberts and Lean, 2008), as summarized by Ebert (2008); an added value index which quantifies the added value of a high resolution product compared to a lower resolution product as introduced by Kanamitsu and

DeHaan (2011); and an object-based approach for specifically verifying precipitation estimates as explained in Li et al. (2015).

As introduced in Sect. 2.5, there will be an ensemble based realization of the regional reanalyses. For the verification of ensemble based products, ensemble based and probabilistic verification procedures have been developed. These procedures cover certain aspects of the uncertainty estimation, so usually many scores and skill scores need to be applied to come to a comprehensive result. One classical way of checking the reliability (or statistical consistency) of the ensemble is to build Talagrand (or rank) histograms (Talagrand et al., 1999; Hamill, 2001). It answers the question how well the ensemble represent the true variability (uncertainty) of the observations. Using a reliability diagram (Hamill, 1997), the reliability, sharpness, and resolution of the ensembles (here: ensembles of regional reanalyses) are evaluated. Another diagram which is widely used is the relative operating characteristics (ROC) plot (Masson, 1982; Candille and Talagrand, 2008). Here, POD is plotted against POFD and it answers the question what the ability of the forecast is to discriminate between events and non-events. And finally, there are different skill scores which test the skill of the probabilistic information in varying ways, such as the Brier (skill) score (Brier, 1950; Candille and Talagrand, 2008), which relates to the user question: "What is the magnitude of the probability forecast errors?" or the (continuous) ranked probability score (Hersbach (2000) and references therein), which relates to the user question: "how well did the probability forecast predict the category that the observation fell into?".

4 Summary

In this paper we summarized methods to characterize a regional reanalysis. Several RRAs are produced in the FP7 UERRA project. Uncertainty characterization is among the main foci of the project. The aim was to arrive at uncertainty estimates which are suitable for end users and allow comparisons between different reanalyses. Following methods have been discussed: (A) feedback statistics, the comparison against (B) station observations, (C) gridded station observations, and (D) satellite data, followed by (E) ensemble based comparisons when the UERRA ensemble become available, as well as (F) evaluating user model output driven by RRA input.

We outlined the benefits and drawbacks of each method. Specifically, the benefit of method (A) feedback statistics is that it can be arranged as output during the reanalysis production, and deals with suitable observation operators to arrive at a scientifically sound comparison between reanalysis output and (independent) observations. The drawback of (A) is that it is not easily comparable between different reanalysis systems. A comparison against station measurements (method B) can be easily performed with different reanalysis systems, the result will depend on model resolution, topographic resolution, and representativeness of the station. Comparisons against gridded observations (method C) have the advantage that they fully cover the area of interest. The drawback of (C) is that the result of the comparison might be influenced by how close the spatial smoothing (or correlation length scales) in the gridding procedure resemble the ones in the reanalyses. The advantage of comparing to (not assimilated) satellite data (method D) is that here again the full area is covered, the disadvantage is that the satellite retrieval itself is a statistical procedure which has to rely on a number of assumptions, thus the uncertainty of the satellite product might be larger than that of the reanalyses in many circumstances. Finally, the advantage of method (E) ensembles is that the random part of reanalysis uncertainty can be estimated (if the ensemble is generated with one model) and additionally assess the model uncertainty by combining several models. We pointed out that method (F) investigates advantages in applications, is user friendly, but does not allow to generalize conclusions about uncertainties. The methods are related to scientific and user questions which might be answered with the respective method.

We discussed for which meteorological parameters which methods are most suitable by focusing especially on: air temperature, humidity, and wind speed at the near-ground levels as these are of primary user interest targeted by the regional reanalysis efforts. For the satellite data, suitable parameters include global radiation, cloud liquid water path, and snow water equivalent. Precipitation and total cloud cover from satellite are harder to interpret. Depending on the meteorological parameter of interest, several skill scores are recommended.

Acknowledgements. This study was supported through the UERRA project (grant agreement no. 607193 within the European Union Seventh Framework Programme). We acknowledge all UERRA WP3 partners for their discussion and scientific input, with special thanks for valuable contributions to all participants of the user's workshop within the UERRA project meeting in June 2014 at DWD, Offenbach, Germany.

