Seventh Framework Programme
Theme 6 [SPACE]

# Project: 607193 UERRA

Full project title:
**U**ncertainties in **E**nsembles of **R**egional **R**e-**A**nalyses

# Deliverable D3.6: Preliminary report of assessment of regional reanalyses – first results

| | |
|---|---|
| WP no: | 3 |
| WP leader: | DWD |
| Lead beneficiary for deliverable : | DWD |
| Name of author/contributors: | Deborah Niermann, Michael Borsche,<br>Andrea Kaiser-Weiss (DWD)<br>Else van den Besselaar, Gerard van der Schrier, Richard Cornes, Ernst de Vreede (KNMI)<br>Cristian Lussana, Ole Einar Tveito, Luca Cantarello (MI)<br>Christoph Frei, Francesco Isotta (EDI)<br>Jemma Davie (MO) |
| Nature: | Other |

| | |
|---|---|
| Dissemination level: | PU |
| Deliverable month: | 45 |
| Submission date: | November, 2017 |

# Report for Deliverable 3.6 (D3.6): Scientific report on assessment of regional analysis against independent data sets

**By Deborah Niermann[1], Michael Borsche[1], Else van den Besselaar[3], Cristian Lussana[2], Francesco Isotta[5], Christoph Frei[5], Andrea Kaiser-Weiss[1], Luca Cantarello[2], Ole Einar Tveito[2], Gerard van der Schrier[3], Richard Cornes[3], Ernst de Vreede[3] and Jemma Davie[4]**

[1] Deutscher Wetterdienst (DWD), Offenbach, Germany
[2] Norwegian Meteorological Institute (MI), Oslo, Norway
[3] Royal Netherlands Meteorological Institute (KNMI), de Bilt, Netherlands
[4] Met Office (MO), Exeter, United Kingdom
[5] Eidgenössisches Departement des Inneren (EDI), Switzerland

# Content

1. Scope of this document ..... 5
2. Method A: Feedback Statistics ..... 7
2.1 Method description ..... 7
2.2 Examples of application ..... 8
3. Method B: Comparison against station observations ..... 9
3.1 Method description ..... 9
3.2 Examples of application – wind speed above ground in 10m to 100m height ..... 10
3.2.1 Final results from the deterministic reanalyses: ..... 10
3.2.2 Final results from the ensemble reanalyses: ..... 24
3.3 Main outcomes ..... 28
4. Method C: Comparison against gridded station observations ..... 28
4.1 Method description ..... 28
4.2 Examples of application – Climate indices ..... 29
4.3 Examples of application – Precipitation ..... 48
4.3.1 Alpine region – Final results ..... 49
Alpine region – Main outcomes ..... 63
4.3.2 Fennoscandia – Final results ..... 64
Fennoscandia – Main outcomes ..... 88
5. Method D: Comparison against satellite data ..... 90
5.1 Method description ..... 90
5.2 Examples of application ..... 90
6. Method E: Ensemble based methods ..... 91
6.1 Method description ..... 91
6.2 Examples of application ..... 92
7. Conclusion ..... 93
8. References ..... 96
9. Supplementary Materials ..... 99
9.1 Verification scores based on the contingency table ..... 99
9.2 Histograms of daily minimum and maximum temperature ..... 100
9.3 Additional material to precipitation analysis over Fennoscandia ..... 102

# 1. Scope of this document

Within work package 3 (WP3), the reanalysis products developed within WP2 have been analysed, evaluated and verified with independent observations. For this, the UERRA partners have identified the adequate scientific methods, which were discussed at the workshop (D3.1), also with input from users. These concepts have resulted in a collection of common evaluation procedures (D3.2), and comprise following methodologies for characterizing the uncertainties of reanalyses:

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, and the fitness for purpose is demonstrated. It builds upon the preliminary report D3.5 [Borsche et al., 2016], updating the results by using the now available final UERRA data. Method A is touched only for completeness, as investigations of feedback statistics are central to WP2 report D2.3 [Jermey et al., 2016]. Comparing them to this document allowed internal checks and a general confirmation of findings. Method F is treated in WP4, so not considered in this document. The focus was on the variables in which users are most interested, i.e., on precipitation, temperature near the ground, solar radiation, and wind speed in heights of relevance for wind energy applications. The reanalyses and comparable datasets investigated here are listed in Table 1.1, together with the spatial resolution and time coverage. Various temporal resolutions (spanning from hourly to interannual) are considered within the individual chapters for each method. For some model systems several abbreviations exist, all the ones used in the subsequent chapters are included in the first column in Table 1.1.

| Name | Domain | CRS | Grid res | Time / ens memb. |
|---|---|---|---|---|
| COSMO-REA6 / COSMO6-REA | EU | rotpol | 0.055deg | 1995-2015 |
| COSMO-REA12 | EU | rotpol | 0.11deg | 2006-2010 |
| COSMO-ENS<br>COSMO ensemble | EU | rotpol | 0.11deg | 2006-2010 / 20 ens |
| HARMONIE V1 / SMHI | EU | lcc | 11Km | 1961-2015 |
| HARMONIE V2 | EU | lcc | 11Km | 2006-2010 |
| MESAN (EURO4M) | EU | rotpol | 0.05deg | 1989-2010 |
| MESCAN | EU | lcc | 5.5Km | 2000-2010 and<br>1981-1990 |
| MESCAN (6 versions)<br>Changing physics/backgrounds | EU | lcc | 5.5Km | 2006-2010 |
| UKMO / UM | EU | rotpol | 0.11deg | 1979-2016 |
| UKMO-ENS<br>UM ensemble | EU | rotpol | 0.33deg | 1979-2016 / 20 ens |
| multi model UERRA ensemble<br>( consisting of UM, HARMONIE,<br>MESCAN and COSMO-REA12) | EU | - | - | 2006-2010 |
| ERAINT / ERA-Interim | EU | latlon | 80Km | 1979-2016 |
| ERA20C | EU | latlon | 125Km | 1970-2010 |
| ERA5 Det | EU | latlon | 0.28deg | 2010-2016 |
| ERA5 ENS | EU | latlon | 0.56deg | 2010-2016 / 10 ens |
| NORA10 | Fenno-<br>scandia | rotpol | 0.1deg | 1958-2016 |
| HCLIM | Norway | lcc | 2.5Km | 2003-2014 |

**Table 1.1:** List of investigated data sets in this evaluation report. The coordinate systems (CRS) are named as follows: latlon=regular latitude/longitude, rotpol=rotated pole, lcc=Lambert conformal conic projection

The used reference observation data sets are dependent on each method, and are thus described individually for each evaluation method. An overview is offered by Table 1.2.

| Name | Domain | CRS | Grid resolution | Time/ens members | variables |
|---|---|---|---|---|---|
| NGCD | fe | laea | 1Km | 1980-2010 | precipitation |
| APGD | al | laea | 5Km | 1971-2008 | precipitation |
| APGD-ENS | al | laea | catchments | 1981-85,2000-08/100 | precipitation |
| E-OBS | EU | latlon | 0.25° | 1950-2010 | precipitation & temperature |
| Station observations | Germany | - | - | Various; longest:1893-2016 | 10m wind speed |
| FINO masts | North-, Baltic Sea | - | - | Since 2004 (FINO2 2007) | wind speed 100m asl |
| Cesar and Lindenberg masts | Germany and Netherlands | - | - | Since 2000 (2001 Lindenberg) | wind speed above ground in 5m to 200m height |

**Table 1.2:** Reference datasets. Domain: EU=E-OBS domain (Europe), fe=NGCD domain (Fennoscandia), al=APGD domain (Alpine Region). CRS (coordinate system): latlon=regular latitude/longitude, laea=ETRS89 Lambert Azimuthal Equal-Area projection coordinate reference system.

# 2. Method A: Feedback Statistics

## 2.1 Method description

The ODB (Observational DataBase - http://www.ecmwf.int/en/elibrary/15080-odb-past-present-and-future) is 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, it 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-Var data 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.

**Advantages**

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

**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.

**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.

## 2.2 Examples of application

The investigation of feedback statistics is part of WP2, highlighted in deliverable D2.3. Further evaluation, including the comparison of modelled 2m-temperature with station data is presented in deliverable D3.5 [Borsche et al., 2016]. The methodologies and proceeding of D3.5 remain valid for the extended dataset spanning the full production interval.

# 3. Method B: Comparison against station observations

## 3.1 Method description

When traditional users of station data consider applying reanalysis data, a natural first question to ask is how both compare. It has to be kept in mind that the station observation is a point measurement, representative for its surrounding, and the respective reanalysis grid cell value is representative for a spatial scale. The answer will depend on the specific characteristics of a location, or a region. It will also depend on the temporal and spatial scale, of the height (or model level) considered, and of course on the variable of interest. As it is not feasible to give a general answer, below we illustrate how to find answers to this question, for example for the region of Germany, for wind speed, for the statistical characteristics of verification scores typical wind energy applications might be interested in.
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, which are amplified in the Appendix, section 9.1. These skill scores for investigation of extreme events 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.

**Advantages**

This method is easy to apply and desired by the users.

**Disadvantages**

Comparison of grid cells of a spatial extend of several tens to hundreds of square kilometres with point measurements is not 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.

**Value of method**

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

## 3.2 Examples of application – wind speed above ground in 10m to 100m height

### 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.

### Used observations

Data which are compared against include tower measurements of Lindenberg, Cabauw, and the FINO platforms, of hourly, daily, and monthly values. More location information can be found in [Borsche et al., 2016]. In addition dependent and independent DWD station data measurements are available from ftp://ftp-cdc.dwd.de/pub/CDC/.

### Investigated reanalyses

Investigated reanalyses include the regional reanalysis COSMO-REA6 covering the time range 1995 to 2014, and the four regional reanalyses developed during UERRA by SMHI, Meteo France, DWD and the UK MetOffice and the two global reanalyses ERA20C and ERA-Interim, as listed in Table 1.1. The evaluation period is set to 2006-2010, based on the maximum overlap of all applied data sets. For the following investigations two time resolutions are discussed separately, motivated by the different temporal resolutions of all reanalysis frameworks. Firstly, the examinations are based on 6-hourly reanalysis data, which are deterministic analysis data sets for COSMO-REA6, COSMO-REA12, HARMONIE, UM and ERA-Interim, but deterministic forecast fields for MESCAN, because the analysis fields are just an interpolation of HARMONIE and will not add any further information. On the other hand, 1-hourly data are used. They come from analysis fields for COSMO-REA6 and COSMO-REA12, but forecast fields for the UERRA products HARMONIE, UM and MESCAN, because these model systems provide analysis fields with a 6-hourly time step, only. Moreover, the probabilistic data sets of the UERRA ensembles UM and COSMO-REA12 are used, named COSMO-ENS and UKMO-ENS in Table 1.1. Both systems possess 20 members.

### 3.2.1 Final results from the deterministic reanalyses:

Comparison of station measurements of 10m wind speed against regional reanalyses is shown exemplarily for the station Hannover, which is situated in a flat surrounding, so that this station can be considered as representative for a larger region.
The evaluation includes all wind speed data from 2006-2010. In Figure 3.1, seven panels of the frequency distribution of 10m wind speed are shown. The numbers of observation points, the mean, median, and the 1st and 99th percentile of measured or calculated wind speed are also given for each plot. The number of measurements is less for COSMO-REA12, as the last two months of the year 2008 were missing due to the production delay, at the time of the analysis. For this particular location (Hannover), the frequency distributions indicate a relatively good match for the COSMO-REA6 reanalysis, whereas MESCAN and ERA-Interim fields tend to underestimate low wind speeds. On the other hand COSMO-REA12 and HARMONIE underestimate the frequency of higher wind speeds (above 5m/s). But it is important to note, that this situation is only valid for station Hannover. For each model one can find stations where a selected model fits best and another worst. This stresses the limits of drawing general conclusions from this kind of evaluation method.

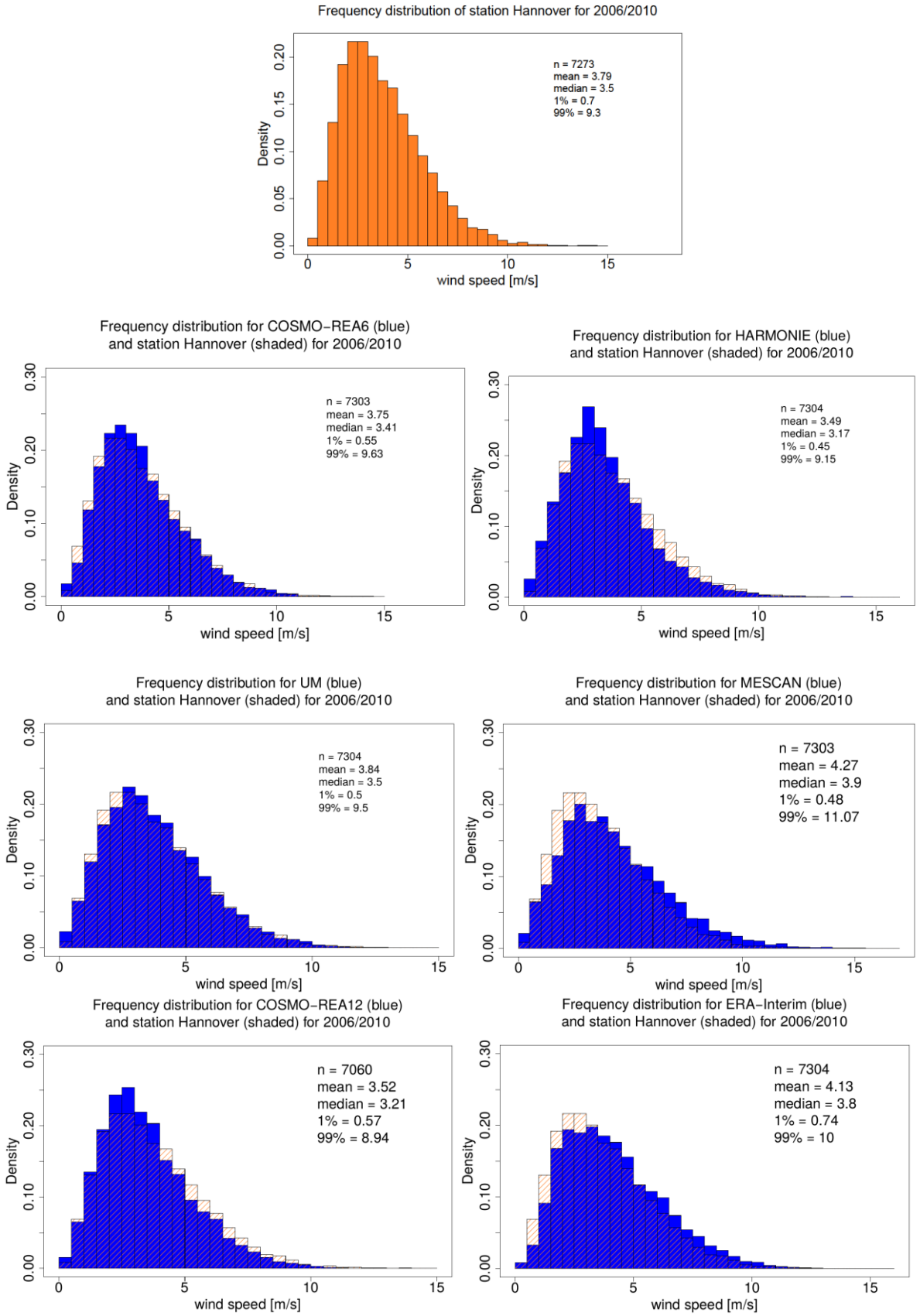


**Figure 3.1:** Frequency distribution at station Hannover (orange and shaded) and the six regional reanalyses (blue) grid cell values at the location of the station for 6-hourly data.

Each location might have its own specialities, depending on its surrounding location, and depending on how well the terrain is modelled.
To facilitate interpretation, bias, MSE and correlation are computed for many stations, and interpreted together. In Figure 3.2 the time dependence of the Pearson correlation is shown for each reanalysis, averaged over all stations, which are located above 100m height. Except UM, none of the reanalyses assimilate wind speed of stations located at heights above 100m above sea level.
The first conclusion is that the highest correlations are reached for every model systems on a weekly timescale, reflecting the fact that some, but not all high frequency variations in the reanalyses fields are reflecting reality, and that some averaging is beneficial.
Secondly, the various reanalysis exhibit similar correlations, which are higher than ERA-Interim, particularly at the shorter time scales.

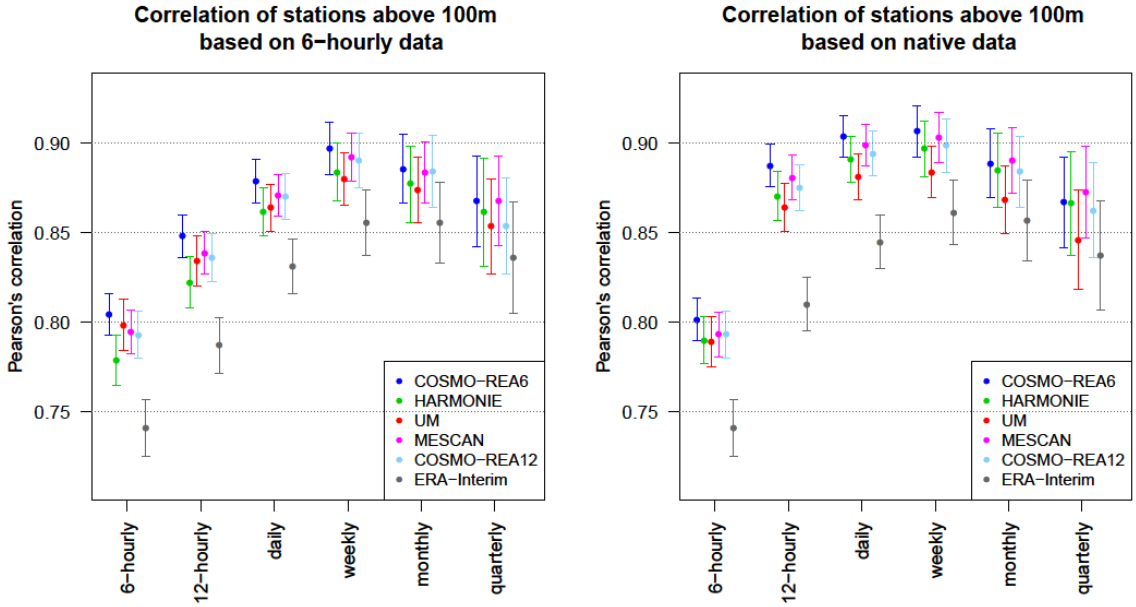


**Figure 3.2:** Pearson correlation based on 6-hourly (left) and native (right) data for all RRA's. Native data is defined by 1-hourly data for COSMO-REA6, UM, COSMO-REA12, HARMONIE and MESCAN, but 6-hourly data for ERA-Interim. The error bars mark the 95% confidence interval

Thirdly, all regional reanalyses have significant better fitness on high time resolutions than the global reanalysis ERA-Interim up to the daily scale, with a diminishing difference at longer time scales. Fourth, here seems to be a slight advantage of COSMO-REA6 (which has the highest spatial resolution, together with the MESCAN reanalysis) on hourly and daily timescale. The comparison between the left and the right panel of Fig. 3.2 indicate that a higher temporal resolution of reanalysis data yields to better results. Furthermore, one can see an improvement of ERA-Interim from the left to the right panel of Fig. 3.2, although in both pictures six-hourly reanalysis output is used for the global reanalysis. But on the left hand side the data is compared to 6-hourly observations and on the right side to hourly observations.
In Figure 3.3 the bias for various reanalyses is shown, binned according to wind percentiles. Here the mean bias for each percentile bin is determined. For instance, to determine the bias of percentile 0.25, all observations falling in the 0.25-0.3 percent quantile are compared with the respective reanalysis value at the time and location of the observation. The lower boundary is included in the bin, the upper one excluded. The resulting difference is averaged over all timesteps. The top panel of Fig. 3.3 shows the results when all station observations are binned together, and the lower panels show the bias for 2 stations: Wittmundhafen (8m

above sea level) and Zinnwald-Georgenfeld (877m above sea level), exemplarily. Averaged over all station locations every model overestimates the lower wind speeds (below 25% quantile) and mainly underestimates the higher wind speeds (above the 75% quantile).
Considering the bias for the station Wittmundhafen (Fig. 3.3, lower left panel) an obvious difference between the various reanalyses attracts attention. However, this should not be generalized as there exist large variations in bias between the single stations, which is illustrated by comparing Wittmundhafen with Zinnwald-Georgenfeld (Fig. 3.3). The effect can be easily understood keeping in mind the observed frequency distribution for wind speed (often described with Weibull shape and scale) might not be captured adequately in the reanalysis, either due to sub-optimal assignment of height levels in case of mountain stations, or due to the station representativity not matching the grid resolution, or due to model deficiencies in capturing the relevant local processes.

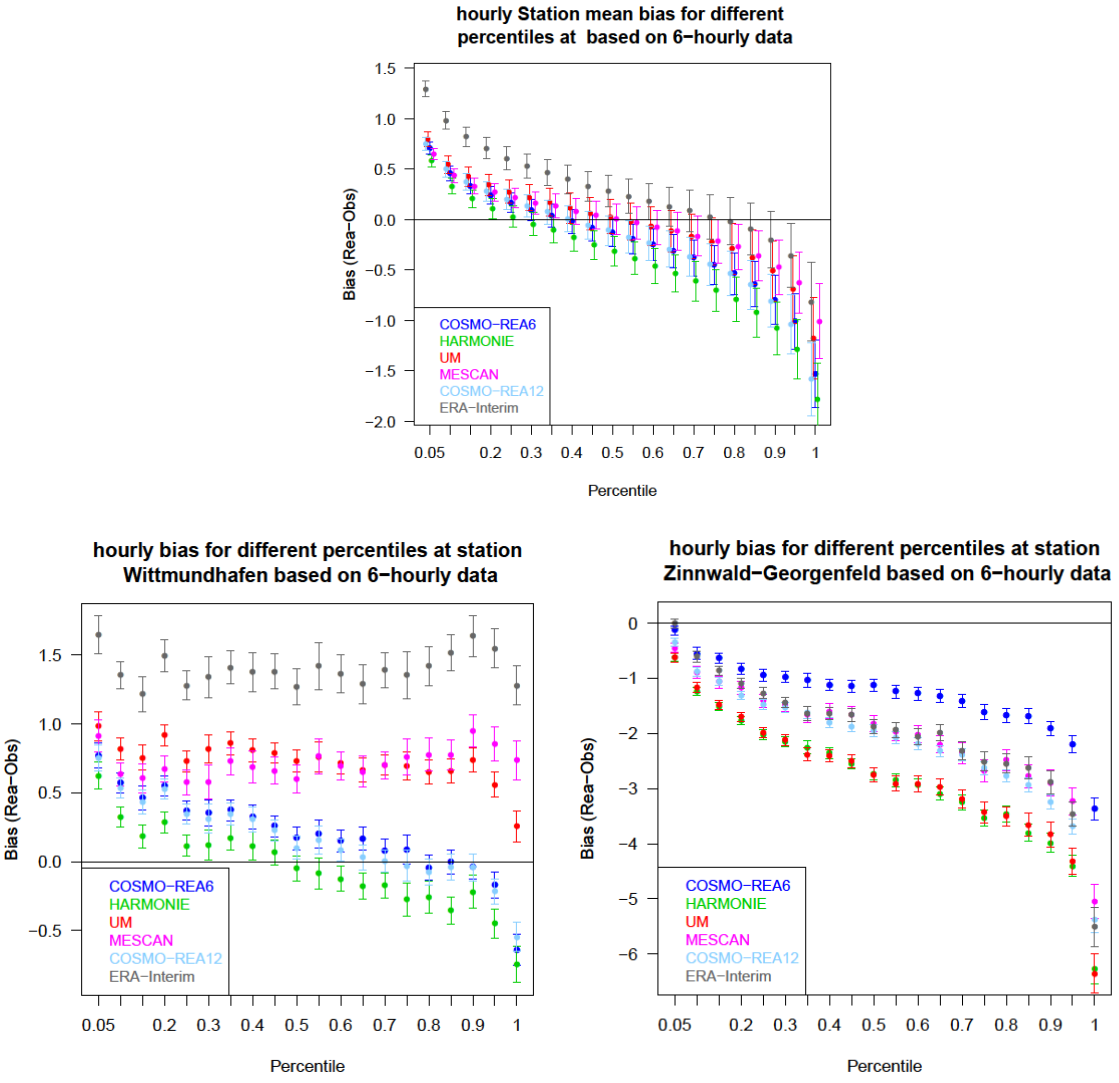


**Figure 3.3:** Bias and 95 percent confidence interval for different wind percentiles and model systems. The upper figure shows averaged bias of all stations. The lower figures are for two arbitrary stations, which differ in local topography.

Thus, the bias changes strongly with model system, wind speed and location.
For higher elevated stations the bias increases due to larger disagreements between station height and model orography. This is shown in Figure 3.4, especially in the mountainous regions of the Alps, the Black Forest, the Erzgebirge or the Harz in Central Germany. For two

stations, marked in the upper left picture of Fig. 3.4 the bias varies strongly, although the stations are close to each other. At station Weinbiet, the bias constitutes -3.33 m/s and the topography varies from 553m for the station height to 269 m for model height. For Mannheim the bias is less than 0.45 m/s and the differences between model and station height is only 14 m. This illustrates that a significant fraction of the local variable bias is related to the difference between model orography and real station height.
The averaged bias over all German stations (also above 500m height) reaches a value of $-0.23 \pm 0.15$ m/s for COSMO-REA6, $-0.42 \pm 0.17$ m/s for HARMONIE, $-0.04 \pm 0.18$ m/s for UM, $-0.06 \pm 0.17$ m/s for MESCAN and $-0.21 \pm 0.16$ m/s for COSMO-REA12, concerning the investigated period 2006-2010. If only stations beneath 500m are considered, the bias reduces to $0.004 \pm 0.09$ m/s for COSMO-REA6, $0.15 \pm 0.09$ m/s for HARMONIE and $0.05 \pm 0.09$ m/s for COSMO-REA12. For UM and MESCAN the averaged bias rises lightly to $0.285 \pm 0.11$ m/s and $0.22 \pm 0.1$ m/s respectively, because the underestimation of stations located in higher areas balance the overestimation of stations in the flat Northern area of Germany. This fact is indicated in Figure 3.4 as well, where more light blue points for UM and MESCAN can be found in the Northern area of Germany. The strong local effects are not only discovered in model bias but also in the correlation between reanalysis and observation. The local dependence of the Pearson correlation is discussed in connection with Figure 3.5.

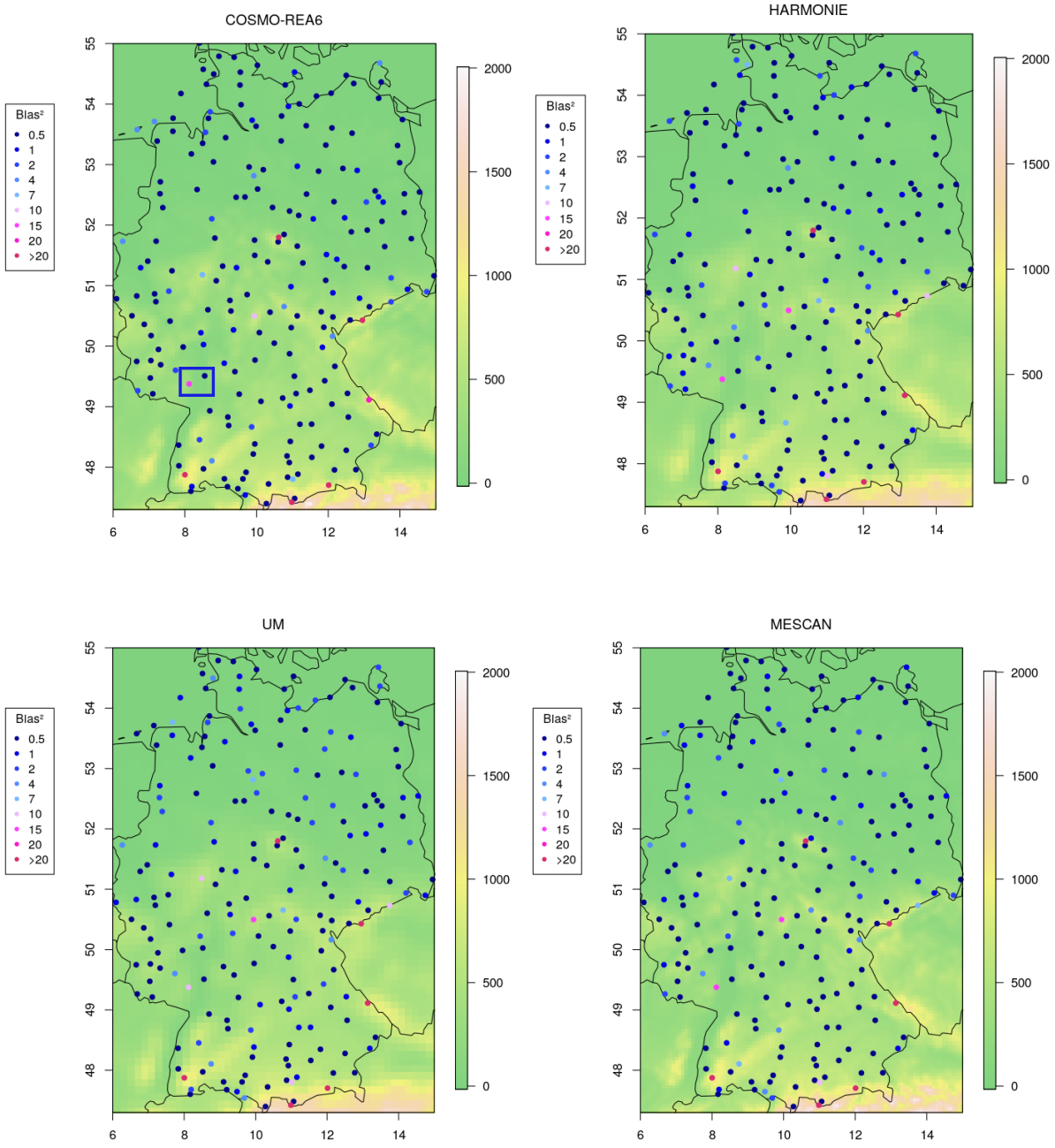


**Figure 3.4:** Squared bias of four regional reanalysis systems for selected station locations in Germany, averaged over the period 2006-2010

For mountain stations the correlations are significantly worse, an effect remaining true for every model system and all temporal resolutions. This includes monthly data as well, which are not shown here. The Pearson correlation between observation and reanalysis varies much more among mountain stations than among low elevation stations.

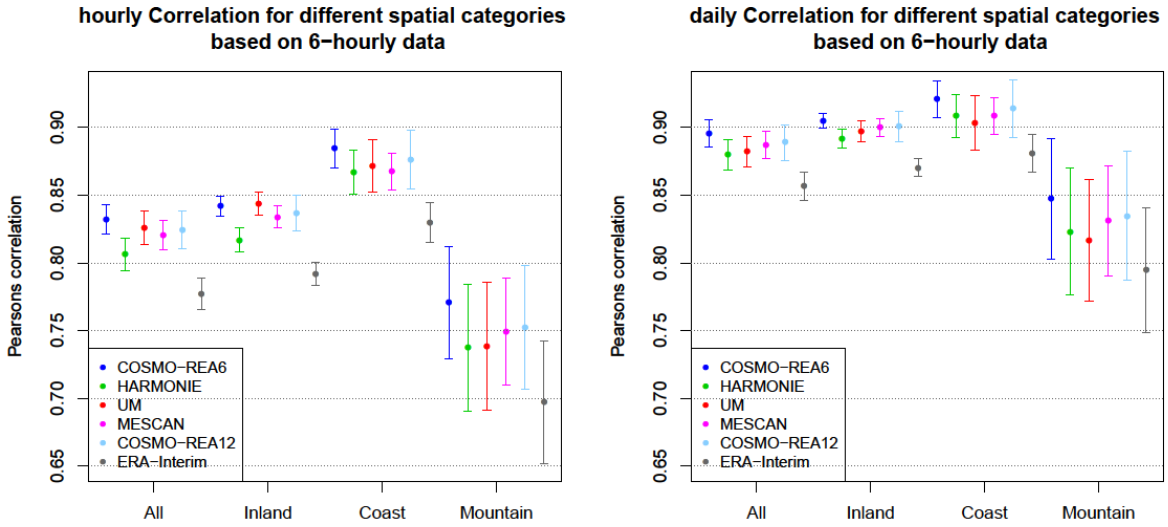


**Figure 3.5:** Pearson correlation and 95% confidence interval averaged according to specific station locations, based on hourly (left) and daily (right) data. Mountain stations include all locations above 500m height and coast stations comprise all measurements that are less than 5km from the coastline. Inland stations are defined as stations beneath 500m height, coast stations excluded.

Coast stations exhibit also a slightly increased variation (i.e., larger range for correlations), compared to inland stations. Based on hourly values, the mean correlations of coast stations is significantly higher than for inland stations. This is valid for all five reanalyses. Considering inland and coast stations a significant advantage of regional reanalyses over ERA-Interim is striking , and in accordance with results shown in Figure 3.2.

In a next step skill scores were calculated for the five regional reanalyses and ERA-Interim. The results are shown in Figure 3.6a and 3.6b for station Hannover, exemplarily. 6-hourly data were used for a fairer comparison. The scores measure different value ranges and properties. Thus examination of several scores is recommended, as is generally the recommended code of practise in verification studies. “Good“ scores does not necessarily allow the conclusion that the model captures all statistics properly, especially when the sample size is varying and influencing the score behaviour, which makes the interpretation more challenging. Especially the behaviour for rare events differs for various scores. As the scores depend on the magnitude of the variable, and higher magnitudes are rarer events, an interpretation of the skills for these is not straightforward. For the interpretation of model fitness for extreme events the verification score should be independent of the base rate or the event frequency. This is guaranteed for EDI and SEDI. For EDI and SEDI only a light drop of skill for higher wind speed is noticeable, which indicates good model fitness for extreme events as well.

One has to keep in mind that these good scores are only reached for a relative measure (wind percentiles). When the absolute value is considered, all model systems lose skill, due to the strong local biases discussed earlier. For all investigated scores the different model systems are similar to each other, though ERA-Interim and HARMONIE have slightly lower and COSMO and UM have somewhat higher scores. In compared to deliverable D3.5 [Borsche et al., 2016], where EURO4M data are used, the UERRA reanalyses by SMHI and UKMO show a great improvement with respect to their EURO4M counterparts.

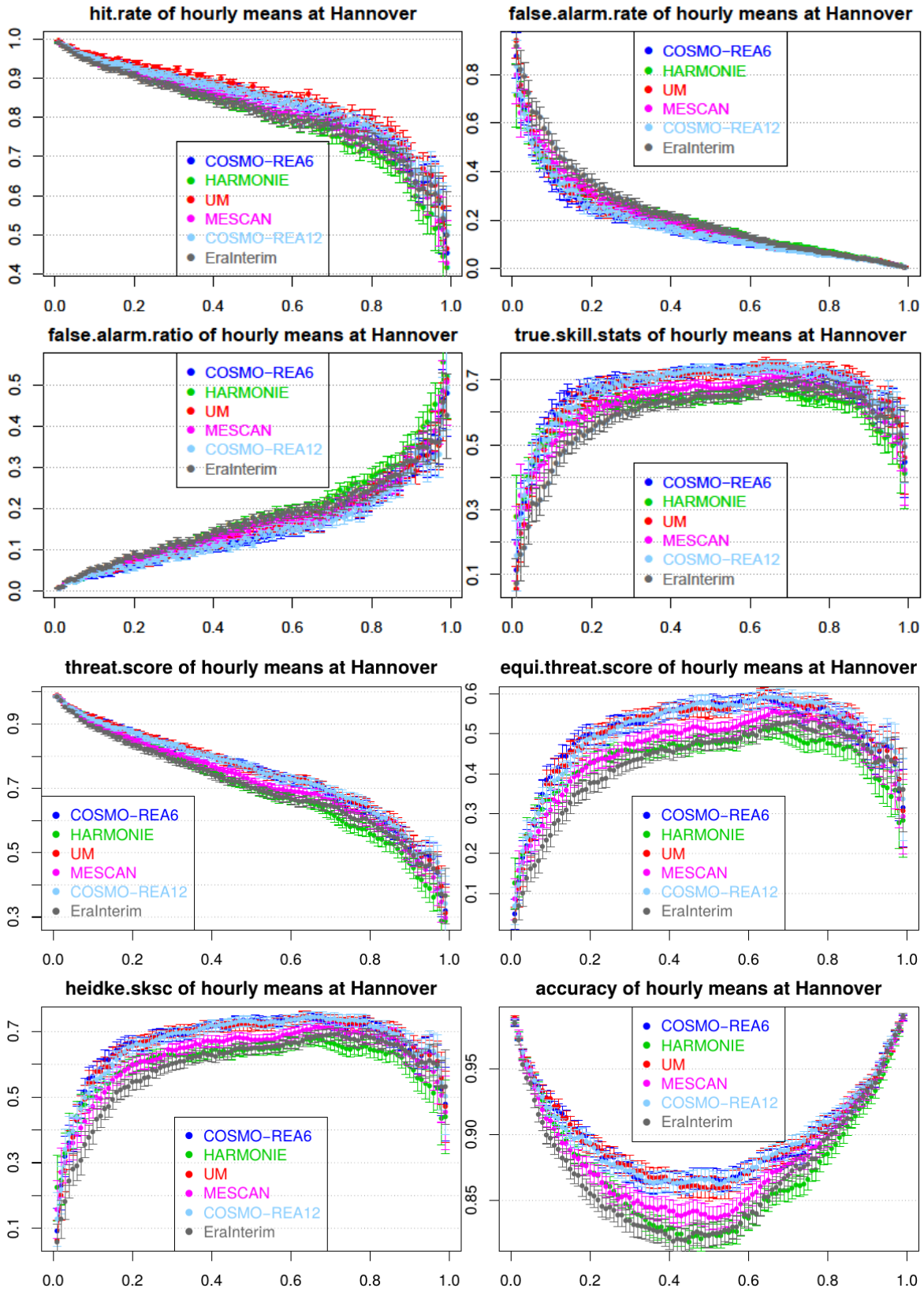


**Figure 3.6a:** Different skill scores of 6-hourly data at station Hannover compared to the various reanalyses, computed for wind percentiles. The error bars mark the double standard error. All observations from 2006 to 2010 are taken into account.

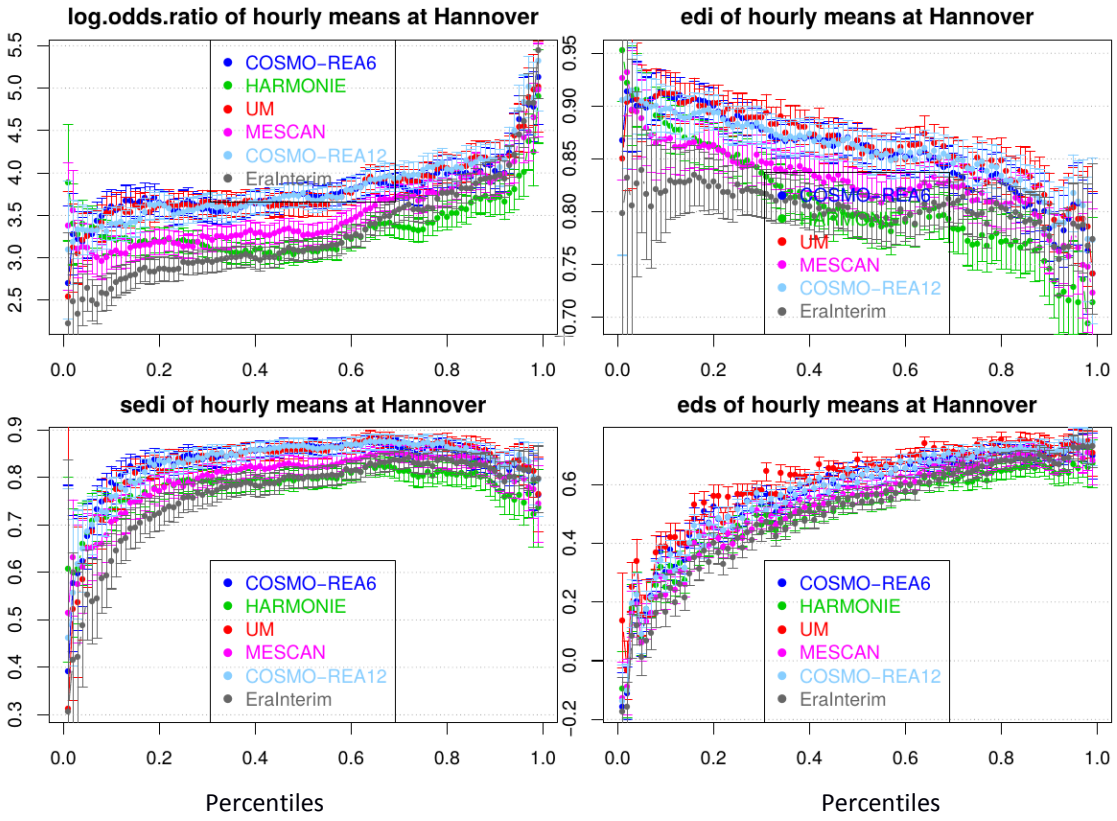


**Figure 3.6b:** Different skill scores of 6-hourly data at station Hannover compared to the various reanalyses, computed for wind percentiles. The error bars mark the double standard error. All observations and reanalysis data from 2006 to 2010 are taken into account.

