Seventh Framework Programme

Theme 6 [SPACE]

Project: 607193 UERRA

Full project title:

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

Deliverable D 3.5

Preliminary report of assessment of regional reanalyses – first results

WP no:	3
WP leader:	DWD
Lead beneficiary for deliverable :	DWD
Name of <u>authors</u>:	Michael Borsche, Andrea Kaiser-Weiss (DWD), Jemma Davie (MO), Gerard van der Schrier (KNMI). Cristian Lussana, Ole Einar Tveito (MI). Christoph Frei, Francesco Isotta (EDI)
Nature:	Report
Dissemination level:	PU
Deliverable month:	34
Submission date: November 28, 2016	Version nr: 1

Deliverable 3.5 (D3.5): Preliminary report of assessment of regional reanalysis – first results from UERRA WP3

By Michael Borsche[1], Jemma Davie[2], Gerard van der Schrier[3], Cristian Lussana[4], Francesco Isotta[5], Christoph Frei[5], and Andrea Kaiser-Weiss[1]

[1] Deutscher Wetterdienst (DWD), Offenbach, Germany

[2] Met Office (MO), Exeter, United Kingdom

[3] Royal Netherlands Meteorological Institute (KNMI), De Bilt, The Netherlands

[4] Norwegian Meteorological Institute (MI), Oslo, Norway

[5] Eidgenössisches Departement des Inneren (EDI), Switzerland

1. Scope of this document

Within work package 3 (WP3), it has been discussed since the start of the project on how to access scientifically the UERRA regional reanalyses. A draft concept has been initiated at the workshop (D3.1) resulting in a collection of common evaluation procedures (D3.2), and agreement on following methodologies for characterizing uncertainties:

Method A: feedback statistics,

Method B: comparison against station observations,

Method C: comparison against gridded station observations,

Method D: comparison against satellite data,

Method E: ensemble based comparison,

Method F: user related models.

In this report, the WP3 activities relating to Method A, B, C, D and E are explained, the applied methods introduced, and the fitness for purpose is demonstrated. The demonstrations rely partly on preliminary data (i.e., either on EURO4M output, or preliminary UERRA output, as UERRA data are still under production). No deviation of methods, only an update of results is expected as soon as the preliminary input can be replaced with final UERRA data. The developed code is linked. Method F is treated in WP4, thus outside the scope of this document.

2. Method A: Use of ODB for Observation Feedback statistics

2.1. Method description

ODB (Observational DataBase - http://www.ecmwf.int/en/elibrary/15080-odb-past-present-and-future) is a format developed by ECMWF to store observation data and metadata, together with useful additional information from an analysis system.

This will typically include model background and analysis values, but can also include other 'feedback' information such as quality control decisions from the observation processing system. There is potential for observation feedback from reanalyses to be useful for many purposes. For instance, they can be used to assess and filter the observation records. Observing sites that report values consistently different from the reanalysis might be regarded as unreliable. Time series of observation minus reanalysis differences can reveal sudden changes at individual stations, possibly due to instrument calibration errors or perhaps the station was relocated.

Here examples are given of feedback information from ODB for a single month (May 1979) from a Met Office reanalysis produced as part of UERRA. The reanalysis uses the UM model at 36-km resolution over the EU-CORDEX domain, using conventional data (surface, upper air and aircraft) together with TOVS radiances in a 4D variational assimilation system. This particular run is the control run ('member 0') for a 20-member ensemble. These examples are to illustrate the potential for ODBs in validation of reanalyses.

2.1.1. Advantages

The observation feedback is produced during the reanalysis production, requiring no extra effort.

2.1.2. Disadvantages

This method is system dependent; observation feedback between different systems can be compared only in connection of understanding the systems. This method is limited to data which are assimilated.

2.1.3. Value of method

Observation feedback from reanalyses can be used to assess and filter the observation records. Observing sites that report values consistently different from the reanalysis might be regarded as unreliable. Time series of observation minus reanalysis differences can reveal sudden changes at individual stations. Observation feedback is calculated to serve the producer of the reanalysis and enhance the production. Users can benefit, if they seek information on to what extend the reanalysis differs from alternative data (in case they are assimilated).

2.2. Example of Application

2.2.1. Parameter

2m temperature fields were investigated.

2.2.2. Investigated spatial and temporal scale

Native spatial and temporal resolution of measurement was investigated.

2.2.3. Used observations

2m temperature from SYNOP stations were used.

2.2.4. Investigated reanalyses

This method was investigated for the regional reanalysis of Met Office.

2.3. Preliminary Results

Figures 2.1 and 2.2 are map plots of the bias and standard deviation of O-B (Observation minus model Background, i.e. 6-hour forecast from reanalysis) 2m temperature at observation stations. These maps reveal which stations have large biases and/or standard deviations, and where they are. This can potentially reveal problems in the observation data or in the model reanalysis. The bias and standard deviation calculations ignore any observations with gross errors as these would be rejected by QC (quality control) and are not necessarily an indicator of the overall quality of observations from the station. However, the percentage of gross errors is considered in deciding whether to use these stations.

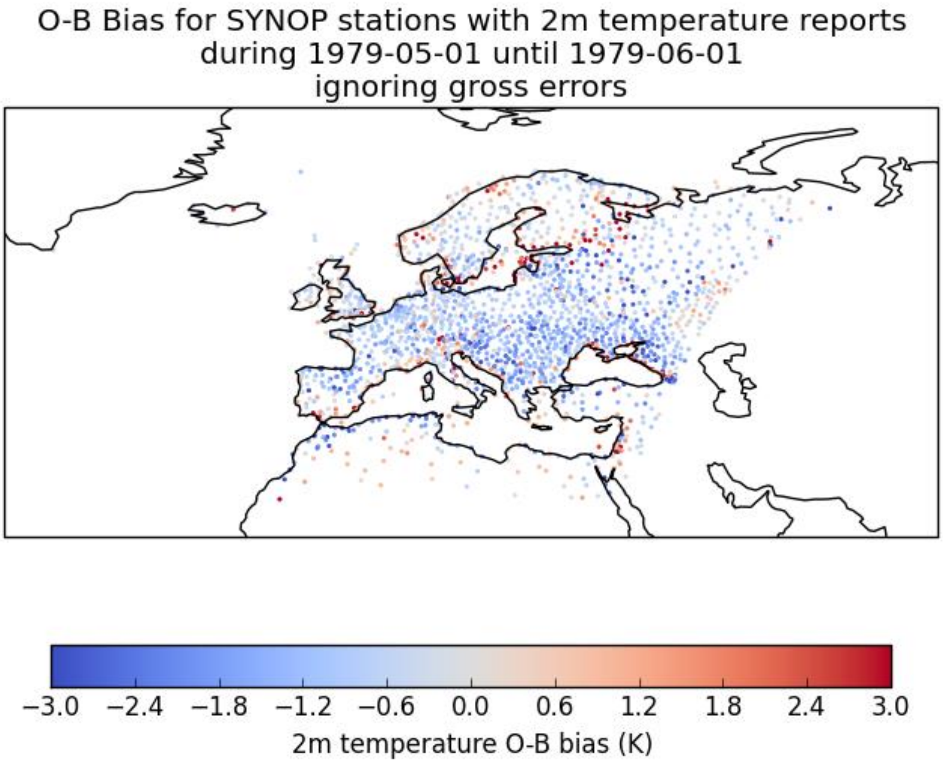

O-B Bias for SYNOP stations with 2m temperature reports
during 1979-05-01 until 1979-06-01
ignoring gross errors
−3.0
−2.4
−1.8
−1.2
−0.6
0.0
0.6
1.2
1.8
2.4
3.0
2m temperature O-B bias (K)

Figure 2.1: O-B bias for 2m temperature, May 1979

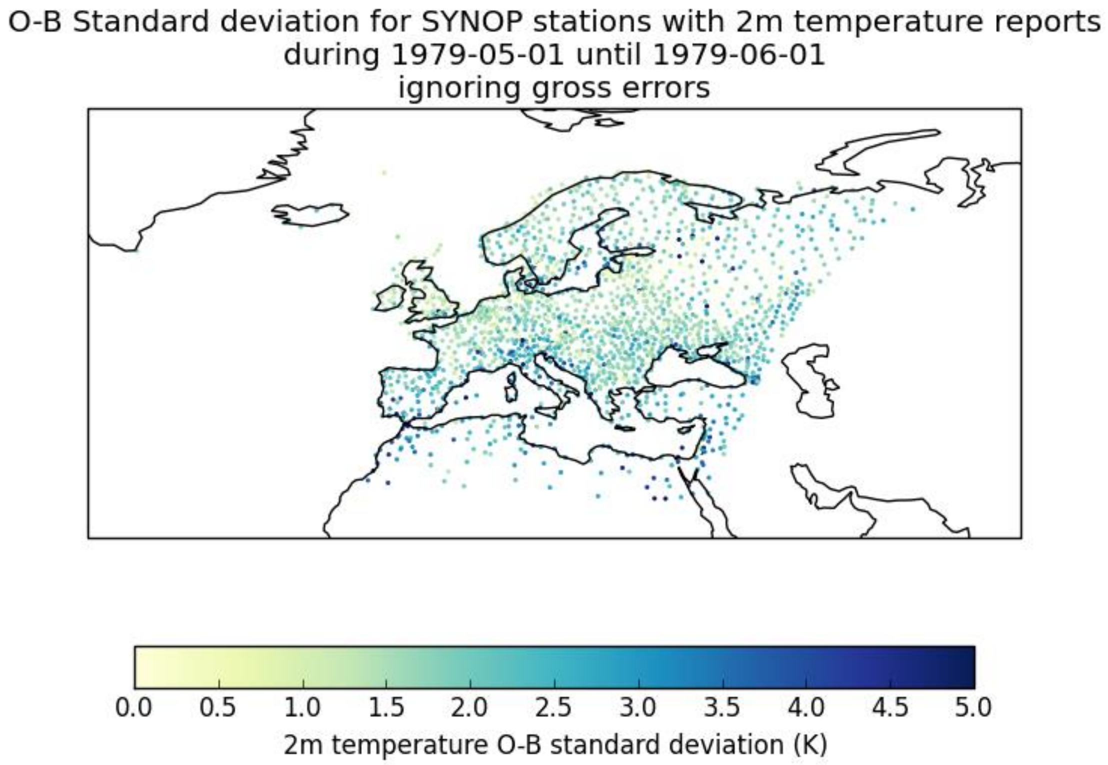

O-B Standard deviation for SYNOP stations with 2m temperature reports
during 1979-05-01 until 1979-06-01
ignoring gross errors
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
2m temperature O-B standard deviation (K)

Figure 2.2: O-B std dev for 2m temperature, May 1979

Histograms for individual stations can be plotted to visualise the distributions of O-B values (Figure 2.3). These can reveal bias, asymmetry or multiple peaks. Figure 2.3 shows one station (06107) with low bias and standard deviation and another (01218) which has exceeded thresholds for bias and standard deviation and so has been rejected from the assimilation system for the month.