Edited by: E. Bazile
Reviewed by: three anonymous referees

References

Berre, L.: Estimation of Synoptic and Mesoscale Forecast Error Covariances in a Limited-Area Model, Mon. Weather Rev., 128, 644–667, doi:10.1175/1520-0493(2000)128<0644:EOSAMF>2.0.CO;2, 2000.

Bollmeyer, C., Keller, J. D., Ohlwein, C., Wahl, S., Crewell, S., Friederichs, P., Hense, A., Keune, J., Kneifel, S., Pscheidt, I., Redl, S., and Steinke, S.: Towards a high-resolution regional reanalysis for the European CORDEX domain, Q. J. Roy. Meteorol. Soc., 141, 1–15, doi:10.1002/qj.2486, 2015.

Brier, G. W.: Verification of forecasts expressed in terms of probability, Mon. Weather Rev., 78, 1–3 , 1950.

Brunet, M., Jones, P. D., Jourdain, S., Efthymiadis, D., Kerrouche, M., and Boroneant, C.: Data sources for rescuing the rich heritage of Mediterranean historical surface climate data, Geosci. Data J., 1, 61–73, doi:10.1002/gdj3.4, 2013.

Bubnova, R., Hello, G., Benard, P., and Geleyn, J.-F.: Integration of the fully elastic equations cast in the hydrostatic pressure terrain-following in the framework of the ARPEGE/ALADIN NWP system, Mon. Weather Rev., 123, 515–535, 1995.

Candille, G. and Talagrand, O.: Impact of observational error on the validation of ensemble prediction systems, Q. J. Roy. Meteorol. Soc., 134, 959–971, doi:10.1002/qj.268, 2008.

Clayton, A. M., Lorenc, A. C., and Barker, D. M.: Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office, Q. J. Roy. Meteorol. Soc., 139, 1455–1461, doi:10.1002/qj.2054, 2013.

Dahlgren, P., Kållberg, P., Landelius, T., and Undén, P.: Comparison of the regional reanalyses products with newly developed and existing state-of-the art systems, EURO4M project report D2.9, http://www.euro4m.eu/ (last access: 23 October 2015), 2014.

Davies, T., Cullen, M. J. P., Malcolm, A. J., Mawson, M. H., Staniforth, A., White, A. A., and Wood, N.: A new dynamical core for the Met Office's global and regional modelling of the atmosphere, Q. J. Roy. Meteorol. Soc., 131, 1759–1782, doi:10.1256/qj.04.101, 2006.

Dee, D. P., Uppala, S. M., Simmons, A. J., et al.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. Roy. Meteorol. Soc., 137, 553–597, doi:10.1002/qj.828, 2011.

Dee, D. P., Balmaseda, M., Balsamo, G., Engelen, R., Simmons, A. J., and Thepaut, J.-N.: Towards a consistent reanalysis of the climate system, Bull. Amer. Meteor. Soc., 95, 1235–1248, doi:10.1175/BAMS-D-13-00043.1, 2014.

Desroziers, G., Berre, L., Chapnik, B., and Poli, P.: Diagnosis of observation, background and analysis-error statistics in observation space, Q. J. Roy. Meteorol. Soc., 131, 3385–3396, doi:10.1256/qj.05.108, 2005.

De Troch, R., Hamdi, R., Vyver, H., Geleyn, J.-F., and Termonia, P.: Multiscale Performance of the ALARO-0 Model for Simulating Extreme Summer Precipitation Climatology in Belgium, J. Climate, 26, 8895–8915, doi:10.1175/JCLI-D-12-00844.1, 2013.

Ebert, E. E.: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework, Meteorol. Appl., 15, 51–64, doi:10.1002/met.25, 2008.

Fischer, C., Montmerle, T., Berre, L., Auger, L., and Stefanescu, S. E.: An overview of the variational assimilation in the ALADIN/France numerical weather-prediction system, Q. J. Roy. Meteorol. Soc., 131, 3477–3492, doi:10.1256/qj.05.115, 2005.

Gandin, L. and Murphy, A.: Equitable Skill Scores for Categorical Forecasts, Mon. Weather Rev., 120, 361–370, doi:10.1175/1520-0493(1992)120<0361:ESSFCF>2.0.CO;2, 1992.