Figure 3.7 shows the time series for the storm event Emma. This period was chosen to compare EURO4M results, which are presented in the deliverable D3.5 [Borsche et al., 2016], with the new UERRA products. The new reanalysis data of SMHI shows an increase in correlation for this period from 0.45 (EURO4M) to 0.77 (UERRA). UM improves the results from 0.75 (EURO4M) to 0.91 (UERRA). The same evaluation for a low wind speed period (21.12.2006-26.12.2006), see Figure 3.8, produces much smaller correlations for all investigated reanalyses. The choice of an arbitrary period also reduces the correlation compared to the correlation obtained during a storm event as depicted in Figure 3.7. The reanalysis systems tend to have fewer problems with the reproduction of strong winds in opposition to low wind speeds.

### Comparison to wind masts

The evaluation with station data has to be considered critically, because some models use the data during their assimilation process. UM assimilates stations beneath 500m height and Cosmo uses stations below 100m height. However, only a limited number of truly independent data exists. Here measurements from meteorological wind masts are used. Further comments can be found in [Borsche et al., 2016]. The analysis of correlation between reanalyses and windmasts is shown in Figure 3.9. For all locations, the regional reanalyses can show significant better consistency with observations than, e.g., significantly lower resolution global reanalysis ERA20C. Over sea, the regional reanalyses are also significant better than the higher resolution global reanalysis ERA-Interim. No general evidence can be found, that the correlation of wind speed reduces or rises with height. This depends on station location and reanalysis system.

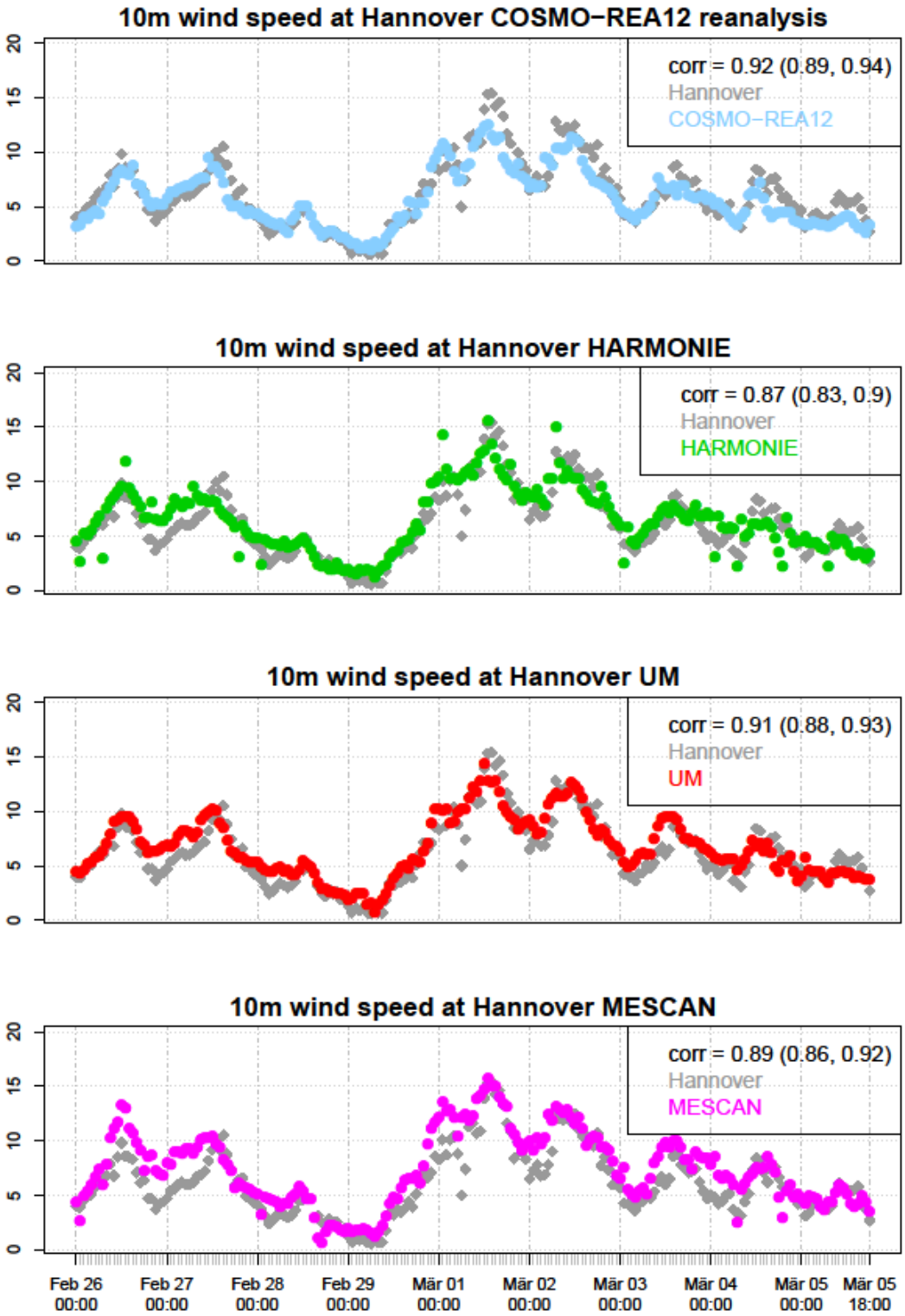


**Figure 3.7:** Time series of storm event Emma between 00 hrs February, 26th 2008 and March, 06th 2008 for the regional reanalyses COSMO-REA12, HARMONIE, UM and MESCAN (upper to lower panels).

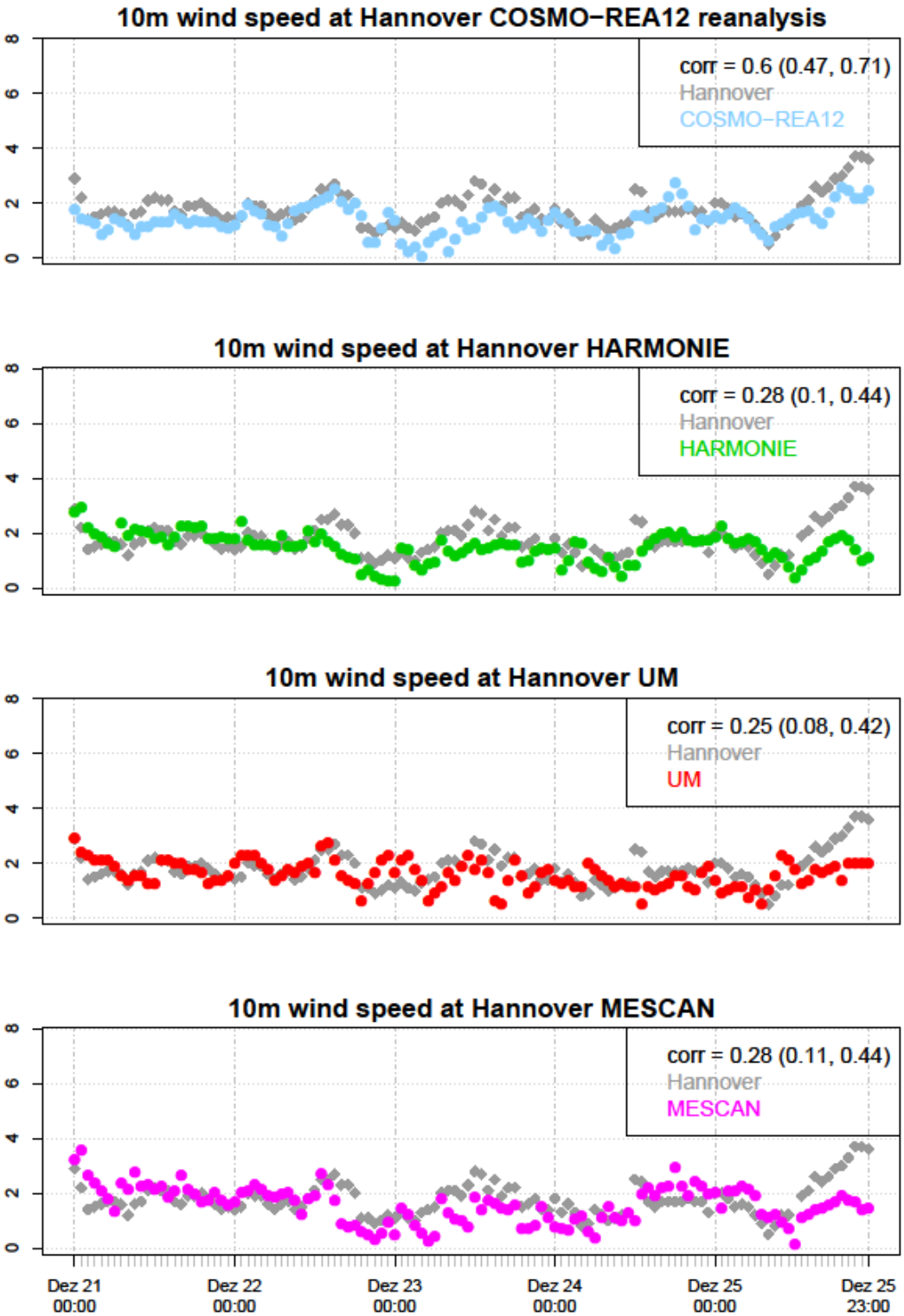


**Figure 3.8:** Time series of storm low wind period between 00 hrs December, 21th 2006 and December, 25th 2006 for the regional reanalyses COSMO-REA12, HARMONIE, UM and MESCAN (upper to lower panels).

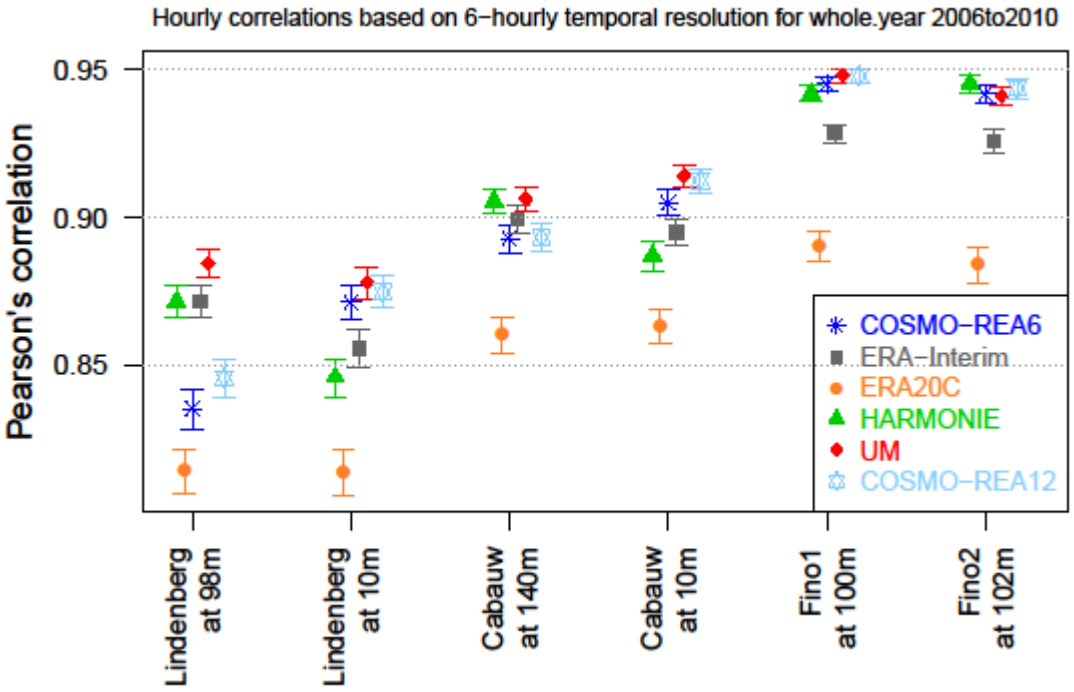


**Figure 3.9:** Pearson correlation with 95% confidence interval for various wind masts and heights, based on 6-hourly model data and measurements.

The comparison between COSMO-REA6 and COSMO-REA12 show some indications of benefits of COSMO-REA12. This could be caused by the different kinds of height attribution employed for both systems.

While COSMO-REA6 provides variables only for model level (10m, 35m, 69m, 116m, 178m, 258m), the reanalysis systems produced in UERRA differentiate between model-, pressure- and height levels (15m, 30m, 50m, 75m, 100m, 150m). Latter are used for Fig. 3.9. Hence the model data sets for COSMO-REA6 and COSMO-REA12 do not match the height of measurements, but also do not match each other.

Figure 3.10 presents the decomposition of the mean squared error into a correlation effect, the effect due to differences in standard deviation and the bias. This method draws conclusions from the causes of MSE. Equation (1) denotes the single effects [Gupta et al., 2009]:

$$MSE = 2\sigma_{rea}\,\sigma_{obs}(1-\rho) + (\sigma_{rea}-\sigma_{obs})^2 + (\mu_{rea}-\mu_{obs})^2 \qquad (1)$$

The upper panel of Figure 3.10 includes all daily means of June, July and August from 2006 to 2010, while the lower panel considers data from December, January and February. On daily timescale in most cases the biggest percentage of MSE is based on a lack of correlation. For Cabauw (140m) one can identify a bigger impact of bias, which is caused by the higher difference of model and height of observation. Differences in standard deviation between model and observation are very low. The correlation effect increases with higher time resolution, because of the lower correlations, see Figure 3.2. The bias stays constant with various time resolutions, so that on monthly scale the MSE is mainly caused by bias effect and on hourly scale mainly by to correlation effects. For most model systems and mast locations the correlation effect increases in wintertime. On the other hand, bias and difference of standard deviation have no strong dependence on season. Figure 3.11 shows the annual cycle in different heights, comparing data from wind mast Lindenberg and

corresponding model data. An annual cycle with the minimum occurring in summer and the maximum in winter is visible for all reanalysis data and every height, as well for the observations. The regional model systems but also the global reanalyses can reproduce this pattern in 10m and 100m quite well. In winter the differences between model systems and mast measurements are larger than in summer, which is also indicated in Figure 3.10 (see the bias at Lindenberg at 98m, comparing the upper and the lower plot). At 100m height, all model systems, except for ERA20C, overestimate the wind speed during winter. UM overestimates the observations for all other level heights as well.

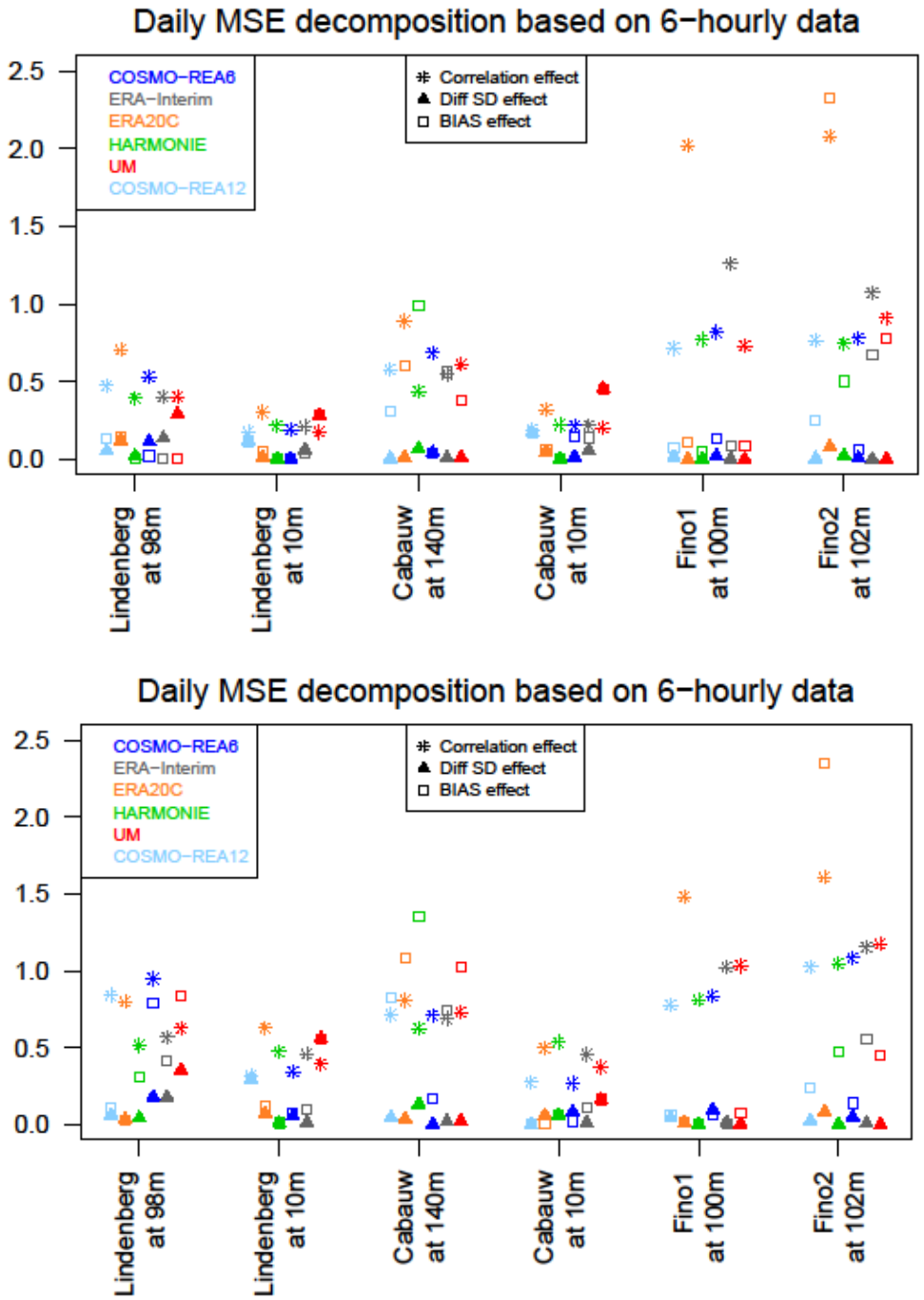


**Figure 3.10:** MSE decomposition for various wind masts and heights, based on 6-hourly data, averaged to daily time scale and considering the winter months (bottom) and the summer months (top) of 2006 to 2010.

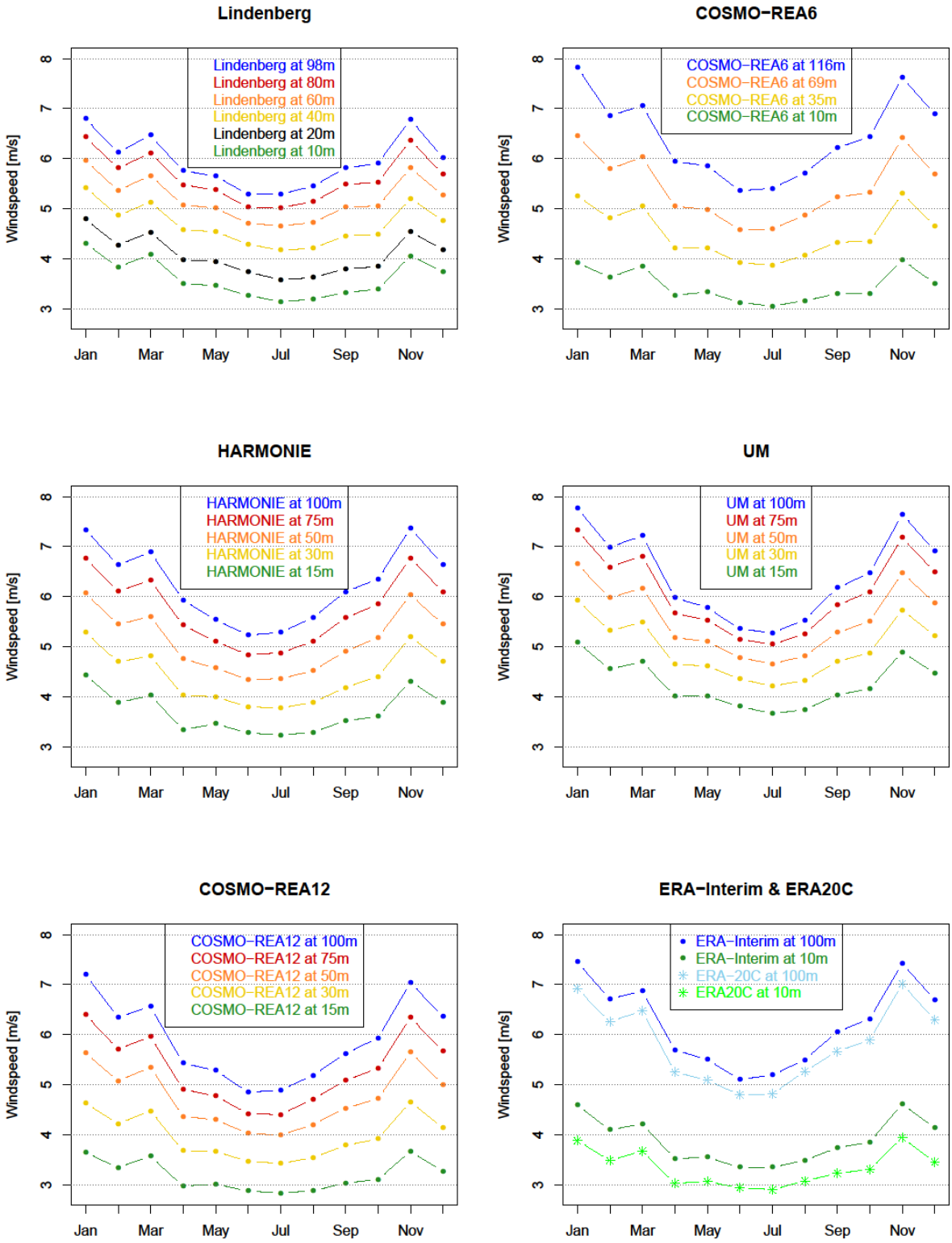


**Figure 3.11:** Annual cycle of wind speed in various heights for mast measurements Lindenberg and model data from COSMO-REA6, HARMONIE, UM, COSMO-REA12, ERA-Interim and ERA20C. The means of calendar months are averaged for the years 2006 to 2010.

### 3.2.2 Final results from the ensemble reanalyses:

In Deliverable D2.14 [Jermey et al.], the assessment of ensemble properties using rank histograms, CRPS, Brier score, reliability diagrams and ROC curves is discussed, using station data for verification of wind speed as well. The following evaluation adds additional information on model fitness for various timescales. Moreover, the multi-model UERRA ensemble is emphasized here. The latter comprises the four deterministic reanalysis products, developed during UERRA, which include COSMO-REA12, MESCAN, HARMONIE and UM.
Figure 3.12 shows the time series of 10m-wind speed at station location Emden for the three ensemble systems. For COSMO-REA12 the deterministic run and the ensemble mean appear quite similar. For the UM reanalysis the difference is slightly larger.. The temporal evolution of wind speed from the deterministic run and of wind speed from the ensemble mean (for COSMO-REA12 and UM) is rather similar. For the chosen period the ensemble spread of both ensemble systems seems too small. This is illustrated especially at times where the deterministic run is far away from the observations (02$^{th}$ March). None of the 20 ensemble members is able to reproduce the station measurements in this case. During summer, when the absolute wind speed is smaller, the spread of the ensembles increases, as indicated in Figure 3.13. This seasonal dependent behaviour of spread is shown for the UM ensemble in deliverable D2.14 and for the COSMO-REA12 ensemble in [Bach, 2016] as well, however, the latter investigation was for precipitation. The spread increases from winter to summer for COSMO-REA12 and UM ensemble systems. In addition to the COSMO-REA12 and UM ensemble, the multi-model UERRA ensemble, based on four members, (the deterministic runs of UM, COSMO-REA12, HARMONIE and MESCAN), is plotted in the bottom panel of Figure 3.13. It covers more variability than the single-model ensembles. The multi-model ensemble can cover the variability of observations for the most timesteps.
The evaluation of ensemble spread is continued in Figure 3.14. The plot depicts the comparison of ensemble spread and RMSE of the ensemble mean. As shown in [Palmer et al., 2005] for a perfect ensemble, spread and RMSE of the ensemble mean should be equal. The formula of spread is given by [Grimit and Mass, 2007]:

$$\text{spread} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(x_m - \bar{x})^2} \qquad (2)$$

M is the number of ensemble members, $\bar{x}$ the ensemble mean and $x_m$ the value of the m$^{th}$ member. Spread and RMSE are computed for the whole period 2006-2010 and averaged over the selected station locations in Germany, which are illustrated in Figure 3.4.
All of the three investigated ensembles show a strong under-dispersiveness, at which the UERRA multi-model ensemble exhibits the largest spread. This was indicated in Figure 3.11 already. Spread and RMSE decreases with coarser temporal resolution, due to the averaging effect. For UM, the root mean square of the ensemble mean is about 0.3 m/s higher than for the other reanalyses, but the discrepancies decline as well with coarser time resolution. For hourly timescale the spread of the multi-model UERRA ensemble is 2.3 times smaller than the corresponding RMSE of ensemble mean, on a daily scale the spread is even 2.7 times smaller. For COSMO-REA12 the factor between RMSE and spread amounts to 3.3 on hourly scale and 8 on daily scale. For UM the factors are 4.6 and 6.6 respectively.
The evaluation of the Brier score (BS) affords the possibility of assessing more attributes of a probabilistic forecasts. It measures the accuracy of an ensemble forecast. It is the mean squared deviation between ensemble probability and binary observations. It can be decomposed into reliability (first term of equation (3)), resolution (second term) and uncertainty (third term) as mentioned in [Murphy, 1973]:

$$BS = \frac{1}{N}\sum_{k=0}^{M} N_k(f_k - \bar{o}_k)^2 - \frac{1}{N}\sum_{k=0}^{M} N_k(\bar{o}_k - \bar{o})^2 + \bar{o}(1-\bar{o}) \quad (3)$$

N denotes the number of data points in the examined time period, $\bar{o}$ is the sample climatology, M the number of forecast possibilities and $\bar{o}_k$ the mean observation for timesteps with forecast $f = f_k$.

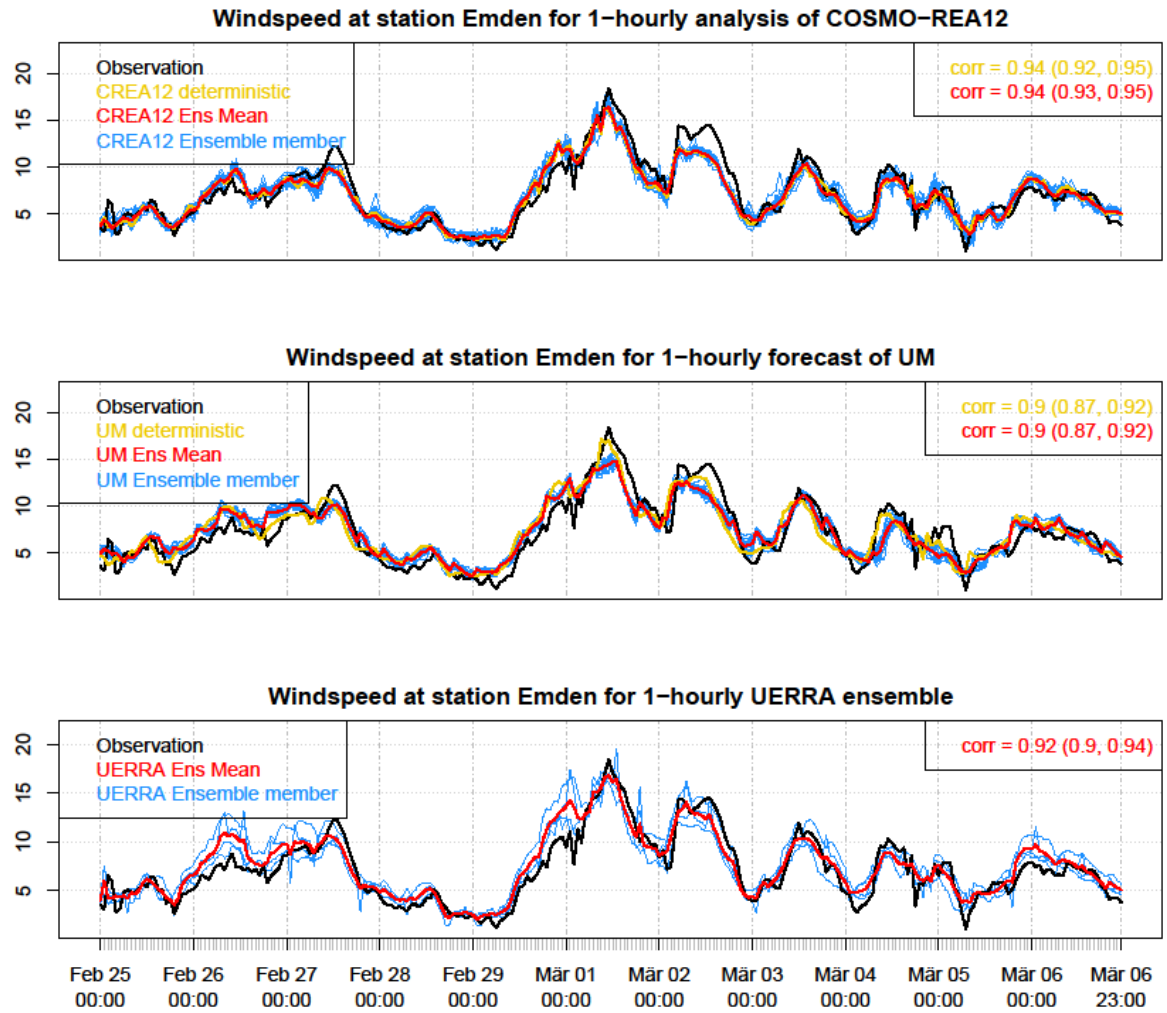


**Figure 3.12:** Timeseries of 10m wind speed for COSMO-REA12 ensemble (top), UM ensemble (middle) and the multi model UERRA ensemble (bottom) at station location Emden. The period includes 25th February to 06th March 2008.

The Brier skill score (BSS) measures the improvement of a probabilistic forecast relative to a reference forecast, mostly the climatology:

$$BSS = 1 - \frac{BS}{BS_{reference}} \quad (4)$$

Figure 3.15 shows the Brier score and the decomposition to reliability, resolution and uncertainty for different absolute thresholds and the three mentioned reanalysis ensembles. The uncertainty varies from 0 to 0.25, where 0 denotes, that the event either always or never occurs and 0.25 means that the event occurs half of the time. For reliability the perfect score is 0. It is reached, if the modelled probability and the observed frequency are equal. The resolution indicates the model property of building forecasts, which significantly differ from

the climatological base rate. Since the resolution term is subtracted in equation (3) this is a preferable situation and denotes the model ability to distinguish between various observed situations.

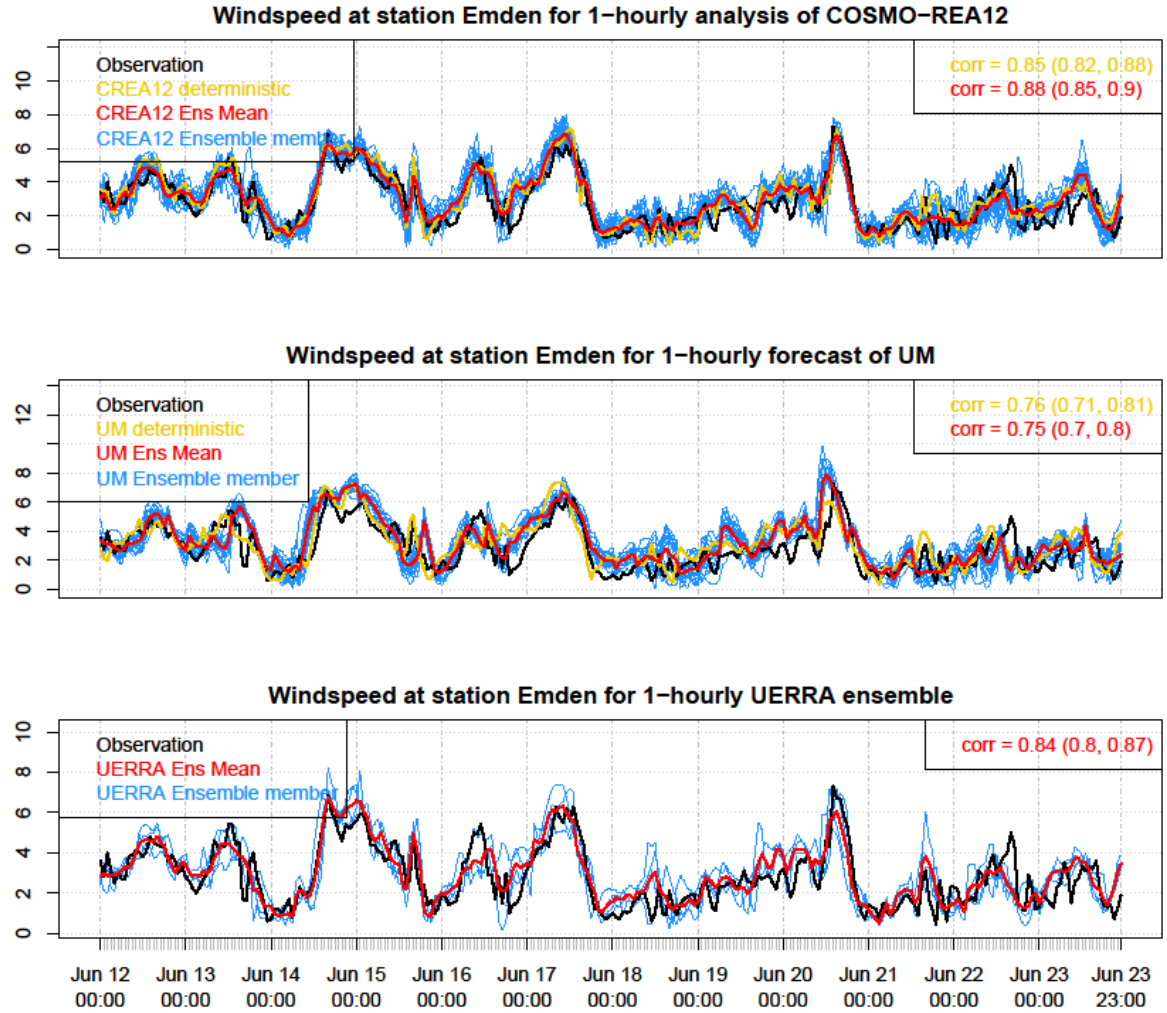


**Figure 3.13:** Timeseries of 10m wind speed for COSMO-REA12 ensemble (top), UM ensemble (middle) and the multi model UERRA ensemble (bottom) at station location Emden. The period includes 12$^{th}$ June to 23$^{th}$ June 2007.

Figure 3.15 indicates a maximal BS at threshold 3m/s for all three ensemble systems. This is mainly caused due to the maximal uncertainty at 3m/s, which indicates, that nearly half of the observations lay below and half above this value. The ensemble of COSMO-REA12 and UERRA are nearly similar, while UM shows worse results. The maximal BS of UM amounts to 0.18, while COSMO-REA12 reaches a value of 0.13 and the multi model UERRA ensemble reaches 0.14. The higher BS of UM is caused by higher factor of reliability and lower resolution. Hence, the agreement between observed frequency and ensemble probability is better for COSMO-REA12. The better results of COSMO-REA12 and the multi model UERRA ensemble in respect to UM ensemble could be caused by the 3 times higher spatial resolution.
The computation of BSS for threshold 3m/s produces $0.37 \pm 0.04$ for COSMO-REA12, $0.16 \pm 0.05$ for UM and $0.36 \pm 0.04$ for UERRA ensemble averaged over all station beneath 500m. Thus COSMO-REA12 and the multi model UERRA ensemble can enhance the accuracy about 37 %, according to the climatology. UM improves the accuracy about 16%.

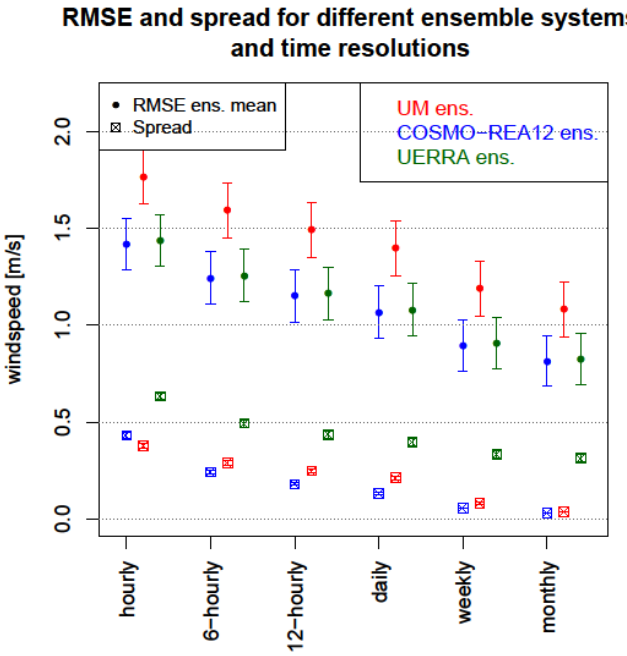


**Figure 3.14:** RMSE and spread of ensemble systems for various time resolutions. The values are averaged over 209 station locations and a period from January 2006 to December 2010. The error bars mark the 95 % confidence interval.

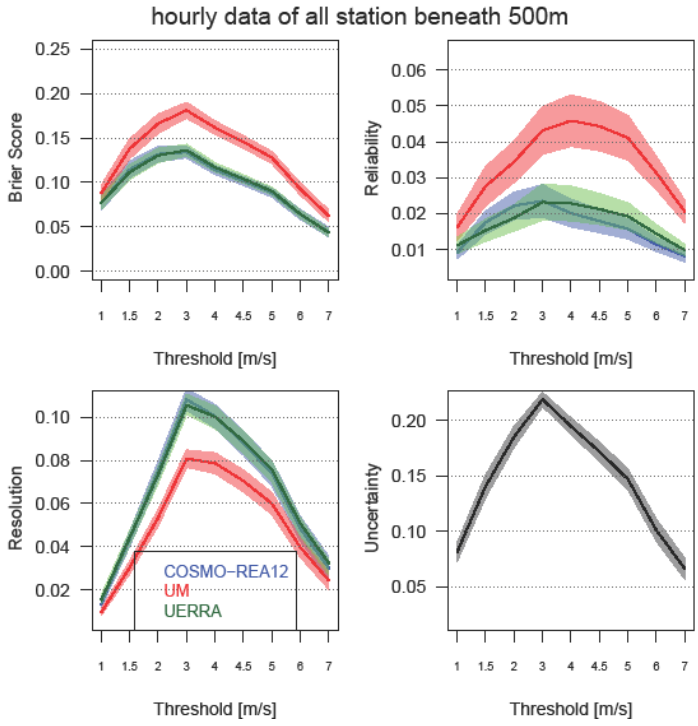


**Figure 3.15:** Decomposition of Brier score (top left) into its three components reliability (top right), resolution (bottom left) and uncertainty (bottom right). The values are calculated for each station separately and then averaged over all German stations beneath 500m height, considering the time period 2006-2010. The uncertainty is given by the 95% confidence interval.