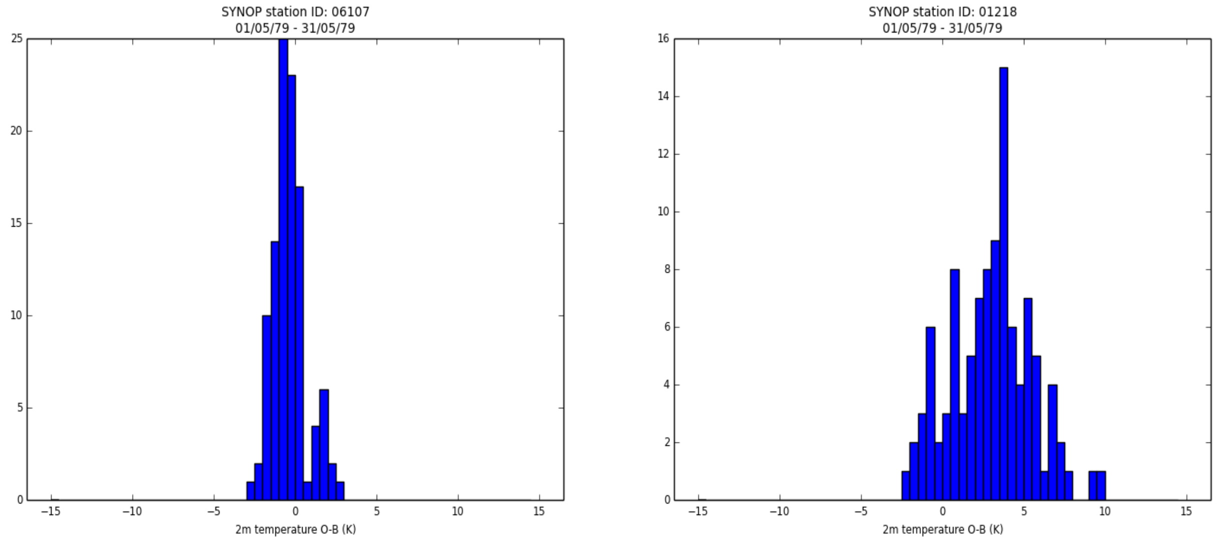


Figure 2.3: O-B histograms for 2m temperature, Stations 06107 (left), 01218 (right)

Time series plots (Figure 2.4) can show any jumps in O-B or trends. Examples of a station which was rejected by the observation monitoring system and one not rejected are shown in Figure 2.4 for May 1979. Station 06107 has reasonably low bias and small spread. Station 01218 has a significant positive bias and larger spread, exceeding the thresholds for rejection. Observations are monitored in this way on a monthly basis to reject stations that exceed O-B thresholds over the month. This provides an additional safeguard to avoid the use of poor quality data.

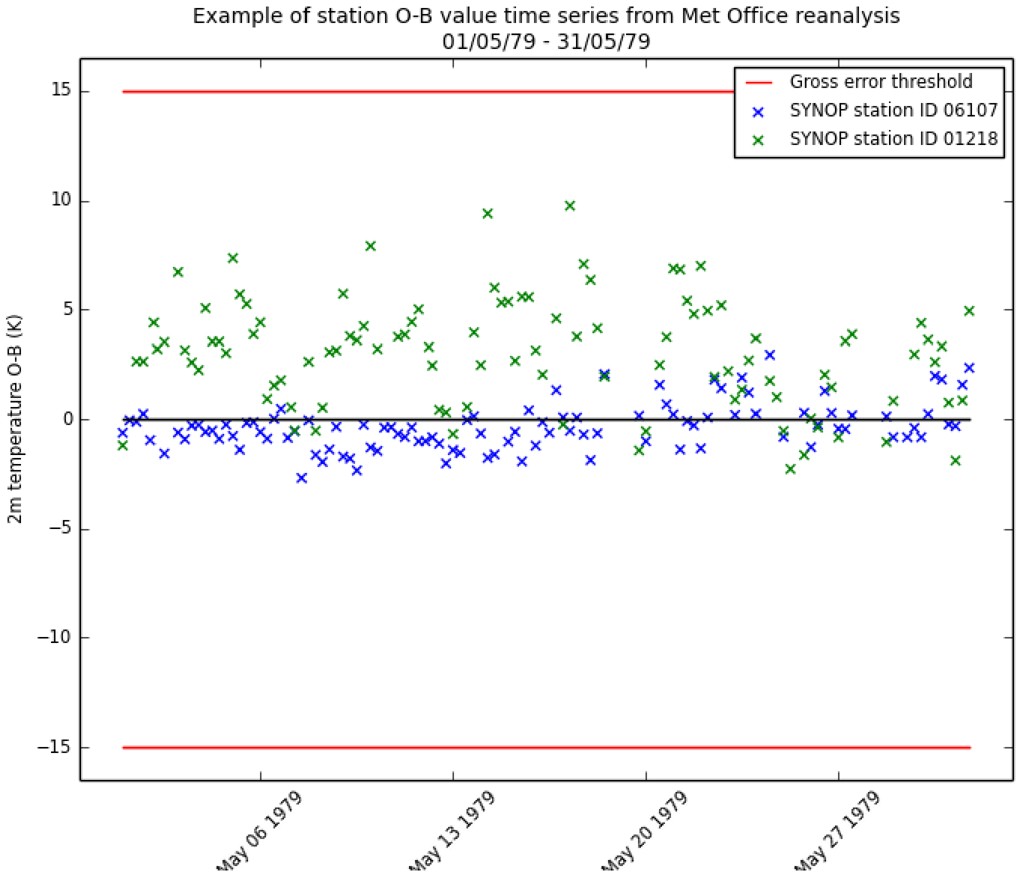


Figure 2.4: Time series of O-B 2m temperature for 2 stations: 06107 (blue), 01218 (green)

Other possible methods of exploiting ODBs for comparing observations with reanalyses are scatter plots of O against B and calculating correlation coefficients between O and B values (Figure 2.6), time series plots (Figure 2.7), and histograms of O and B values (Figure 2.5). As well as looking at groups of observations, information can be extracted about whether individual observations have been assimilated or rejected. For upper air observation types, vertical profiles of O-B bias and standard deviation on model levels for each station can be calculated.

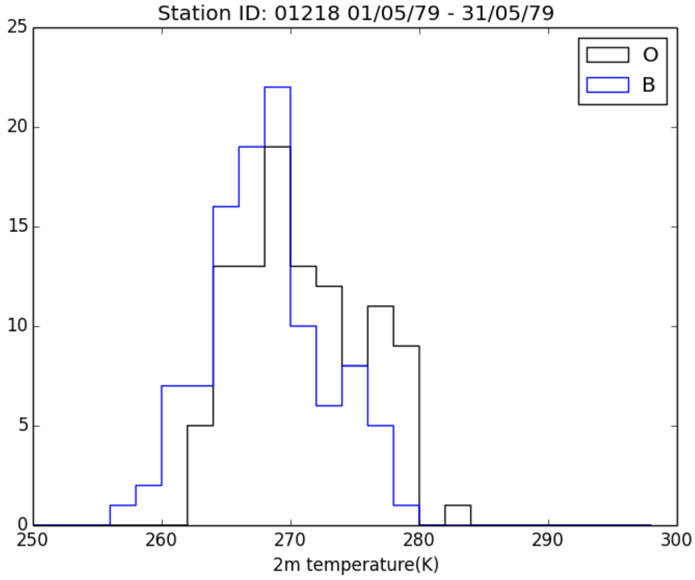


Figure 2.5: Histogram of O and B 2m temperature for station 01218

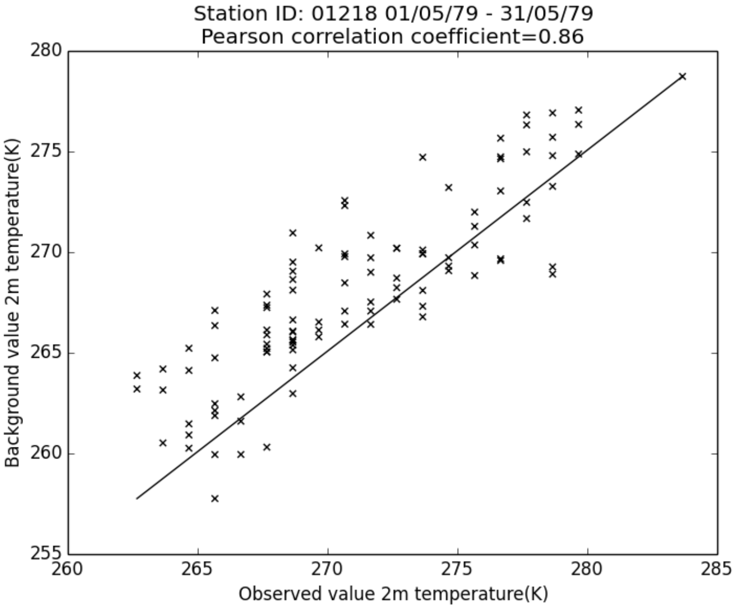


Figure 2.6: Scatterplot of O, B 2m temperature for station 01218

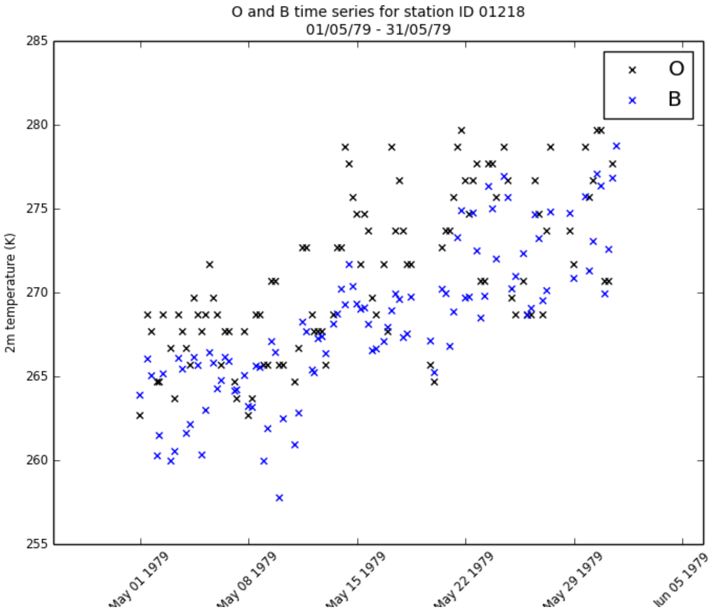


Figure 2.7: Time series of O (black), B (blue) for station 01218

3. Method B: Comparison against station observations

3.1. Method description

Grid cell values of regional reanalyses are compared against point measurements of either operational station data over Germany operated by DWD or measurements taken by tall meteorological towers. Whereas station observations are limited to one height near the ground, tower measurements are taken at different heights up to hundreds of meters above the ground. These measurements can be compared against values in corresponding model level heights of the reanalyses.

Statistics of different temporal scales ranging from hourly to inter-annual observations were calculated and include correlation, bias, RMSE, anomalies, PDF-score, and frequency distribution. In addition, skill scores based on a 2x2 contingency table are calculated. These enable investigation of extreme events and include the hit rate, false alarm rate, false alarm ratio, Heidke skill score (HSS), threat score (TS), equitable threat score (ETS) or Gilbert skill score, frequency bias index, accuracy, odds ratio, extremal dependence index (EDI), and symmetric extremal dependence index (SEDI), the latter two introduced by [Ferro and Stephenson, 2011].

When comparing absolute values between station data and regional reanalyses it needs to be kept in mind that point measurements are compared with grid cell values. Differences could be caused by insufficient representativity, mismatching surface roughness, and (as is especially the case with tower measurements) by mismatching heights. For these reasons, a relative comparison is pursued here for the determination of the contingency table based skill scores. The benchmark for which to calculate the values of the contingency table is based on percentiles of the station and reanalysis time series instead of their absolute values.

3.1.1. Advantages

This method is easy to apply.

3.1.2. Disadvantages

Comparison of grid cells of a spatial extend of several tens to hundreds of square kilometres with point measurements is far from comparing like-with-like. Keeping in mind that the station measurement is often treated as representative for a certain region, this method is still justified.

3.1.3. Value of method

This method helps users who traditionally rely on station measurements to understand the potential of using reanalysis data.

3.2. Developed and shared code

The procedures implementing the evaluation have been developed within this project by using the R-language (R. Core Team (2105). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/) and are being shared on github at https://github.com/UERRA-EVA/EVA_stationobs (licensed under GPL version 2 or any later version).

3.3. Example of Application

3.3.1. Parameter

The method was applied to wind speed near the ground, covering 10m to 100m height.

3.3.2. Investigated spatial and temporal scale

The evaluation was performed on grid cell values of the regional reanalyses versus point measurements for hourly, daily, monthly, annual and inter-annual time scales.

3.3.3. Used observations

Data which are compared against include a) tower measurements of Lindenberg, Cabauw, and the FINO platforms, of hourly, daily, and monthly values and b) DWD station data measurements available from ftp://ftp-cdc.dwd.de/pub/CDC/.