Gerard, L., Piriou, J.-M., Brožková, R., Geleyn, J.-F., and Banciu, D.: Cloud and Precipitation Parameterization in a Meso-Gamma-Scale Operational Weather Prediction Model, Mon. Weather Rev., 137, 3960–3977, doi:10.1175/2009MWR2750.1, 2009.

Häggmark, L., Ivarsson, I., Gollvik, S., and Olofsson, O.: Mesan, an operational mesoscale analysis system, Tellus A, 52, 2–20, doi:10.3402/tellusa.v52i1.12250, 2000.

Hakuba, M.Z., Folini, D., Sanchez-Lorenzo, A., and Wild, M.: Spatial representativeness of ground-based solar radiation measurements – Extension to the full Meteosat disk, J. Geophys. Res.-Atmos., 119, 11760–11771, doi:10.1002/2014JD021946, 2014.

Hamill, T. M.: Reliability diagrams for multicategory probabilistic forecasts, Weather Forecast., 12, 736–741, 1997.

Hamill, T. M.: Interpretation of rank histograms for verifying ensemble forecasts, Mon. Weather Rev., 129, 550–560, doi:10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2, 2001.

Harnisch, F. and Keil, C.: Initial conditions for convective-scale ensemble forecasting provided by ensemble data assimilation, Mon. Weather Rev., 143, 1583–1600, doi:10.1175/MWR-D-14-00209.1, 2015.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., and New, M.: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, J. Geophys. Res., 113, D20119, doi:10.1029/2008JD010201, 2008.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather Forecast., 15, 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Hollingsworth, A. and Lönnberg, P.: The statistical structure of short-range forecast errors as determined from radiosonde data, Part I: The wind field, Tellus A, 38, 111–136, doi:10.1111/j.1600-0870.1986.tb00460.x, 1986.

Hollmann, R., Merchant, C. J., Saunders, R., Downy, C., Buchwitz, M., Cazenave, A., Chuvieco, E., Defourny, P., de Leeuw, G., Forsberg, R., Holzer-Popp, T., Paul, F., Sandven, S., Sathyendranath, S., van Roozendael, M., and Wagner, W.: The ESA Climate Change Initiative: Satellite Data Records for Essential Climate Variables, B. Am. Meteorol. Soc., 94, 1541–1552, doi:10.1175/BAMS-D-11-00254.1, 2013.

Horányi, A., Ihász, I., and Radnóti, G.: ARPEGE/ALADIN: a numerical weather prediction model for Central-Europe with the participation of the Hungarian Meteorological Service, Időjárás, 100, 277–301, 1996.

Hunt, B. R., Kostelich, E. J., and Szunyogh, I.: Efficient data assimilation for spatiotemporal chaos: A local ensemble transformation Kalman filter, Physics D, 230, 112–126, doi:10.1016/j.physd.2006.11.008, 2007.

Isotta, F. A., Frei, C., Weilguni, V., Perčec Tadič, M., Lassègues, P., Rudolf, B., Pavan, V., Cacciamani, C., Antolini, G., Ratto, S.M., Munari, M., Micheletti, S., Bonati, V., Lussana, C., Ronchi, C., Panettieri, E., Marigo, G., and Vertačnik, G.: The climate of daily precipitation in the Alps: development and analysis of a high-resolution grid dataset from pan-Alpine rain-gauge data, Int. J. Climatol., 34, 1657–1657, doi:10.1002/joc.3794, 2013.

Isotta, F. A., Vogel, R., and Frei, C.: Evaluation of European regional reanalyses and downscalings for precipitation in the Alpine region, Meteorol. Z., 24, 15–37, doi:10.1127/metz/2014/0584, 2015.

Jansson, A., Persson, C., and Strandberg, G.: 2D meso-scale reanalysis of precipitation, temperature and wind over Europe-ERAMESAN: Time period 1980–2004, SMHI Rep. Meteorol. Clim., 112, 44, 2007.

Kaiser-Weiss, A. K., Kaspar, F., Heene, V., Borsche, M., Tan, D. G. H., Poli, P., Obregon, A., and Gregow, H.: Comparison

of regional and global reanalysis near-surface winds with station observations over Germany, Adv. Sci. Res., 12, 187–198, doi:10.5194/asr-12-187-2015, 2015.

Kanamitsu, M. and DeHaan, L.: The Added Value Index: A new metric to quantify the added value of regional models, J. Geophys. Res., 116, D11106, doi:10.1029/2011JD015597, 2011.