## 3.3 Main outcomes

- The correlation between reanalyses and station data is dependent on the time scale, with a maximum at the weekly time scale.
- A significant advantage of regional reanalyses with respect to global reanalyses is achieved for high temporal resolutions, (on the hourly to daily scale).
- The reanalyses show better correspondence with observations, if hourly reanalysis fields instead of 6-hourly reanalysis fields are used.
- The reanalysis fields compare better with averaged instead of instantaneous measurements.
- All investigated reanalyses have problems with high elevated stations, due to higher deviations between modelled and real topography.
- The bias can depend strongly on local effects (mainly for mountain stations, and stations not positioned to be representative for an area spanning several tens of km)
- The bias may be a function of wind speed (especially if height attribution is difficult).
- The bias depends on the model system and its spatial resolution.
- All reanalyses underestimate high wind speeds and overestimate low wind speeds.
- The evaluation of different scores (Hit rate, False alarm ratio, threat score, odds ratio, EDI and SEDI) shows a good performance of all regional reanalyses, in particular for extreme events.
- The new products of SMHI and UKMO show a significant improvement compared to the data sets from EURO4M.
- To avoid strong local effects, the regional reanalysis data sets could be employed to investigate relative wind speeds instead of absolute values (for example using wind percentiles), this would improve the performance.
- The comparison with tower measurements show significant better results of the regional systems compared to the global reanalyses ERA-Interim and ERA20C at tower locations over sea.
- The annual cycle of wind speed is reproduced well by all regional reanalysis systems for different heights.
- The analysis of the ensembles COSMO-REA12 and UM show an increased spread in summer for both models.
- Averaged over 5 years and all selected stations, the spread of the ensemble systems UERRA, UM and COSMO-REA12 is on daily scale of factor 2.7, 6.6, 8 smaller than the RMSE of the ensemble mean.
- Analysis of Brier score shows an advantage for COSMO-REA12 over the UM ensemble, due to a better reliability and resolution of COSMO-REA12.

# 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 (E-OBS) based on a dense network of stations covering Europe is used to assess reanalysis results for their similarity. The comparison is made by aggregating results of the observational dataset and the reanalysis in space and time, providing both maps where time-averaged quantities are compared and graphs showing the

temporal evolution of quantities averaged over selected regions. Using estimates of the uncertainty associated with the gridding of observed data, 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.

**Advantages**

The aggregation of data over selected regions in 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 reanalyses. 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.

**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.

**Value of method**

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

## 4.2 Examples of application – Climate indices

### Investigated spatial and temporal scale

Daily minimum, mean and maximum temperature data are used. The reanalyses data fields over Europe are considered, but only land based gridcells, concerning the land-sea mask of E-OBS, are included for the evaluation.

### Used observations and investigated reanalyses

For each day in the common period, reanalysis data sets from SMHI and UKMO have been downloaded from the ECMWF MARS archive. These data sets hold the first 6 forecast steps

for each 6-hourly run and together make 24 hourly values. Daily minimum temperatures are derived by taking the minimum value from the parameter 'minimum 2-meter temperature since previous post-processing' from the 24 values between 0-0 UT for each day. Similarly for maximum temperature by taking the maximum value from the parameter 'maximum 2-meter temperature since previous post-processing'. Daily mean temperature is derived by taking the mean value from 24 hourly values of the parameter '2-meter temperature'.
The daily values are regridded to a common 0.25˚x0.25˚ regular latitude-longitude grid (using bilinear interpolation). The land-sea mask from E-OBS is applied as well to compare only land-based grid cells, since over large water bodies no observational data is available in E-OBS.
A range of Climate Impact Indices based on minimum and maximum temperature are calculated using a mixture of the python icclim package.
At the time of evaluation, the common continuous period between the SMHI, UKMO and COSMO reanalysis and the E-OBS data spans the period 20050101-20101231, although unforeseen errors in extracting data from Mars (which was noticed too late to fix in time for this deliverable) limited the comparison to only the years 2005 and 2010. With only 2 years of data, the number of indices is limited since no climatology can be determined for indices based on percentile values. The comparisons in this part of the report are based on daily data for these two years only. Available temperature indices are:

- Frost days (FD, number of days with minimum temperature < 0°C)
- Tropical nights (TR, number of days minimum temperature > 20°C)
- Maximum number of consecutive frost days (CFD)
- Summer days (SU, number of days with maximum temperature > 25°C)
- Ice days (ID, number of days with maximum temperature <0°C)
- Maximum number of consecutive summer days (CSU)

With the limited timespan of data available for this comparison, a view on the probability distribution of extreme values is difficult to establish. However, what can be done is by lumping-in data (averaged over a selected area) for the summer or the winter season. This approach is used here. Another option is to simple select, for each yearday, the maximum or minimum value per grid box. The comparison of these extremes from the reanalysis data with observational data gives some indication of how extreme values might behave. This approach is also included. The indices are determined per year, for winter (DJF), summer (JJA) and, where appropriate, summer half-year (AMJJAS).
The version of E-OBS used for the analysis is the ensemble-mean of E-OBS v15.0e_beta.

## Final results:

The assessment and quantification of uncertainties is crucial for the interpretation of the regional reanalysis products which are produced in the UERRA project. The proper use in applications and downstream services hinges on the knowledge of the quality of the reanalysis and the representation of uncertainties. In this deliverable, the information content of the regional reanalyses and their uncertainties are statistically assessed by comparison against observation-based independent datasets at the user relevant scales. The reference dataset used here is the gridded dataset based on the pan-European high-density station series E-OBS. In the UERRA project, an ensemble-based uncertainty of E-OBS is computed. Where applicable, the spread in the ensemble will be used to compare against the spread in reanalysis ensemble.
The comparison between observation-based E-OBS and the reanalysis products of SMHI, UKMO and DWD will be done using the perspective of the so-called Climate Impact Indices.

These are indices which are defined aiming to quantify impacts of weather and climate rather than average quantities. An often used subset of these indices follows the definitions recommended by the CCl/CLIVAR/JCOMM Expert Team on Climate Change Detection and Indices (ETCCDI). Nevertheless, more common metrics like simple averages are used where appropriate. The UKMO and COSMO reanalyses provide an ensemble of realizations for the reanalysis. Using the similar Climate Impact Indices, the spread in the reanalysis ensembles is compared against the spread in the E-OBS ensemble.

**A critical view on E-OBS**

Before a comparison against the observational E-OBS dataset makes sense, quality and reliability issues with E-OBS need to be made clear. The reliability of any observational dataset relates to the amount of stations and the stations density which is used in the gridding process. For E-OBS, the number of stations which have daily maximum temperature data spanning the complete period of 20050101 to 20101231 is 2992 (sourced from the 'blended' stations which is what E-OBS is based on). The distribution of these stations over Europe is very inhomogeneous. Figure 4.2.1 shows a map of the station availability. The number of stations with daily minimum temperature is a little higher at 3015.
Figure 4.2.1 shows that there are hardly any stations over Poland, Belarus, Bulgaria, the mainland of Turkey, Greece, the Middle East and northern Africa for this period. Density is low for the Ukraine, Russia (especially north western Russia), Iceland, Denmark and Italy. Differences between the reanalysis and E-OBS over these regions should be treated with some caution.

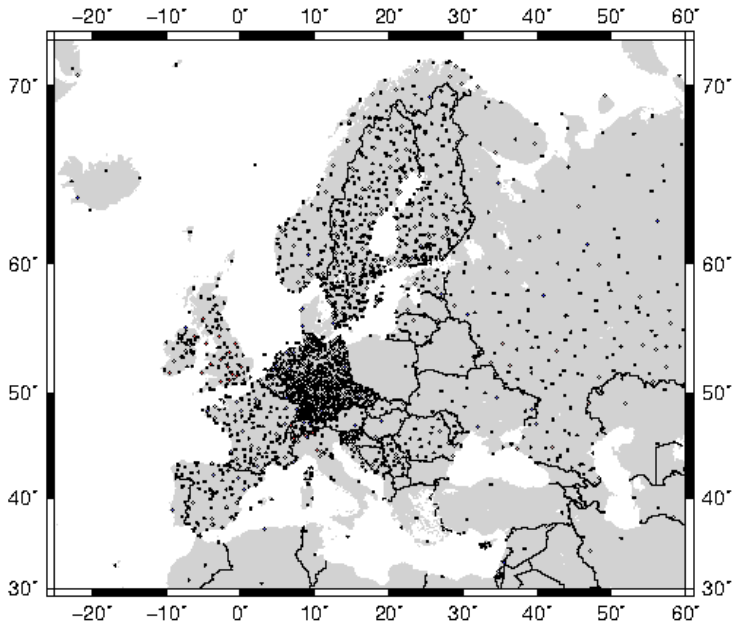


**Figure 4.2.1:** Map showing the stations which have daily maximum temperature series for the complete period January 2005 to December 2010.

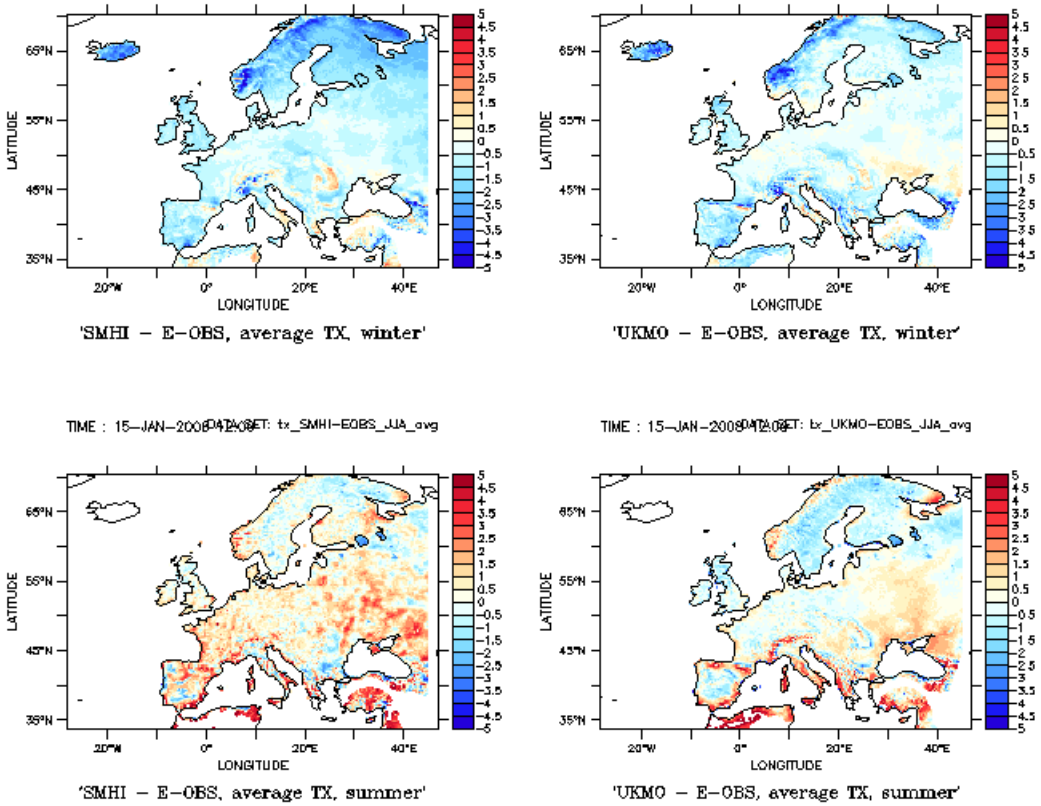


**Figure 4.2.2:** Simple averages over winter (top row) and summer (bottom row) of the difference in daily maximum temperature between the SMHI reanalysis (left column) and E-OBS and the UKMO reanalysis (right column) and E-OBS. Units are °C.

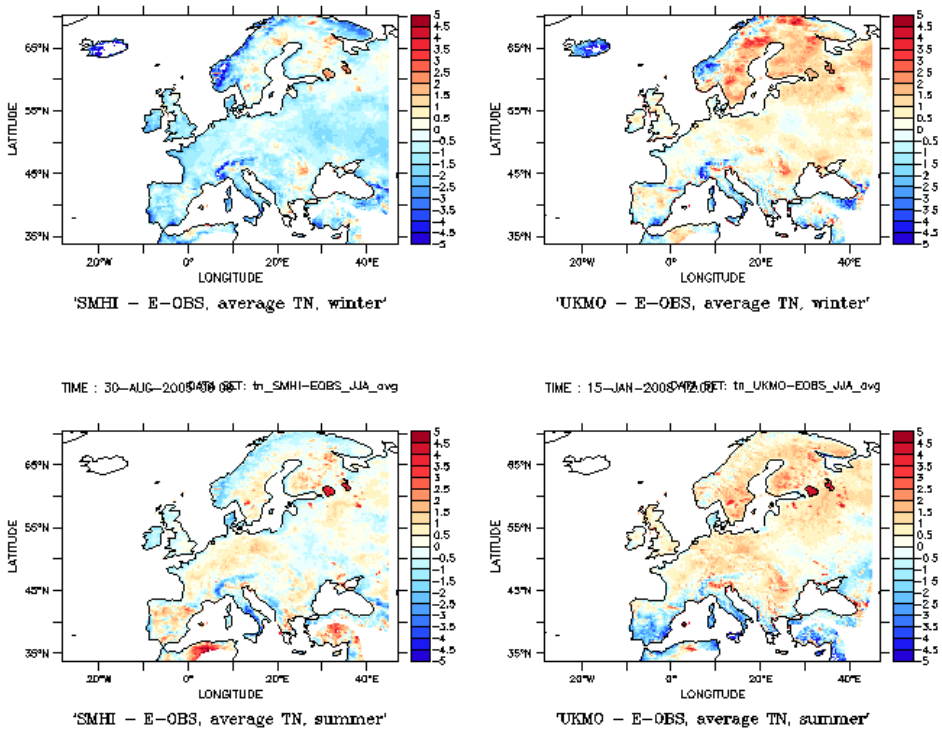


**Figure 4.2.3:** Simple averages over winter (top row) and summer (bottom row) of the difference in daily minimum temperature between the SMHI reanalysis (left column) and E-OBS and the UKMO reanalysis (right column) and E-OBS. Units are °C.

A first comparison between reanalyses and E-OBS is shown in Figure 4.2.2 and Figure 4.2.3. The large differences in e.g. Turkey are related to the station density. E-OBS has barely any stations in Turkey making the dataset less reliable in that area. The same is true for northern Africa.

In these simple winter and summer means of the difference between reanalysis and observations, we expect that 'bulls-eyes' of rather large but localized offsets relate to an issue with the observations rather than a problem with the reanalysis. Such areas are the south coast of Spain (visible in winter TN and TX) and bulls-eyes over Turkey, Romania and the Ukraine. Clearly, the interpolation of temperatures over the lakes NE of St. Petersburg in E-OBS is not to be trusted. Note also that in the maps of Figure 4.2.2 and 3, strong topographic features are recognizable (Alpine region, Pyrenees, Norway, Carpathians) which is partially be related to the use of different topography maps in E-OBS and the reanalysis.

Quality issues with the station data and low station density which relate to localized areas where E-OBS should not be trusted are evident in maps of the standard deviation of daily values for daily maximum and minimum temperature (Figure 4.2.4). The rationale for this criterion is that localized areas of large standard deviation – not coinciding with complex topography – suggest a lack of homogeneity with surrounding areas. Due to its nature of being dynamically consistent, the reanalysis is unlikely to have a localized region that has a poor relation to surrounding areas. The E-OBS lacks any dynamical constraints on spatial homogeneity, making it the more likely candidate to show a 'bulls-eye' at the place where a station has poor quality data.

For daily maximum temperature, the high station density over south Sweden, Ireland, Germany, the Czech Republic, Slovenia and the Netherlands relate to low values for this difference. Over most of Europe, the standard deviation is low. Exceptions are areas with complex topography (Alpine region, Pyrenees, Norway). Areas where stations are located at the coast and not in the interior (Iceland, Sardinia) have large standard deviation too. Areas with low station density (Turkey, northern Africa) show-up as areas with high standard deviation. For daily minimum temperature, the situation is somewhat more complex. Again we see that station density and standard deviation are related, but the picture is quite different from what we see for daily maximum temperature. Standard deviation for minimum temperature is over-all more noisy with high values over all of Scandinavia and eastern Europe.

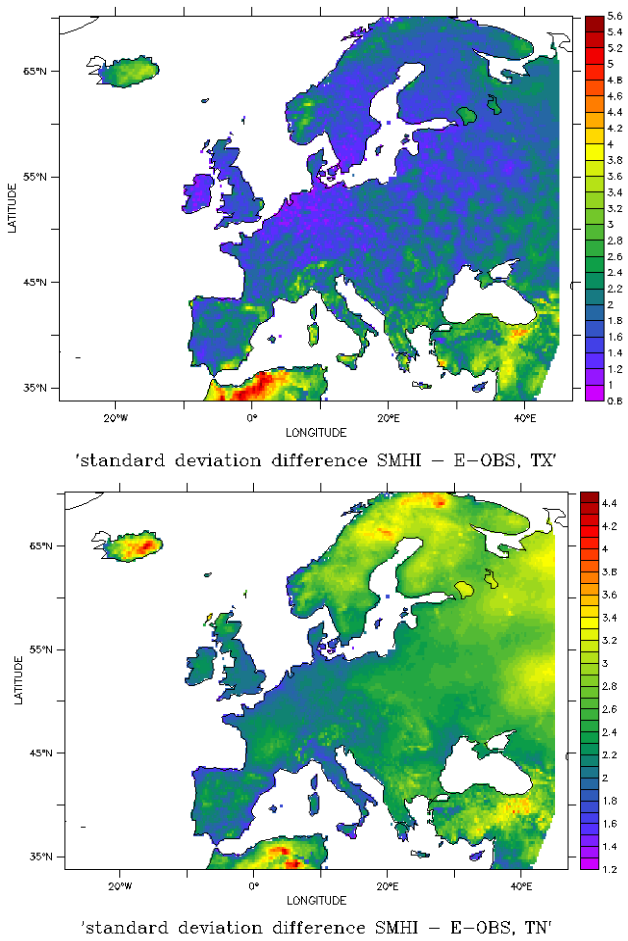


**Figure 4.2.4:** Maps of the standard deviation of the difference between daily maximum (left) and minimum (right) temperature between the SMHI reanalysis and E-OBS. Units are °C.

### Comparing the annual cycle

Figure 4.2.5 and 4.2.6 show for four selected areas (Balkan, Eastern Europe, Iberia and Scandinavia) the area-averaged annual cycle for daily minimum and maximum temperature respectively, based on the common period for the two reanalyses and E-OBS. The mean value is shown plus one standard deviation.

Figure 4.2.5 shows that the E-OBS and UKMO reanalysis agree to a fairly high degree for the Balkan and Eastern European region. The similarity with the SMHI reanalysis is high for summer, but deviates more clearly for the other seasons. For the Iberian Peninsula, UKMO is cooler than E-OBS in summer and the SMHI reanalysis is very similar to E-OBS while for the colder seasons, the similarity between UKMO and E-OBS is strong. For Scandinavia, the UKMO and SMHI reanalysis are both very similar to E-OBS. In general, the standard deviation of the seasonal cycle in daily minimum temperature is comparable between the reanalyses and E-OBS.

For daily maximum temperature (Figure 4.2.6), the similarity between the reanalyses and E-OBS is strong, for all regions and all seasons except for the Iberian Peninsula where E-OBS is warmest in the cold season and coldest in the warm season. For daily maximum temperature, the differences between E-OBS and the UKMO and SMHI reanalyses are largest in the cold season.

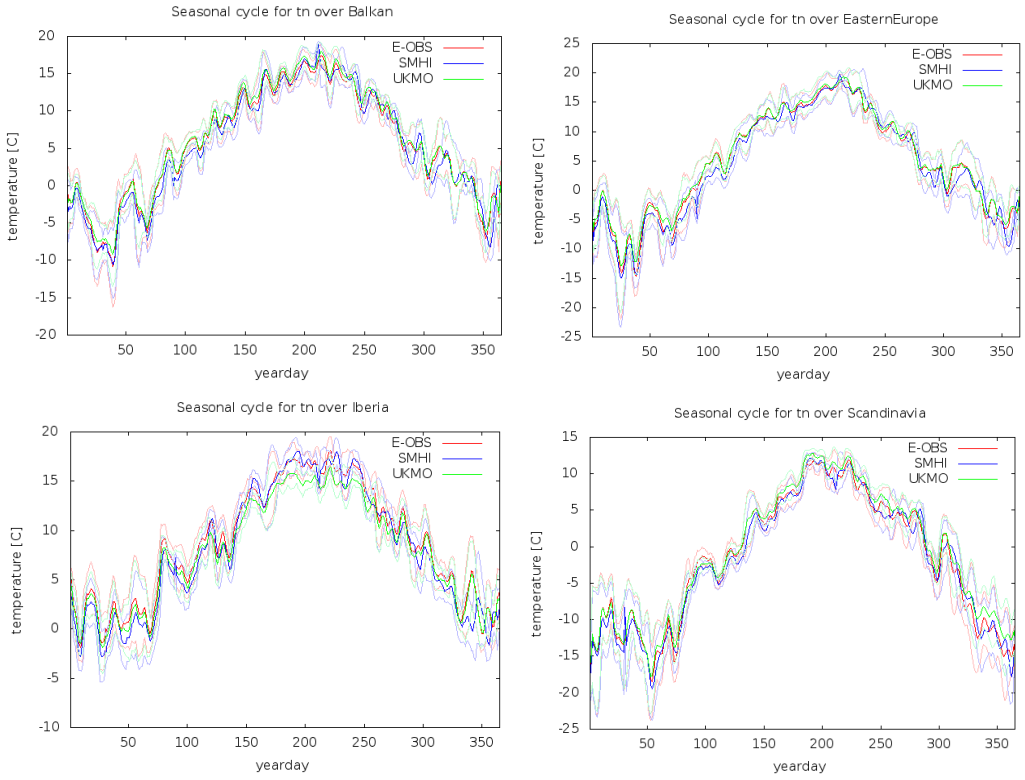


**Figure 4.2.5:** Graphs of the annual cycle in daily minimum temperature over selected areas, including one standard deviation in faint colours. Red colours denote the E-OBS data, blue and green denote the SMHI and UKMO reanalyses respectively.

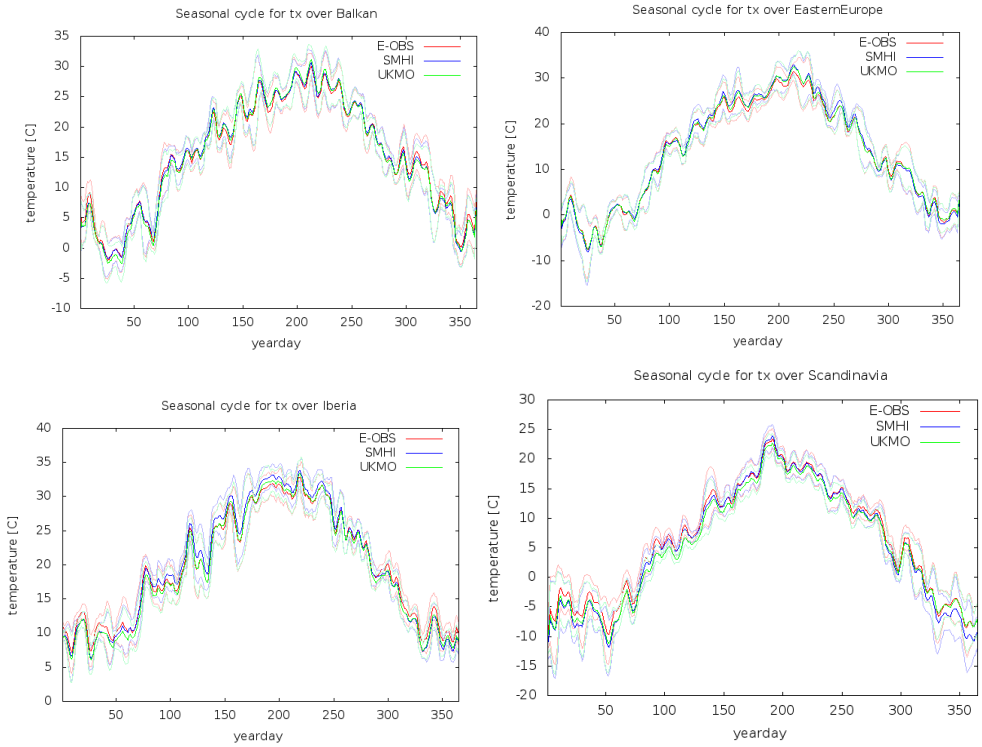


**Figure 4.2.6:** Graphs of the annual cycle in daily maximum temperature over selected areas, including one standard deviation in faint colours. Red colours denote the E-OBS data, blue and green denote the SMHI and UKMO reanalyses respectively.

### Comparing the probability distribution

In order to get a view on the distribution of daily maximum and minimum temperatures in the reanalysis datasets, daily minimum and maximum temperature are averaged over the four selected areas, stratified by season, and a simple histogram is made. The bins are chosen to be 1˚C wide to obtain a smooth distribution. A further smoothing of the distribution is introduced by a straightforward running mean with a window of 3˚C. Figure 4.2.7 shows these histograms for daily minimum temperature in winter (DJF). It shows that the reanalyses and E-OBS over these regions are remarkably similar. The negatively skewed distributions over the Balkan, Eastern Europe and, to a lesser extent, the Scandinavian region are reproduced quite well, while the positively skewed distribution of the Iberian Peninsula is reproduced by the reanalyses as well. Especially the SMHI reanalysis has the distribution shifted to the colder end of the spectrum in comparison with UKMO and E-OBS.
Figure 4.2.8 shows the histograms for daily maximum temperature for summer. Similarly to the histograms in Figure 4.2.7, the general similarity in the shape of the distributions is quite good. Noteworthy is that over Eastern Europe, E-OBS is coldest while SMHI reanalysis is the warmest. Over the Iberian Peninsula, the SMHI reanalysis has the peak of the distribution some 2˚C warmer than E-OBS and UKMO which is the largest deviation shown in these plots. Over Scandinavia, the UKMO reanalysis is coldest.
Figures 9.1 and 9.2 in the Supplementary Information show the histograms for daily minimum temperature in summer and daily maximum temperature in winter respectively.
In order to show a spatially more detailed picture of how the extremes behave the maximum and minimum value of daily maximum temperature is calculated for each yearday and for each grid box, over the two years used in this comparison (2005 and 2010).

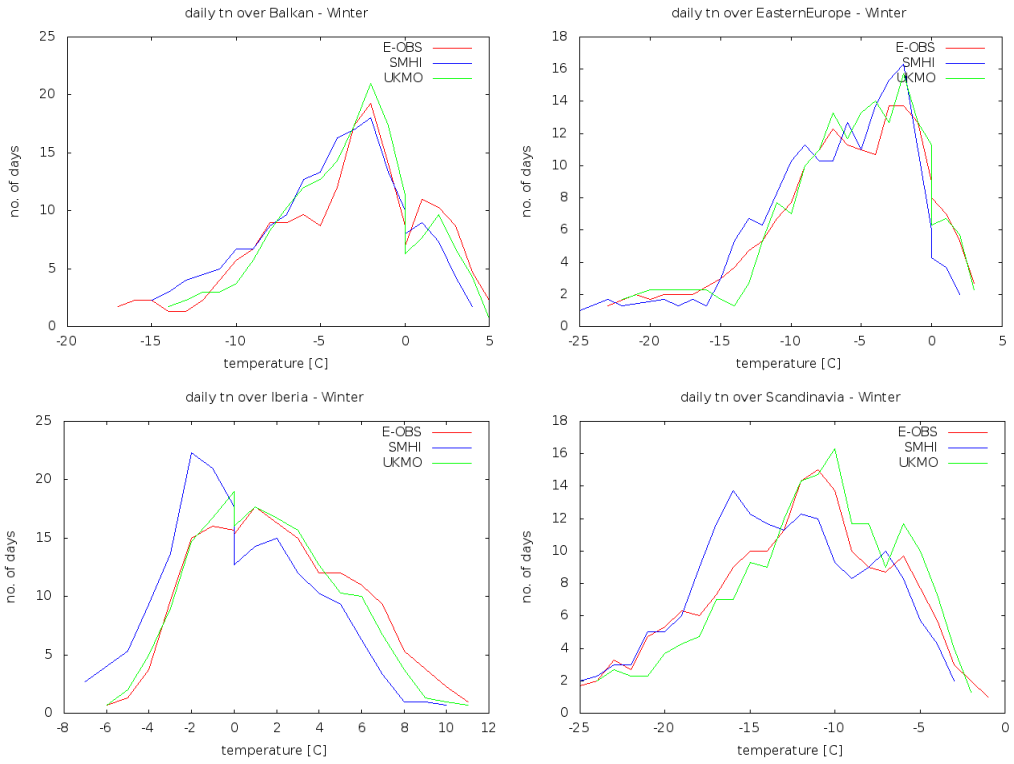


**Figure 4.2.7:** Histograms of daily minimum temperature during winter over selected areas. Red colours denote the E-OBS data, blue and green denote the SMHI and UKMO reanalyses respectively.

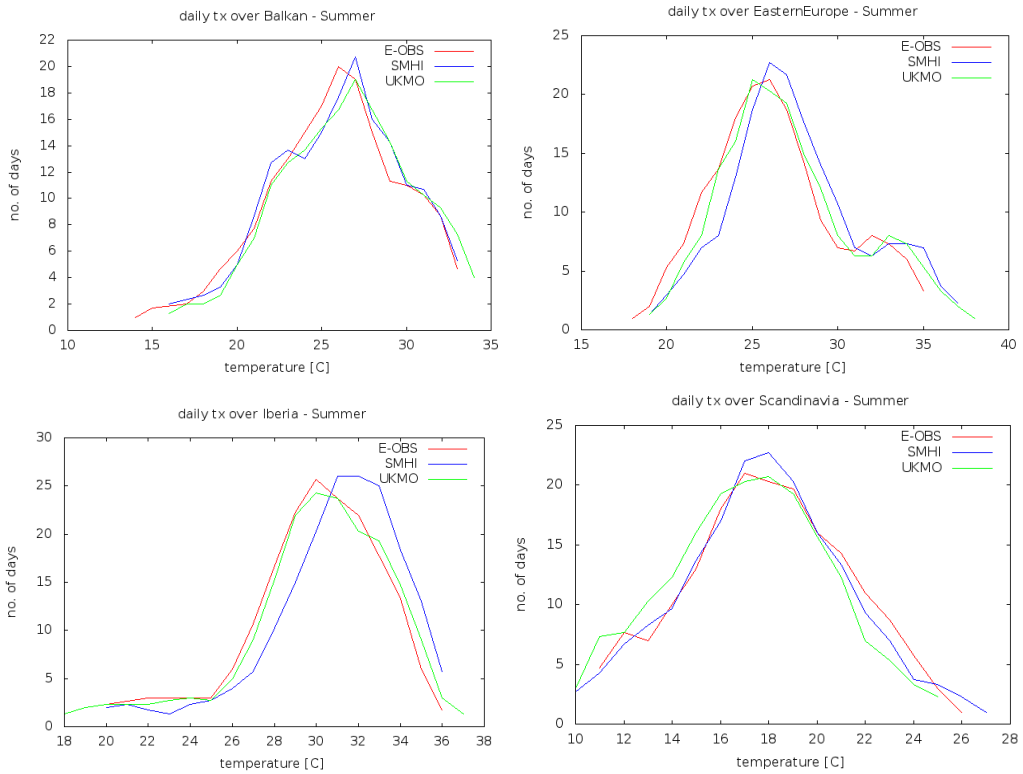


**Figure 4.2.8:** Histograms of daily maximum temperature during summer over selected areas. Red colours denote the E-OBS data; blue and green denote the SMHI and UKMO reanalyses respectively.

These values are then averaged over the winter and the summer season. The difference in the minimum value of daily maximum temperature between reanalyses and observations is shown for winter and summer in Figure 4.2.9. It shows that the minimum in the SMHI reanalysis in winter is colder than E-OBS, but generally not more than 1˚C. The UKMO reanalysis is somewhat warmer but generally not more than 1˚C - 2˚C. In summer, the minimum in the reanalyses is warmer than E-OBS, with the UKMO reanalysis having the smallest deviation from E-OBS.

For daily minimum temperature, the picture is quite different. In winter, the minimum value for daily minimum temperature (Figure 4.2.10) in the SMHI reanalysis is warmer than observations over Sweden and Finland and colder elsewhere – especially over Norway. The UKMO reanalysis is warmer than the observations with the highest differences in NE Europe of up to 5˚C. For summer, the SMHI reanalysis is closer to the observations than the UKMO reanalysis – although the difference can be as large as 3˚C. The minimum value for the UKMO reanalysis is generally warmer than observations, except for the Mediterranean region where UKMO is colder than the observations.

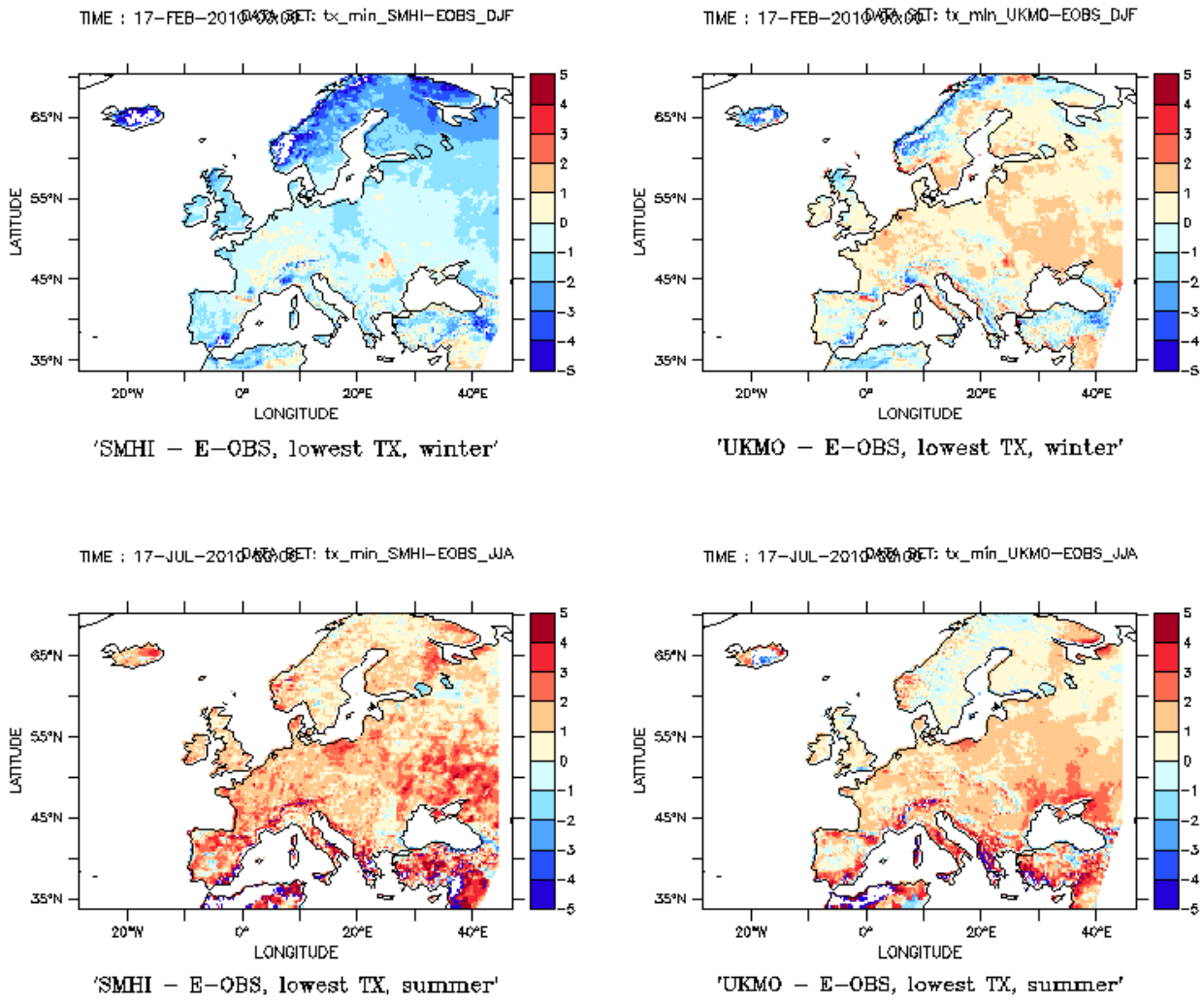


**Figure 4.2.9:** Maps of the difference in minimum value of daily maximum temperature for winter (top row) and summer (bottom row) between the SMHI reanalysis (left column), the UKMO reanalysis (right column) and E-OBS. Units are °C

The maximum values of daily minimum (Figure 4.2.11) and maximum (Figure 4.2.12) temperatures in the reanalyses generally compares well to the observations. The maximum value of daily minimum temperatures in winter is colder over Europe than the observations in the SMHI reanalysis, especially in southern Italy, Greece and Norway. The 90$^{th}$ percentile in the UKMO reanalysis in winter is warmer than the one from E-OBS, especially over Sweden and Finland where UKMO is up to 3˚C warmer than E-OBS. Similar to the SMHI reanalysis, southern Italy, Greece and Norway stand out as regions where the maximum of minimum temperatures are colder than observations. For summer, the SMHI reanalysis generally compares very well to observations. The UKMO reanalysis is also generally warmer than E-OBS over Europe in summer, except the Mediterranean region where it is up to 3˚C cooler than the E-OBS.