3.3.4. Investigated reanalyses

Investigated reanalyses include preliminary data, namely the regional reanalysis COSMO-REA6 covering the time range 1995 to 2014, and the two regional reanalyses developed during EURO4M by SMHI and the UK MetOffice covering the years 2008 and 2009, and the two global reanalyses ERA20C (1901 to 2010) and ERA-Interim (1979 to 2010).

3.4. Preliminary Results

Comparison of station measurements of 10m wind speed against regional reanalyses is shown here exemplarily for the station Hannover. In Figure 3.1, four panels of the frequency distribution of the 10m wind speed for the station and the three reanalyses are shown. Indicated are also the numbers of hourly and six-hourly measurements in the distribution plots, the mean, median, and the 1st and 99th percentile of measured or calculated wind speed. The number of measurements is less for SMHI because the last three months of the two year time series are missing in the preliminary EURO4M. The statistical measures provided indicate a good match for the COSMO-REA6 reanalysis, whereas the preliminary EURO4M SMHI fields tend to have too high wind speed and the preliminary EURO4M MetOffice fields tend to have too low wind speed at this station.

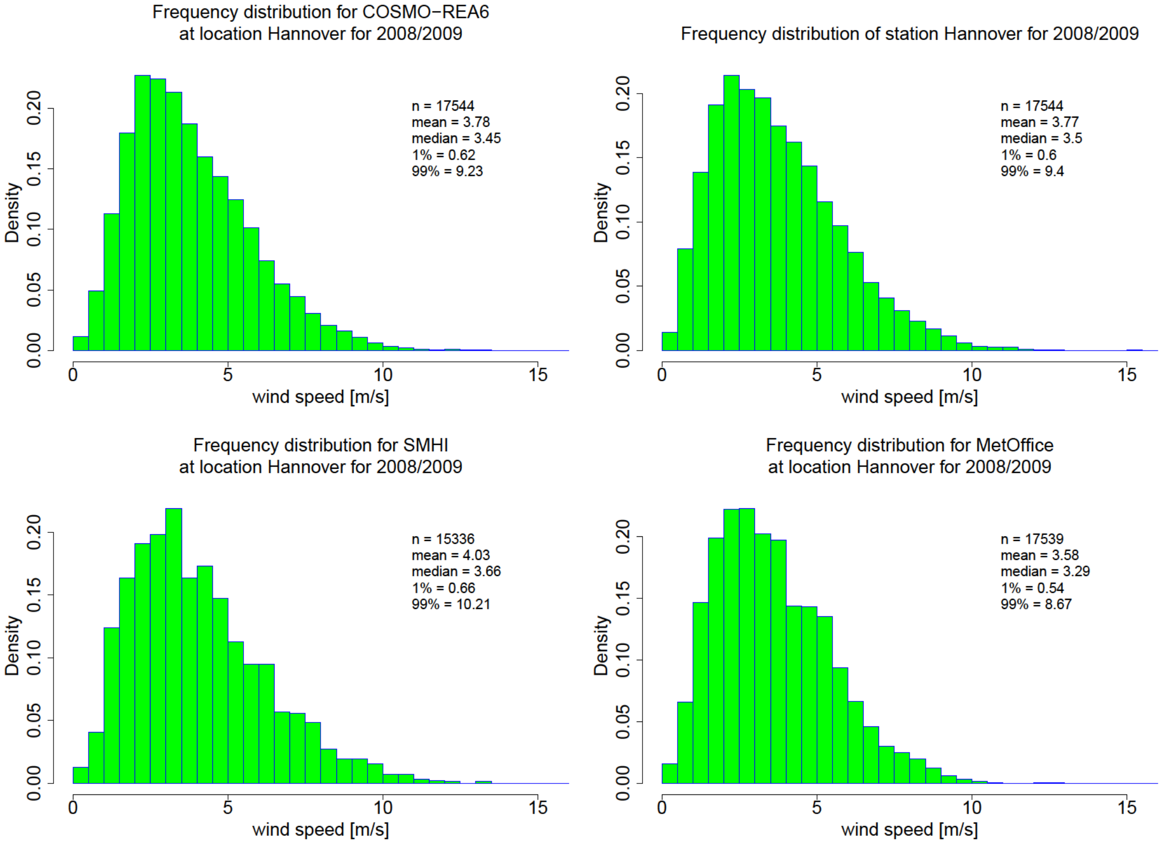
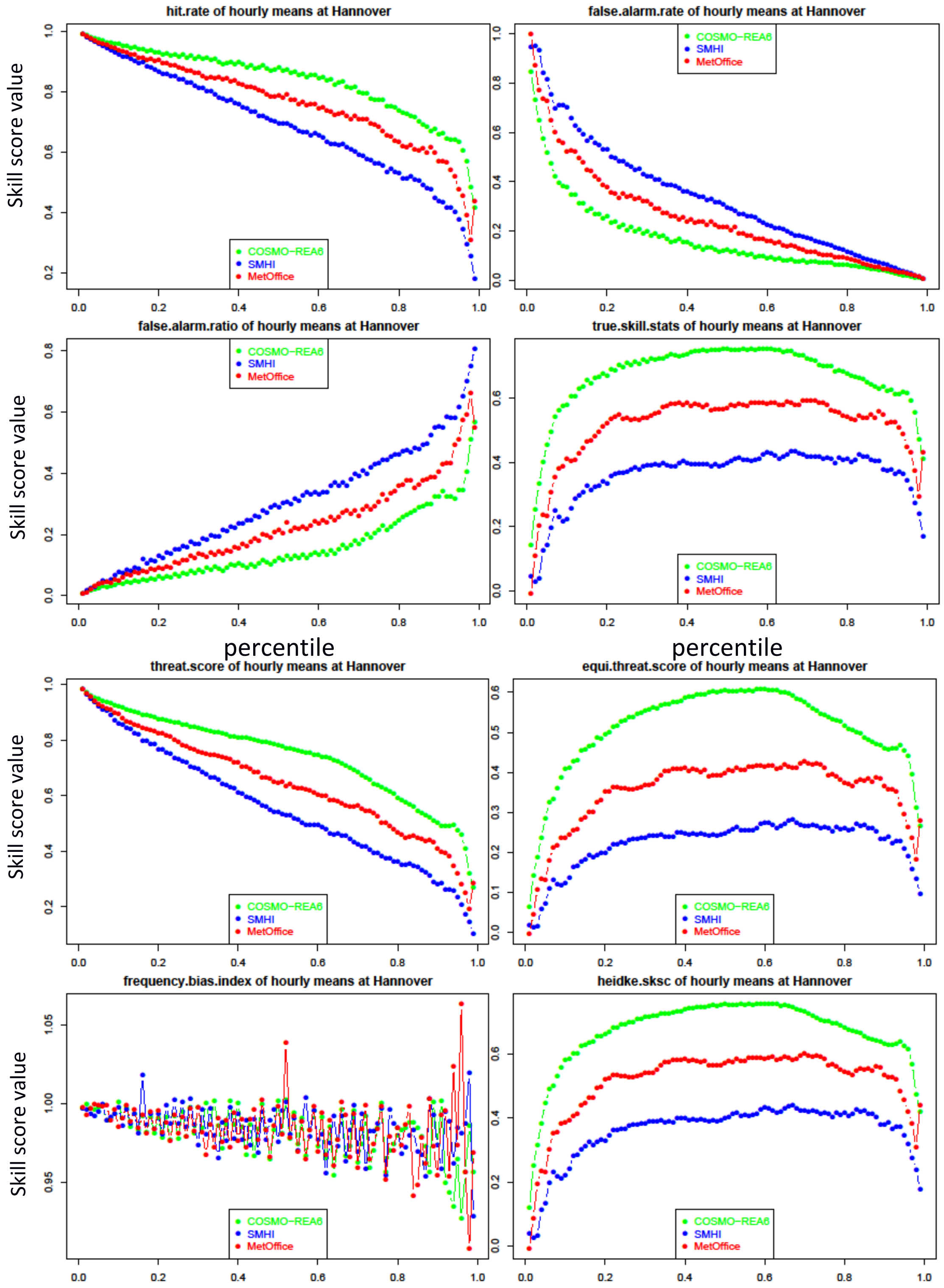


Figure 3.1: Frequency distribution at station Hannover (top left) and the three regional reanalyses at the location of the station for COSMO-REA6 (top right), EURO4M SMHI (bottom left), and EURO4M MetOffice (bottom right).

Skill scores were calculated for the three regional reanalyses as shown in Figure 3.2. Exemplarily, station Hannover was used and hourly mean data was chosen for the calculation. The output of COSMO-REA6 is hourly, whereas the output of the SMHI and MetOffice regional reanalyses is six-hourly only. So, for the calculation of the skill scores, every sixth hourly mean value of the station data was used to compare against that of the reanalyses.

The skill scores shown in Figure 3.2 feature different value ranges. Additionally, the value of perfect score varies throughout. For instance, for the hit rate the range of scores lies between (0,1) and the perfect score is 1. That is, the higher the score for each benchmark (here: percentile) the better the reanalysis performs. On the contrary, the opposite is true for the false alarm ratio: the range still lies between (0,1) but here the perfect score is 0, i.e., the lower the score for each benchmark the better the reanalysis. A third case is also possible, in which the score cannot distinguish between the different reanalyses and stays (almost) constant throughout the complete range of the benchmark as is true for the frequency bias index. Here, the perfect score is 1 and the value range of the score lies between (0, ∞).

For all skill scores shown, COSMO-REA6 performs best which might be due to its hourly output compared to the six-hourly output of the EURO4M SMHI and EURO4M MetOffice reanalyses.


hit.rate of hourly means at Hannover
false.alarm.rate of hourly means at Hannover
false.alarm.ratio of hourly means at Hannover
true.skill.stats of hourly means at Hannover
threat.score of hourly means at Hannover
equi.threat.score of hourly means at Hannover
frequency.bias.index of hourly means at Hannover
heidke.sksc of hourly means at Hannover
Skill score value
percentile
● COSMO-REA6
● SMHI
● MetOffice
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5
0.95
1.00
1.05

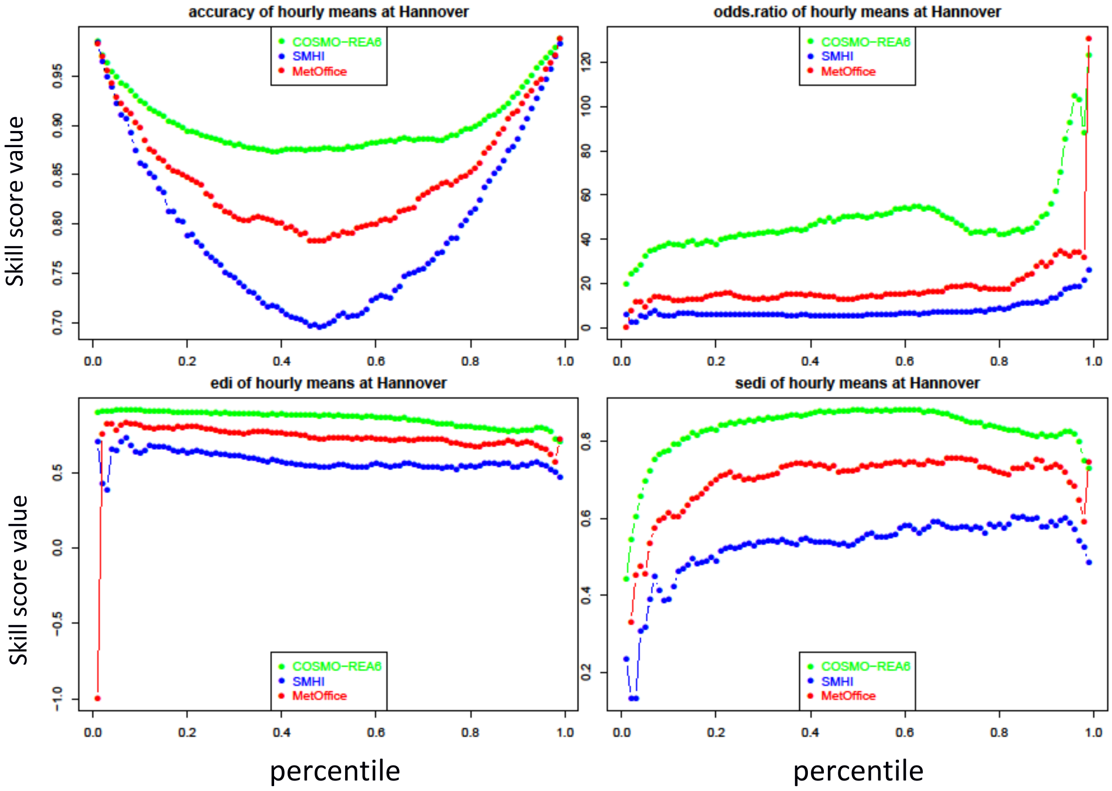


Figure 3.2: Twelve different skill scores of hourly means at station Hannover compared to the three regional reanalyses COSMO-REA6 (green), EURO4M SMHI (blue), and EURO4M MetOffice (red).