Kidd, C., Bauer, P., Turk, J., Huffman, G. J., Joyce, R., Hsu, K.-L., and Braithwaite, D.: Intercomparison of High-Resolution Precipitation Products over Northwest Europe, J. Hydrometeorol., 13, 67–83, doi:10.1175/JHM-D-11-042.1, 2012.

Li, J., Hsu, K., AghaKouchak, A., and Sorooshian, S.: An object-based approach for verification of precipitation estimation, Int. J. Remote Sens., 36, 513–529, doi:10.1080/01431161.2014.999170, 2015.

Lönnberg, P. and Hollingsworth, A.: The statistical structure of short-range forecast errors as determined from radiosonde data Part II: The covariance of height and wind errors, Tellus A, 38, 137–161, doi:10.1111/j.1600-0870.1986.tb00461.x, 1986.

Luojus, K., Pulliainen, J., Takala, M., Derksen, C., Rott, H., Nagler, T., Solberg, R., Wiesmann, A., Metsämäki, S., Malnes, E., and Bojkov, B.: Investigating the feasibility of the GlobSnow snow water equivalent data for climate research purposes, IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Honolulu, HI, USA, 4851–4853, doi:10.1109/IGARSS.2010.5741987, 2010.

Masson, I.: A model for assessment of weather forecasts, Aust. Meteorol. Mag., 30, 291–303, 1982.

Masson, V., Le Moigne, P., Martin, E., Faroux, S., Alias, A., Alkama, R., Belamari, S., Barbu, A., Boone, A., Bouyssel, F., Brousseau, P., Brun, E., Calvet, J.-C., Carrer, D., Decharme, B., Delire, C., Donier, S., Essaouini, K., Gibelin, A.-L., Giordani, H., Habets, F., Jidane, M., Kerdraon, G., Kourzeneva, E., Lafaysse, M., Lafont, S., Lebeaupin Brossier, C., Lemonsu, A., Mahfouf, J.-F., Marguinaud, P., Mokhtari, M., Morin, S., Pigeon, G., Salgado, R., Seity, Y., Taillefer, F., Tanguy, G., Tulet, P., Vincendon, B., Vionnet, V., and Voldoire, A.: The SURFEXv7.2 land and ocean surface platform for coupled or offline simulation of earth surface variables and fluxes, Geosci. Model Dev., 6, 929–960, doi:10.5194/gmd-6-929-2013, 2013.

Oki, T. and Sud, Y. C.: Design of Total Runoff Integration Pathways (TRIP) – A global river channel network, Earth Interact., 2, 1–37, doi:10.1175/1087-3562(1998)002<0001:DOTRIP>2.3.CO;2, 1998.

Poli, P., Hersbach, H., Tan, D., Dee, D., Thépaut, J.-N., Simmons, A., Peubey, C., Laloyaux, P., Komori, T., Berrisford, P., Dragani, R., Trémolet, Y., Hólm, E., Bonavita, M., Isaksen, L., and Fisher, M.: The data assimilation system and initial performance evaluation of the ECMWF pilot reanalysis of the 20th-century assimilating surface observations only (ERA-20C), ERA Rep. Ser., 14, 59, 2013.

Rawlins, F., Ballard, S. P., Bovis, K. J., Clayton, A. M., Li, D., Inverarity, G. W., Lorenc, A. C., and Payne, T. J.: The Met Office global four-dimensional variational data assimilation scheme, Q. J. Roy. Meteorol. Soc., 133, 347–362, doi:10.1002/qj.32, 2007.

Renshaw, R.: New state-of-the-art NAE-based regional atmospheric data assimilation reanalysis system, EURO4M project report, D2.1, D2.2, http://www.euro4m.eu/ (last access: 23 October 2015), 2013.

Roberts, N. M. and Lean, H. W.: Scale-selective verification of rainfall accumulations from high-resolution forecast of convective events, Mon. Weather Rev., 136, 78–97, doi:10.1175/2007MWR2123.1, 2008.

Saetra, Ø., Hersbach, H., Bidlot, J.-R., and Richardson, S.: Effects of Observation Errors on the Statistics for Ensemble Spread and Reliability, Mon. Weather Rev., 132, 1487–1501, doi:10.1175/1520-0493(2004)132<1487:EOOEOT>2.0.CO;2, 2004.