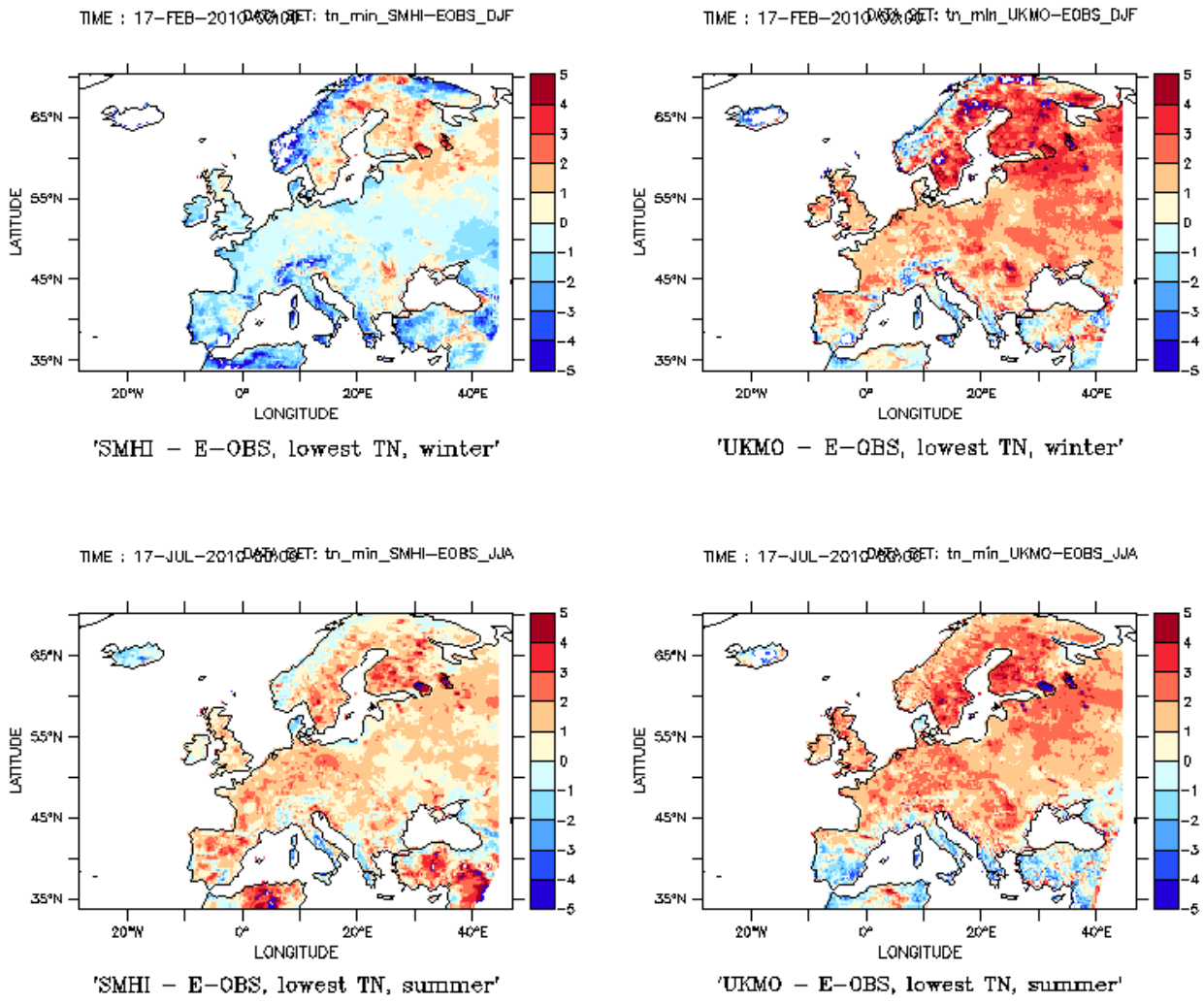


**Figure 4.2.10:** Maps of the difference in minimum values of daily minimum temperature for winter (top row) and summer (bottom row) between the SMHI reanalysis (left column), the UKMO reanalysis (right column) and E-OBS. Units are °C.

The maximum value of daily maximum temperatures (Figure 4.2.12) is colder than the observations for the SMHI reanalysis in winter, while the UKMO reanalysis shows a more mixed picture with vast areas within 0.5˚C of the observations. In summer, which is the more interesting season for this quantity, the situation is a bit more complex. The maximum value of the SMHI reanalysis is generally warmer than observations (except over northern Scandinavia). The difference map is very noisy. Similar to the SMHI reanalysis, the difference in maximum value of TX between the UKMO reanalysis and observations is rather noisy, where the UKMO reanalysis is somewhat colder over the western part of Europe. Between the Black Sea and NE Germany a separation in how the UKMO reanalysis compares to E-OBS is observed. West of this line, the difference with observations is noisy. East of this line, UKMO reanalysis is warmer and the difference is rather smooth.

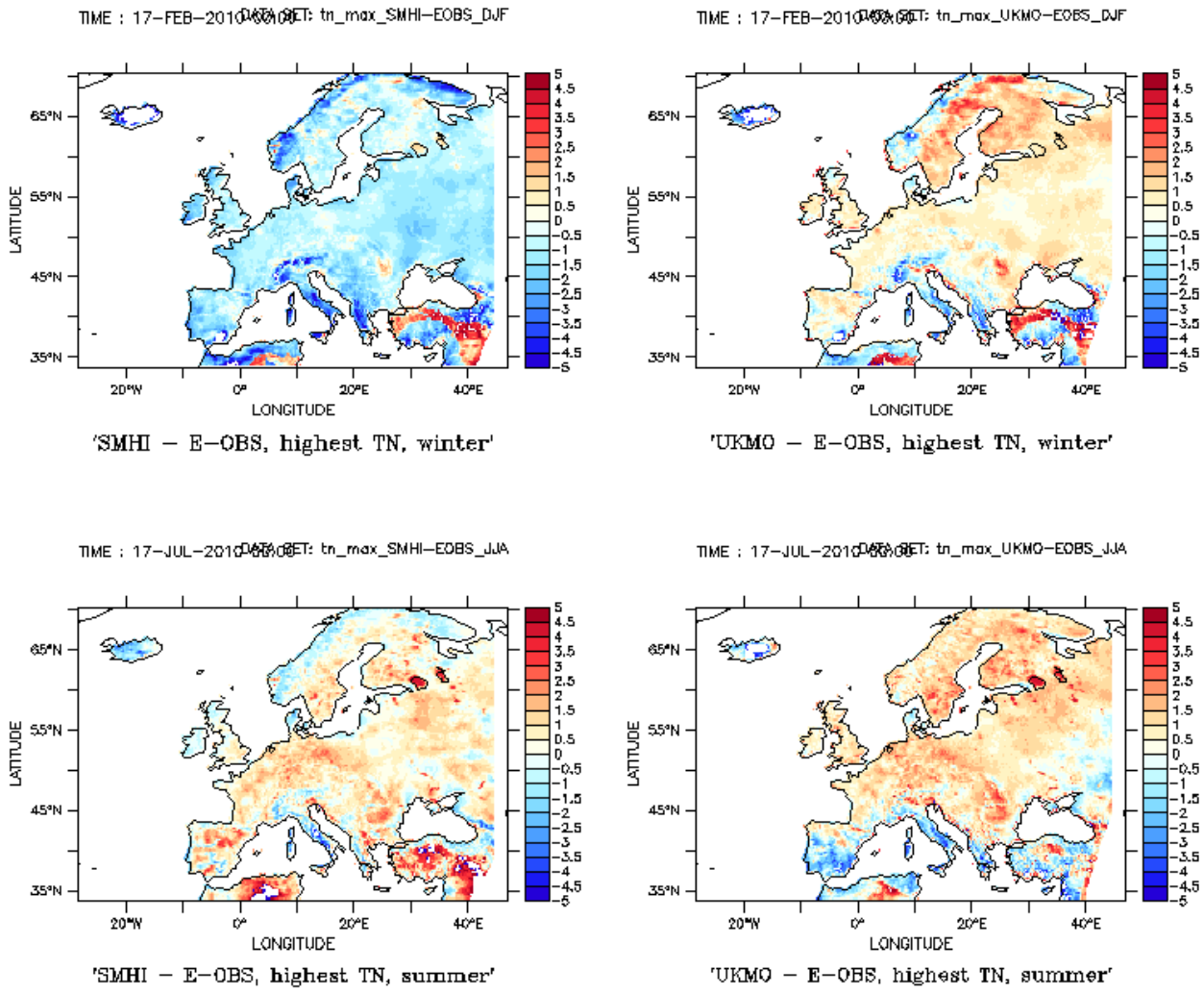


**Figure 4.2.11:** Maps of the difference in 90$^{th}$ percentile of daily minimum temperature for winter (top row) and summer (bottom row) between the SMHI reanalysis (left column), the UKMO reanalysis (right column) and E-OBS. Units are °C.

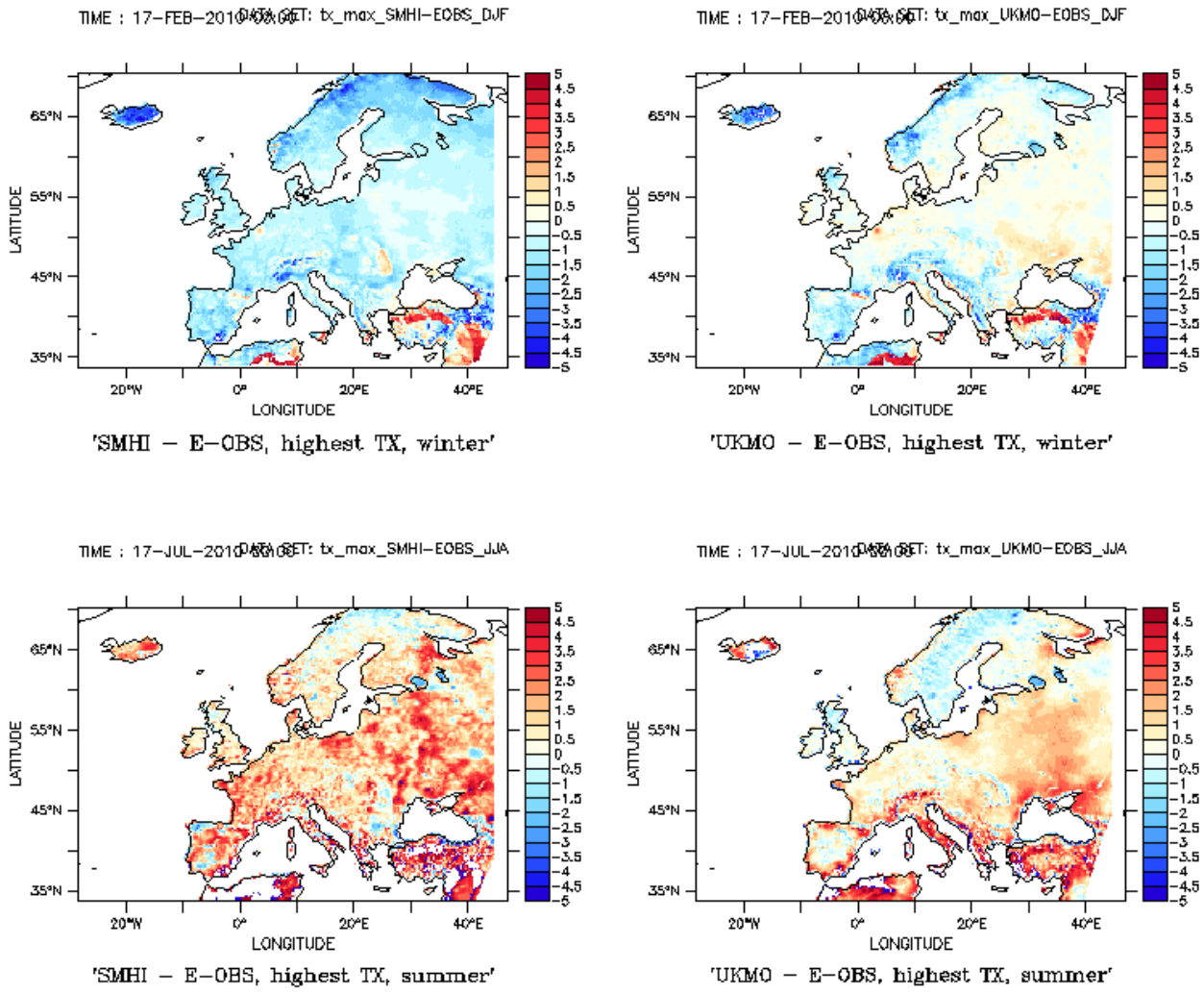


**Figure 4.2.12:** Maps of the difference in highest values of daily maximum temperature for winter (top row) and summer (bottom row) between the SMHI reanalysis (left column), the UKMO reanalysis (right column) and E-OBS. Units are °C.

### Comparison using Climate Impact Indices

For the years 2005 and 2010, the difference in a selection of Climate Impact Indices between the UKMO and SMHI reanalyses and E-OBS is plotted. These maps are available digitally and can be visualized through:
http://euro4mvis.knmi.nl/adagucviewer/?srs=EPSG%3A4326&bbox=-25,7.4122807017543835,75,107.58771929824562&service=http%3A%2F%2Feuro4mvis.knmi.nl%2Fcgi-bin%2Fuerra.cgi%3F&layer=SMHI%2Fyear%2Ffd_smhi_eobsens_year%24image%2Fpng%24true%24temperatureanom%2Fnearest%241%240&selected=0&dims=time$2010-07-01T00:00:00Z&baselayers=world_polygons$world_line
Other indices or periods can be selected by changing the layer using the dropdown menu.

### Frost days

The SMHI reanalysis has, in general, more frost days than E-OBS in most parts of Europe. The difference is largest in northern and Eastern Europe, along the coasts and over the Alpine region. The difference can be as large as 40 days accumulated over the complete year. An example is shown in the left panel of Figure 4.2.13. This is in line with the observation that the minimum temperature and the 10$^{th}$ percentile in minimum temperature in the SMHI reanalysis are lower than in E-OBS in the winter season. The UKMO reanalysis show a more balanced picture with a lower number of frost days in several areas (Figure 4.2.13, right panel).

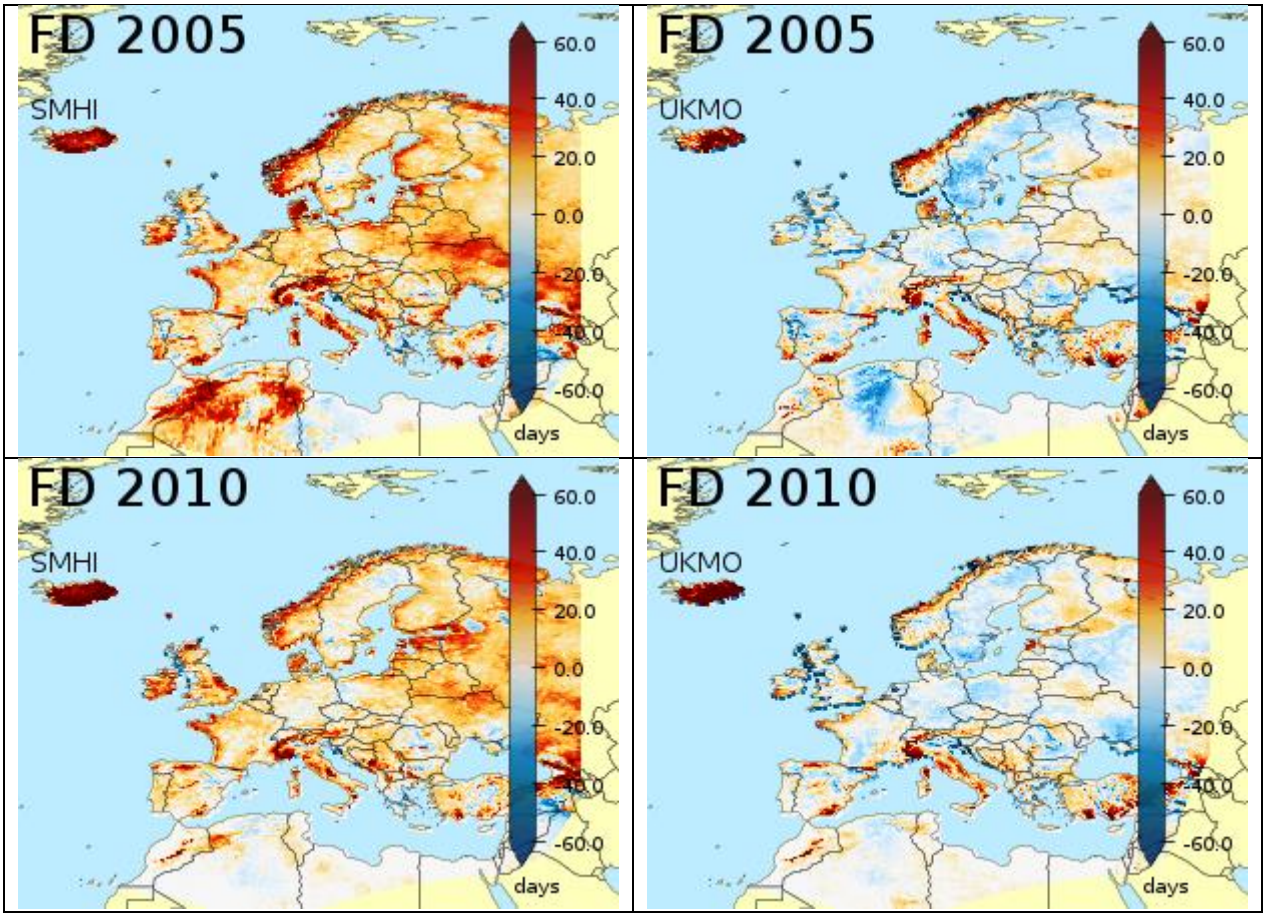


**Figure 4.2.13:** Difference in number of frost days for 2005 and 2010. Left: SMHI reanalysis – E-OBS, right: UKMO reanalysis – E-OBS.

### Tropical nights

The SMHI reanalysis has, in general, more tropical nights than E-OBS with the exception of the southwestern coast of Italy (Figure 4.2.14). The differences over northern Africa are large, but the station density of E-OBS in that area is such that the trust in this result is limited. The situation for the UKMO reanalysis is different with a lower number of tropical nights, except for (north) eastern Europe (Figure 4.2.14, right panel). Interesting is that a sharp contrast exists in the difference maps between Mediterranean coastal grid boxes and more inland areas for the UKMO reanalysis.

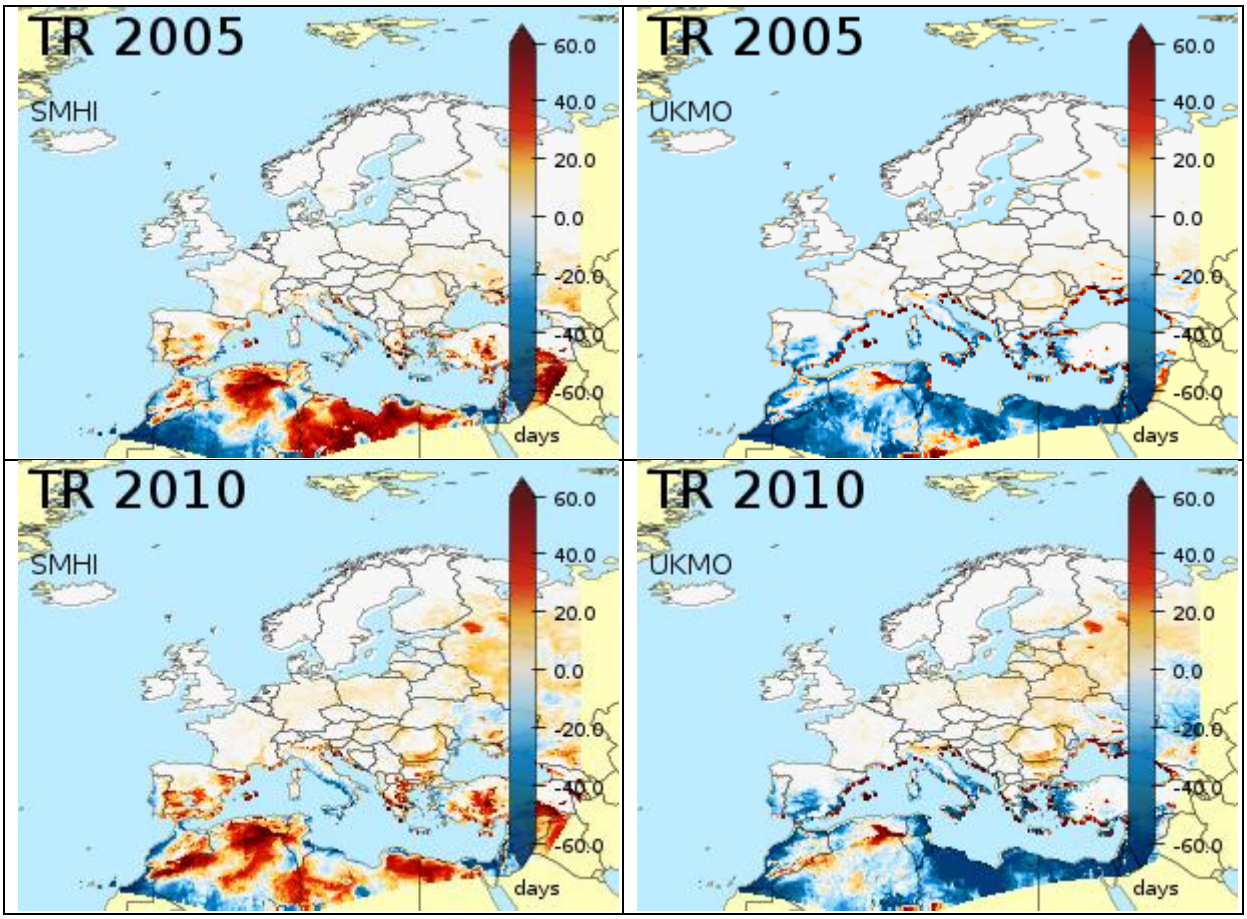


**Figure 4.2.14:** Difference in number of tropical nights in 2005 and 2010. Left: SMHI reanalysis – E-OBS, right: UKMO reanalysis – E-OBS.

### Consecutive frost days

The result for the maximum number of consecutive frost days is similar to frost days for SMHI. This number is larger in SMHI for most of the years and domain (Figure 4.2.15, left), indicating lower minimum temperatures in the winter period where the maximum number of consecutive frost days is expected. The UKMO reanalysis has in general a larger value for the maximum number of consecutive frost days in northern and Eastern Europe (see Figure 4.2.15, right panel) compared to E-OBS, but the difference is smaller than what is observed for the SMHI reanalysis.

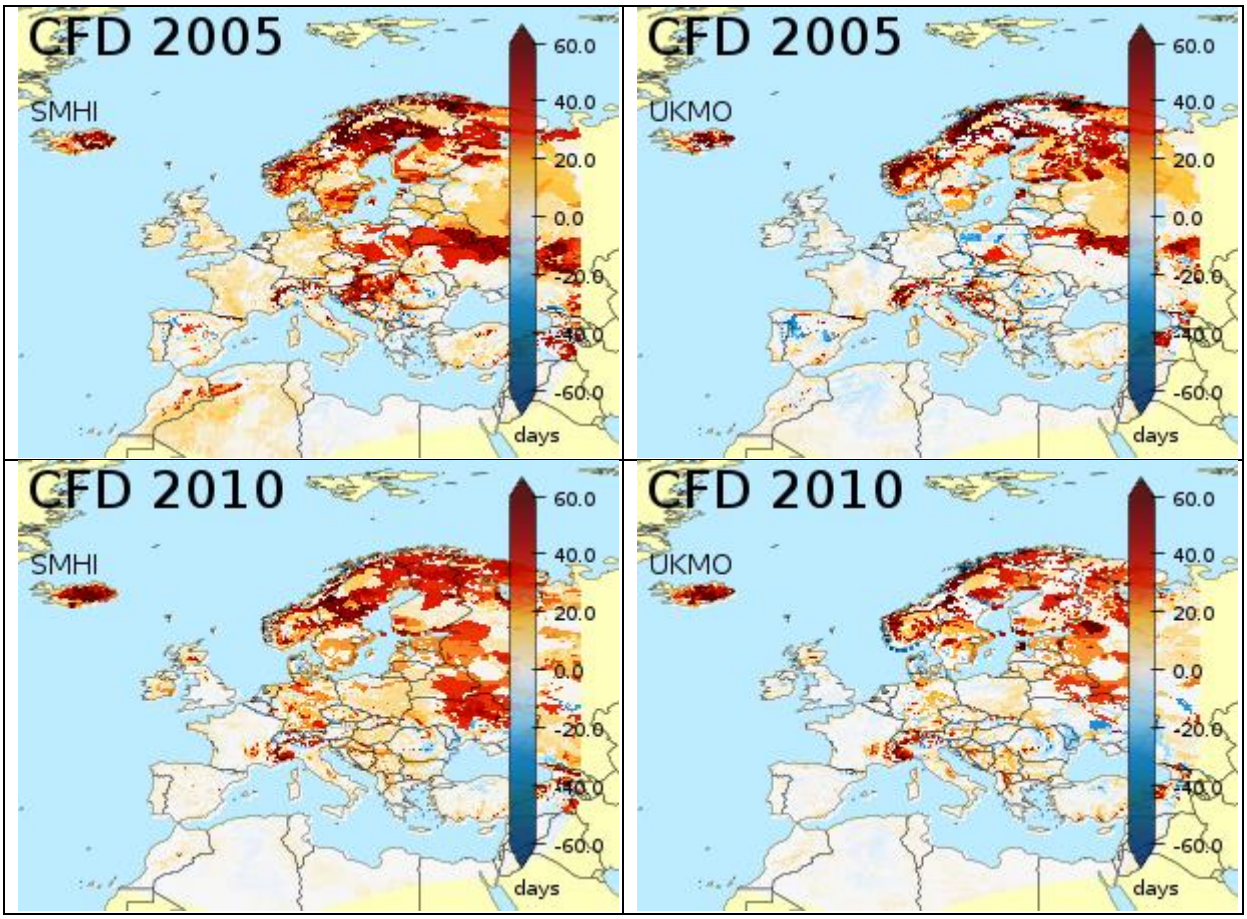


**Figure 4.2.15:** Difference in consecutive number of frost days for 2005 and 2010. Left: SMHI reanalysis - E-OBS, Right: UKMO reanalysis - E-OBS

### Summer days

The situation for the difference in number of summer days is a bit mixed over Europe for the SMHI reanalysis, where the difference map with E-OBS is rather noisy (Figure 4.2.16).

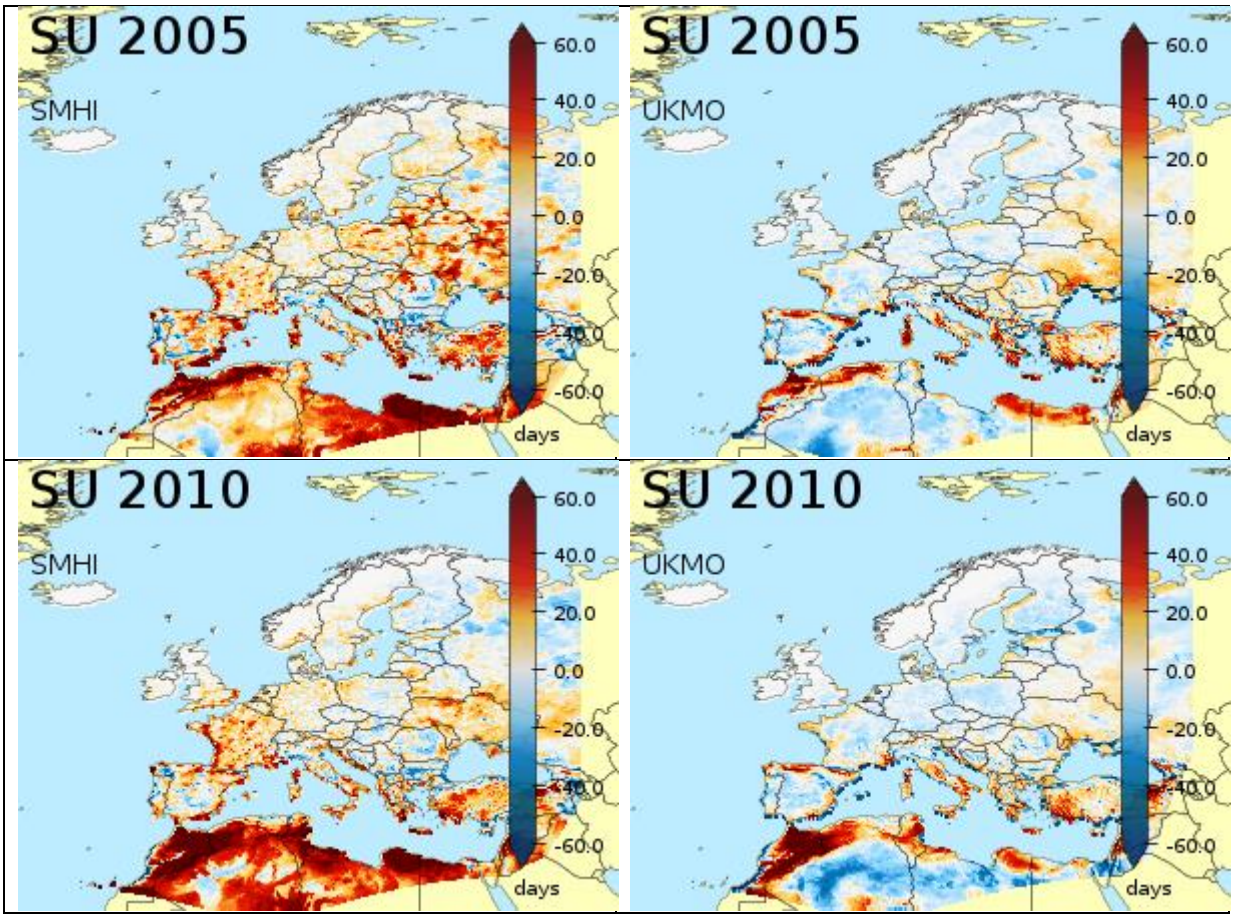


**Figure 4.2.16:** Difference in number of summer days for 2005 and 2010. Left: SMHI reanalysis – E-OBS, Right: UKMO reanalysis – E-OBS.

In general, the SMHI reanalysis shows a higher number of summer days. This contrasts with the UKMO reanalysis, which has some smaller areas where the number of summer days is higher than in the observations, but generally a lower number is seen.

### Ice days

The northern part of Europe sees more ice days in SMHI reanalysis compared to E-OBS (Figure 4.2.17). For the UKMO reanalysis, the number of ice days over the complex topography of Norway is much larger than E-OBS or the SMHI reanalysis. Interesting is that the number of ice days over the UK is overestimated in both the SMHI and UKMO reanalysis, in comparison to E-OBS in 2010, with between 10-20 days.

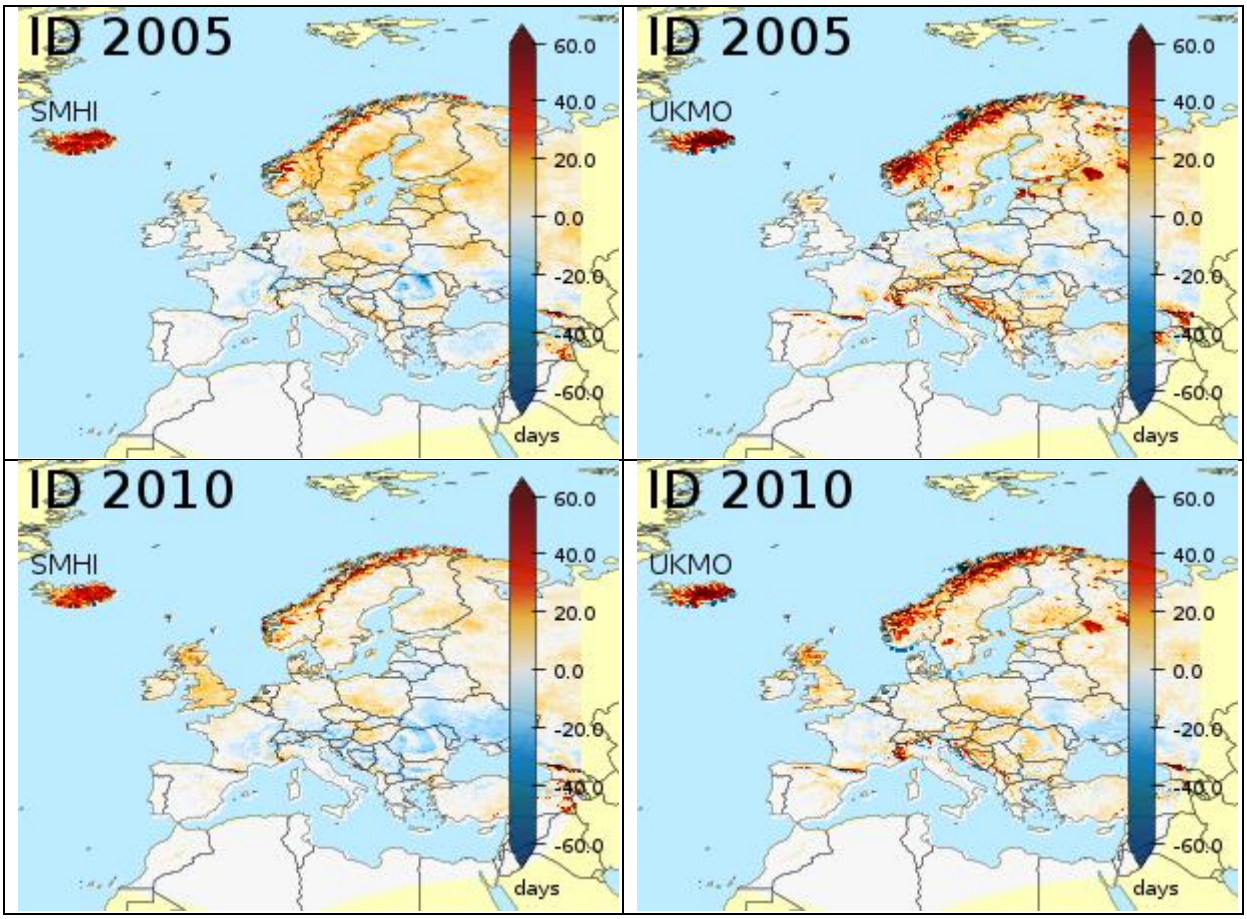


**Figure 4.2.17:** Difference in number of ice days for 2005 and 2010. Left: SMHI reanalysis – E-OBS, Right: UKMO reanalysis – E-OBS.

### Consecutive summer days

The comparison between the reanalyses for the number of consecutive summer days gives a very noisy pattern (Figure 4.2.18). Large, but localized differences with E-OBS are found along the Mediterranean.

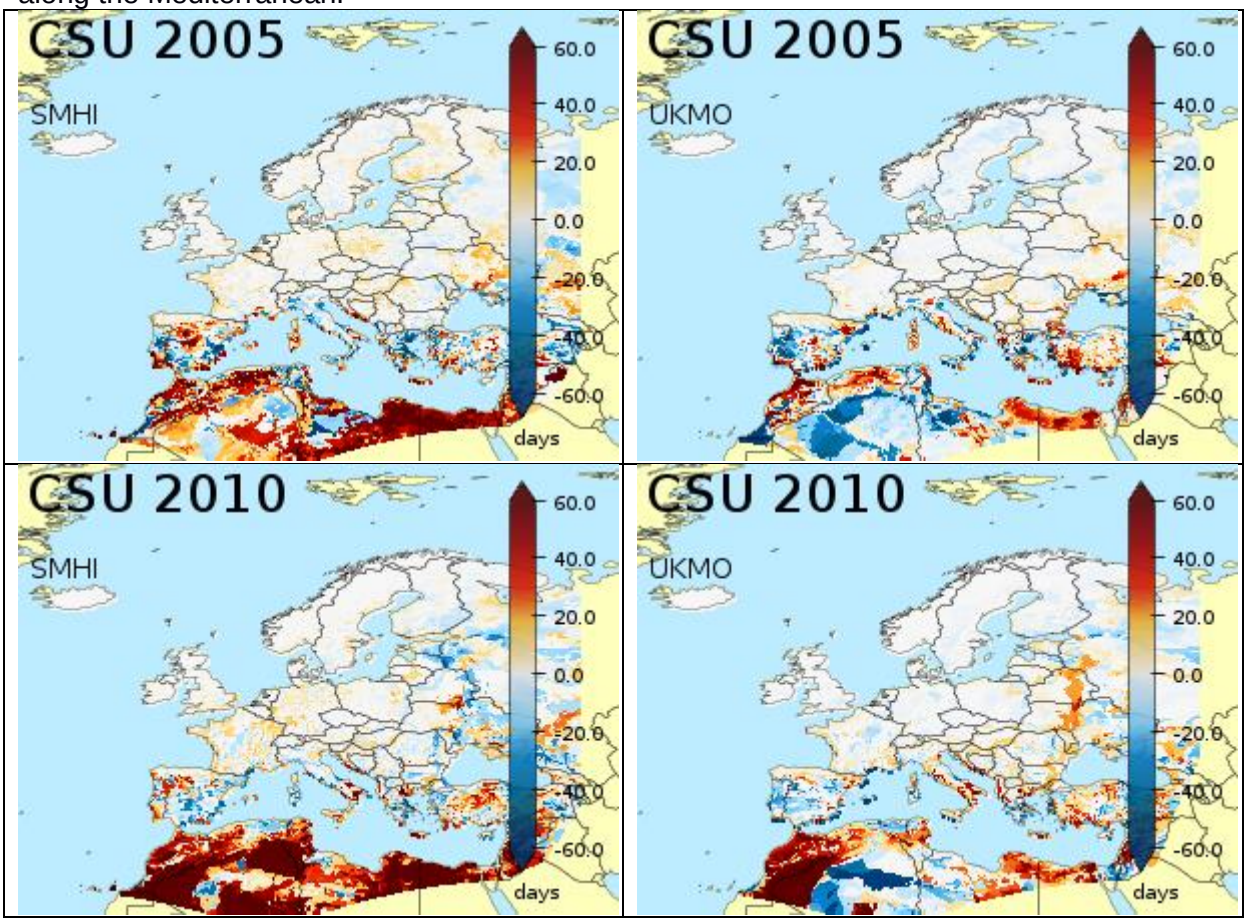


**Figure 4.2.18:** Difference in consecutive number of frost days for 2005 and 2010. Left: SMHI reanalysis - E-OBS, Right: UKMO reanalysis - E-OBS

### Comparison of the ensembles

The UKMO and COSMO reanalyses provide a 20-member ensemble. The information in these ensembles is compared against the 100-member ensemble provided by E-OBS. Comparisons are made in terms of a selected set of Climate Impact Indicators. This set is frost days (FD) and tropical nights (TR) (both based on daily minimum temperature) and ice days (ID) and summer days (SU) (both based on daily maximum temperature). In order to make a meaningful comparison of the ensemble, area-averaged quantities for each ensemble member are calculated and a histogram is produced.
FD and ID are averaged over Sweden, TR and SU are averaged over Spain. The motivation to select these areas is that the station density is good (for Spain) to superb (for Sweden). Furthermore, there are strong gradients in the indices over the countries, which suggest that the spread in the ensemble may be strong too.
The histograms are shown in Figure 4.2.19. These panels make clear that there is no overlap between the histograms neither of the reanalyses nor between E-OBS and the reanalyses.

The difference in terms of the averaged number of the index between the UKMO, COSMO and E-OBS reanalyses is much larger than the spread within the ensemble. This suggests that the spread in the reanalysis ensemble is insufficient to provide a realistic estimate of the uncertainty in the reanalysis.

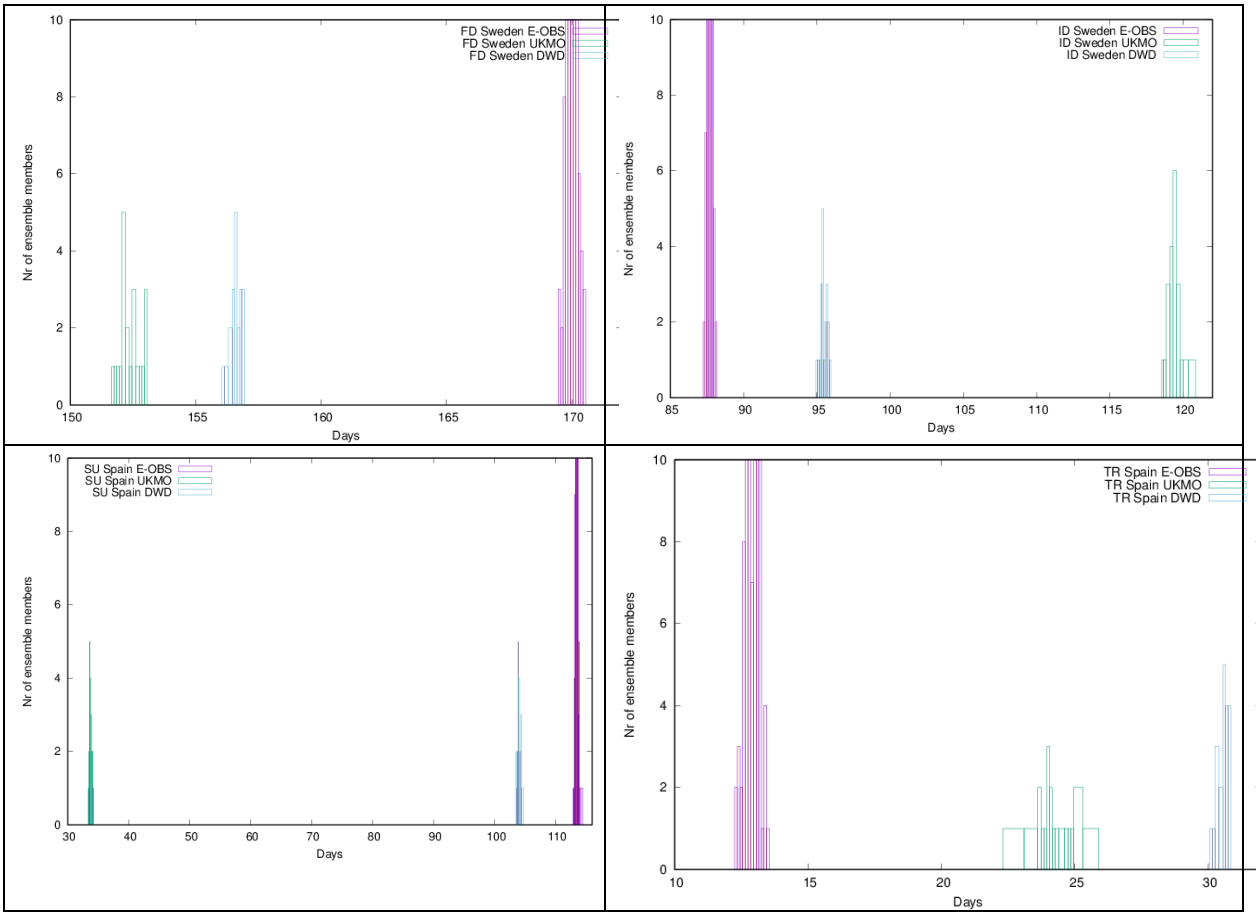


**Figure 4.2.19:** Histograms of the number of ensemble members for the indices FD and ID averaged over Sweden (top row) and SU and TR averaged over Spain (bottom row).