Figure 3.3 shows the combination of two skill scores, here the hit rate and false alarm ratio. The interpretation of this combination of scores is that a particular reanalysis has skill up to that value of the benchmark at which both lines of the hit rate and false alarm ratio cross each other. The hourly availability of the COSMO-REA6 regional reanalysis seems favourably for achieving a skilful analysis throughout most (97th percentile) of the measured value range, whereas for the EURO4M SMHI (85th percentile) and EURO4M MetOffice (90th percentile) reanalyses the skill does not cover such a wide range.

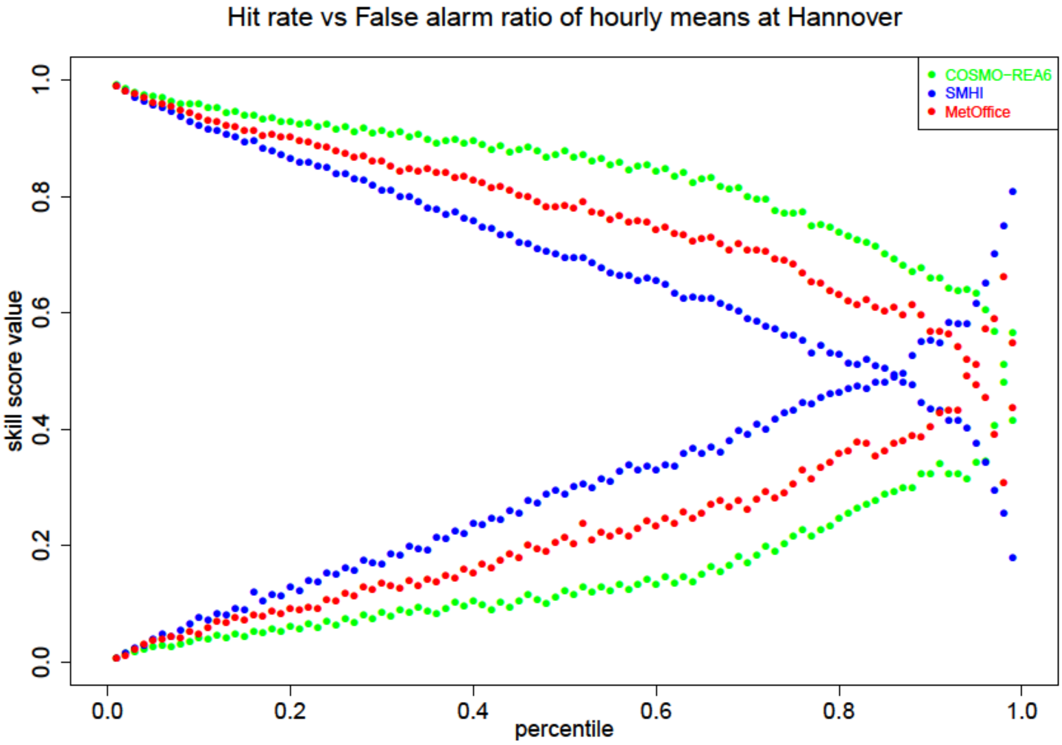


Figure 3.3: Hit rate versus False alarm ratio of hourly means at station Hannover for three different regional reanalyses.

Figure 3.4 shows the storm event Emma which officially lasted from February, 29th 2008 to March, 2nd 2008 and hit central Europe. In the three upper panels, regional reanalysis data for COSMO-REA6, EURO4M SMHI, and EURO4M MetOffice, respectively, together with station data are shown. Values are hourly means and station data is shown at the same time steps as available from reanalyses. In the legend, correlation values with the 95% confidence interval between station and reanalysis data are provided. The correlation between the COSMO-REA6 reanalysis and station measurements is significantly higher than the correlation between the two other reanalysis and station data.

Note all these results are only valid for the EURO4M and COSMO-REA6 output. This investigation has to be repeated with available UERRA output, before drawing any conclusions. The method’s potential though is demonstrated with this example.

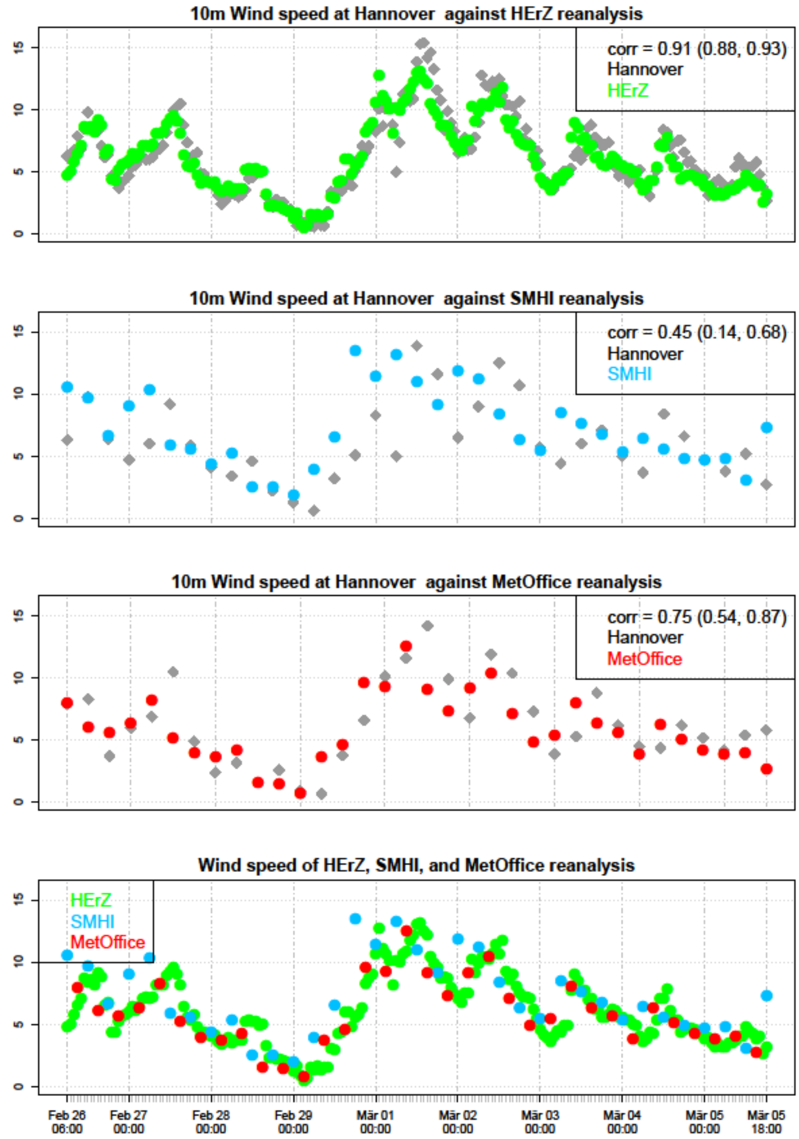


Figure 3.4: Time series of the storm event Emma between 00 hrs February, 26th 2008 and March, 06th 2008 for the regional reanalyses COSMO-REA6 (top panel), EURO4M SMHI (2nd panel), and EURO4M MetOffice (3rd panel) together with station data (grey) are shown. The bottom panel depicts all three regional reanalyses.

A recent study by Borsche et al., 2016 investigated wind speed from tall meteorological tower measurements and compared these measurements against regional (COSMO-REA6) and global (ERA20C and ERA-Interim) reanalyses. Below, the results of the Lindenberg tower are shown exemplarily. Figure 3.5 shows box plots of monthly wind speed at different heights for the mast measurements at Lindenberg and corresponding values of the regional (COSMO-REA6) and the global (ERA-Interim and ERA20C) reanalyses. The range of the box plot whiskers indicates 1.5 times the interquartile range. Median values of measurements and reanalyses are of comparable values throughout the height range, but the COSMO-REA6 reanalysis overestimates variability – given here as the range of the box plot whiskers – at Lindenberg. The wind speed increases with height as expected and median values between measurements and regional reanalyses are nearly the same. The variability at the

Lindenberg mast at 10m as derived from COSMO-REA6 is 10% lower than observed. However, in all heights above, the COSMO-REA6 variability is systematically larger than observed. The variability of the global reanalyses at 100m is also larger than observed.

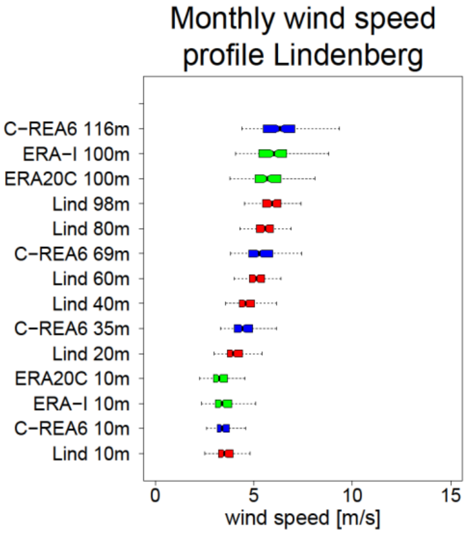


Figure 3.5: Box plot of monthly mean wind speed at Lindenberg at different heights between 10m and around 100m (Borsche et al., 2016). Mast measurements are shown in red, regional reanalysis data (COSMO-REA6) in blue, and global reanalysis data (ERA-Interim and ERA20C) in green.

A direct comparison of absolute values between mast measurements and COSMO-REA6 output on a specific height level is not recommended because these are influenced by biases which could be caused by insufficient representativity, mismatching heights, and mismatching surface roughness. This is especially true for comparisons over land where height mismatch might be large due to differences between model and real orography. Note further that surface roughness is kept constant with time in COSMO-REA6. For these reasons it is recommended to use anomalies when working with reanalysis data, as presented in Figure 3.6. Here, the time series of the reanalysis versus observed monthly wind speed anomalies is shown at 10m and around 100m for Lindenberg.