Santos, C. and Ghellie, A.: Observational probability method to assess ensemble precipitation forecasts, Q. J. Roy. Meteorol. Soc., 138, 209–221, doi:10.1002/qj.895, 2012.

Schättler, U., Doms, G., and Schraff, C.: A description of the nonhydrostatic regional COSMO-model – Part VII: User's guide, Technical report, Deutscher Wetterdienst, Offenbach, Germany, http://www.cosmo-model.org/ (last access: 23 October 2015), 2014.

Simmer, C., Adrian, G., Jones, S., Wirth, V., Göber, M., Hohenegger, C., Janjic, T., Keller, J., Ohlwein, C., Seifert, A., Trömel, S., Ulbrich, T., Wapler, K., Weissmann, M., Keller, J., Masbou, M., Meilinger, S., Riß, N., Schomburg, A., Vormann, A., and Weingärtner, C.: HErZ – The German Hans-Ertel Centre for Weather Research, B. Am. Meteorol. Soc., doi:10.1175/BAMS-D-13-00227.1, in press, 2015.

Soci, C., Landelius, T., Bazile, E., Undén, P., Mahfouf, J.-F., Martin, E., and Besson, F.: Comparison of existing ERAMESAN with SAFRAN downscaling, EURO4M project report D2.10, http://www.euro4m.eu/ (last access: 23 October 2015), 2011.

Soci, C., Bazile, E., Besson, F., Landelius, T., Mahfouf, J.-F., Martin, E., and Durand, Y.: Report describing the new MESAN-SAFRAN downscaling system, EURO4M project report D2.6, http://www.euro4m.eu/ (last access: 23 October 2015), 2013.

Stengel, M., Kniffka, A., Meirink, J. F., Riihelä, A. Trentmann, J., Müller, R., Lockhoff, M., and Hollmann, R.: CLAAS: CM SAF CLoud property dAtAset using SEVIRI – Edition 1 – Hourly/Daily Means, Pentad Means, Monthly Means/Monthly Mean Diurnal Cycle/Monthly Histograms, Satellite Application Facility on Climate Monitoring, Offenbach, Germany, doi:10.5676/EUM_SAF_CM/CLAAS/V001, 2013.

Stengel, M., Kniffka, A., Meirink, J. F., Lockhoff, M., Tan, J., and Hollmann, R.: CLAAS: the CM SAF cloud property dataset using SEVIRI, Atmos. Chem. Phys., 14, 4297–4311, doi:10.5194/acp-14-4297-2014, 2014.

Takala, M., Luojus, K., Pulliainen, J., Derksen, C., Lemmetyinen, J., Karna, J. P., Koskinen, J., and Bojkov, B.: Estimating northern snow water equivalent for climate research through assimilation of space-borne radiometer data and ground-based measurements, Remote Sens. Environ., 115, 3517–3529, doi:10.1016/j.rse.2011.08.014, 2011.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluations of probabilistic evaluation systems, Proceedings, ECMWF Workshop on Predictability, October 1997, Reading, UK, 1–25, 1999.

Uppala, S. M., Kållberg, P. W., Simmons, A. J., et al.: The ERA-40 re-analysis, Q. J. Roy. Meteorol. Soc., 131, 2961–3012, doi:10.1256/qj.04.176, 2005.

van den Besselaar, E. J. M., Haylock, M. R., van der Schrier, G., and Klein Tank, A. M. G.: A European daily high-resolution observational gridded data set of sea level pressure, J. Geophys. Res., 116, D11110, doi:10.1029/2010JD015468, 2011.

Whitaker, J. S. and Loughe, A. F.: The Relationship between Ensemble Spread and Ensemble mean Skill, Mon. Weather Rev., 126, 3292–3302, doi:10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2, 1998.

Whitaker, J. S., Compo, G. P., and Thépaut, J.-N.: A Comparison of Variatinal and Ensemble-Based Data Assimilation Systems for Reanalysis of Sparse Observations, Mon. Weather Rev., 137, 1991–1999, doi:10.1175/2008MWR2781.1, 2009.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, 3rd Edn., International Geophysics, Academic Press, Oxford, UK, 704 pp., 2011.