## 4.3 Examples of application – Precipitation

### Investigated spatial and temporal scale

In this evaluation we are focusing on precipitation (daily, i.e. 06-06 UTC) over the Alpine Region and Fennoscandia. All datasets have been post-processed so to have daily precipitation (06-06 UTC) on two grids:

(a) a coarse-resolution grid. 0.25° regular grid (same as the "original" E-OBS grid), which covers Europe

(b) a fine-resolution grid. 5 Km Lambert Azimuthal Equal Area covering the Alpine Region and Fennoscandia

### Used observations

The used observations are listed in Table 1.2 at page 6-7 and include the gridded datasets NGCD, APGD, APGD-ENS and E-OBS (v14.0).

### Investigated reanalyses

All reanalyses data sets of Table 1.1 are investigated in this section.

### 4.3.1 Alpine region – Final results

The study region extends from 2-17.5°E and 43-49°N, comprising the entire mountain range of the European Alps as well as adjacent flatland and smaller hill ranges. The complex topography is a challenge for climate modelling and gridded observational datasets.
During the FP7 EURO4M project (European Reanalysis and Observations For Monitoring) a high resolution dataset based on more than 6000 rain-gauges observations every day was developed, covering the period 1971-2008. This datasets, called Alpine Precipitation Gridded Dataset (APGD), is used here as reference and is analysed in detail in [Isotta et al., 2014]. In the same project, regional reanalyses, downscaling products and a global reanalysis were evaluated [Isotta et al., 2015].
In UERRA, we compare new or further developed datasets within the project and already existing datasets commonly used. A special focus of the evaluation is on uncertainties in regional reanalyses and their scale dependencies. The scale separation is obtained using an additional dataset as reference, named “APGD-ENS” hereafter. Starting from the same observations as APGD, a probabilistic spatial analysis of daily precipitation that is capable of quantifying uncertainties was developed. Instead of a regular grid, daily precipitation is represented for hydrological units of different sizes [Frei et al., 2017].
In the following chapters different indices and scores are discussed. The evaluation period is 2006-2008 only, corresponding to the maximum overlap of all datasets and the reference. The datasets are rescaled to the E-OBS grid (0.25°) and the reference grid (5km ETRS-LAEA). Notice that the UK Met Office ensemble reanalysis is mostly not shown in the evaluation due to a strong overestimation of precipitation amounts and frequencies. The problem was solved but the new dataset was not ready for the evaluation.

#### Mean annual precipitation

Figure 4.3.1.1a shows the mean annual precipitation for datasets rescaled to 0.25° regular grid. The two downscaling products (MESAN and MESCAN) are very detailed, especially MESCAN with a strong topography signal only over the Alps, which stems from the driving model HARMONIE (same pattern). MESAN is closer to the reference regarding the precipitation pattern over the Alps. The performance of the downscaling is dependent on the rain-gauges density. As a consequence, France and Germany are very close to the reference while in Italy, where the station density available for the downscaling is lower, discrepancies are more evident (e.g. Dolomites and Julian Alps).
Regional reanalyses (UKMO, HARMONIE and COSMO6-REA) tend generally to overestimate precipitation but capture well the precipitation pattern. COSMO6-REA is closer to the reference than UKMO and HARMONIE. There are only moderate differences in the Apennine, Dinaric Alps and Massif Central. The higher resolution and the use of non-hydrostatic dynamics may be the main reason for the performance.

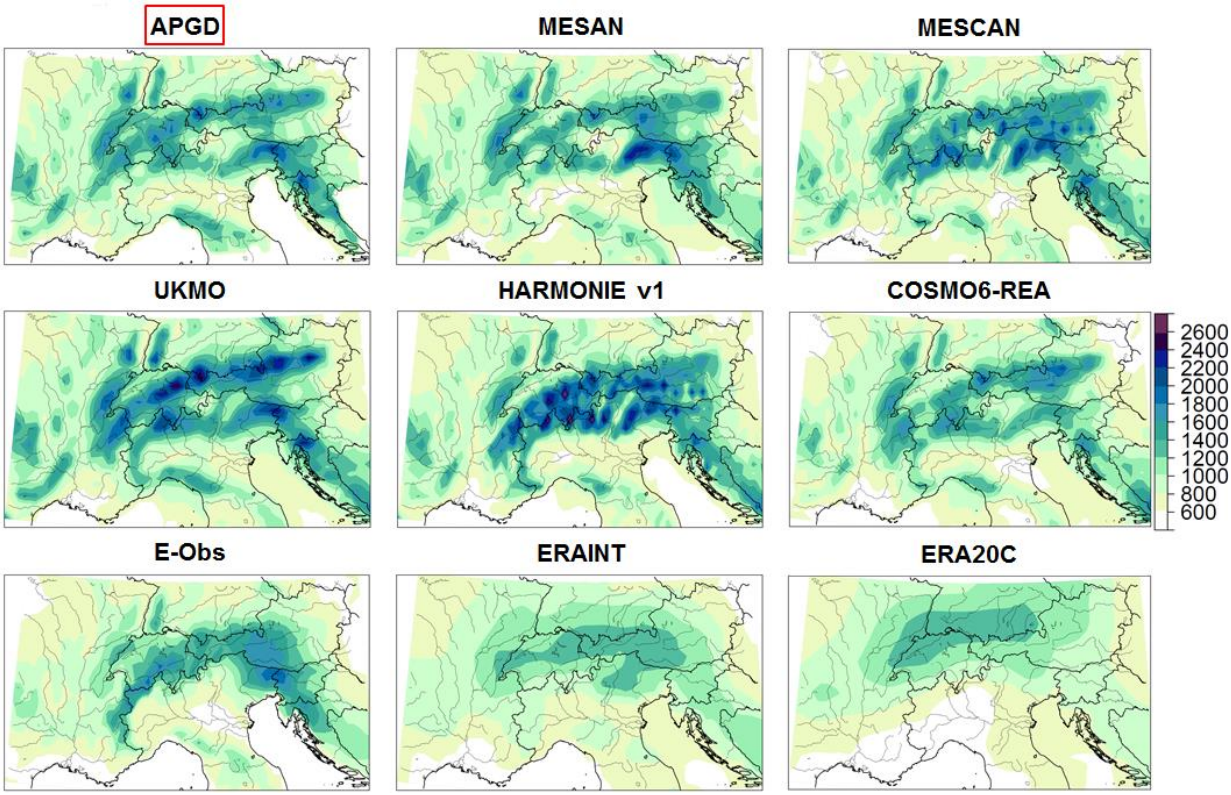


**Figure 4.3.1.1a:** Mean annual precipitation (mm per year, 2006-2008). Datasets rescaled to 0.25° regular grid. Reference: APGD.

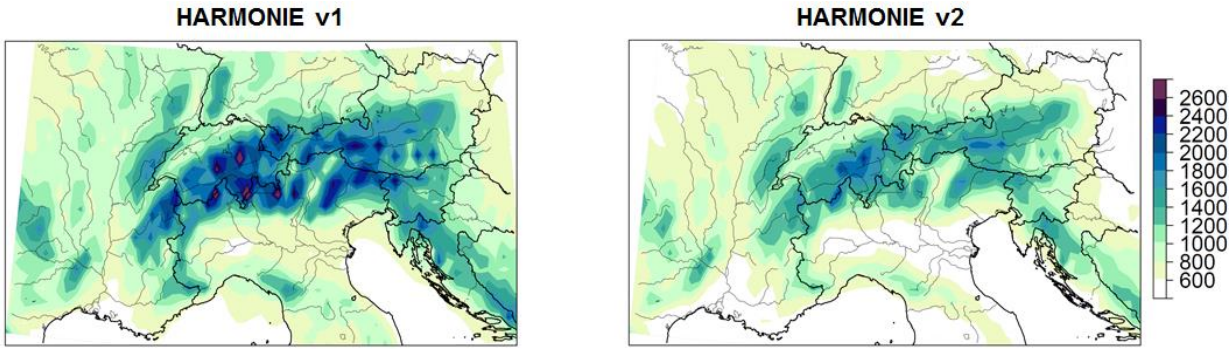


**Figure 4.3.1.1b:** Mean annual precipitation (mm per year, 2006-2008). Datasets rescaled to 0.25° regular grid.

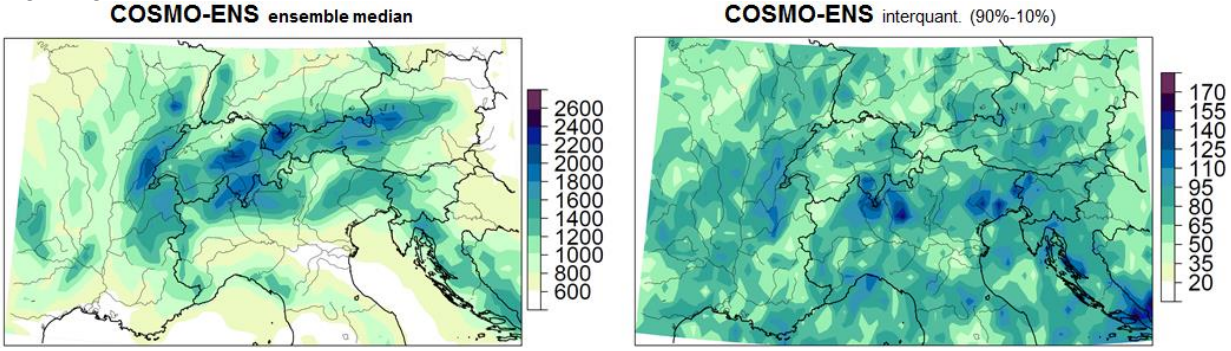


**Figure 4.3.1.1c:** Mean annual precipitation (mm per year, 2006-2008), ensemble mean (left) and interquantile (between the 10% and the 90% quantile). Datasets rescaled to 0.25° grid.

E-OBS, the observational gridded dataset, is not correctly representing the two moist bands at the Alpine rims, placing the moist region over the main ridge. The low station density over the Alps and the interpolation method are the main reasons for this behaviour and are responsible for the coarse resolution of the dataset. For the Alps, regional reanalysis seems to have an added value compared to E-OBS. They better represent precipitation amount and pattern. The advantage of using regional reanalysis is even more evident when compared to global reanalyses (ERAINT and ERA20C), which are of much coarser resolution, resulting in low precipitation amounts and missing of the main pattern.

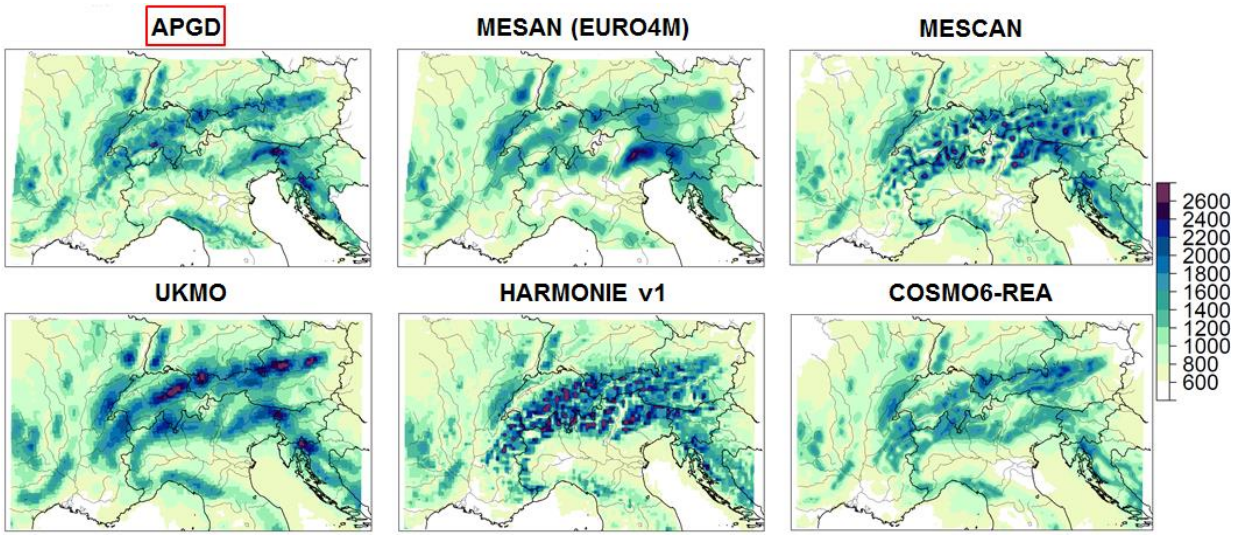


**Figure 4.3.1.2a**: Mean annual precipitation (mm per year, 2006-2008). Datasets rescaled to 5km ETRS-LAEA coordinate system. Reference: APGD.

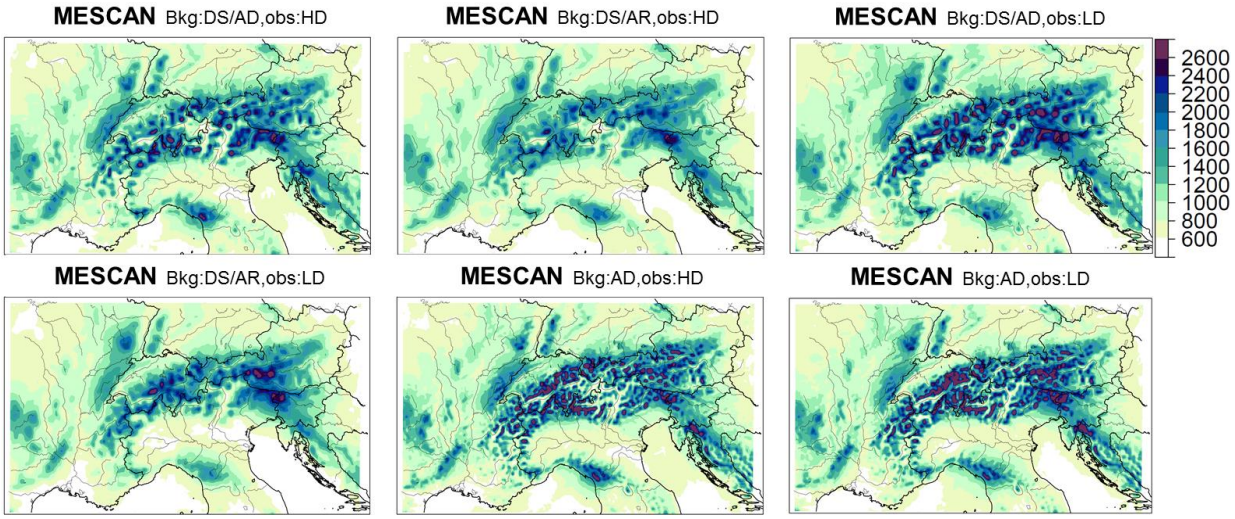


**Figure 4.3.1.2b:** Mean annual precipitation (mm per year, 2006-2008). Datasets rescaled to 5km ETRS-LAEA coordinate system. Bkg=background (AD=ALADIN, AR=ALARO), HD/LD=High Density/Low Density network used. DS=downscaling from HARMONIE at 11km to 5.5km (without DS=background from ALADIN model at 5.5km without downscaling).

A comparison of the two HARMONIE versions (Figure 4.3.1.1b, v1 uses ALADIN physics and v2 ALARO) shows precipitation amounts over the Alps closer to the reference in v2 but too low amounts over the adjacent hills and flatlands, where v1 performs better. Not for all

indices the differences to APGD between the two versions are so clear. For example the 95% quantile is overestimated over the Alps in v1 whereas v2 underestimate it (not shown). If not specified, we are always referring to v1 here.
Figure 4.3.1.1c depicts the ensemble mean and the interquantile of COSMO-ENS. We find the same good pattern reproduction of COSMO6-REA also in the coarser COSMO-ENS ensemble median. The precipitation amounts are slightly overestimated. The interquantile is around 5% of the respective precipitation values.
The mean annual precipitation rescaled to 5km ETRS-LAEA coordinate system is represented in Figure 4.3.1.2a. The differences between MESAN, which is closer to the reference, and MESCAN in the spatial variance are more evident here compared to Figure 4.3.1.1a (both downscalings have nearly the same spatial resolution as the 5km ETRS-LAEA grid). COSMO6-REA shows a slightly less detailed pattern than APGD but still an impressive performance. Also the pattern of UKMO is near to the reference although precipitation amount is overestimated.
For a limited period of five years (2006-2010), six MESCAN versions (see Figure 4.3.1.2b) were calculated based on two background physics (ALADIN or ALARO), differing network density (high or low density) and a version without downscaling where the background is directly derived from a high resolution run of ALADIN at 5.5km.
The strong topographic signal already described for Figure 4.3.1.1a and 4.3.1.2a originating from HARMONIE is especially visible in the simulations with ALADIN as background. The higher spatial variance of ALADIN compared to ALARO (especially for the versions where the background is directly derived from the ALADIN model at 5.5km and the high density network is used) has no confirmation in the reference. The high density network improves precipitation amounts (e.g. no overestimation of precipitation maximum over the Julian Alps).

**Wet-day frequency and 95% quantile**

A correct representation of the wet-day frequency (Figure 4.3.1.3) is challenging. The benefit of a dense network in downscaling datasets is more evident than for the mean annual precipitation (Figure 4.3.1.1a). Both MESAN and MESCAN are too wet over the eastern Alps, but are in good agreement in the surroundings, where the regional reanalyses have too much wet-days, beside near the Mediterranean coasts.
Again, COSMO6-REA shows a remarkably good performance. In contrast to the mean annual precipitation, the direct use of station observation in E-OBS is, as for the downscaling, a clear advantage compared to the regional reanalyses (except COSMO6-REA). The patterns of E-OBS suggest a coarser effective resolution than the 0.25° of the native coordinate system. Evaluating the 95% quantile of the three years period from 2006 to 2008 (Figure 4.3.1.4) leads to similar results as for the mean annual precipitation. The regional reanalyses overestimation is reduced in the downscalings, especially in stations dense regions. E-OBS and the global reanalysis are not able to capture the general precipitation pattern.

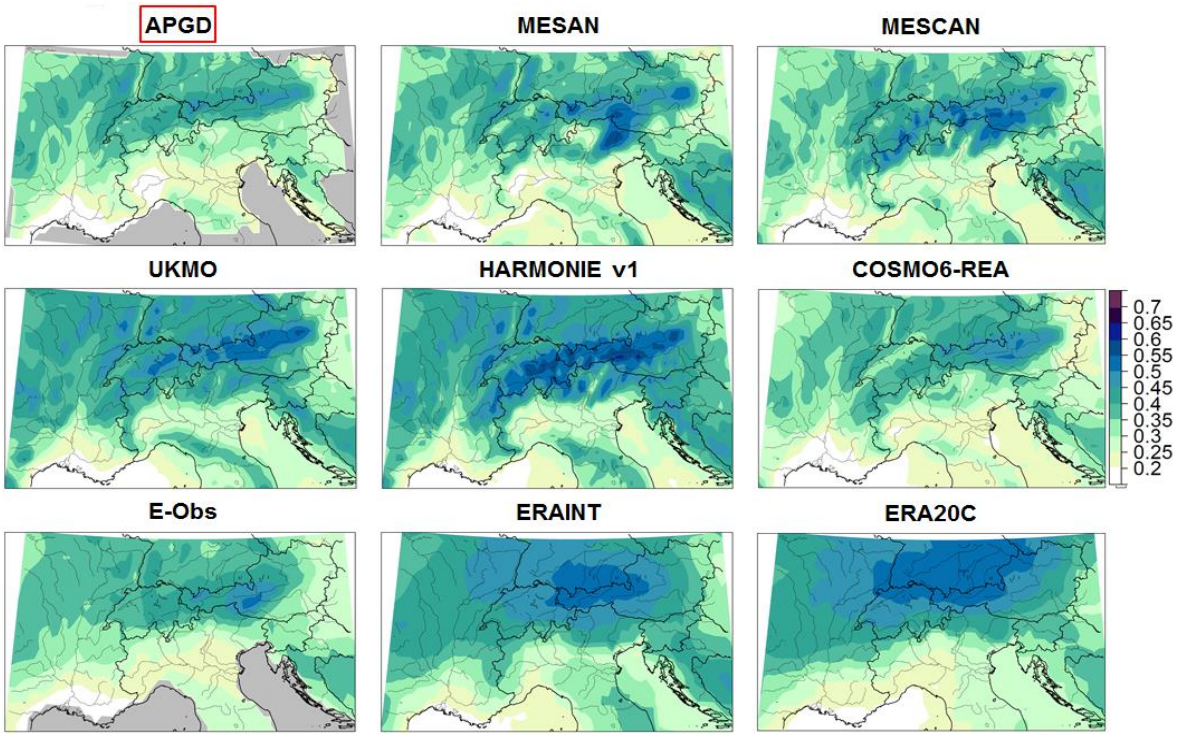


**Figure 4.3.1.3:** Annual frequency of wet days (≥ 1mm/d, fraction, 2006-2008). Datasets rescaled to 0.25° regular grid. Reference: APGD.

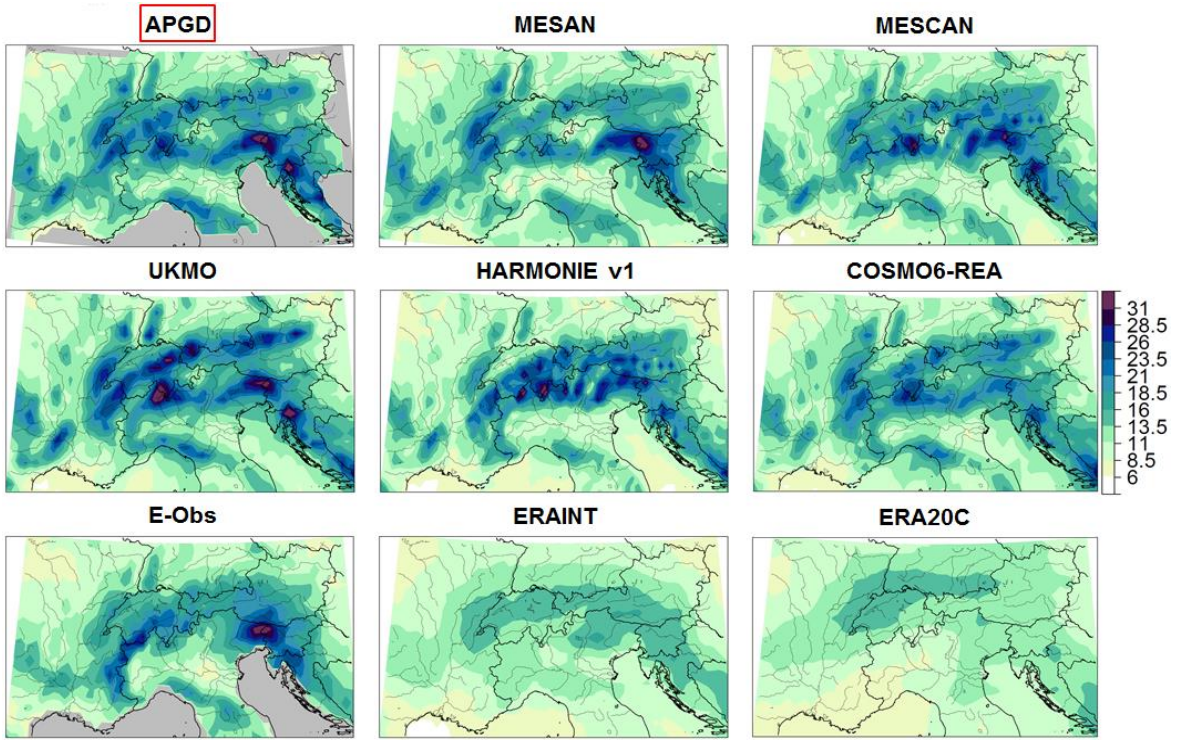


**Figure 4.3.1.4:** 95% quantile of daily precipitation (mm/d, 2006-2008). Datasets rescaled to 0.25° regular grid. Reference: APGD.

### Root Mean Square Error (RMSE)

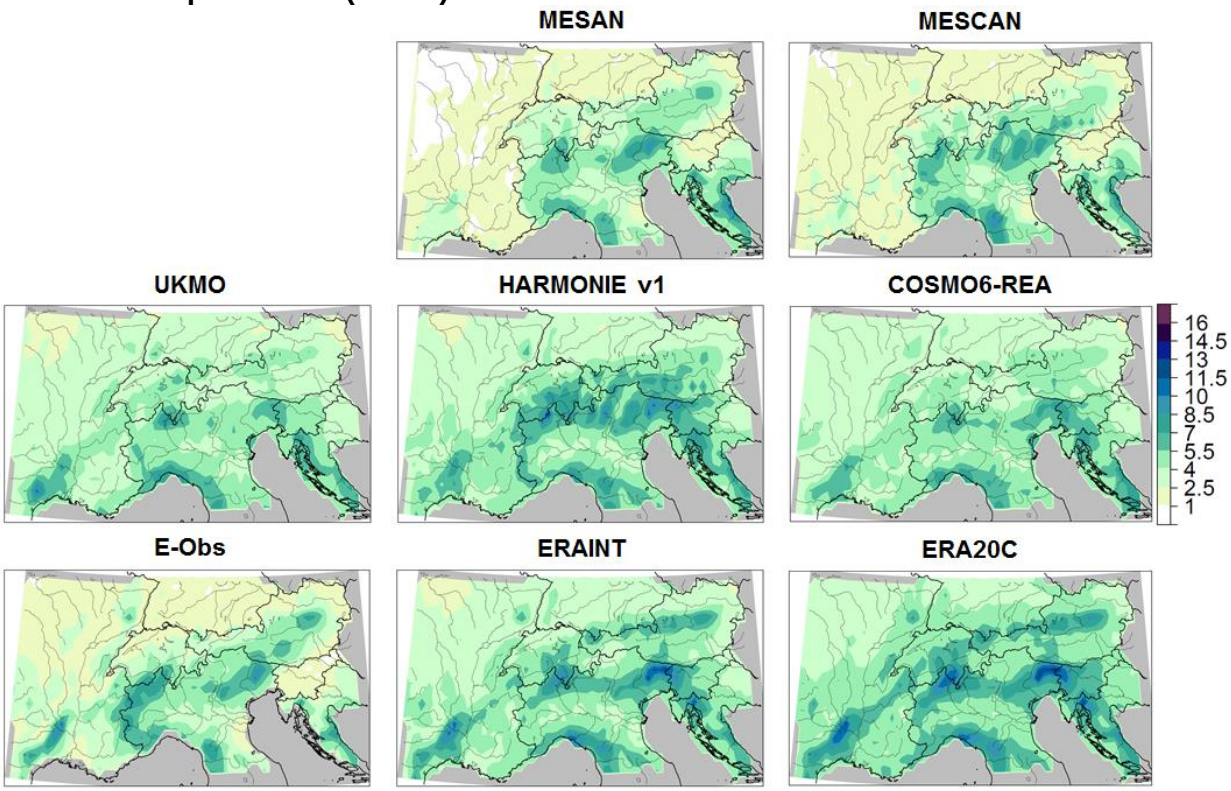


**Figure 4.3.1.5:** Root Mean Square Error [mm/d] of daily precipitation (mm/d, 2006-2008). Datasets rescaled to 0.25° regular grid. Reference: APGD.

The influence of station density in MESAN and MESCAN is evident in the root mean square error score (see Figure 4.3.1.5). The error is lower over France, Germany and Slovenia compared to the other regions. The errors of UKMO and COSMO6-REA are comparable and generally low. HARMONIE has higher values over the Alps and the southern rim. As for the downscaling, the effect of stations density and the complexity of the topography are reflected in E-OBS.

### Time series of daily precipitation

Figure 4.3.1.6 depicts the daily precipitation in the lower part of the Aare river catchment (Switzerland) in April 2008. UKMO-ENS strongly overestimates precipitation compared to the references (APGD-ENS or APGD). This behaviour is reconducted to an incorrect handling of the spin up phase which request a new calculation of the whole dataset (not possible during the UERRA project). Thus, the UKMO-ENS is not further commented here.
The probabilistic reference APGD-ENS permits to discern between datasets that are in its uncertainty range and the one where the difference to the observation cannot be explained only by uncertainties due to interpolation errors (not measurement errors) in the reference datasets.
In most days with precipitations below 10 mm/d all models (except UKMO-ENS) are very close to the reference. As the precipitation increases, the differences become more pronounced. All datasets, and in particular global reanalyses, underestimate precipitation. The very good performance of COSMO-ENS is confirmed also here by the overlap almost every day with the reference.

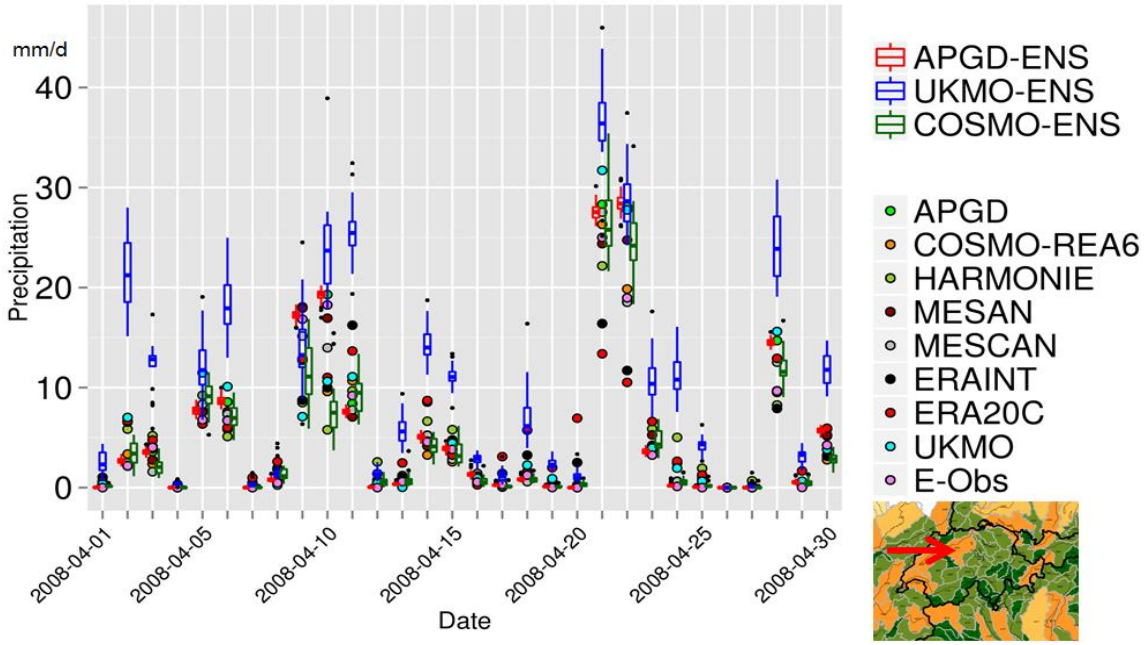


**Figure 4.3.1.6:** Daily precipitation [mm/d] for the lower part of the Aare catchment (Switzerland).

### Catchment dependent 95% quantile

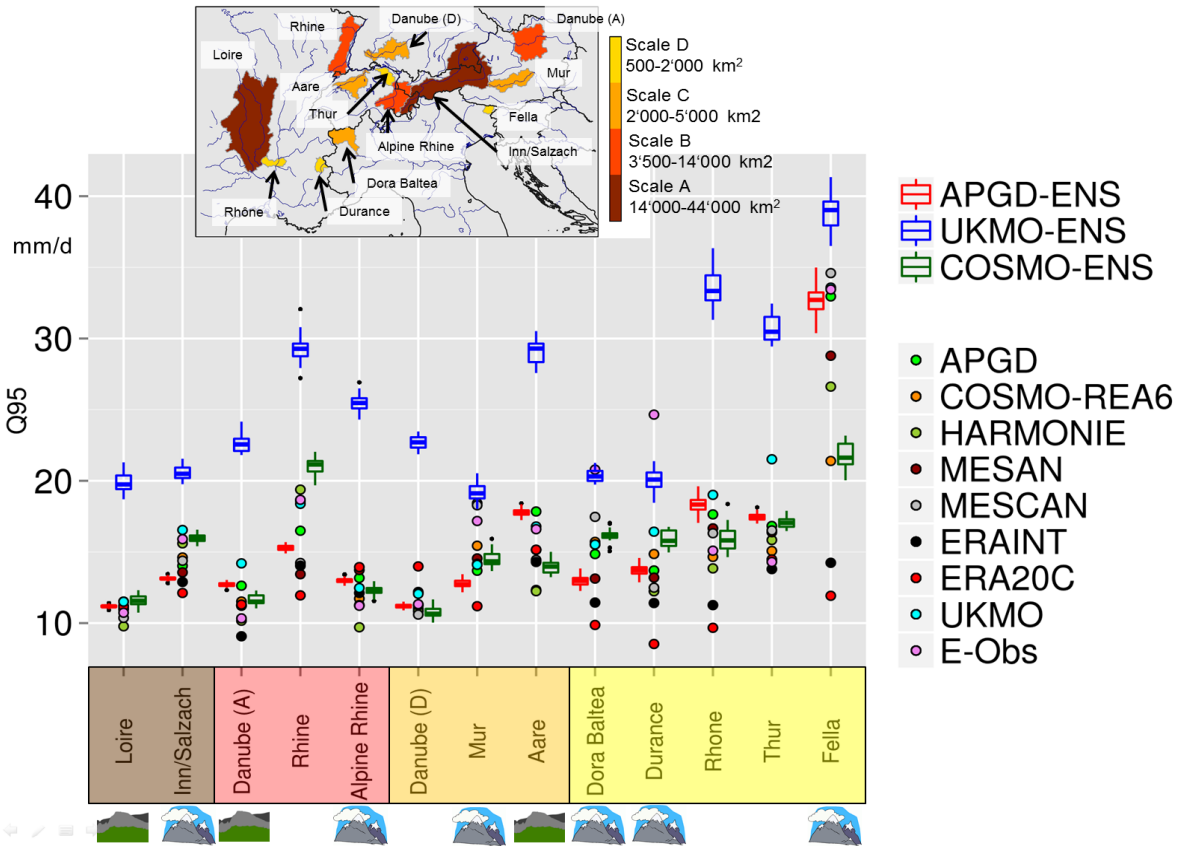


**Figure 4.3.1.7:** 95% quantile [mm/d] for different catchments (see map) in the period 2006-2008. The colour in the abscissa axis denotes the size class of the catchment and the icons below the topography (rather flat land, hills or mountains).

Figure 4.3.1.7 shows several catchments differing in size (from large in brown to small in yellow) and position (in the Alps, hills ranges or flatter areas). As expected, in more complex topography the differences to the reference tend to be higher (for example compare Loire vs. Inn/Salzach or Thur vs. Fella). Equally, for smaller catchments the agreement between the datasets decreases, thus from the left to the right side of Figure 4.3.1.7. The biggest differences can be seen for the Fella catchment, a region of complex topography, high precipitation amounts and events of high intensity. There, only MESCAN, UKMO and E-OBS are in the uncertainty range of the reference APGD-ENS, which extends over 5 mm/d (nearly 20%).

**Catchment dependent Brier skill score**

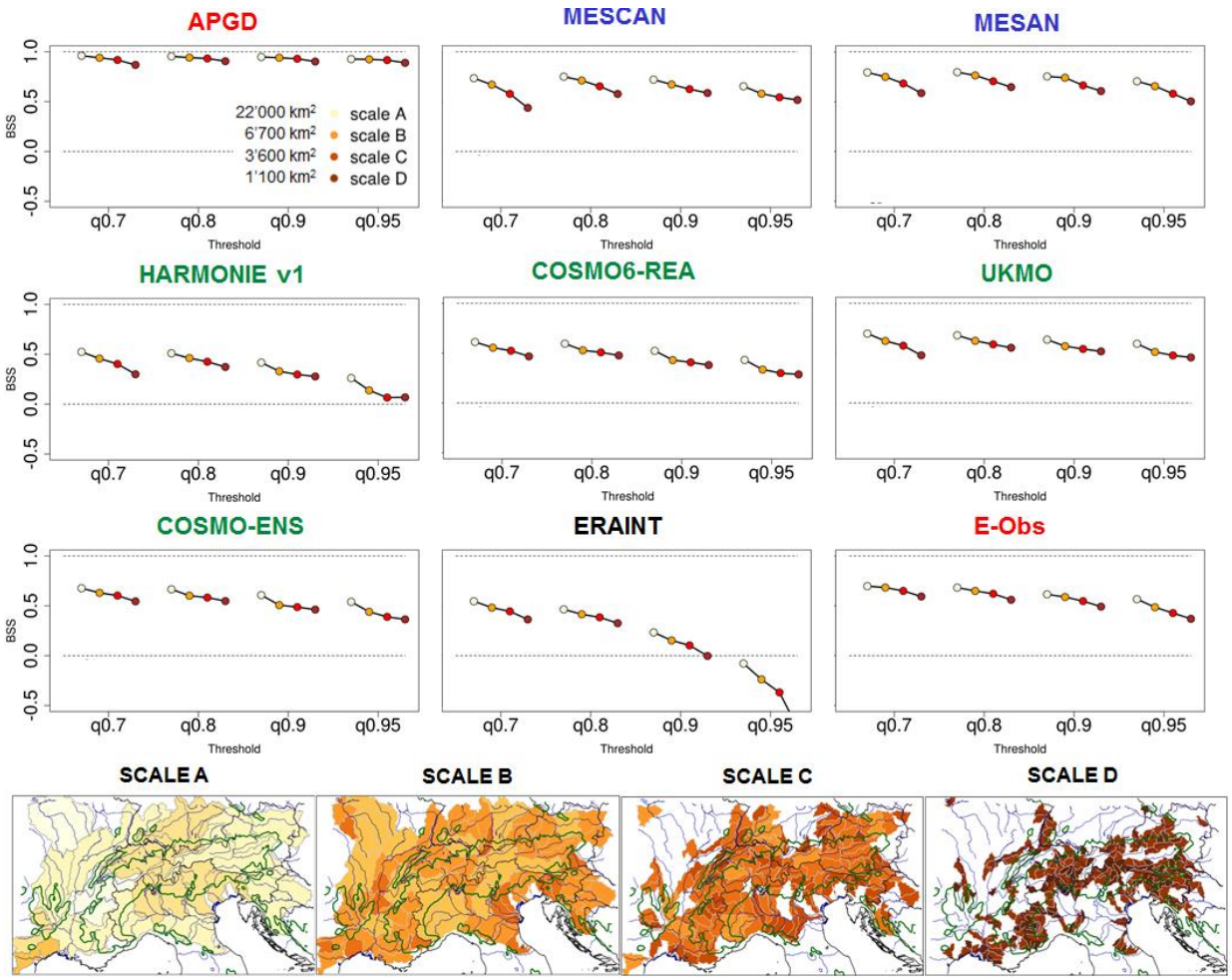


**Figure 4.3.1.8:** Brier Skill Score in the period 2006-2008. Each panel is a dataset. The abscissa axis is the threshold used for the score, corresponding to the 70%, 80%, 90% and 95% quantile (for each catchment). The point chain represents the score for different scales (from A, the biggest catchments to D, the smallest, see legend in the panel of the APGD dataset). Each scale is composed of several catchments, as illustrated in the bottom panels. Reference: APGD-ENS.