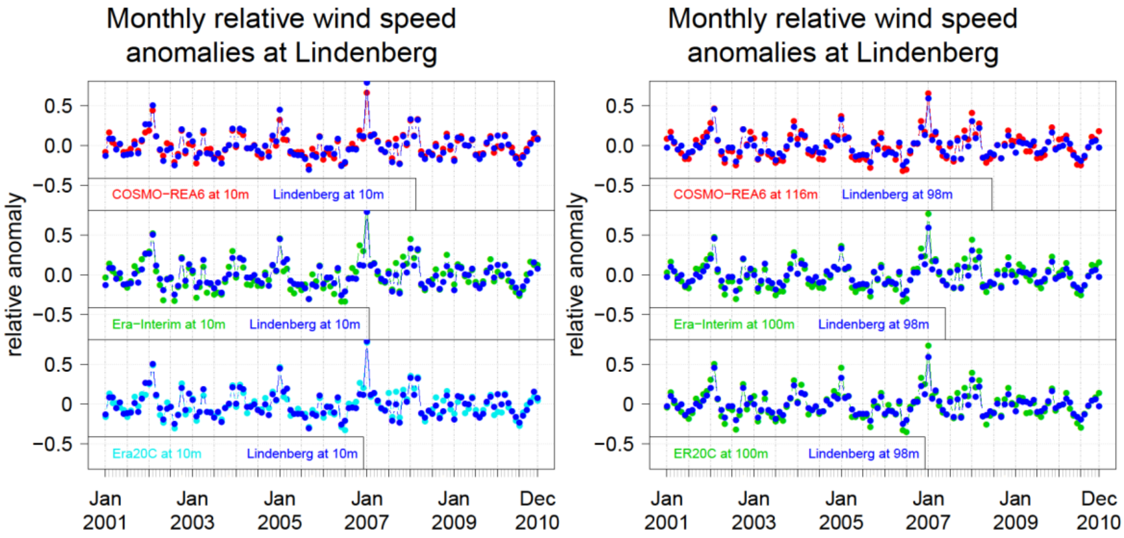


Figure 3.6: Time series of relative anomalies in monthly mean wind speed of Lindenberg mast measurements at 10m (left) and around 100m (right) height against regional reanalysis (COSMO-REA6) and global reanalyses (ERA-Interim and ERA20C) (Borsche et al, 2016).

Figure 3.7 shows the diurnal cycle of mean hourly wind speed at each Lindenberg measurement level and the COSMO-REA6 model level output. Measurements and reanalyses show a diurnal cycle with a maximum in the early afternoon for the lower levels up to 40m. From 60m height above ground onwards, the diurnal cycle degrades in the reanalysis. The observed reversal with height is not captured. Closer to the ground, COSMO-REA6 reproduces the temporal evolution of the diurnal cycle qualitatively very well, albeit with a lower amplitude. For instance, at 10m, the reanalysis captures the mean diurnal cycle of about 33% with respect to the minimum, whereas the mast measurements record about 50%. The relatively good match at the ground and mismatch above can be explained by the parametrizations of the boundary layer and the sub-grid scale orography which were particularly optimized with respect to the observed 10m wind speed 26 statistics (Schulz, 2008). The global reanalysis ERA20C with its three hourly output captures the diurnal cycle at 10m very well with an amplitude similar to the regional reanalysis. Also ERA-Interim with only six-hourly output indicates a diurnal cycle at 10m. However, at 100m, both global reanalyses show almost the same wind speed during the day in contrast to what the measurements show.

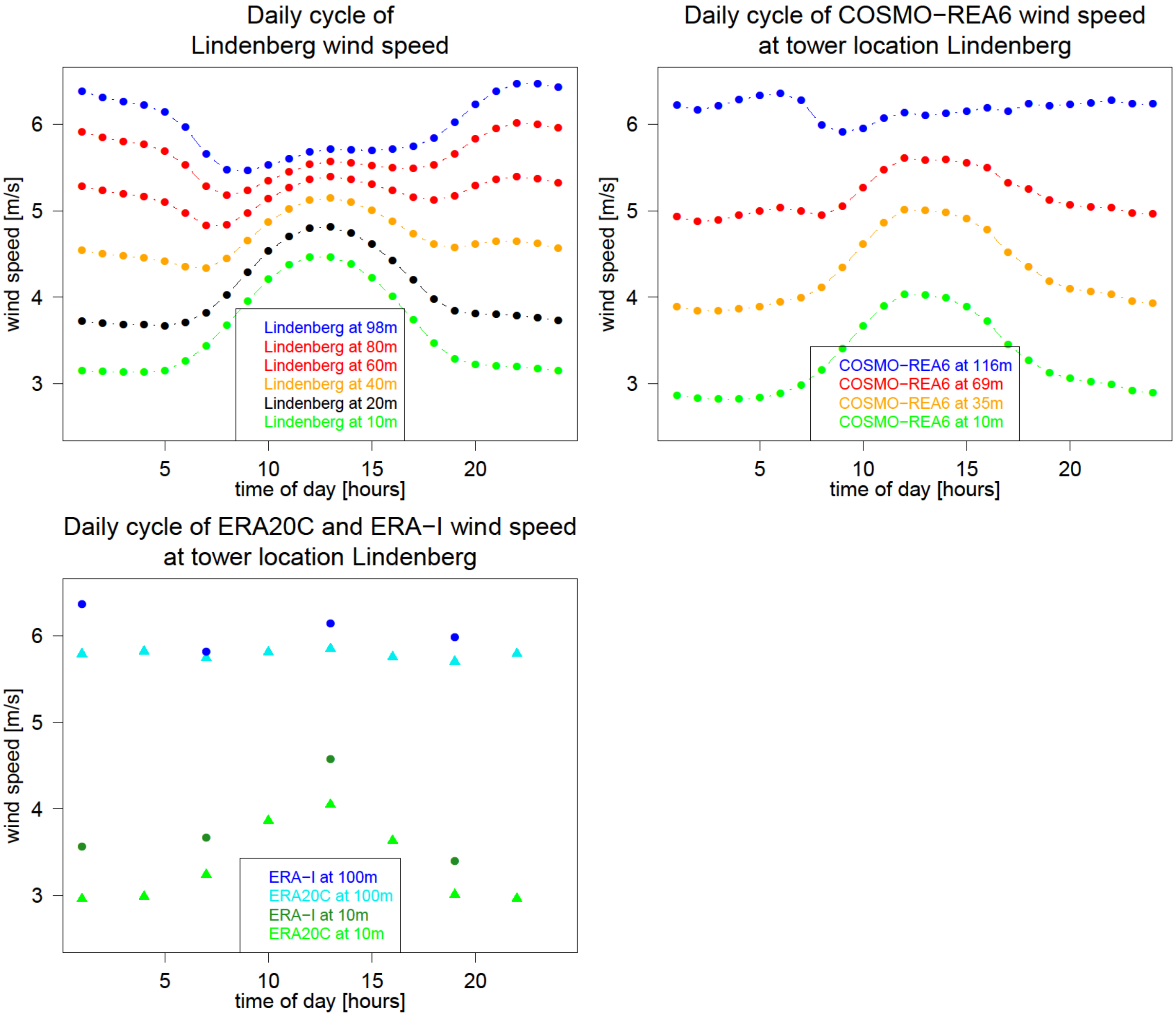


Figure 3.7: Diurnal cycle of wind speed at the location of Lindenberg. Mast measurements are shown at the top left, COSMO-REA6 values at the top right, and global reanalyses in 10m and 100m height at the bottom (Borsche et al., 2016).

4. Method C: Comparison against gridded station observations

4.1. Method description

Two approaches are used. One focuses on the whole of Europe (by KNMI and CRU), the other is targeted at a sub-region within Europe (by MetNo). For the pan-European approach, gridded observational data based on a dense network of stations covering Europe is used to assess reanalysis results for their similarity. The comparison is made by either aggregating results of the observational dataset and the reanalysis in space and time, providing monthly means of daily averaged temperature for Europe. Using estimates of the uncertainty associated with the gridding of observed data, the uncertainty in European-averaged temperature related to the use of inhomogeneous station data and an estimate of the Urban Heat Island-related warming in Europe, an uncertainty estimate on the observational data is produced. This uncertainty is used in the comparison against reanalysis data. Another comparison is made by comparing the replication of trends in extremes of surface temperature across Europe between reanalysis and observations.

For the comparison over a smaller part of Europe, a scale-separation spatial verification method similar to the Intensity-Scale Technique (*Casati*, 2010) has been applied to compare the reanalysis (or hindcast) surface fields against observational gridded datasets. Reanalyses and reference precipitation fields are decomposed into the sum of orthogonal wavelet components each characterized by a different spatial scale. The scale-dependence of the bias and the capability of the forecast to reproduce the observed scale structure are then assessed by comparing the wavelet component power spectrum. The scale-separation verification can be applied both to original or precipitation values truncated at a threshold: the latter enables to focus on low versus high precipitation intensities, and bridges the scale-separation verification to traditional categorical scores.

4.1.1. Advantages

The aggregation of data over the European domain and over time provides one simple index which can be used as an easy-to-interpret metric for the overall quality of the reanalysis. Contrasting with this is the more elaborate method based on wavelets, which has as advantage that it provides specific information on which parts of the wavelet component power spectrum are reproduced and which parts not. In this sense the two methods are complementary. The focus on the replication of trends in extreme temperatures offers the advantage that it specifically targets a known weakness in the reanalysis.

4.1.2. Disadvantages

The spatial aggregation obscures possible local problems. Furthermore the comparison may show a good resemblance between gridded observations and reanalysis for the ‘wrong’ reason due to cancellation of deviations. The disadvantage of the wavelet-based comparison is that it is a complex method and results are more difficult to interpret.

4.1.3. Value of method

The two comparison methods are complementary; one is simple and effective but lacks (spatial) detail; the other is complex but highlights differences in scale structure between reanalysis and observations.

4.2. Developed and shared code

The procedures implementing the evaluation are currently under development by using the R-language (Team, R. Core. "R: A language and environment for statistical computing." (2013): 409) and it will be shared on github (project name: UERRA-EVA, is licensed under GPL version 2 or any later version).

4.3. Example of Application

4.3.1. Parameter

Daily averaged temperature and daily precipitation sums

4.3.2. Investigated spatial and temporal scale

Daily temperatures aggregated to the monthly level and aggregated (spatially) to a European average for one method.

4.3.3. Used observations

E-OBS, seNorge2

4.3.4. Investigated reanalyses

ERA Interim, ERA-40 and 20th Century Reanalysis, NORA10, AROME-MetCoOp

4.4. Preliminary Results

The annual averaged EurAvg temperature is compared to European averages of 2 m temperatures as calculated in popular reanalysis data sets. These are the 20th Century Reanalysis [*Compo et al.*, 2011], which starts in January 1871 and is based on a reanalysis using surface pressure data only. The spatial resolution of this reanalysis data set is 2° × 2°. The ERA-40 reanalysis [*Uppala et al.*, 2005] spans the period September 1957 to August 2002 at a spatial resolution of 1.5° × 1.5° and the ERA-Interim reanalysis [*Dee et al.*, 2011] spans the period January 1979 to December 2011 with 0.7° × 0.7° resolution.

The comparison shows that the European averaged temperature and the reanalysis temperatures generally agree, with the exception of 1992 for which all three reanalysis temperatures are below the E-OBS uncertainty estimate. Interesting is that 1989 is the warmest year in the 20th Century Reanalysis (outside the E-OBS uncertainty) and that from 1994 onward, the temperatures with respect to the 1961–1990 climatology in this reanalysis have been structurally lower than those in E-OBS and the other reanalysis products. In fact, the difference increases in time and even falls outside the E-OBS uncertainty from 1994 onward with the exception of only 2 years. Apparently, the warming over Europe evident in E-OBS and the other reanalysis products over the last decade fails to be captured in the 20th Century Reanalysis.

A comparison between monthly values of the E-OBS data and reanalysis data from ERA40 and ERA-Interim shows that E-OBS deviates most strongly from the reanalysis data in June 1992 (with the reanalysis data being colder). A relatively large area with large differences is found in the desert of Algeria where temperatures in E-OBS are 1°C−2°C higher. Next to this larger region, clusters of a few grid squares with differences of up to 5°C are found south of the Alps, between the Black and Caspian Sea, in Iraq, Israel, and southern Iceland. Given the location of most differences, data scarcity is suspected as the reason for the deviation. Note that E-OBS deviates less strongly from the global data sets in 1992 than it does for the reanalysis data. Apparently, the reanalysis data sets produce a temperature anomaly for June 1992, which is dynamically consistent with the data sets assimilated in the models, but which are not evident in the in situ data sets.