In Figure 4.3.1.8, the Brier Skill Score (BSS) is shown for the period 2006-2008 for different thresholds, namely the 70%, 80%, 90% and 95% quantile value of each of the 399 catchments separately. The catchments are subdivided into four dimensions scales, represented in the lower panels. For each threshold, the chain of four points illustrates the mean BSS of all catchments in the respective scale (each point is a scale). The reference is APGD-ENS.

BSS is decreasing for smaller catchment sizes (lowering of BSS in the same point chain). Interestingly, in all regional reanalyses the tendency to decrease is reduced from scale B to D. Regional reanalyses keeps a more stable skill for smaller catchments size compared to downscalings and global reanalyses.
BSS is decreasing for higher quantiles, thus for more extreme precipitation events (lowering of BSS between the different point chains of the same dataset). This effect is much lower in downscalings and the UKMO reanalysis. HARMONIE has a strong BSS lowering for the 95% quantile instead, where ERAINT has no skill anymore. E-OBS has a comparable/slightly lower BSS to MESAN and MESCAN.

**Catchment dependent yearly cycle**

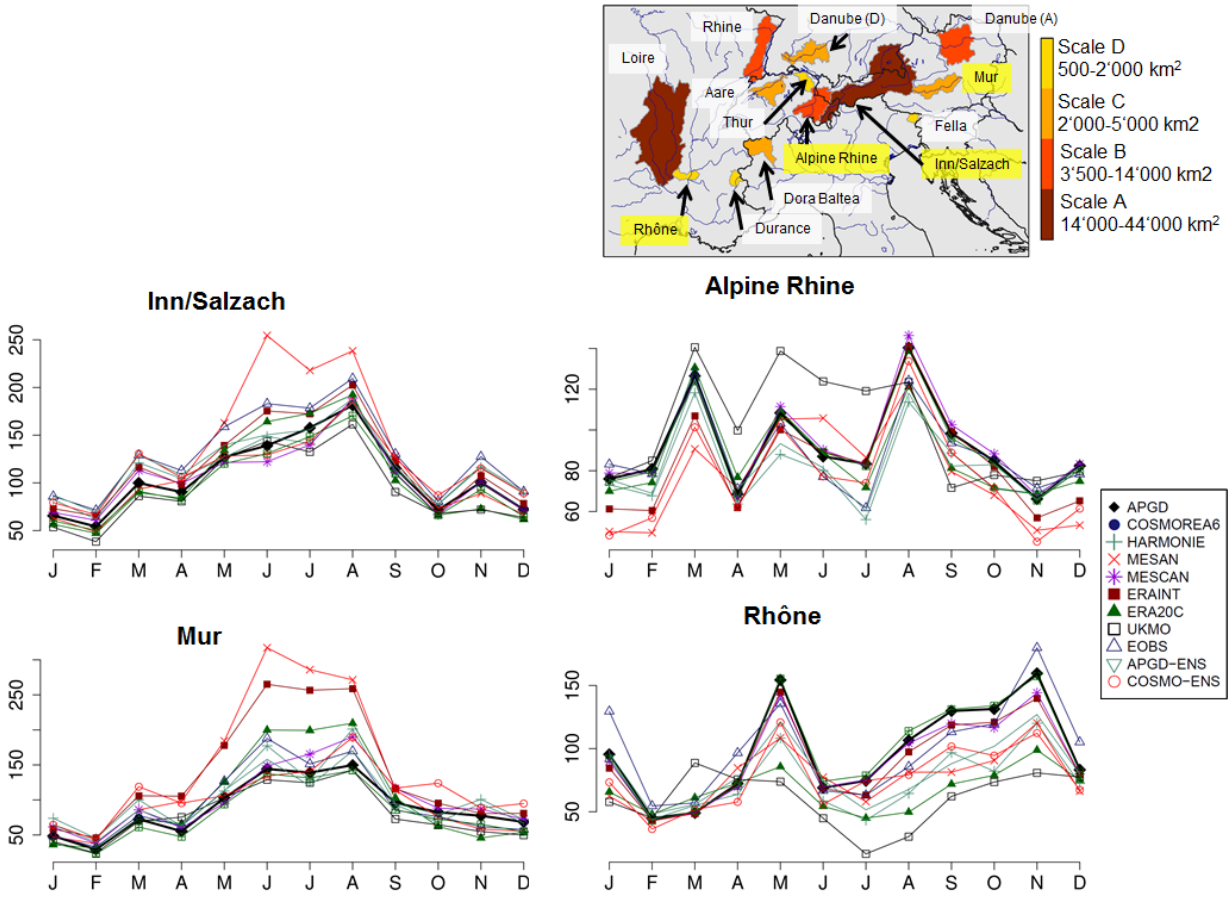


**Figure 4.3.1.9:** Mean monthly precipitation (2006–2008) for four catchments (yellow background in the map). The domain-mean values are derived from the 0.25° versions of all datasets.

The yearly cycle (Figure 4.3.1.9) is mostly well reproduced in all datasets. The regional reanalysis UKMO, the downscaling MESAN and the two global reanalysis ERAINT and ERA20C are the datasets that show most differences from the reference and are for some catchments not able to correctly reproduce the shape (the comparison has been done for all catchments highlighted in the upper right panel in Figure 4.3.1.9 but only four of them are shown).

## Catchment dependent frequency distribution function

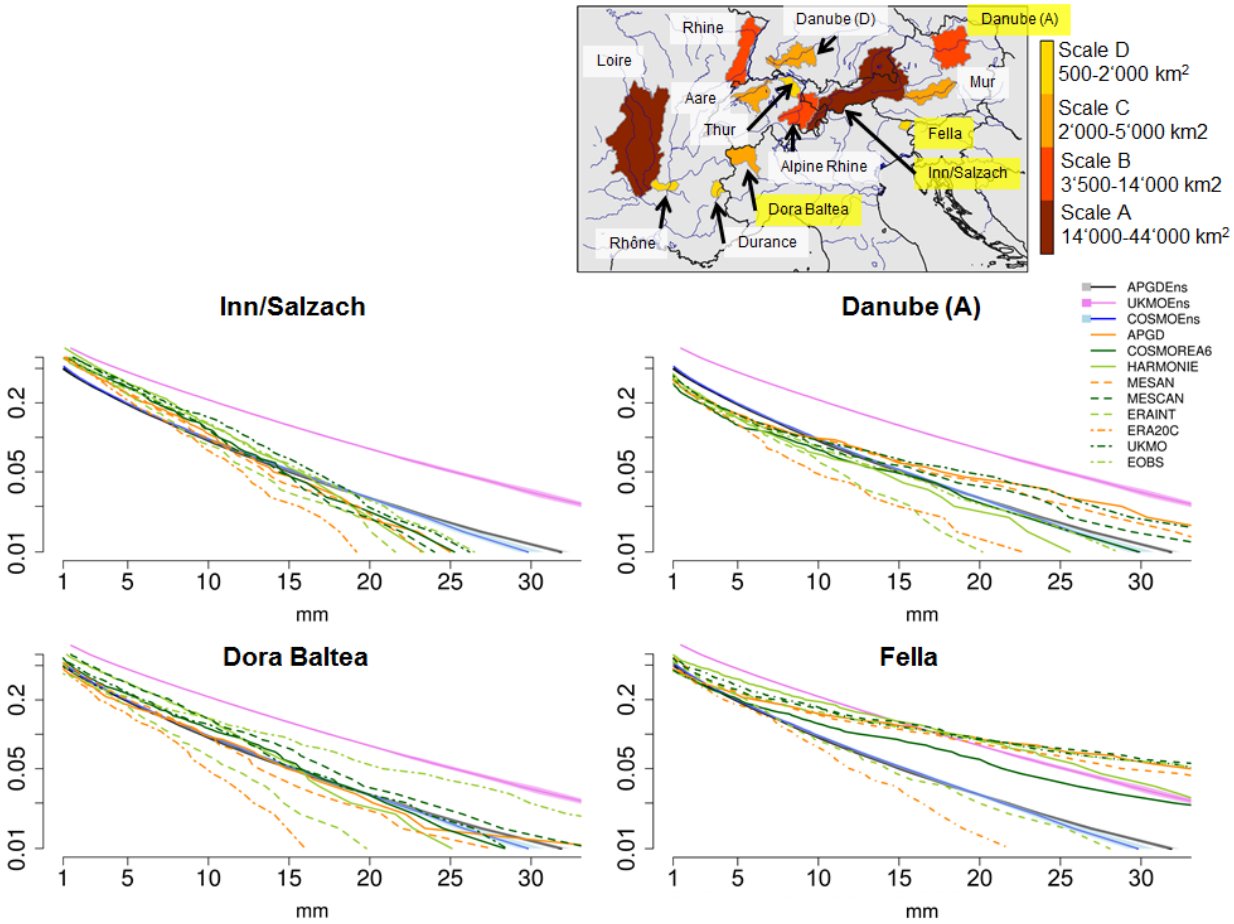


**Figure 4.3.1.10:** Frequency (y-axis, fraction of days, log-scale) at which daily precipitation amounts (values at 0.25° resolution grid points) exceed a threshold (x-axis, mm). Period 2006-2008, four catchments (yellow background in the map). Ensemble datasets are shown as coloured area with a line for the median.

Four typical results of the frequency at which daily precipitation amounts exceed a threshold are shown in Figure 4.3.1.10, where a catchment of each scale is chosen. COSMO-ENS is nearly overlapping the reference, with a marginal tendency to underestimate the strongest events. UKMO-ENS is constantly (notice the logarithmic ordinate axis) overestimating the frequency whereas the global reanalysis underestimate precipitation amounts increasingly for higher thresholds. The behaviour of regional reanalysis is not uniform but depends on the catchment. For the Inn/Salzach small precipitations are overestimated and larger events underestimated. In the small catchment Fella all regional reanalysis strongly overestimate the frequency. In general the regional reanalyses tend behave similarly, overestimating or underestimating frequency in the same way.

## Scale dependent mean annual precipitation

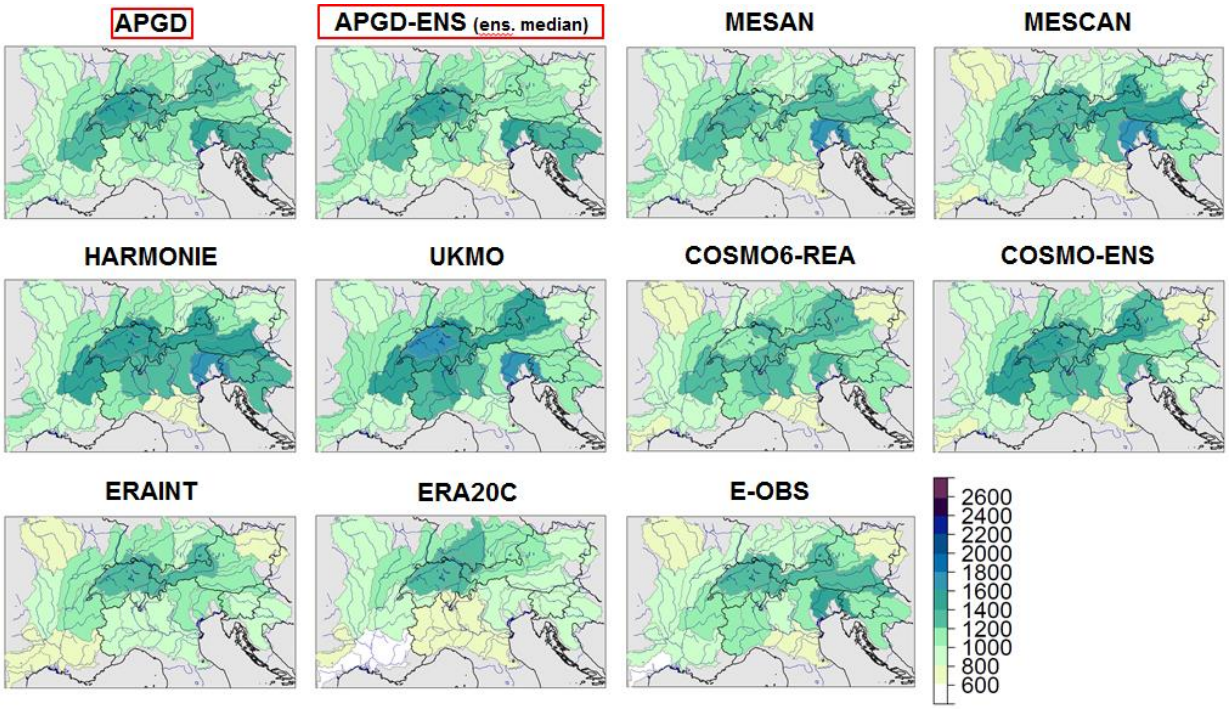


**Figure 4.3.1.11a:** Mean annual precipitation [mm/y] for catchments of dimension between 14000-44000 km$^2$ (scale A), reference: APGD-ENS. Period 2006-2008.

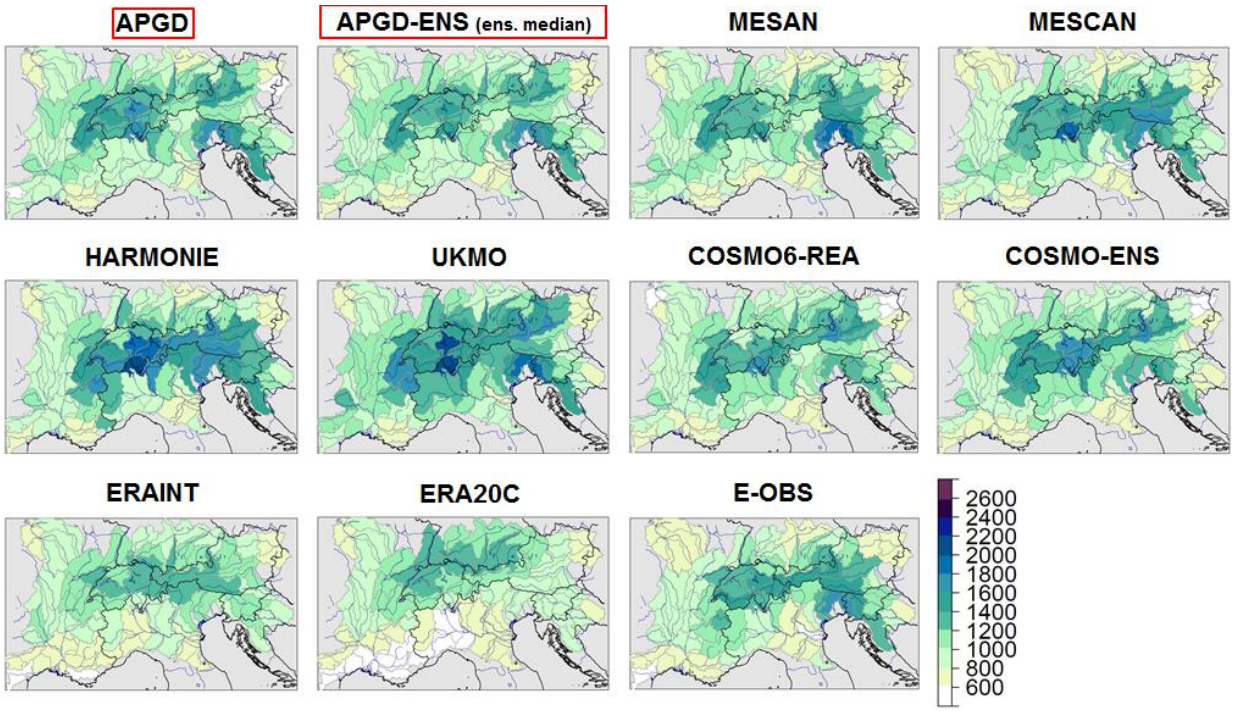


**Figure 4.3.1.11b:** Mean annual precipitation [mm/y] for catchments of dimension between 3500-14000 km$^2$ (scale B), reference: APGD-ENS. Period 2006-2008

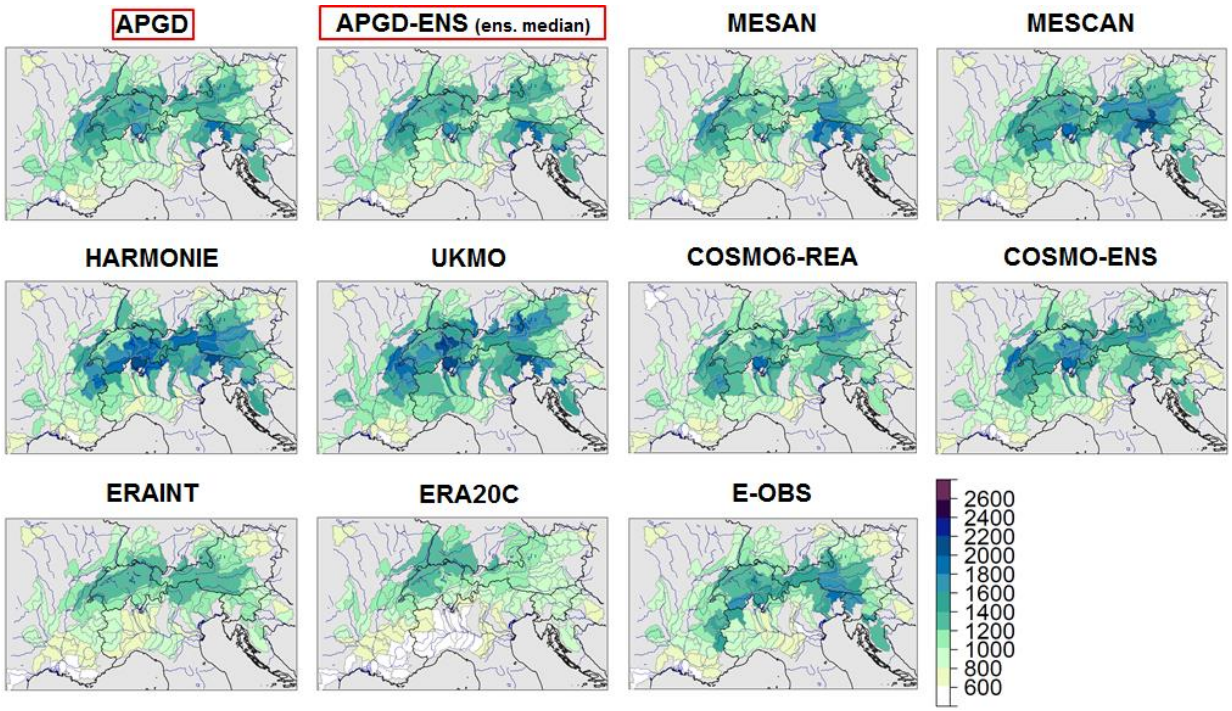


**Figure 4.3.1.11c:** Mean annual precipitation [mm/y] for catchments of dimension between 200-5000 km$^2$ (scale C), reference: APGD-ENS. Period 2006-2008

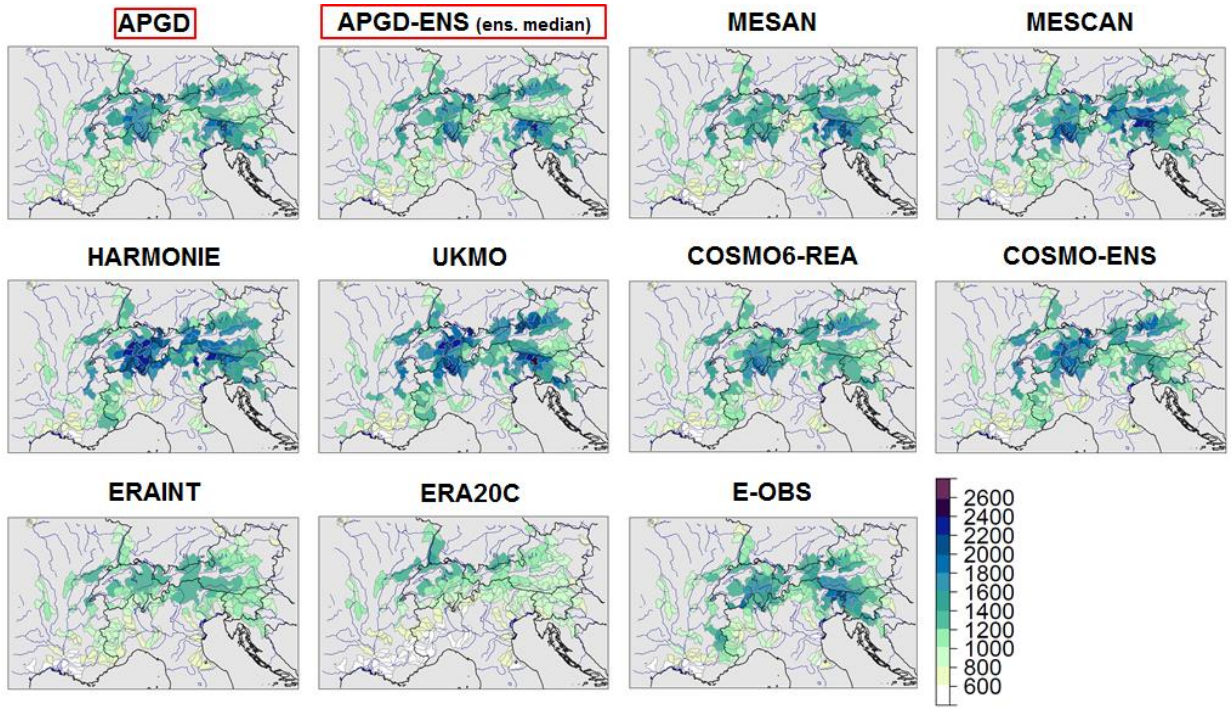


**Figure 4.3.1.11d:** Mean annual precipitation [mm/y] for catchments of dimension between 500-200 km$^2$ (scale D), reference: APGD-ENS. Period 2006-2008

The panels a)-d) of Figure 4.3.1.11 represent the mean annual precipitation on a catchment scale (from A to D, see Figure legend for details). Global reanalysis are quite different from reference already in the biggest scale. As the scale becomes smaller, also regional reanalyses and downscalings are increasingly differing from reference (MESAN, MESCAN and COSMO less than UKMO and HARMONIE). For users it is essential to be aware of the uncertainty of the datasets, which is dependent on the scale.

**Rank histogram**

Figure 4.3.1.12 depicts the rank histograms of the deterministic datasets, which are compared to APGD-ENS. In an “usual” rank histogram discussion the ensemble would be characterized as underdispersed or overconfident. As here the ensemble corresponds to the reference, we conclude that the datasets are mostly outside the uncertainty range of APGD-ENS. Only downscalings, in particular MESAN, are visibly more often in the range of the reference (notice the varying y-axis).

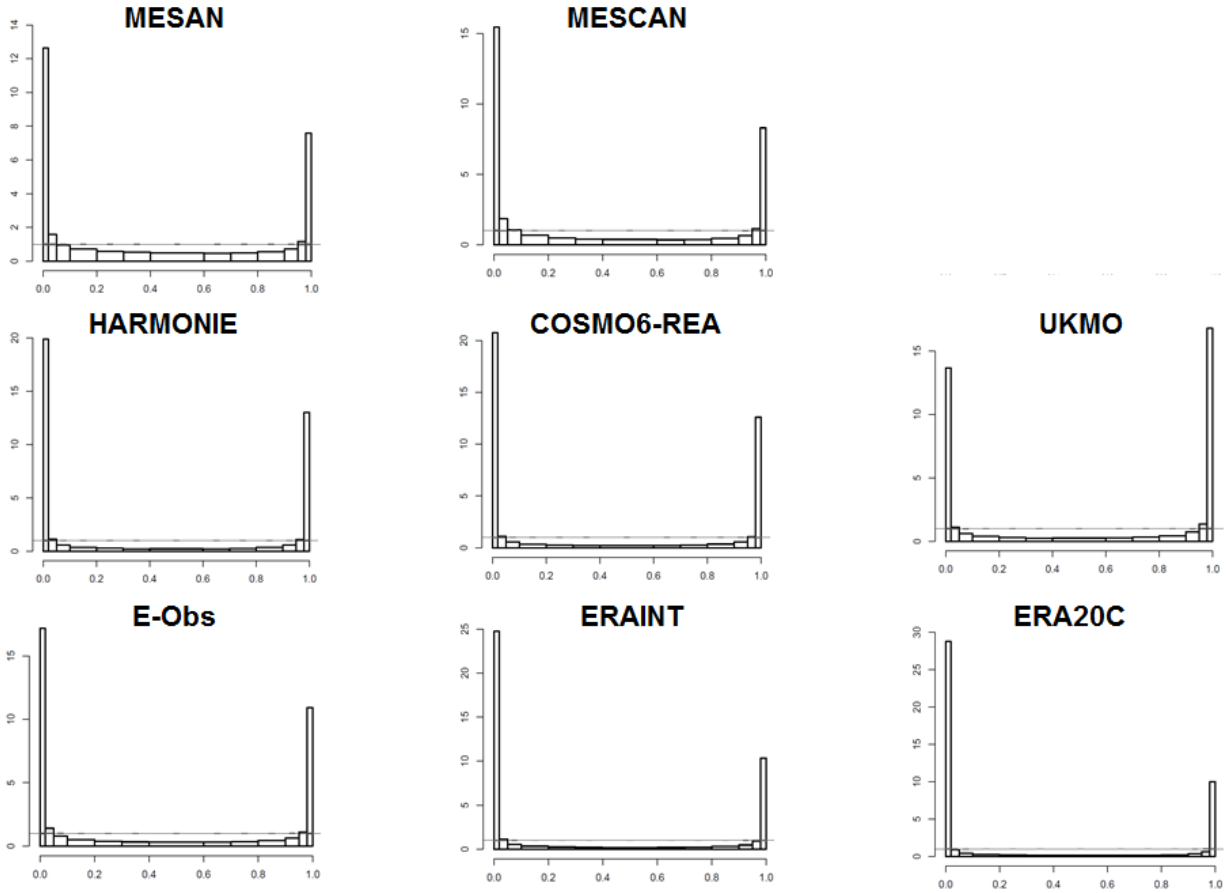


**Figure 4.3.1.12:** Rank histogram. The reference is APGD-ENS. Period 2006-2008

**Case study of an extreme precipitation event**

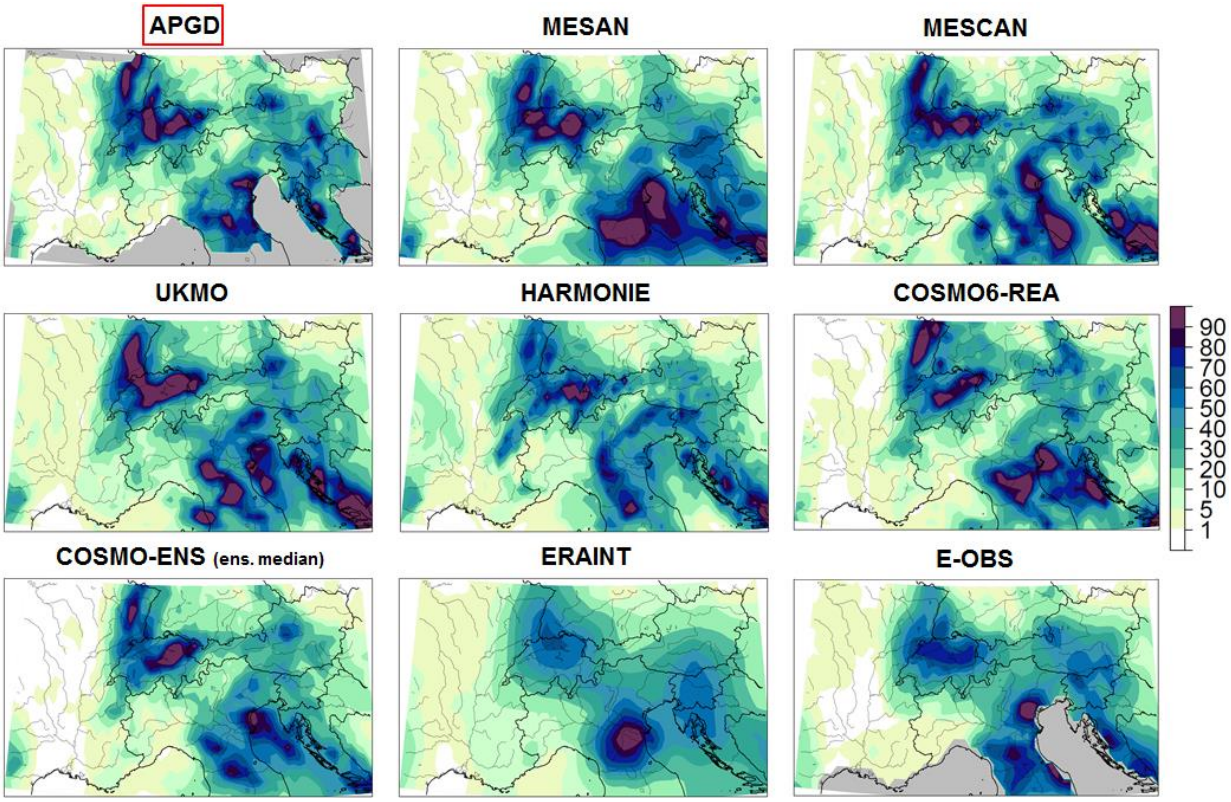


**Figure 4.3.1.13:** Precipitation sum [mm] for a 3-days period from 16.9.2006 to 18.9.2006.

During the 3-days period between 16.9.2006 and 18.9.2006, intense precipitation was measured in Northern Switzerland, Vosges and Po Valley. A precipitation band was present over mid Austria, Slovenia and Croatia (see Figure 4.3.1.13). Downscalings capture well the precipitation pattern, except for northern Italy, where the effective resolution seems to be quite coarse, especially for MESAN. Regional reanalysis are also able to represent the general precipitation patterns. UKMO is overestimating the spots of intense precipitation whereas HARMONIE underestimate the maximum in the Vosges. The two COSMO reanalyses matches well the reference. ERAINT and E-OBS shows a coarse pattern and captures only the main wet areas without the details.

## Alpine region – Main outcomes

Regional reanalyses:

- Added value compared to global reanalyses, which are not able to resolve the topographic complexity of the Alps and correctly represent precipitation patterns and amounts.
- Tend to overestimate precipitation amounts and frequency, especially in complex terrain.
- Regional reanalyses often shows a better performance with much more detailed precipitation structures than observational gridded datasets as E-OBS in region of low station density. An exception is the wet-day frequency, where the datasets using directly precipitation measurements performs better than regional reanalyses apart of COSMO datasets.
- COSMO6-REA and COSMO-ENS show the best performance of all regional reanalyses. The higher resolution and the use of non-hydrostatic dynamics are the main reason for the performance.

Downscaling datasets:

- Additional value in regions with dense station network compared to regional reanalyses (their performance is strongly dependent on the station density). Improvement is most evident in the precipitation frequency (fraction of wet days).
- Where the station density is high, downscaling datasets reach a very high detail of the precipitation pattern.

General comments:

- In most days, the datasets have an error bigger than the uncertainty range of the reference dataset. Especially for days with more than around 10 mm/d, the spread of results from the different models is huge, with the global reanalyses having the biggest under- and overestimations.
- The biggest differences from the reference and the lowest Brier skill score are found in complex topography, small catchment sizes and for higher precipitation amounts. Regional reanalyses have a less pronounced decrease of the Brier skill score as the catchment size is decreasing compared to global reanalyses and downscaling datasets.
- Annual cycle is mostly well reproduced in all datasets.
- User should be aware of the effective resolution of datasets, which for datasets as E-OBS and the reference APGD is coarser than the grid resolution.

### 4.3.2 Fennoscandia – Final results

The reanalysis and downscaling datasets have been evaluated to assess their ability to model precipitation for climatological applications over Fennoscandia, which is the area shown in Figure 4.3.2.0. The evaluation has been carried out on two grids: a high-resolution grid with grid-spacing of 5 km and a low-resolution grid with spacing of 0.25° (approximately 12 Km over Fennoscandia). The time period considered in the evaluation ranges from 2006 to 2010 and for some of the reanalyses we have considered also the period 1986-1990 (shown in section 9.3 as supplementary material). The summary statistics considered consist of: annual precipitation mean; mean value of the annual 95th percentile; frequency of wet days (>1mm/day) annual mean; Root-Mean-Square Error of daily precipitation. For the COSMO ensemble reanalysis, the Brier Skill Score has been computed. Furthermore, the reanalyses skill in modeling daily precipitation have been investigated by means of a verification method based on scale-separation [Casati et al., 2004] and [Casati, 2010], where a single-band spatial filter has been used to separate forecast and observation fields into spatial components (e.g. wavelets), and then each spatial component has been verified separately (e.g. with traditional continuous, categorical scores).

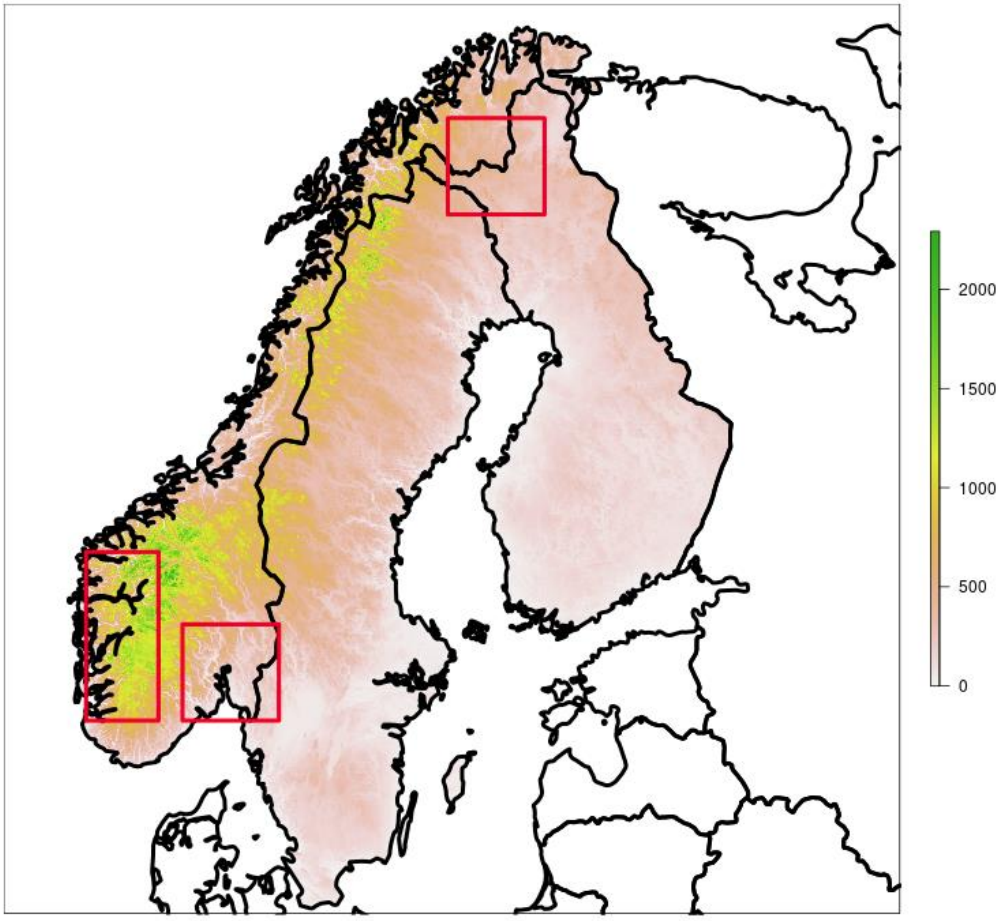


**Figure 4.3.2.0:** Fennoscandia, elevation (m a.m.s.l.) is shown for the domain considered. Red boxes indicate regions where local evaluations have been performed (cfr. §...).

The Nordic Gridded Climate Dataset (NGCD), described in [Lussana et al., 2017] and [Gisnås et al., 2017] has been used as reference in most cases. The NGCD is a high-resolution (1 $km^2$) gridded climate data sets for daily accumulated precipitation for the period 1981-2010. The input data used in NGCD are data from ECA&D database (www.ecad.eu) and the Norwegian Climate Database (eklima.met.no). Original non-homogenized time series were used. The number of stations used for the interpolation varies with time due to data availability. For more than 80% of the time steps the interpolation is based on data from more than 1100 precipitation stations and 371 temperature stations. These are distributed over the three countries with approximately 25% of the stations in each of Norway and Finland and 50% in Sweden. All data are open and publically available. The spatial interpolation methods adopted are inspired by the ones developed for Norway by Met Norway. The NGCD has been also presented in the FP7 project UERRA, as an in-kind contribution from MET Norway.
In addition to the UERRA reanalysis dataset, we have considered other datasets in our evaluation, such as: global reanalysis (see the website reanalysis.org for an overview of current atmospheric reanalyses); European regional reanalysis not developed within UERRA but in the FP7 EURO4M project (euro4m.eu), such as MESAN [Soci et al. 2016]; NORA10 [Reistad et al., 2011], hindcast based on a downscaling of ECMWF global analysis developed and used at MET Norway. By considering these datasets, it is possible to assess the quality of the UERRA datasets not only in absolute terms but also with reference to the current standard of available reanalyses (and hindcast) datasets.
This section is organized as follows. First, we evaluate the available deterministic reanalyses. Second, we assess the two different set of ensemble reanalyses systems and the MESCAN downscaling dataset that is provided to the users as an ensemble of precipitation fields. In the third paragraph, we will evaluate how the deterministic models perform over three subdomains on a monthly basis. In the fourth paragraph, the scale-separation evaluation is presented. In the conclusions, the key messages of our evaluation are summarized.
Supplementary materials are included in section 9.3, where: the results are shown on the low-resolution grid; a comparison of 2006-2010 and 1986-1990 precipitation statistics is reported; the comparison of MESCAN and HARMONIE modelling system is shown by means of several Figures.

## DETERMINISTIC REANALYSIS

### Mean annual precipitation (2006-2010)

The datasets considered in the evaluation are able to represent the mean annual precipitation pattern, with maxima along the Norwegian coast and a minimum over Lapland.
With reference to Figures 4.3.2.1 and 4.3.2.2: global reanalyses tend to underestimate the precipitation totals, as expected; the downscaling datasets MESAN and MESCAN better match the NGCD precipitation field, with MESCAN displaying features at a finer scale compared to MESAN; UKMO and COSMO6-REA are not too different from MESAN and MESCAN, though they overestimate the precipitation along the west coast, with maxima greater than 4000 mm; UKMO shows greater amount of precipitation also over Finland and Lapland. HARMONIE v1 overestimate the precipitations over Lapland, but in general it performs fairly well over the rest of the domain; HARMONIE v2 presents smaller precipitation amounts with respect to HARMONIE v1; the observational gridded dataset EOBS resembles a lot NGCD, although with a coarser resolution; NORA10 overestimates the precipitation over the entire domain and the precipitation maxima are located inland compared to the other regional reanalyses (this effect, might be related to choices on the processing of the NORA10 model topography).
Figure 4.3.2.3 shows the quantile-quantile plots for all the UERRA datasets. Only grid points

within the NGCD area have been considered, so to assess the reanalyses in their ability to reproduce the distribution of precipitation values observed over the domain. In each panel, the diagonal represents the perfect agreement between the reference (NGCD, on the x axis) and the model. COSMO6-REA (top-left panel) and MESCAN (black dots, bottom-right panel) show good agreement up to ~2000 mm of mean annual precipitation and they overestimate the precipitation in the upper quantiles. UKMO (top-right panel, red line) overestimates the precipitation, especially over 2000 mm. HARMONIE v1 and v2 behave in a similar way (bottom-left panel) up to 2000 mm, underestimating the precipitation; beyond that, HARMONIE v1 tend to overestimate the annual totals, whereas v2 stays gradually catches up with the reference. Both versions 1 and 2 overestimate the totals in the highest quantiles. MESCAN results are discussed in the following, where a section has been dedicated to the downscaling dataset.

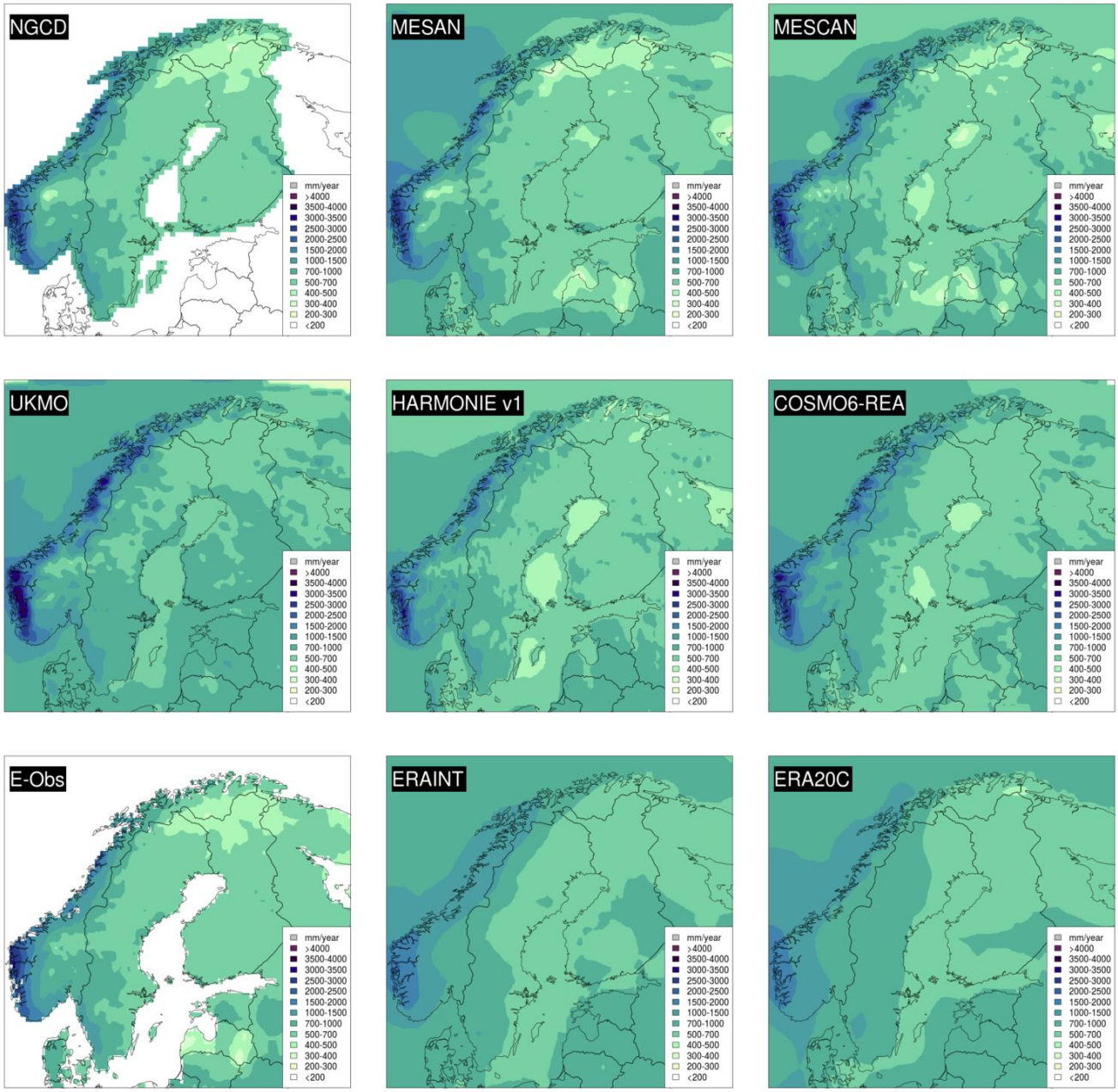


**Figure 4.3.2.1:** Mean annual precipitation (mm per year, 2006-2010). Rescaled to 0.25° regular grid. Reference: NGCD (top-left panel).

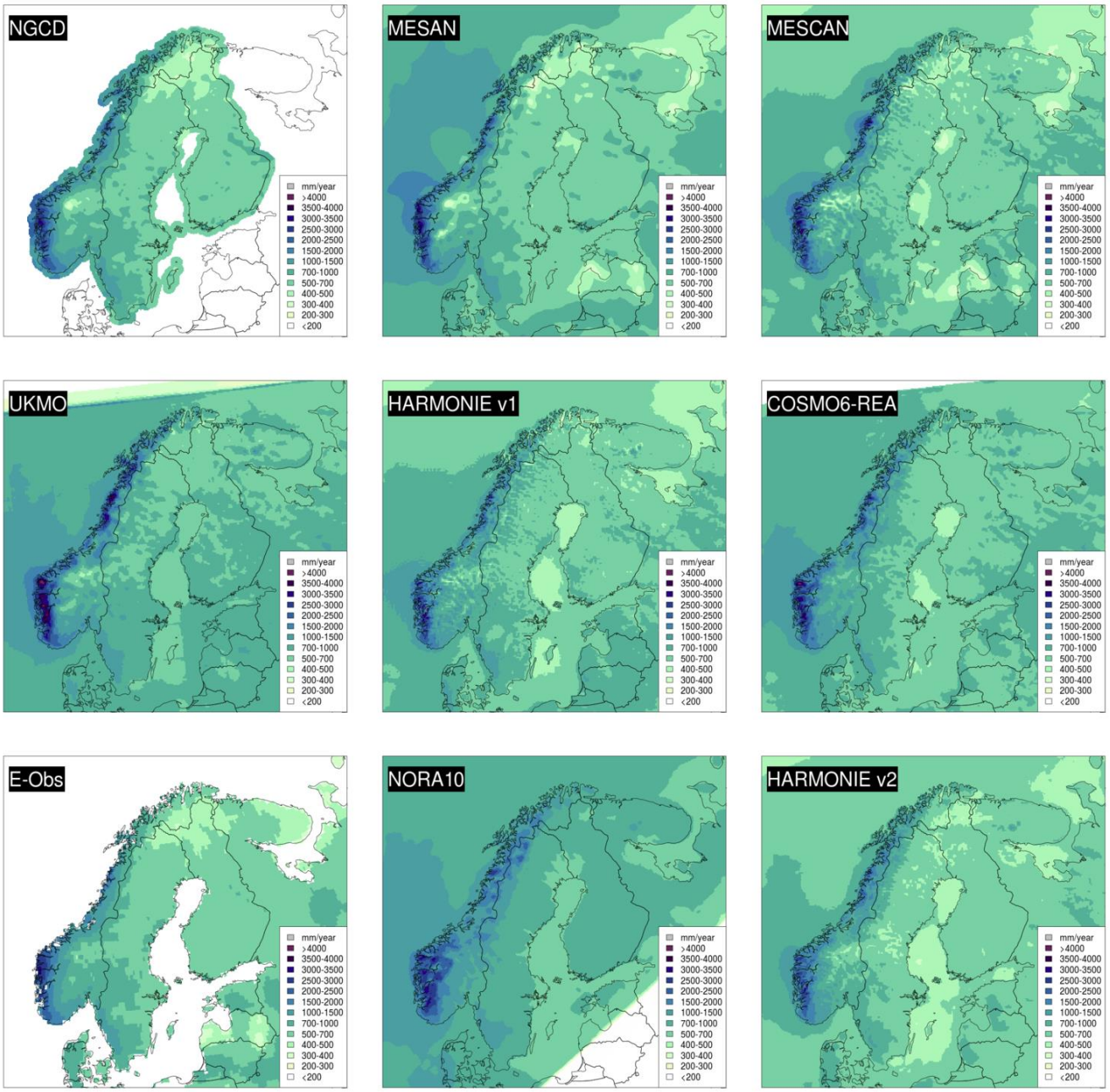


**Figure 4.3.2.2:** Mean annual precipitation (mm per year, 2006-2010). Rescaled to 5km ETRS-LAEA coordinate system. Reference: NGCD (top-left panel).

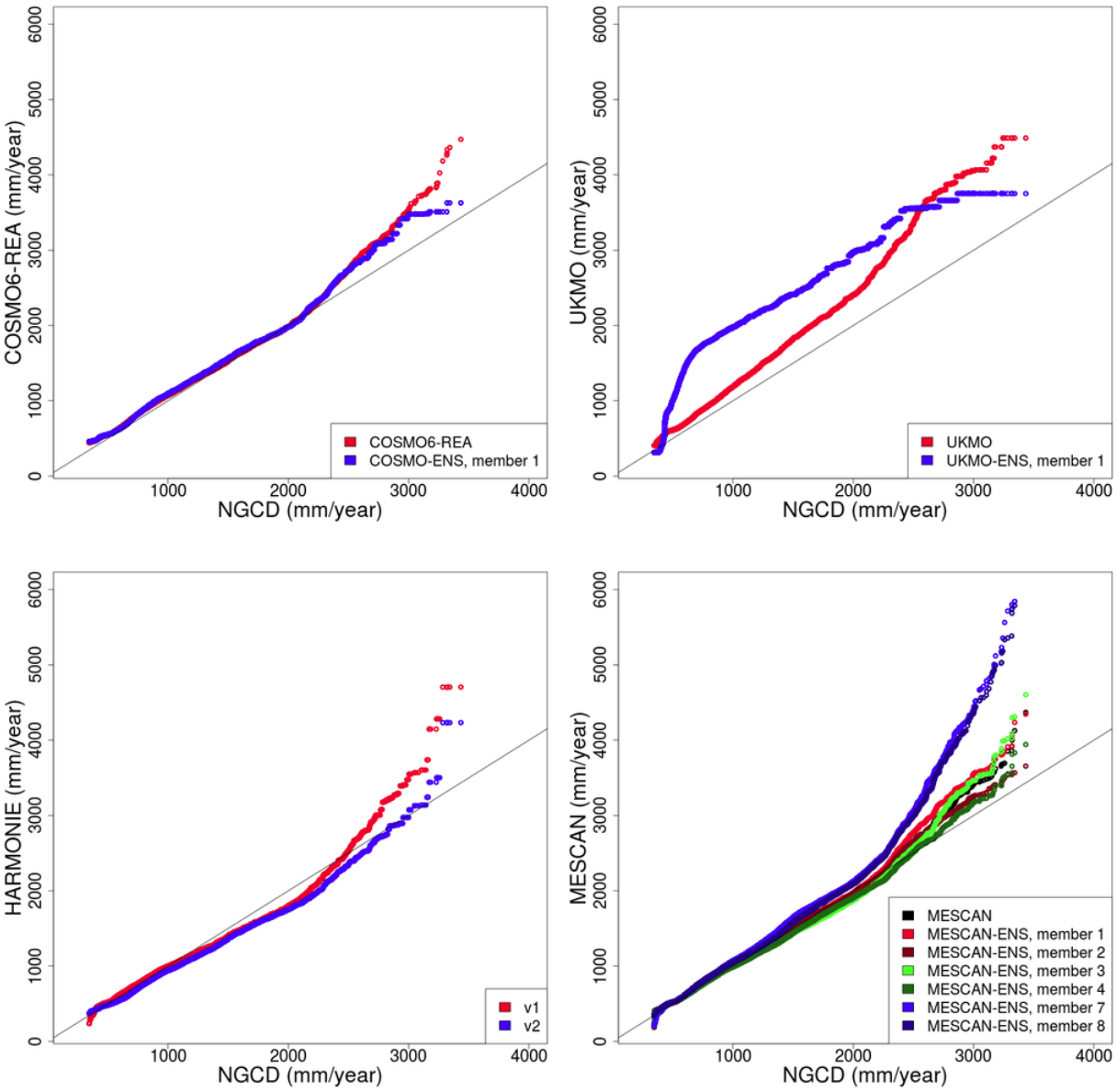


**Figure 4.3.2.3:** Quantile-quantile plot for the mean annual precipitation. The reference (x-axis) is NGCD. Only UERRA datasets are shown.

### Wet-day frequency and 95% quantile (2006-2010)

The results reported refer to the high-resolution grid. Similar results hold for the low resolution grid (see Supplementary material). As shown in Figure 4.3.2.4, all regional reanalyses provide satisfactory results for the wet-day frequency. HARMONIE v2, MESAN and MESCAN are the ones closest to NGCD. HARMONIE v1 and UKMO generally overestimate the frequency of wet days, especially on southern Sweden and in the central/northern Norwegian coast. In general, models which do not make use the observed precipitation values tend to overestimate the frequency of wet days along the coastal areas in Norway.

With respect to the 95% quantile shown in Figure 4.3.2.5, MESAN, MESCAN, COSMO6-REA and both HARMONIE v1 and v2 are the models showing the closest patterns and values to NGCD, with values between 4-8 mm in the northern Fennoscandia, around 8-12

mm in most of the southern Finland and Sweden, and values gradually larger (up to maximums of 30-40 mm) along the Norwegian coast. HARMONIE v1 and v2 display the same pattern with smaller values (around 4-8 mm) in both Finland and Sweden. NORA10 is similar, apart from the overestimation in the inland Norway. UKMO is generally similar to NGCD but it overestimates the values along the coast (up to 95th percentile larger than 40 mm).

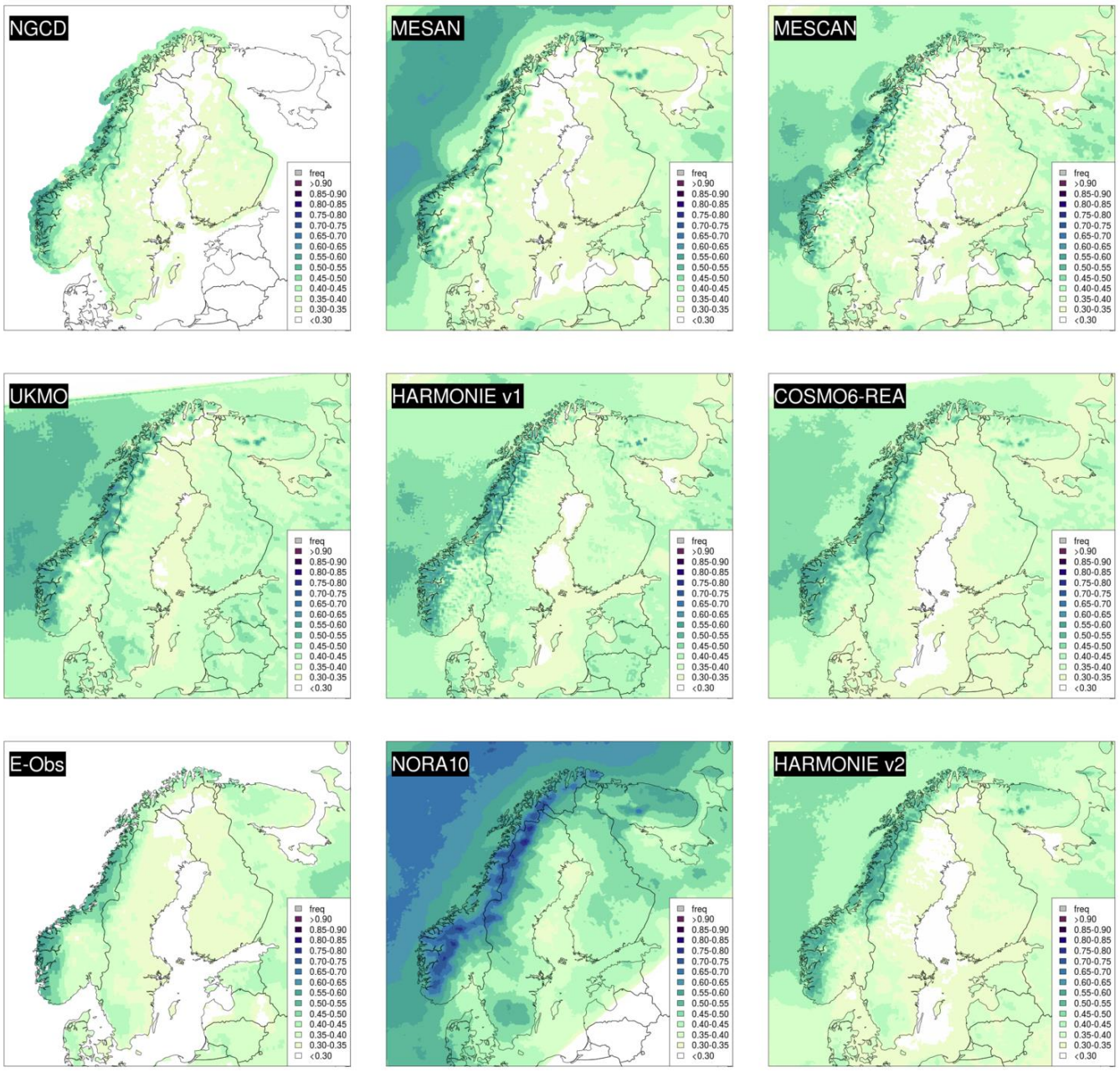


**Figure 4.3.2.4:** Annual frequency of wet days (≥ 1mm/d, fraction, 2006-2010). Rescaled to 5Km ETRS-LAEA coordinate system. Reference: NGCD.

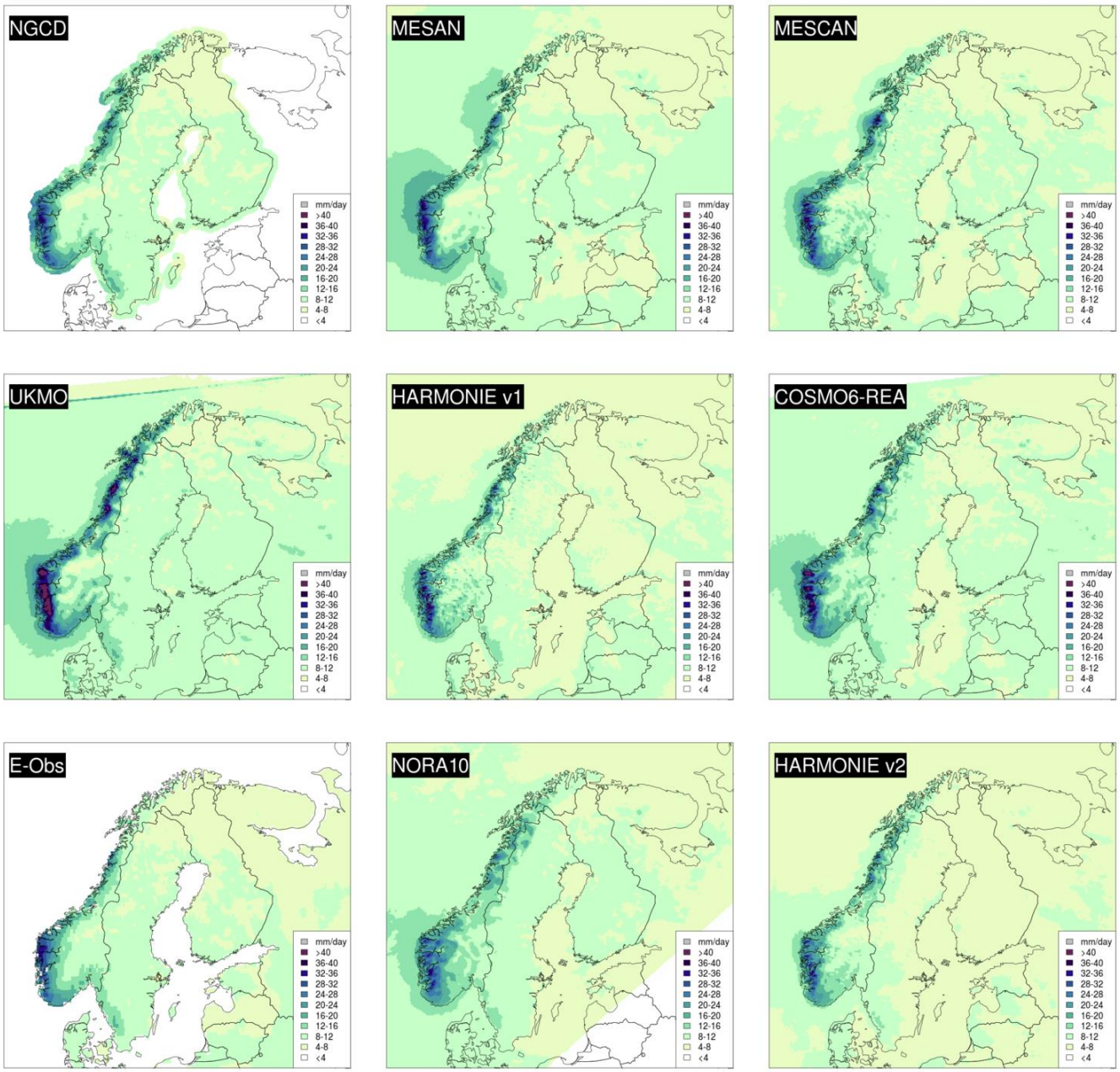


**Figure 4.3.2.5:** 95% quantile of daily precipitation (mm/d, 2005-2008). Rescaled to 5Km ETRS-LAEA coordinate system. Reference: NGCD.

### Root-Mean-Square Error (2006-2010)

The (gridpoint-by-gridpoint) RMSE has been computed by taking into account the whole period 2006-2010, so that the dataset considered for the mean is composed by the differences between simulated and observed daily precipitation.
With reference to Figure 4.3.2.6, MESAN and MESCAN have the smallest RMSE values (between 1-3 mm everywhere except from the Norwegian coast where peaks up to 6-7 mm are present). COSMO, UKMO, HARMONIE v1 and HARMONIE v2 show all the same pattern and values (around 2-3 mm over large part of Sweden and Finland, over 5 mm along the Norwegian coast). UKMO has the largest values along the Norwegian coast (up to more than 12 mm). NORA10 performs well in Sweden and Finland (values smaller than 3 mm), but

present larger values (up to 8-9 mm) over Norway. The observational gridded dataset EOBS, which is based on pretty much the same observational network as NGCD, tend to stay close to NGCD.
Bias patterns (not shown here) are similar to those of the root mean squared error, with only positive values (all the models overestimate the precipitation over long-term). All the models show small values (<1mm/day) over Sweden and Finland, apart from NORA10 and UKMO. Values are generally larger over the Norwegian inland and coast. NORA10 and UKMO are the models most affected by bias (up to over 4.5 mm/day).

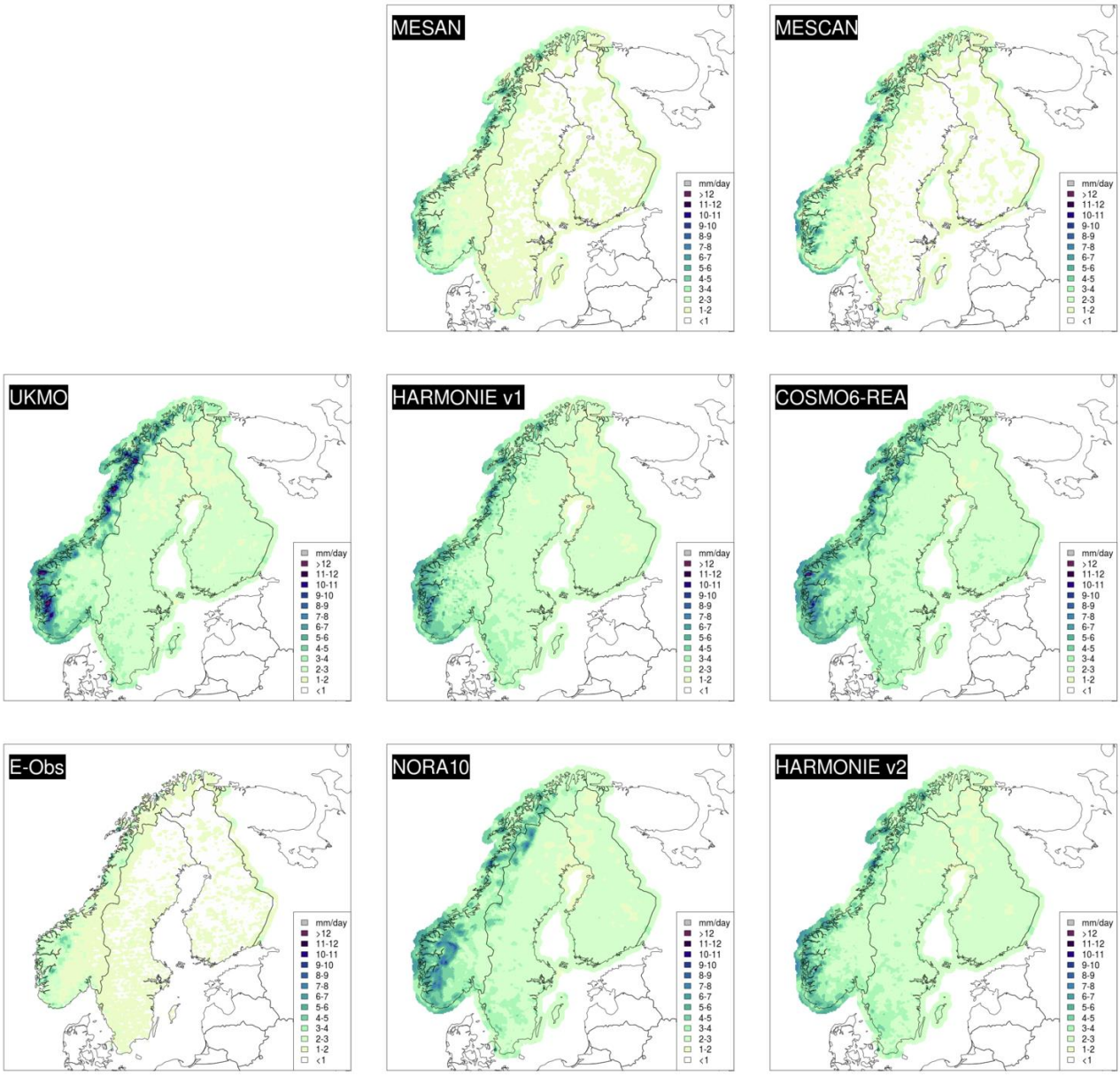


**Figure 4.3.2.6:** root Mean Square Error of daily precipitation (mm/d, 2006-2010). Rescaled to 5Km ETRS-LAEA coordinate system. Reference: NGCD.

## REANALYSIS ENSEMBLES and DOWNSCALING DATASETS

Within the UERRA project, the UK Met Office, the Deutscher Wetterdienst as well as MeteoFrance provided gridded datasets as ensemble of model fields. Two of this datasets are reanalysis ensembles:

- UKMO-ENS, 20 members
- COSMO-ENS, 21 members

While MeteoFrance provides an ensemble of downscaled precipitation fields by means of different downscaling strategies:

- MESCAN-ENS, 6 members

In this section we want to assess how these datasets perform with respect to their deterministic counterpart.

### MESCAN-ENS

The MESCAN-ENS reanalyses system is designed to provide the end-users information on the uncertainty of the surface analysis fields. As described by Bazile et al. 2017 in the UERRA report D2.9, MESCAN-ENS for precipitation consists of 6 different members, resulting from different combinations of perturbed observation networks and background. Hereafter, we will refer to each of the ensemble member with the same nomenclature defined in the table at page 17 of the above-mentioned report.
The results are reported in Figures 4.3.2.7 - 4.3.2.10. MESCAN ensemble members 7 and 8 (with ALADIN model at 5.5 km as background) are able to resolve the precipitation at finer spatial scales, especially in the mountains and along the Norwegian west coast. Members 1 and 3 (HIRLAM-ALADIN as background) show a finer resolution than members 2 and 4 (HIRLAM-ALARO as background). The use of a high (members 1, 2 and 7) or low (members 3, 4 and 8) density observational network does impact on the quality of the final analysis, as it is shown by the RMSE. However, the evaluation for the long term climate indexes, such as annual mean precipitation, shows that all the members provide satisfactory results.
In Figure 4.3.2.3, it is possible to compare the behaviour of the various ensemble members in reproducing the distribution of the mean annual precipitation over the domain: all members maintain themselves generally close to the reference (diagonal line) up to 2000 mm, then they all overestimate the precipitation. Members 7 and 8 are those which depart the most from the reference. On the contrary, members 2 and 4 are the ones closer to the reference.

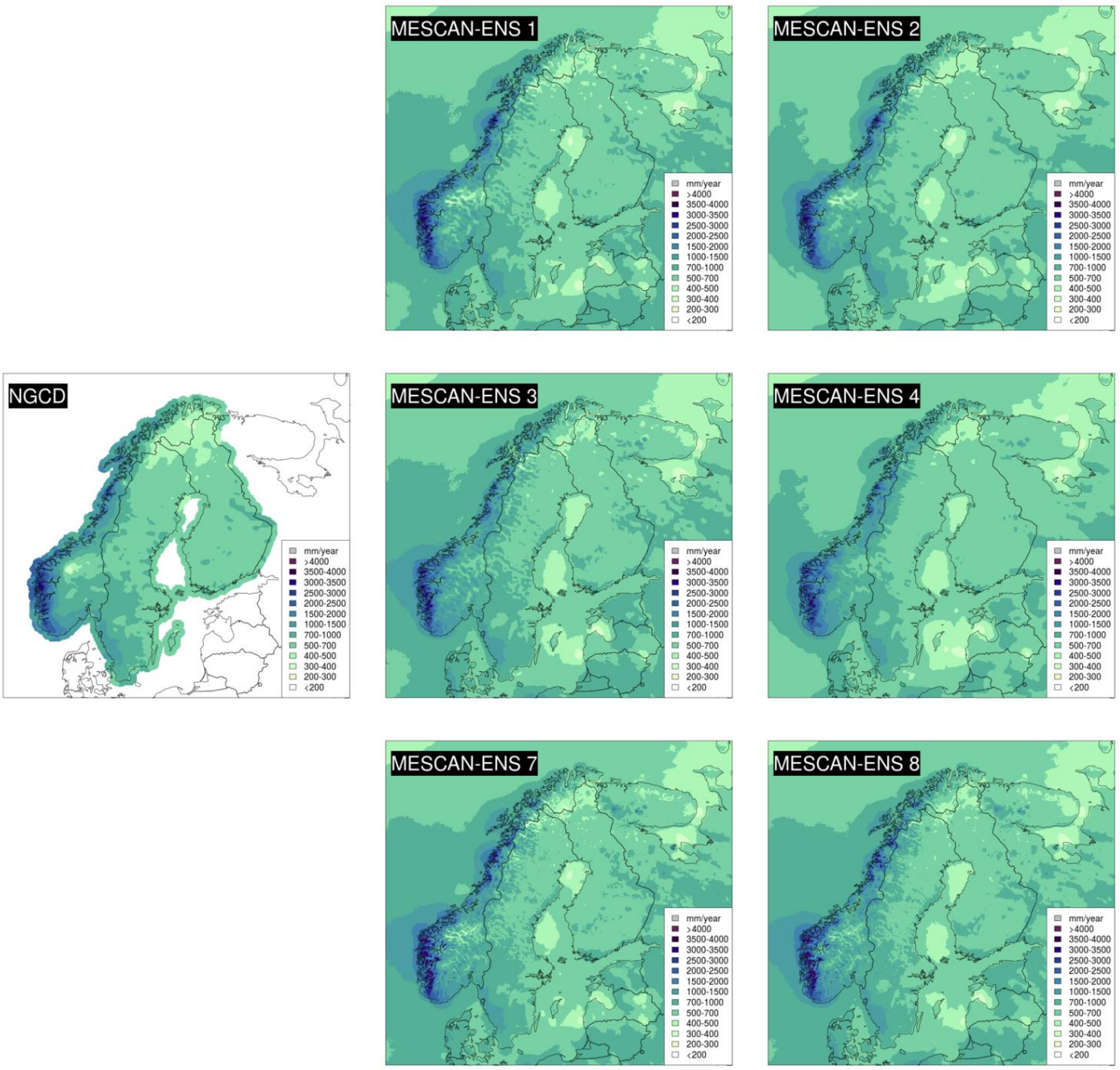


**Figure 4.3.2.7:** Mean annual precipitation (mm per year, 2006-2010). Rescaled to 5km ETRS-LAEA coordinate system. Reference: NGCD.

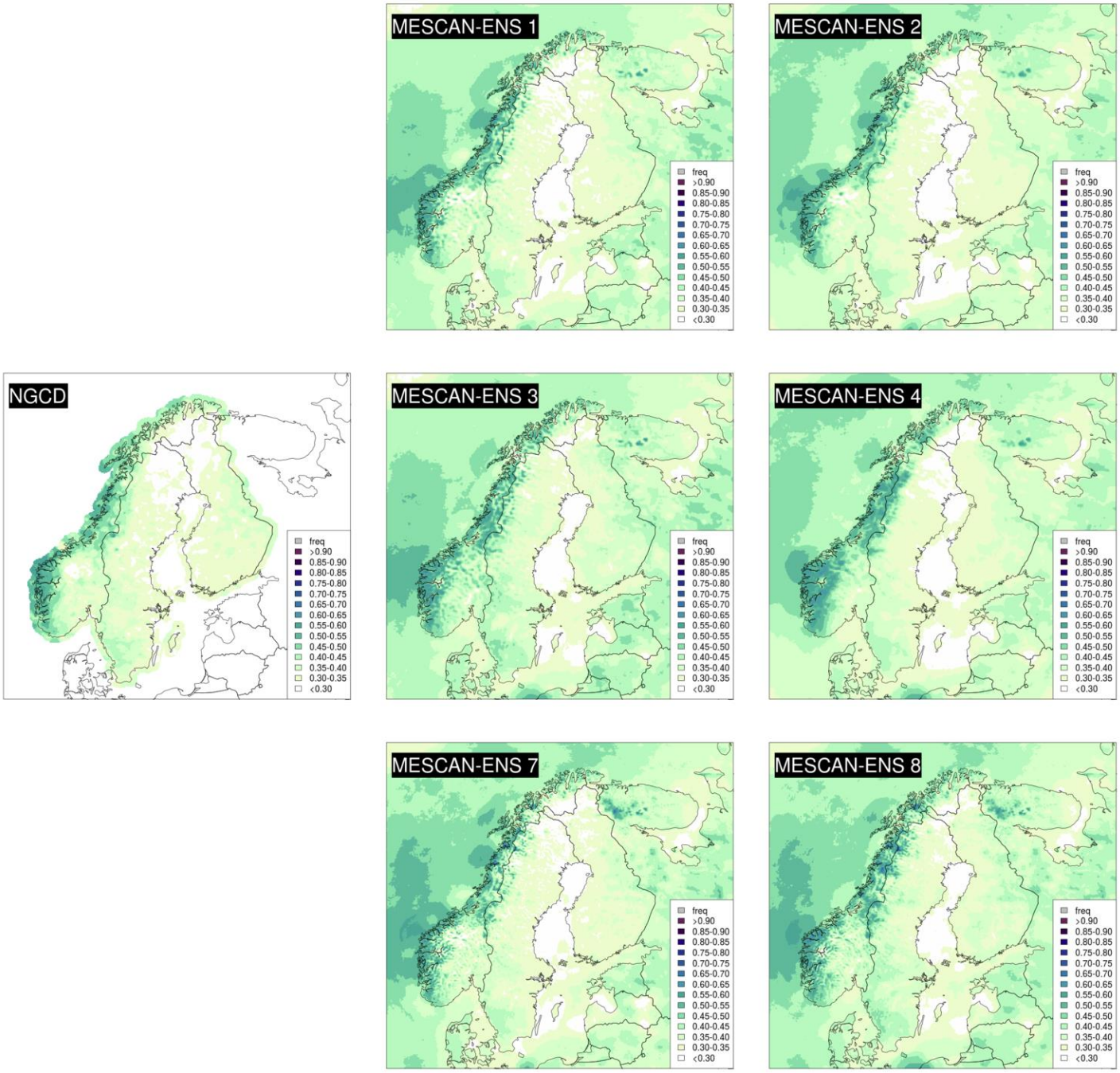


**Figure 4.3.2.8:** Annual frequency of wet days (≥ 1mm/d, fraction, 2006-2010). Rescaled to 5Km ETRS-LAEA coordinate system. Reference: NGCD.

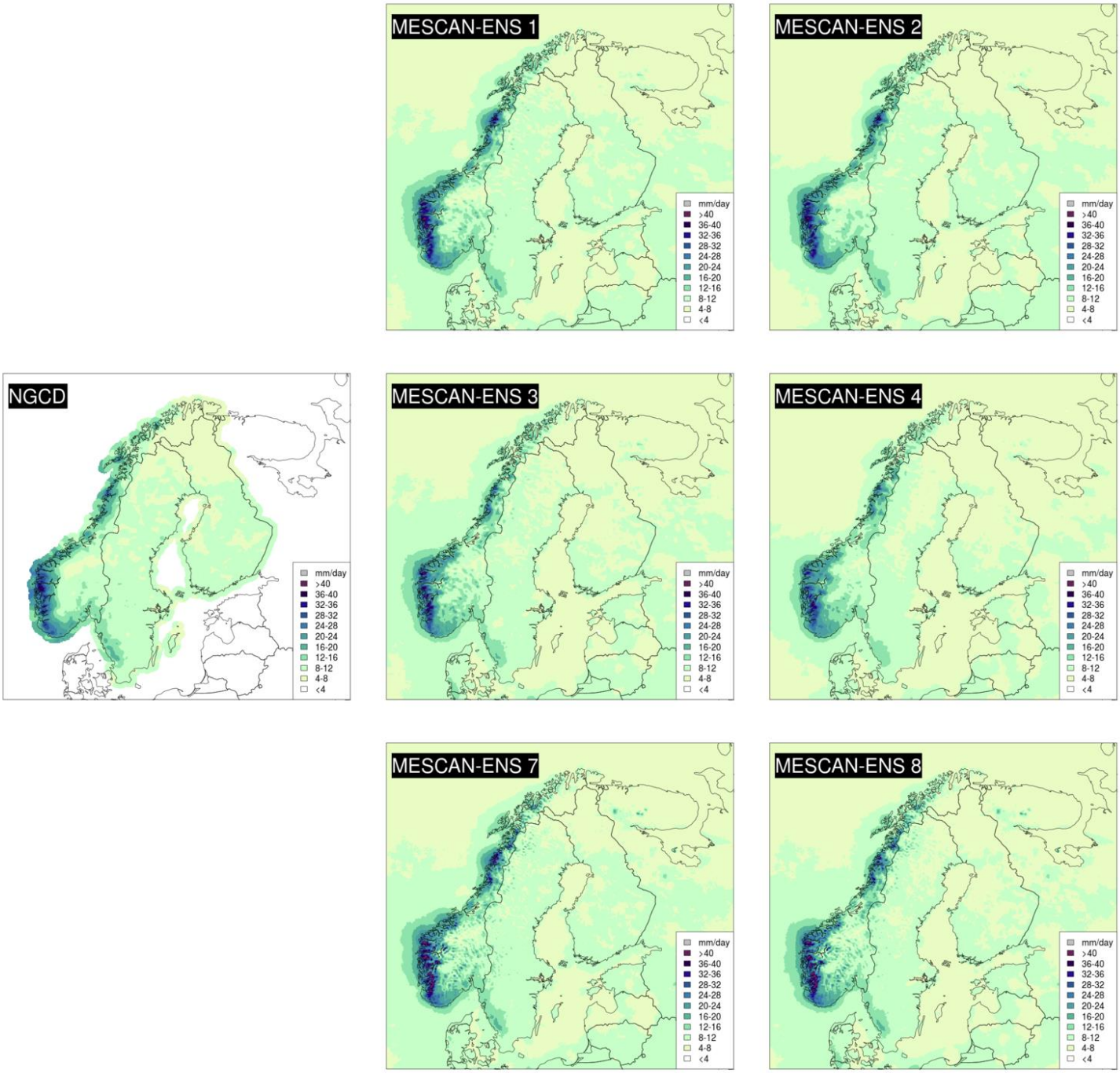


**Figure 4.3.2.9:** 95% quantile of daily precipitation (mm/d, 2005-2008). Rescaled to 5Km ETRS-LAEA coordinate system. Reference: NGCD.