A difference in trends between the 20th Century Reanalysis and E-OBS over the period 1979–2008 is observed, with the largest differences in central and eastern Europe and northern Scandinavia. A provisional explanation for the divergence in estimates of the warming in Eastern Europe is related to the use of observed Sea Surface Temperatures (SST) as boundary conditions for the 20th Century Reanalysis and the use of pressure observations only in the assimilation scheme. Apparently, the further away from the sea and ocean, the less the warming signal present in the SSTs impacts on the surface temperatures. The absence of any land-based temperature measurements in the assimilation scheme of the 20th Century Reanalysis means that no temperature adjustment is made. In the comparison of trends in extreme temperature, the period 1980–2011 is considered and both station series and the gridded E-OBS data are used to determine the success of the ERA Interim reanalysis, with indices of the numbers of days above or respectively below the 90th and 10th percentiles of daily maximum and minimum temperature used as metrics. It is observed that the ERA-Interim reanalysis data are generally very good at replicating both the seasonally and spatially varying trends in the indices across Europe. At the station level, the reanalysis data are also able to depict the observed trends remarkably well. However, the success of the reanalysis data depends on the season and the particular index considered. The reanalysis is least successful in replicating trends in the number of days exceeding the 90th percentile of maximum temperature, particularly during the summer season. The success of the reanalysis is also somewhat dependent on the time step of the reanalysis data used. Daily maximum and minimum temperatures calculated from the 3-hourly time step reanalysis data tend to be more reliable than those derived from the 12-hourly data.The scale-separation assessment has been tested on numerical model output datasets available within MET Norway and covering the Norwegian mainland, such as NORA10 (atmospheric downscaling based on ERA40 from 1957 to 2002 and on ECMWF operational analyses from 2002 onwards, 10 Km grid spacing) and the high-resolution forecasting numerical weather prediction model AROME-MetCoOp (2.5 Km grid spacing). As a reference, the seNorge2 observational gridded datasets have been used. The objectives were: regarding NORA10 our interest is in the description of the precipitation climatology; while for AROME-MetCoOp, we assess the added value of the enhanced resolution.

5. Method D: Comparison against satellite data

5.1. Method description

Reanalysis fields of global radiation are directly compared against satellite data of the EUMETSAT Satellite Application Facility on Climate Change (CM SAF). Both data sets are available for the CORDEX-EU domain, whereas the far northern parts of this domain are not covered by the satellite data during wintertime. The satellite data is provided on a regular longitude latitude grid of 0.05° spatial resolution. In order to facilitate a fair comparison, the reanalysis and satellite data need to be re-projected onto the same grid and the same spatial resolution, which is determined by the coarser native resolution of the two data sets. The temporal resolution of the satellite data ranges from 30min instantaneous measurements to aggregated hourly, daily, and monthly values. Also for the temporal resolution, a fair comparison can only be performed at the same resolution which is, again, determined by the coarser native resolution of the two data sets.

Comparison is performed on the complete CORDEX-EU domain, as well as on selected land areas over Germany and the Iberian Peninsula. Relative and absolute differences as well as frequency distributions and scatter plots are calculated on the annual, monthly, and daily scale. These measures enable to investigate the spatial and temporal distribution of agreement and disagreement between the two data sets. The scatter plots also allow for the determination of correlation and bias between the reanalysis and satellite measurements.

5.1.1. Advantages

This method allows for comparison of reanalysis data spatially over the complete domain against independent and spatially homogeneous measurements, which have undergone a thorough quality check and qualify as a climate data record. The satellite data are provided in a high spatial and temporal resolution which matches or even exceeds that of available regional reanalyses.

5.1.2. Disadvantages

The quality of the reference satellite data is not of equal quality throughout the domain depending on the ability of the retrieval to generate radiation estimates over different surfaces, i.e., snow covered regions, mountainous regions, different land covers, the ocean. For instance, it is known that the satellite data set is not the best estimate for snow covered (Trentmann, personal communication, 2016) and mountainous regions (Buffat and Grassi, 2015).

5.1.3. Value of method

Allows for evaluation against independent reference data over a large domain.

5.1.4. Developed and Shared code

The procedures implementing the evaluation are currently under development by using the R-language (R. Core Team (2105). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/) and will be shared on github (project name: UERRA-EVA, is licensed under GPL version 2 or any later version).

5.2. Example of Application

5.2.1. Parameter

The method was applied to global radiation.

5.2.2. Investigated spatial and temporal scale

The evaluation was performed on a common regular grid of 0.1° spatial resolution covering the CORDEX-EU domain for annual, monthly, and daily means for the year 2008.

5.2.3. Used observations

The CM SAF satellite data of the parameter surface incoming solar (SIS) radiation are based on the Meteosat Second (for the year 2008) Generation (MSG) Spinning Enhanced Visible and Infrared Imager (SEVIRI).

5.2.4. Investigated reanalyses

The comparisons are performed with the regional reanalysis COSMO-REA6 (produced by Uni Bonn together with DWD) and preliminary UERRA data set HARMONIE (produced by SMHI).

5.3. Preliminary Results

Global radiation of the regional reanalyses COSMO-REA6 (Bollmeyer et al, 2014) and HARMONIE as developed by SMHI within WP2 of this project are evaluated against CM SAF SIS data for the year 2008 covering the complete CORDEX-EU domain. On an annual mean, relative differences as shown in Figure 5.1 depict an overall good agreement. For COSMO-REA6 in general there is a negative bias throughout. Only the mountainous regions, especially over the Alps, there is a strong positive bias which is probably due to the underestimation of radiation of the CM SAF SIS data over these regions (Trentmann, personal communication, 2016; Buffat and Grassi, 2015). For HARMONIE, there are more heterogeneous features apparent over the domain: Over most of the land area the difference to the satellite data is within ±5%. There is also a strong positive bias over mountainous regions. There is a negative bias of around -10% over the Mediterranean and the covered parts of Fennoscandia. And there is a positive bias of around 10% to 20% over the British Isles and most parts of the North Atlantic.

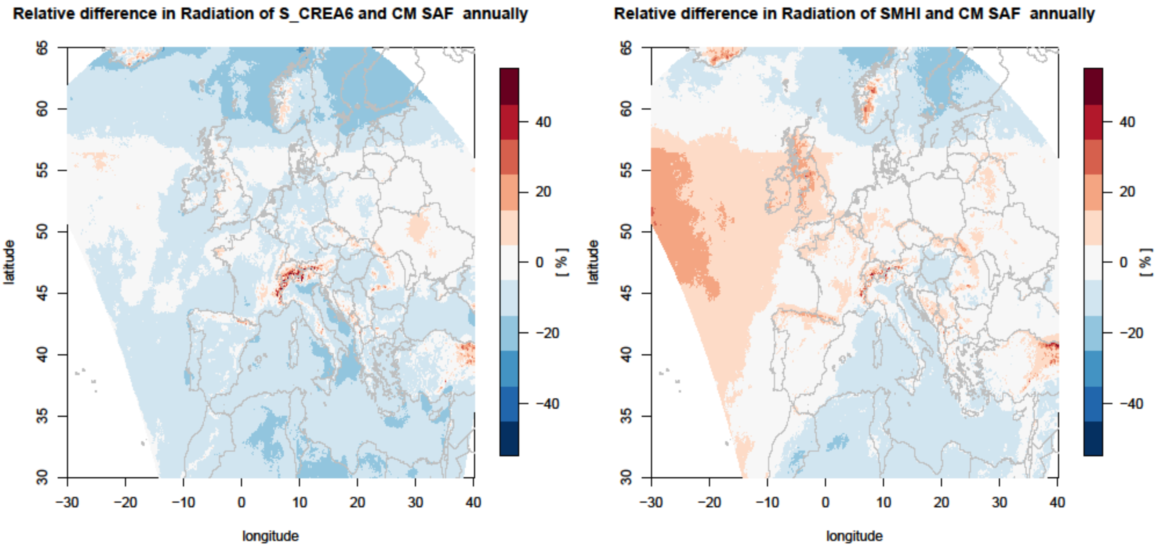

Relative difference in Radiation of S_CREA6 and CM SAF annually
Relative difference in Radiation of SMHI and CM SAF annually
latitude
65
60
55
50
45
40
35
30
longitude
−30
−20
−10
0
10
20
30
40
[%]
−40
−20
0
20
40

Figure 5.1: Relative difference of the annual mean 2008 between COSMO-REA6 and CM SAF SIS (left) and between HARMONIE and CM SAF SIS (right).

Figure 5.2 shows the frequency distribution of the annual mean for COSMO-REA6, HARMONIE, and the CM SAF SIS data. For the COSMO-REA6 values, an overestimation of low radiation values is apparent as well as an underestimation of high radiation values. For the HARMONIE regional reanalysis there is a close fit of low radiation values but also a slight underestimation of high radiation values. Figure 5.3 shows the relative differences of monthly means for July 2008. Compared to the annual features of Figure 5.1, the monthly differences exhibit more heterogeneous features but the overall patterns remain. For COSMO-REA6, there is still an overall negative bias but the variability of the difference is higher. For HARMONIE, the distinct overestimation over the North Atlantic reaches values as high as 50%. Additionally, there is a strong underestimation over the Atlas Mountains of up to -40%. The hexagon plots as shown in Figure 5.4 are a special scatter plot which provides the amount of values within a given hexagon. Additionally, the correlation between the regional reanalysis and the CM SAF SIS values can be calculated. Figure 5.4 reveals that for COSMO-REA6 most of the radiation values are closely aligned around the regression line and that the negative bias is not constant. There are two areas of clustering of the radiation values which are located at around 200W/m^2 and 300W/m^2 representing the morning and afternoon as well as the midday values, respectively. For HARMONIE, a clear cut-off at 200W/m^2 can be seen which mirrors the overestimation over the North Atlantic and may be caused by issues of the cloud scheme within the HARMONIE model.

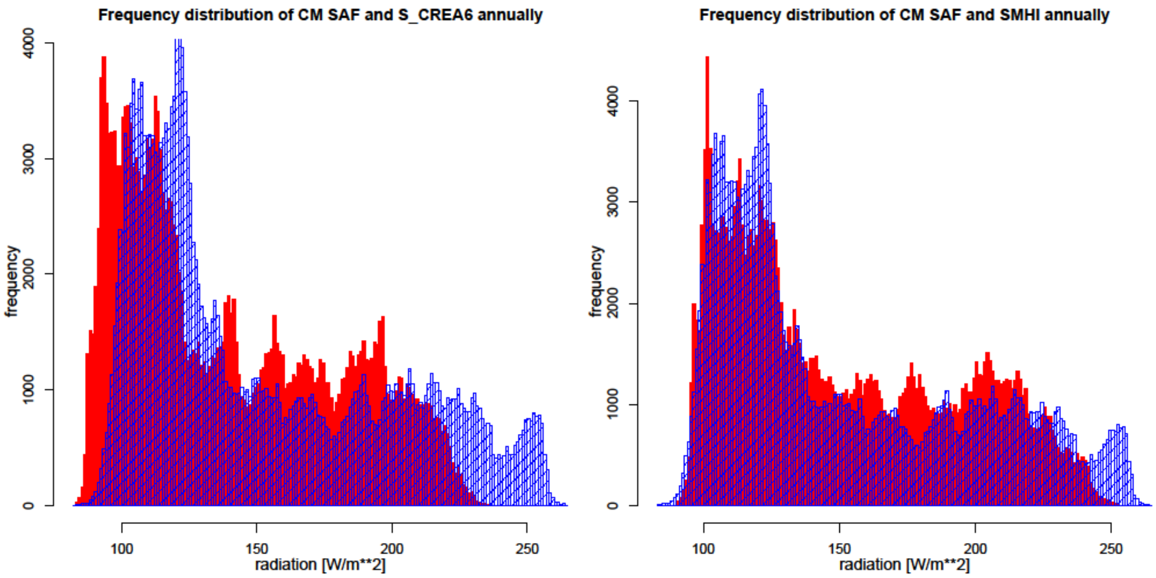


Figure 5.2: Frequency distribution of annual mean radiation values between COSMO-REA6 (red) and CM SAF SIS (blue) on the left side and between UERRA HARMONIE (red) and CM SF SIS (blue) on the right side.

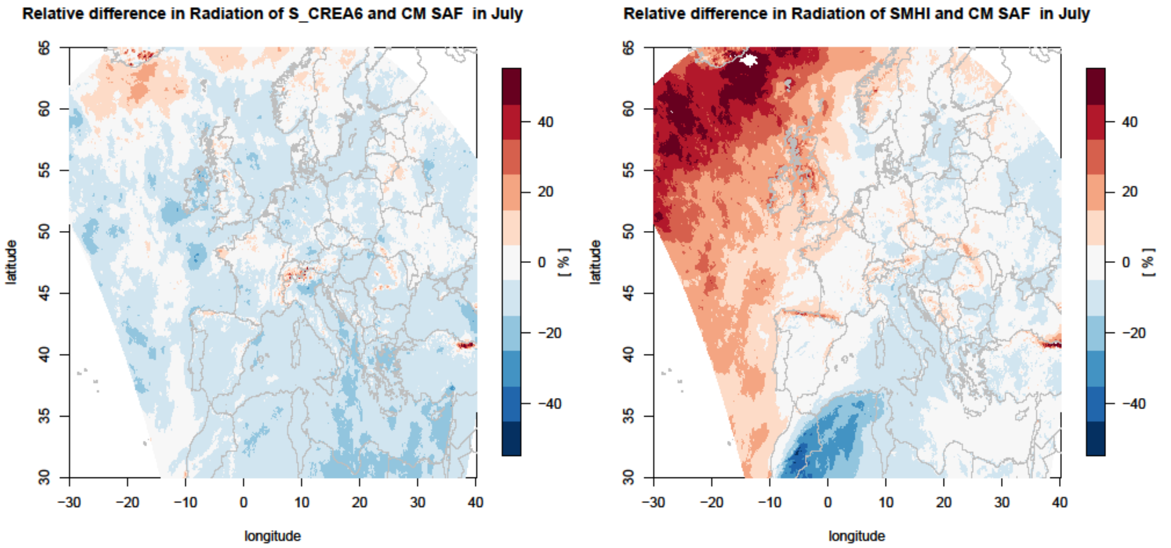

Relative difference in Radiation of S_CREA6 and CM SAF in July
Relative difference in Radiation of SMHI and CM SAF in July
latitude
30
35
40
45
50
55
60
65
longitude
−30
−20
−10
0
10
20
30
40
[%]
−40
−20
0
20
40

Figure 5.3: Relative difference of the monthly mean of July 2008 between COSMO-REA6 and CM SAF SIS (left) and between UERRA HARMONIE and CM SAF SIS (right).

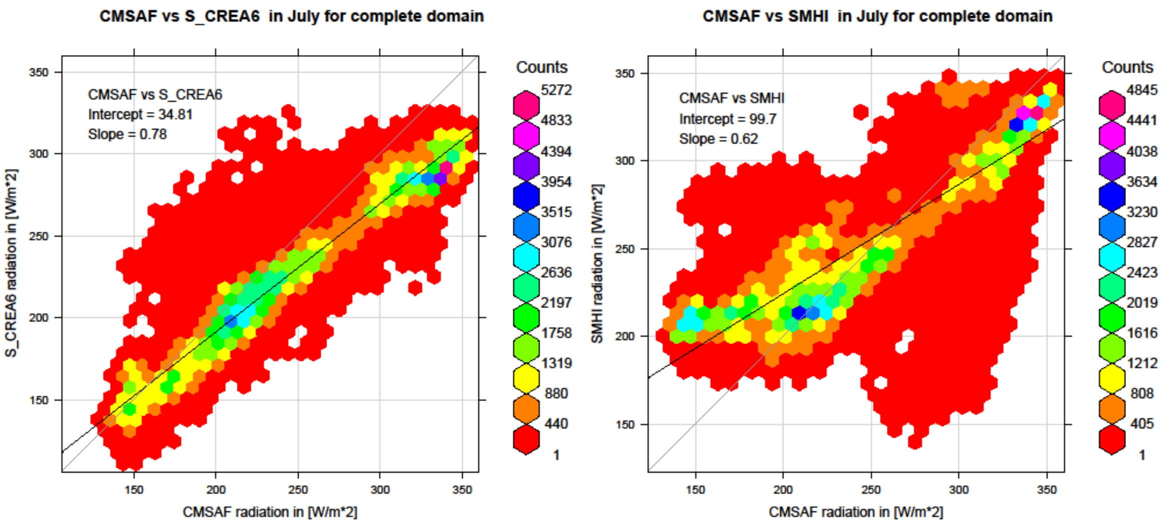


Figure 5.4: Hexagon plot of monthly means in July 2008 between COSMO-REA6 and CM SAF SIS (left) and between UERRA HARMONIE and CM SAF SIS (right).

The annual cycle as reproduced by daily mean radiation values is depicted in Figure 5.5. Daily means were calculated for area means over two land areas for Germany and the Iberian Peninsula. The daily means are shown for CM SAF SIS in blue, UERRA HARMONIE in green, and COSMO-REA6 in red for the year 2008. The correlation of UERRA HARMONIE and COSMO-REA6 daily means as compared to CM SAF SIS is very high with values between 0.98 and 0.99 for both regions.

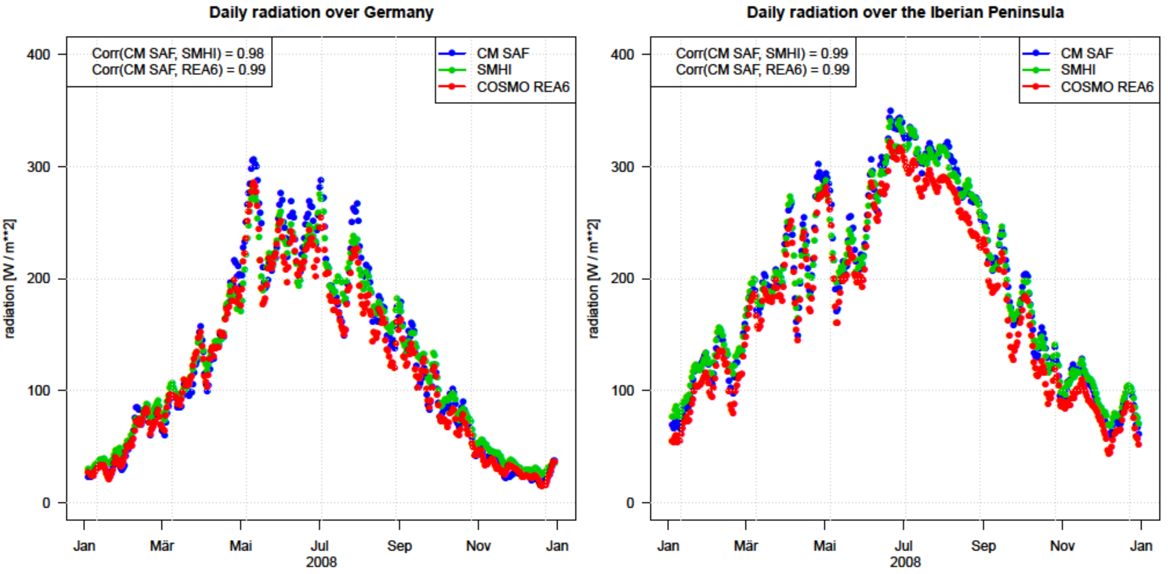


Figure 5.5: Annual cycle in 2008 of daily means for the area mean over Germany (left) and the Iberian Peninsula (right) for CM SF SIS (blue), UERRA HARMONIE (green), and COSMO-REA6 (red).

6. Method E: Ensemble based methods

6.1. Purpose

An ensemble system of regional reanalysis, such as the one developed in UERRA, provides predictions that inform users not only about the most likely state of the atmosphere but also about the level of uncertainty of this prediction. The ensemble mean should be a more accurate estimate of the atmospheric state than the one provided by a deterministic system and the ensemble spread estimates uncertainty in the ensemble mean.

In order for an ensemble reanalysis to be useful in applications, the predictions themselves and the pertinent uncertainty ranges need to be in a balance (consistent). An ensemble (or probabilistic) prediction that is consistent is denoted as “reliable”. If there are several fully reliable ensemble predictions the one with the smaller uncertainty range on average (commonly quantified by sharpness) is more useful in applications. The purpose of ensemble-based verification is to test the reliability of ensemble reanalyses and to comparatively assess which of those that are reliable exhibit higher sharpness.

Ensemble reanalyses offer a fundamentally different usage of climate data in applications, because they allow to trace uncertainties thoroughly to the end result. The success of this promising procedure is less sensitive to systematic and random errors, which is the primary focus of deterministic evaluation, but on the consistency of the ensemble spread with these errors. Ensemble verification not only informs users about classical error components, but also about how literally he/she can take ensemble spread as the range within which the truth is. This knowledge will change the way users deal with uncertainties of the ensemble system.

6.2. Method Description

Empirical evaluation of ensemble reanalyses is not fundamentally different from traditional comparisons for deterministic reanalyses. Specific extreme events, long-term averages, climate indices, etc. can be compared to the observed analyses, yet the results of an ensemble reanalysis have uncertainty ranges attached to them. The magnitude of discrepancy can then be assessed against these ranges. A considerable part of the ensemble evaluation in UERRA is following simple extensions of classical deterministic evaluation, with the advantage of results being directly comparable between deterministic and ensemble reanalyses.

A more formal framework for evaluating ensemble predictions (here reanalyses) is provided by the formalisms of “ensemble forecast evaluation” and “probabilistic forecast evaluation” (Jolliffe and Stephenson, 2012). It compares the ensemble against deterministic observations and uses a number of graphical diagnostics (e.g. reliability diagrams, Talagrand histograms, relative-operating characteristics (ROC) curves) and numerical summary measures (e.g. ranked probability skill-score (RPSS), continuous ranked probability skill-score (CRPSS), Brier Score, ROC curve areas). These describe the nature of deficiencies of the ensemble system in detail with separate contributions from (conditional) biases, the reliability of ensemble spread and the sharpness. Examples of such a formal comparison will be provided in UERRA.

As a source of many verification references see the web page maintained by the WMO Joint Working Group on Forecast Verification Research (JWGFV) http://www.cawcr.gov.au/projects/verification/

A special case arises if the observations themselves are subject to uncertainty in which case these should be formally accounted for in the comparison. The problem is mathematically complex but extensions of some of classical probabilistic forecast evaluation have been made to deal with this complication. Of particular mention is the extension of the Brier Score (Candille and Talagrand), which is one of the procedures that will be utilized in UERRA.

The evaluation of ensemble system of regional reanalysis will start once the first datasets will be available, which is planned for the end of 2016. Currently, the evaluation procedures are under development by the UERRA WP3 partners. In particular, the evaluation will focus on total precipitation and two-meter temperature.

6.3. Developed and shared code

The procedures implementing the verification statistics, scores and skill-scores of an ensemble system of regional reanalysis against observational gridded datasets are currently under development by using the R-language (Team, R. Core. "R: A language and environment for statistical computing." (2013): 409) and it will be shared on github (project name: UERRA-EVA, is licensed under GPL version 2 or any later version).

6.4. Example of application

The added value of considering the observation uncertainty estimates in our evaluation is shown in Figures 6.1 and 6.2 where an Ensemble version of the pan-Alpine high-resolution grid dataset of daily precipitation (APGD_Ens) is used as a reference for the comparison with the daily precipitation from HIRLAM 3D-VAR (Dahlgren and Gustafsson 2012) and the E-OBS observational gridded dataset (Haylock *et al.* 2008). Both the HIRLAM and the E-OBS datasets presented here have been obtained within the EURO4m EU-FP7 project (see www.euro4m.eu). The APGD (Frei and Schär 1998, Isotta *et al.* 2014) has also been developed as part of EURO4m and the APGD Ensemble version is currently under development.

In Figure 6.1, the APGD_Ens estimates of the 95% quantile of average daily precipitation in 2008 for three different catchment areas in Switzerland is shown as a boxplot summarizing the underlying distribution of values: the bottom and top of each box are the 25% and 75% quantiles (of the 95% quantile of average daily precipitation), the line near the middle of the box is the median, the ends of the whiskers represent the distribution tails and the black dots marks outliers. The estimates for E-OBS and the HIRLAM model are shown on the same graph. The availability of an estimate for the uncertainty in the APGD_Ens makes evident the significance of the difference between the estimates under evaluation and the reference.

The same concept is illustrated in Figure 6.2 for the time series of daily precipitation in a catchment area in Ticino (Switzerland).

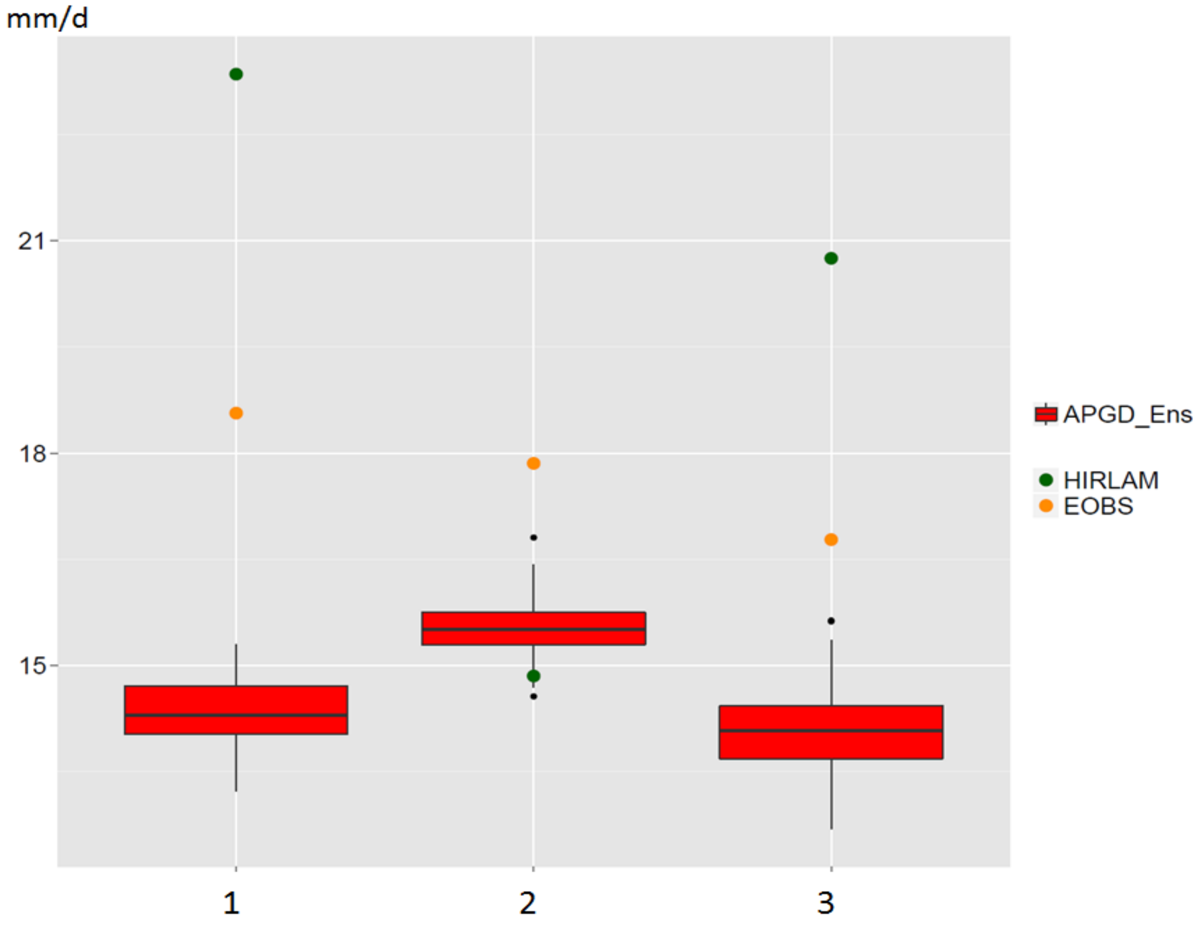


Figure 6.1: 95% quantile for three different catchment areas (1: Dora Bàltea, Aosta Valley, 2: part of the Aare river, Switzerland, 3: part of the Salzach, Austria) for the year 2008. The boxplots correspond to the probabilistic observational dataset, the green dots to the HIRLAM model (EURO4M) and the orange dots to E-Obs.

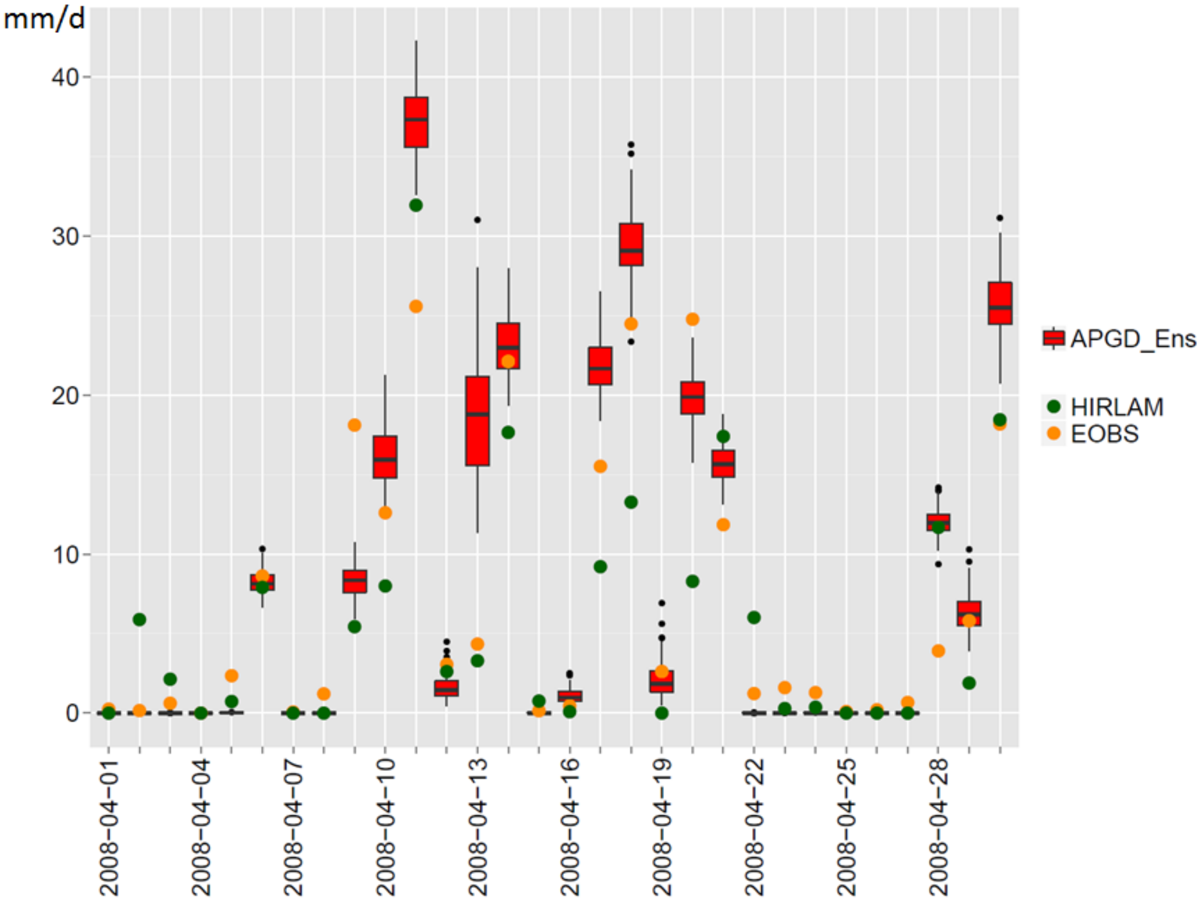


Figure 6.2: Daily precipitation in April 2008 in a part of the Ticino catchment area. The boxplots correspond to the probabilistic observational dataset, the green dots to the HIRLAM model (EURO4M) and the orange dots to E-Obs.

7. Summary

This report lists five methods to evaluate regional reanalyses. These include the evaluation based on feedback statistics (Method A), comparison against station observations (Method B), comparison against gridded station observations (Method C), comparison against satellite data (Method D), and an ensemble based comparison (Method E).

Each of these methods has its own advantages, disadvantages and value for users and producers of regional reanalysis. Method A is particularly valuable for producers and users who wish a comparison against conventional data. Method B is easy to perform, but harder to interpret scientifically, still user friendly. Method C is desirable for users who already use the gridded product, but again harder to interpret. Method D is limited to parameters provided by satellite retrievals. Probabilistic approaches to evaluation (Method E) become essential for the assessment of ensemble reanalyses. They inform users about aspects of data reliability that deterministic evaluation does not cover, notably the reliability of uncertainties conveyed by the ensemble. For a professional quantitative application of ensemble reanalyses this knowledge is indispensable. Probabilistic evaluation can be considered a generic concept that can be adopted in comparisons with feedback statistics, station data, grid data and satellite data.

No single method alone will be able to give a concise characterization of the complexity of uncertainty of regional reanalyses. Much can be learned, for both reanalyses and reference data sets, through a variety of methods complementing each other.

Note the shown results are only valid for preliminary data (used in Method A), or EURO4M data (used in Method B) and COSMO-REA6 output (used in Method B). This investigation has to be repeated with available UERRA output before drawing any conclusions on the current UERRA system components.

Here it is demonstrated that there is considerable potential in applying methods A,B,C,D, and E, preferably in combination or complementing each other, for purpose of both scientific intercomparison and user friendly uncertainty estimates of regional reanalyses.

8. References

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. R. Meteorol. Soc., 141, 1–15, doi: 10.1002/qj.2486, 2015.

Borsche, M., Kaiser-Weiss, A.K., and Kaspar, F.: Wind speed variability between 10 and 116 m height from the regional reanalysis COSMO-REA6 compared to wind mast measurements over Northern Germany and the Netherlands. Adv. Sci. Res., 13, 151–161, doi: 10.5194/asr-13-151-2016, 2016.

Buffat, R. and Grassi, S.: Validation of CM SAF SARAH solar radiation datasets for Switzerland, IEEE Xplore, doi: 10.1109/IRSEC.2015.7455044, 2015.

Candille, G., and Talagrand, O.: Impact of observational error on the validation of ensemble prediction systems. Q. J. R. Meteorol. Soc., 134, 959-971, 2008.

Casati, B.: New developments of the intensity-scale technique within the Spatial Verification Methods Intercomparison Project. Weather and Forecasting, 25(1), 113-143, 2008.

Compo, G. P., et al.: The Twentieth Century Reanalysis Project, Q. *J. R. Meteorol. Soc.*, *137*, *1–28*, doi: 10.1002/qj.776, 2008.

Dahlgren, P. and Gustafsson, N.: Assimilating Host Model Information into a Limited Area Model. Tellus A, 64, 15836, doi: 10.3402/tellusa.v64i0.15836, 2012.

Dee, D. P. *et al.:* The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc., 137, 553–597, doi:10.1002/qj.828, 2011.

Ferro, C.A.T and Stephenson, D.B.: Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events. doi: 10.1175/WAF-D-10-05030.1, 2011.

Frei, C., and Schär, C.: A precipitation climatology of the Alps from high-resolution rain-gauge observations. International Journal of climatology, 18(8), 873-900, 1998.

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

Isotta, F. A., Frei, C., Weilguni, V., Perčec Tadić, 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(5), 1657-1675, doi: 10.1002/joc.3794, 2014.

Jolliffe, I. T., and Stephenson, D. B.: Forecast Verification: A Practitioner's Guide in Atmospheric Science, Second Edition, doi: 10.1002/9781119960003.ch1, 2012.

Uppala, S. et al.: The ERA-40 re-analysis, Q. J. R. Meteorol. Soc., 131, 2961–3012, doi:10.1256/qj.04.176, 2005.

Schulz, J.-P.: Revision of the Turbulent Gust Diagnostics in the COSMO Model, COSMO Newsletter, 8, 17-22, 2008.