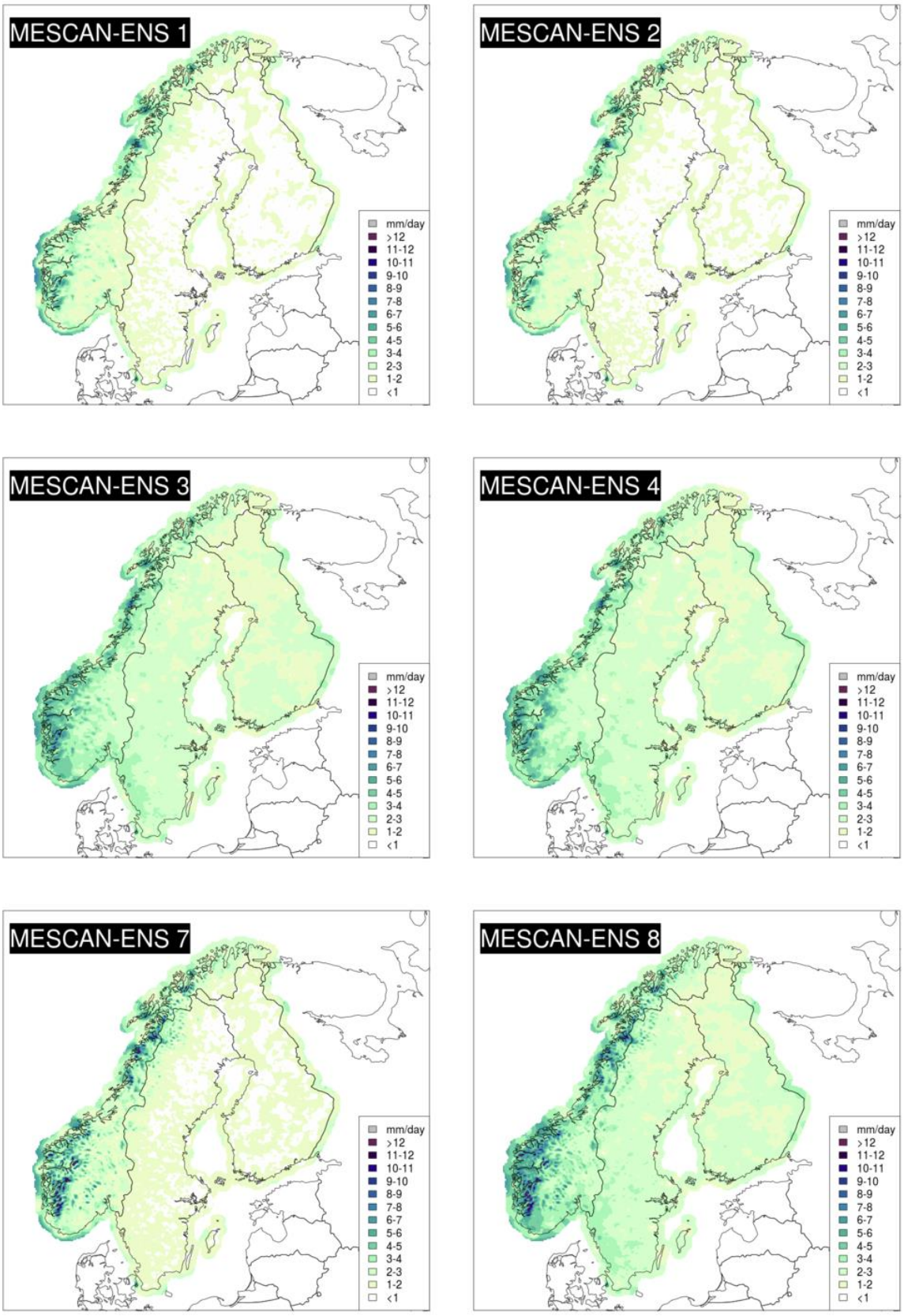


**Figure 4.3.2.10:** Root Mean Square Error of daily precipitation (mm/d, 2006-2010). Rescaled to 5Km ETRS-LAEA coordinate system. Reference: NGCD

**UKMO-ENS**

The UK Met Office provided a 20-members ensemble reanalysis system (UKMO-ENS) along with the deterministic reanalysis (UKMO). As shown in Figure 4.3.2.11, the UKMO-ENS model struggles with the representation of precipitation over Fennoscandia.
Both the mean annual precipitation and the frequency of wet days are overestimated by the UKMO-ENS members (only the first member is shown here, the other members show similar results). The differences are particularly evident over Finland and Sweden, where both the totals and the wet-day frequency is the double with respect to the reference. In Figure 4.3.2.3 the departure of the ensemble (blue line) from the reference is evident, with a large overestimation of yearly totals even in the lower quantiles (up to 1000 mm).
The UKMO-ENS dataset used for our evaluation consists of the first 6 hours of model output after the 3DVAR assimilation cycle, which is performed every 6 hours. Those fields are affected by a systematic overestimation of precipitation probably caused by the adaption of boundary conditions from the global model to the local area model configuration (so-called "spin-up" issue. Peter Jermey, personal communication). This issue can be solved by considering a selection of output fields with forecast lead time greater than 6 hours but this data was not available for download via MARS (to the best knowledge of the Authors).

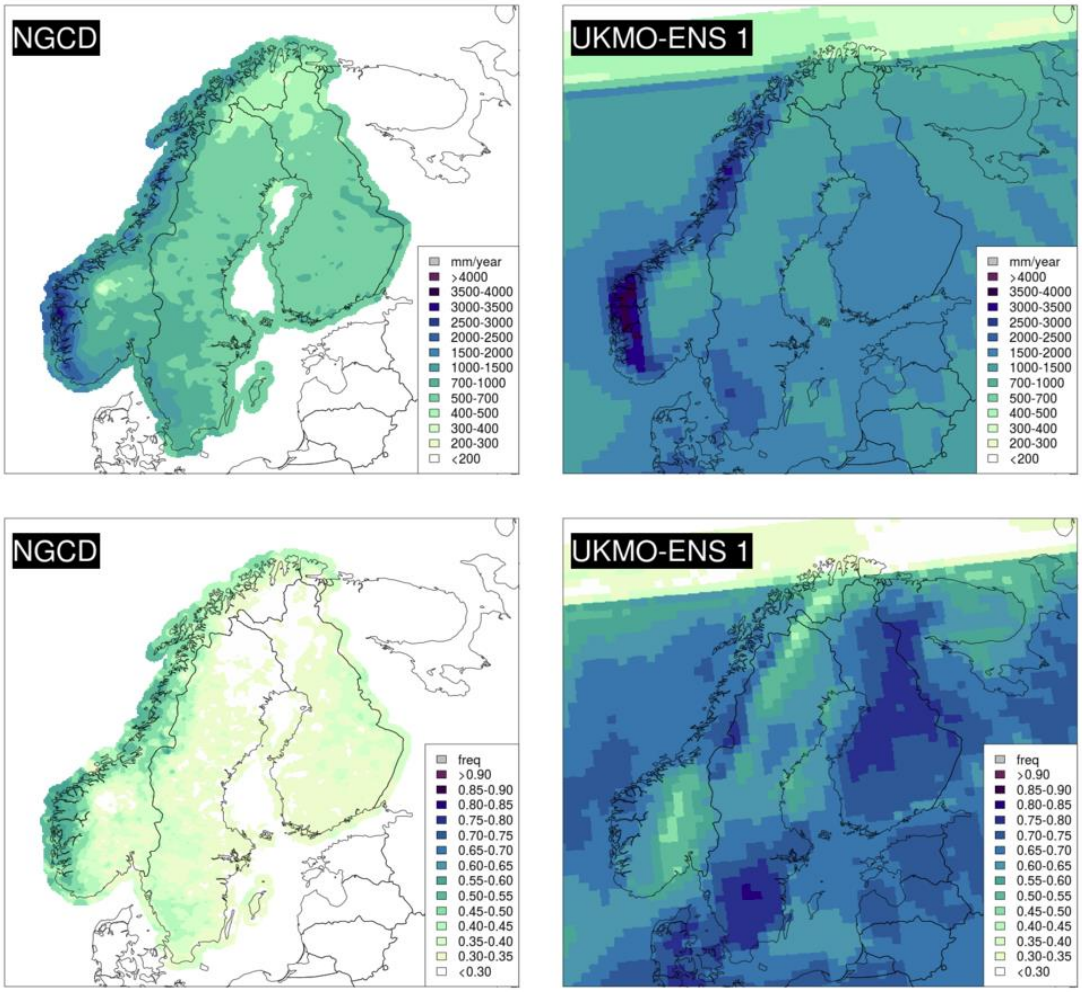


**Figure 4.3.2.11:** Comparison between NGCD and UKMO-ENS (member 1) for the mean total annual precipitation (top panel) and the mean annual frequency of wet days (bottom panel). Rescaled to 5Km ETRS-LAEA coordinate system.

**COSMO-ENS**

The Deutscher Wetterdienst and the University of Bonn has produced a probabilistic regional reanalysis system in the form of a 20+1-members ensemble model at a 12-km grid spacing [Bach, 2016]. The performances of such a model in representing the precipitation over Fennoscandia are here discussed.
In Figure 4.3.2.12, an evaluation of the probabilistic skills of COSMO-ENS is shown by means of the Brier Skill-Score (BSS) for the threshold of 1 mm of daily precipitation. Values are generally bigger than 0.5 over most of the domain, with the exception of the Finnmark region and the mountains in southern Norway, where the reliability component (top-right panel) displays great values and the resolution one shows small values.
The patchy pattern visible in the Figure, though, could be due to the observational nature of the reference used (NGCD). In order to understand if this is the case, we computed the BSS for the same threshold and period using E-OBS as a reference, obtaining generally better results, i.e. smaller values of BBS in the above-mentioned problematic areas (Figure 4.3.2.13).
Figure 4.3.2.14 summarize the general behaviour of COSMO-ENS, displaying both the mean value of the ensemble (center panels) and the interquantile range (the difference between 10th to 90th quantiles, right panels) of - from the top to the bottom - mean annual precipitation, mean annual frequency of wet days and mean annual 95th quantile. The panels on the left show the reference.
The quantile-quantile plot in Figure 4.3.2.3 shows how COSMO-ENS (and in particular the control member) reproduces the same distribution as COSMO6-REA over the reference domain, apart from departing from it in the upper quantiles (over 3000 mm), where it shows smaller maxima of mean annual precipitation with respect to the deterministic version of the model.

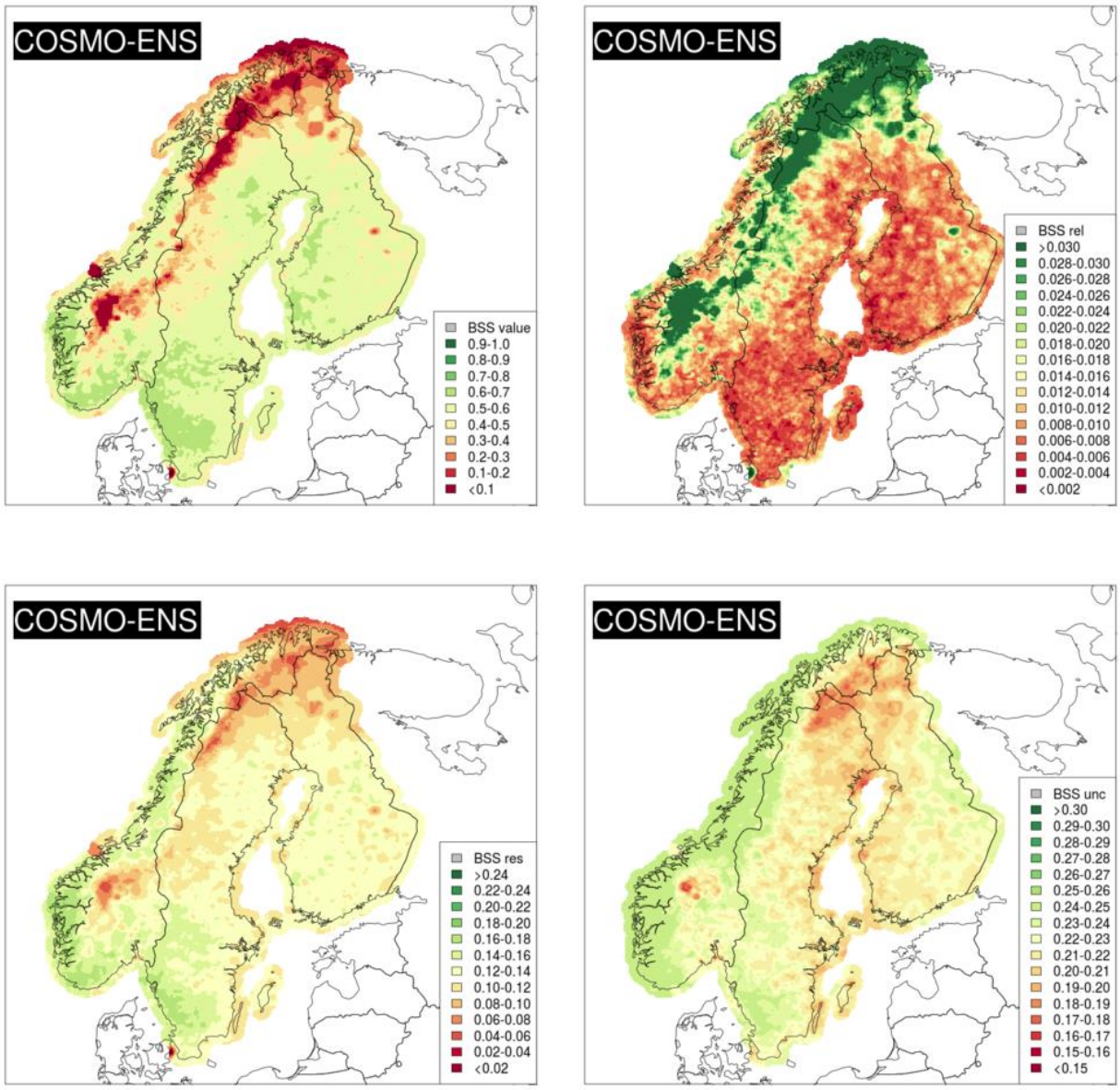


**Figure 4.3.2.12**: Brier Skill Score computed for the threshold daily rain > 1mm (top-left panel). Reliability component (top-right panel), resolution component (bottom-left panel), uncertainty component (bottom-right panel). NGCD as reference. Rescaled to 5Km ETRS-LAEA coordinate system.

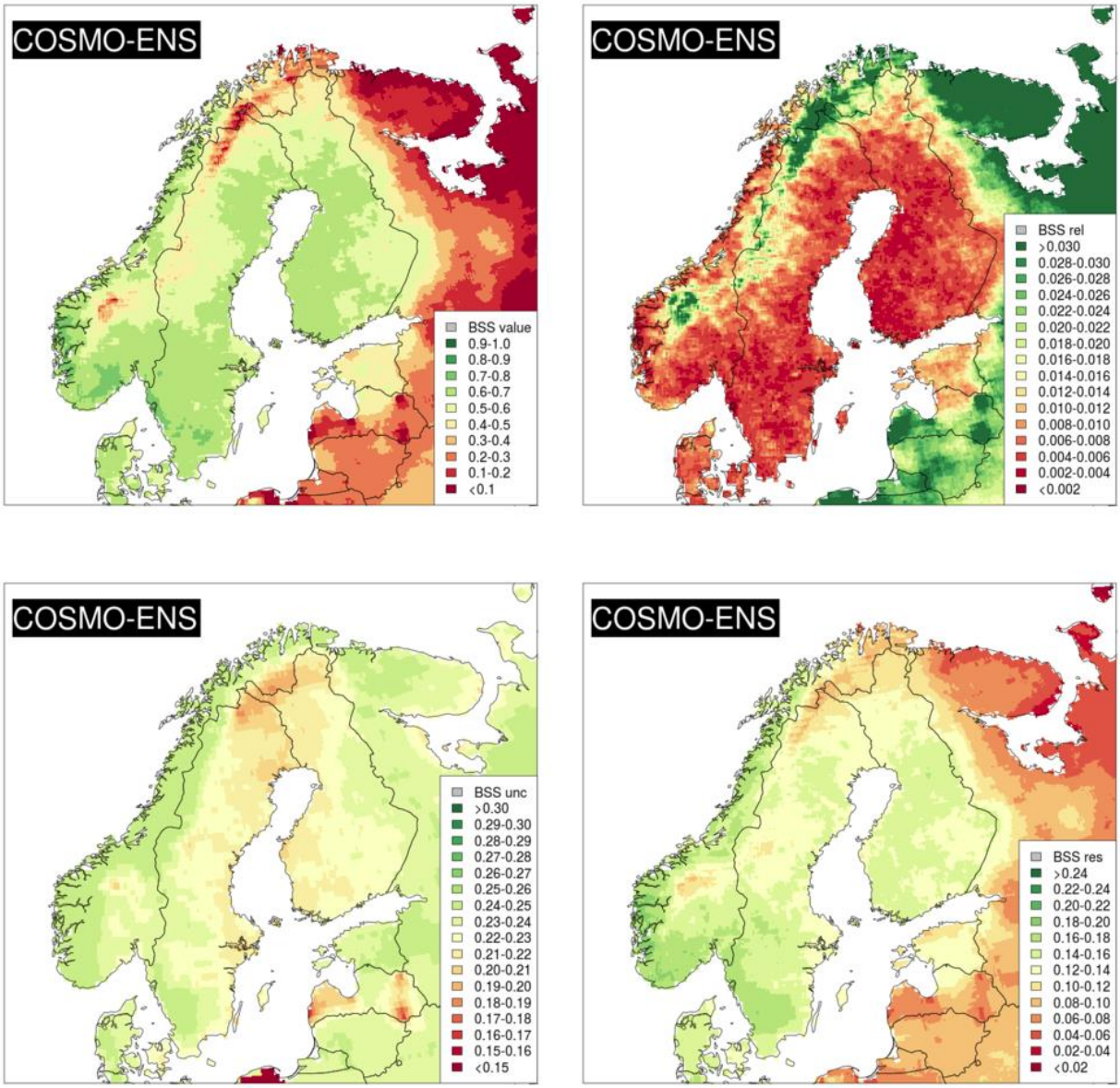


**Figure 4.3.2.13:** Brier Skill Score computed for the threshold: daily rain > 1mm (top-left panel). Reliability component (top-right panel), resolution component (bottom-left panel), uncertainty component (bottom-right panel). E-OBS as reference. Rescaled to 5Km ETRS-LAEA coordinate system.

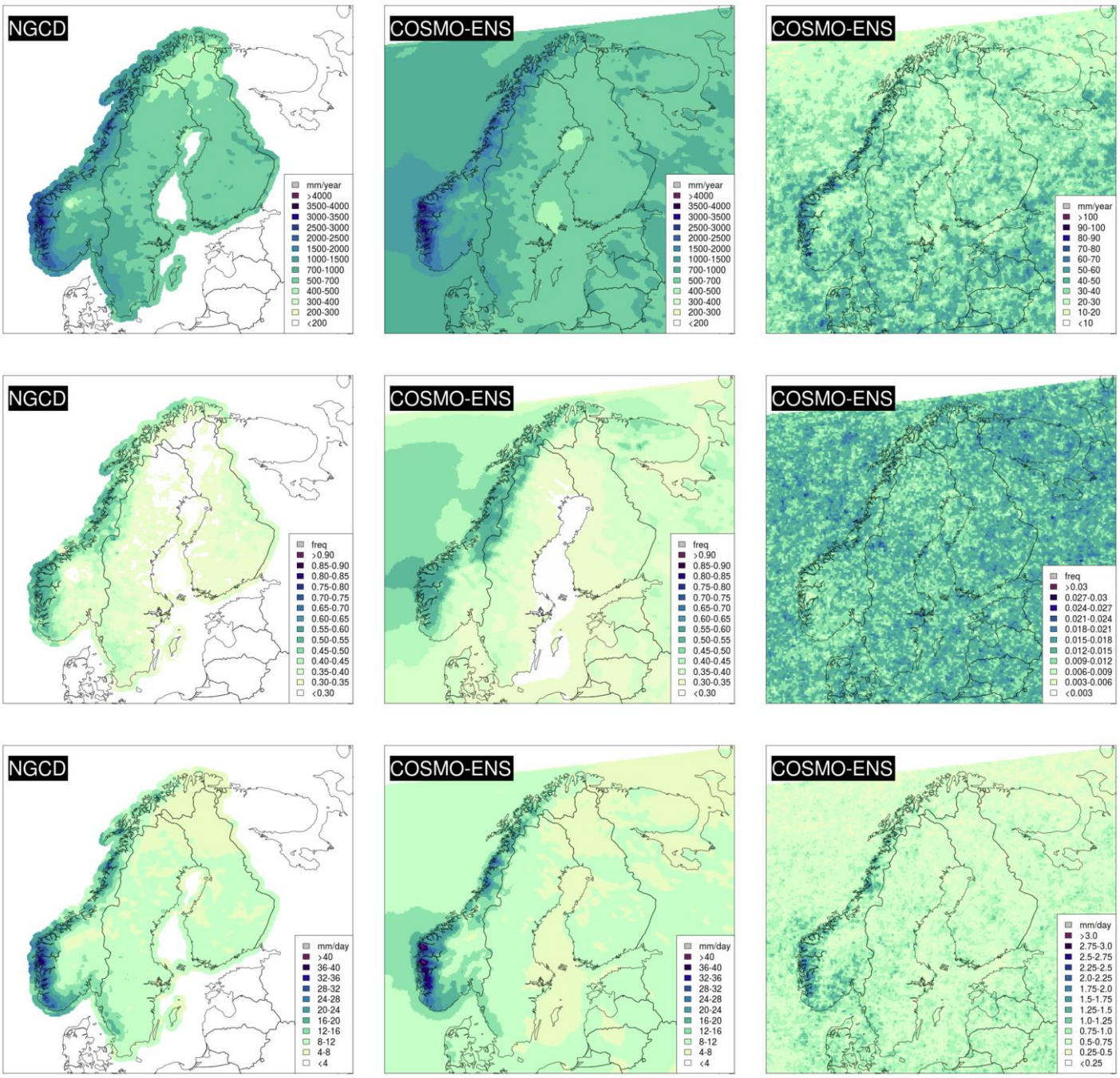


**Figure 4.3.2.14:** Mean total annual precipitation (top panels), mean annual frequency of wet days (middle panels), mean annual 95th percentile (bottom panels). Left panels are NGCD reference, central panels are COSMO-ENS ensemble mean and right panels are COSMO-ENS interquantile range (10 to 90th percentile).

# Evaluation over subdomains

This section investigates how different models perform in reproducing the annual cycle of precipitation on a monthly basis over significant subdomains (see the red boxes in Figure 4.3.2.1). Figures 4.3.2.15 and 4.3.2.16 show the mean monthly precipitation and the monthly frequency of wet days for the period 2006-2010, respectively.
With reference to the mean monthly precipitation timeseries, all models seem to satisfactory reproduce the annual cycle over all the three areas, while they tend to overestimate the precipitation (NGCD=reference, black like), especially along the west coast and in Lapland. NORA10 and UKMO tend to overestimate the precipitation over throughout the year. HARMONIE v1 shows higher amounts of precipitation during summer both in the Oslo area and in Lapland, whereas it seems in line with the reference during winter. HARMONIE v2 is the only reanalysis which underestimates the precipitation in the Oslo area throughout the entire year. The datasets better matching the mean monthly precipitation over the three considered areas are MESCAN and MESAN.
With reference to the monthly frequency of wet days, all models show the same annual cycle, apart from HARMONIE v1, which overestimates the frequency of wet days only during the warm season and especially over Lapland and the Oslo area. The models, which best reproduce the wet-day frequency on a monthly basis over the three considered areas, are MESCAN and MESAN.

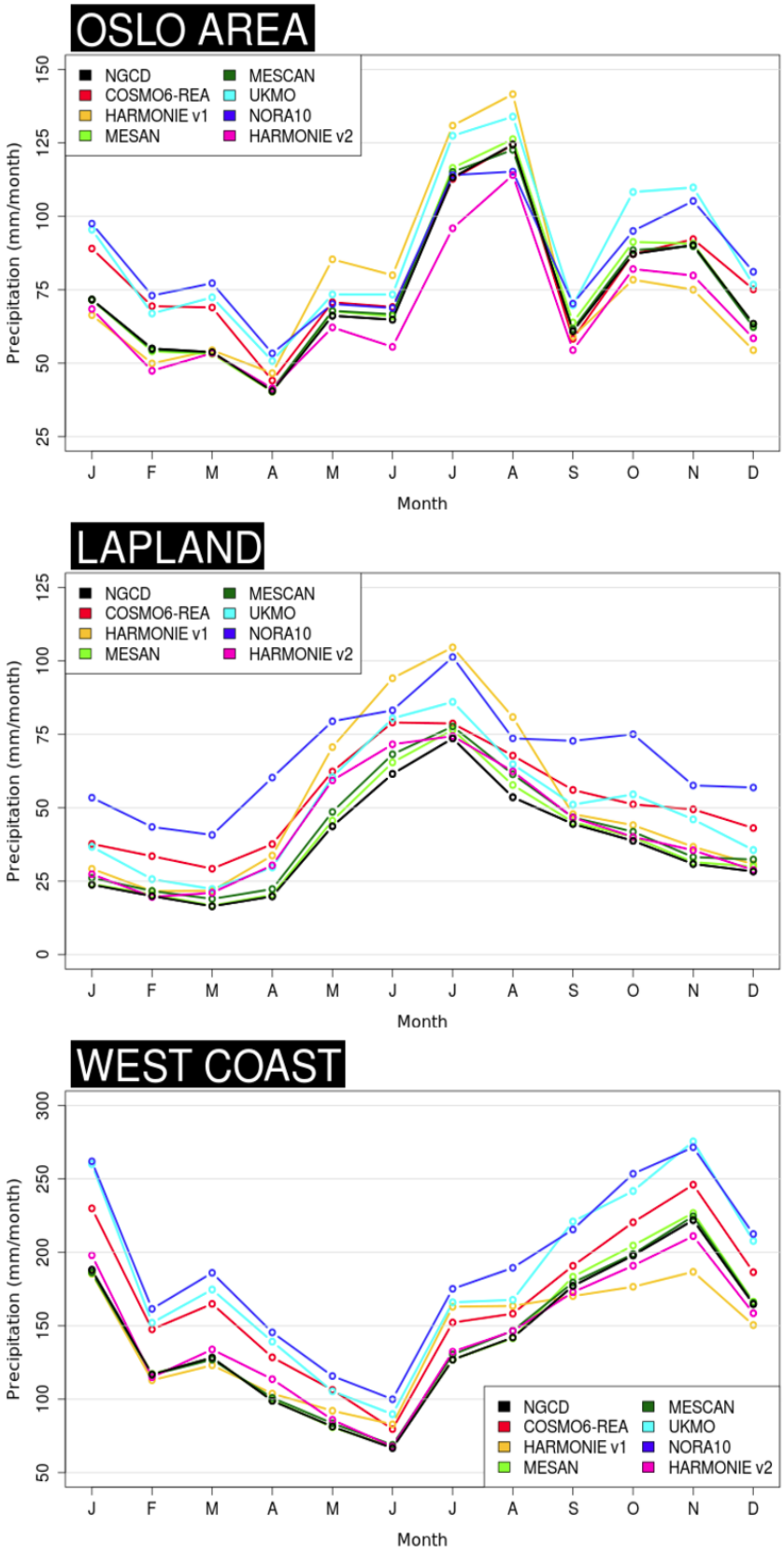


**Figure 4.3.2.15:** Timeseries of the (spatially averaged) monthly mean total precipitation over three different areas: Oslo (top panel), Lapland (center panel), west coast (bottom panel). Values from the rescaled to 5km ETRS-LAEA coordinate system grid.

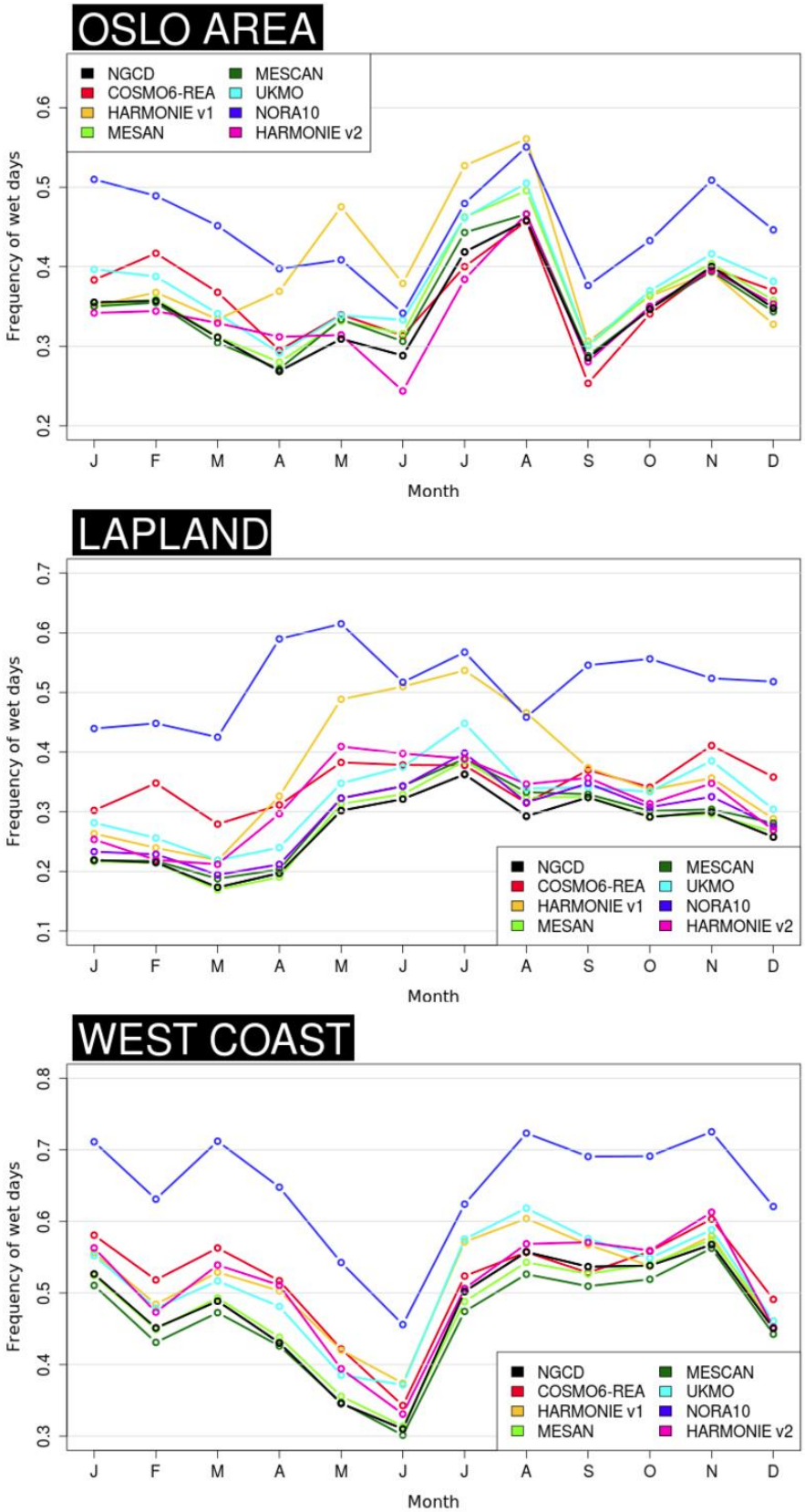


**Figure 4.3.2.16:** Timeseries of the (spatially averaged) monthly mean frequency of wet days over three different areas: Oslo (top panel), Lapland (center panel), west coast (bottom panel). Values from the rescaled to 5km ETRS-LAEA coordinate system grid.

### Verification of daily precipitation with a scale-separation method

The wavelet-based scale-separation Mean Squared Error (MSE) skill-score [Casati et al., 2015], which is based on [Casati et al., 2004] and [Casati, 2010] has been used in this section to assess the added value of enhanced resolution in UERRA regional reanalysis. The Haar wavelet filter has been used to decompose the precipitation field (reanalysis and observations) into the sum of spatial components on different scales, then the verification is done separately on those spatial components. In general, the wavelet decomposition depends on the size of the domain chosen, for this reason the results presented in this section are an average over 25 slightly different subdomains (i.e. the south-western corner of the domain has been placed in different positions). Besides, we are presenting time averages and as a consequence the results obtained can be considered representative of the average situation over the domain.

In Figure 4.3.2.17, the squared energy of daily precipitation as a function of the spatial scale is shown for most of the datasets considered in our evaluation. The squared energy is an indication of the average amount of precipitation represented by the model at a particular spatial scale. The precipitation fields are considered over their original grids before the application of any post-processing procedure (e.g. regridding). The time interval considered cover the years from 2006 to 2010 (ERA5, 2010 to 2016 that is the only period available) and only days with more than 5% of the domain with precipitation greater than 1 mm/day have been considered. In general, the spatial structure of regional reanalyses includes a broader range of scales than global reanalyses and the mode of the energy distribution is shifted towards smaller spatial scales, moreover regional reanalyses simulates higher amounts of precipitation across their ranges of spatial scales. In particular, UKMO and COSMO simulate more precipitation than the others, while the MESCAN-SURFEX downscaling datasets and the two versions of HARMONIE presents squared energies not too different from the global reanalyses, though spanning a wider range of scales. The MESCAN-SURFEX compared to MESAN has less energy on the larger spatial scales. Note that the spatial structures of each ensemble within an ensemble dataset are all (almost) identical, this is not the case for MESCAN-SURFEX that is a different form of ensemble dataset (i.e. a collection of precipitation fields derived from different setup of post-processing systems instead of being the result of perturbations on the initial state of a model) and this can be clearly seen from Figure 4.3.2.17.

The wavelet-based MSE skill-score is shown in Figures 4.3.2.18 and 4.3.2.19 for the lower and the higher resolution grids, respectively. In Figure 4.3.2.18, the reference dataset is EOBS and the spatial domain considered for evaluation is the entire EOBS domain (i.e. most of continental Europe), so to include in our evaluation a wider range of scales; the datasets have been rescaled to the 0.25° regular grid and global reanalyses have been considered together with the regional ones. In Figure 4.3.2.19, the reference dataset is NGCD and the spatial domain is Fennoscandia, the evaluation focus on the local scales; the datasets have been downscaled to the 5Km grid. Both reference datasets do not have data over the sea surface and this may have an impact on the evaluation on the larger scales because of the breaks in the spatial continuity of the synoptic precipitation systems.

The results show that the downscaling datasets MESAN and MESCAN-SURFEX (both deterministic and ensemble) have the spatial structure of precipitation more similar to the reference datasets. The performances of the other regional reanalyses are similar: on the coarser grid, it is not possible to distinguish between regional and global reanalyses (only ERA20C performs worse than the others) and this may be due to the definition of the random component in the MSE skill-score; on the finer grid, the COSMO ensemble performs worse than the other regional reanalyses. Figure 4.3.2.20 refers to the finer grid and it shows only MESCAN-SURFEX MSE skill-scores. The deterministic version scores better than the others. The behaviour of the MSE skill-scores for the ensemble members allows us to divide the members in three “clusters”.

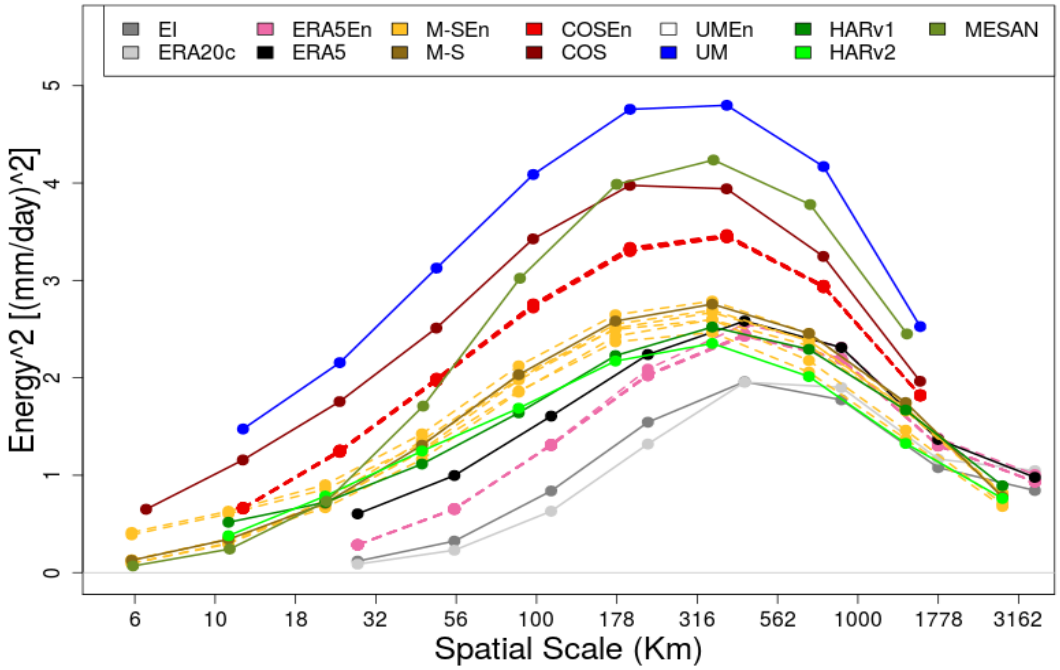


**Figure 4.3.2.17:** Squared Energy of the scale components of daily precipitation fields computed for each model considering output on its original grid. Time interval considered is 2006-2010 (ERA5, 2010-2016).

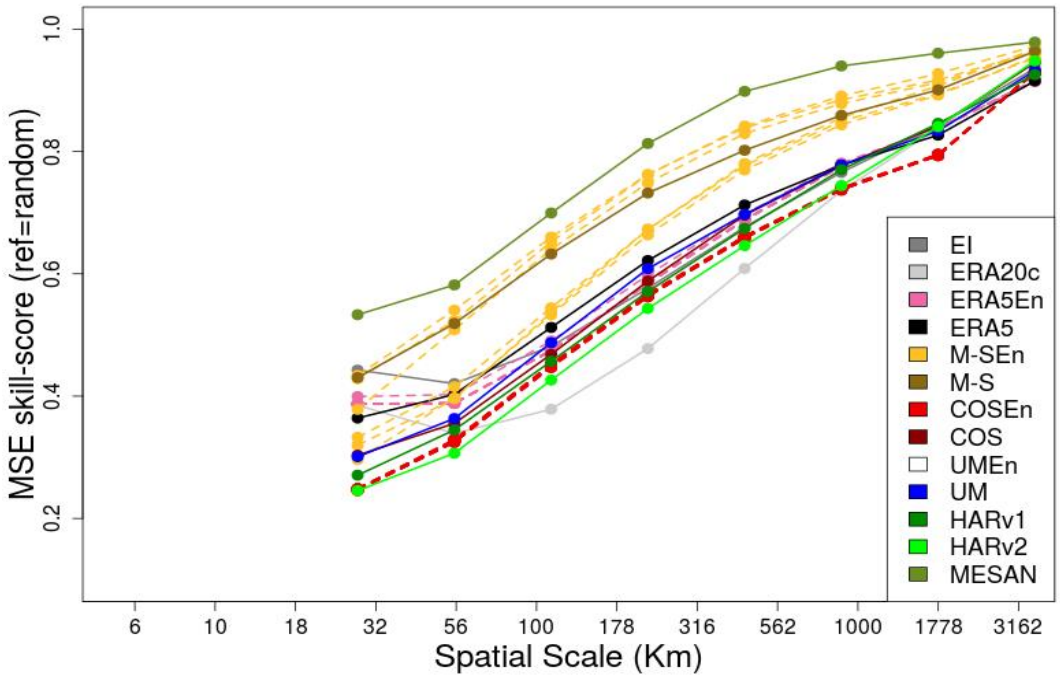


**Figure 4.3.2.18:** Scale-separation MSE skill-score, the datasets have been rescaled over the 0.25° regular grid. Time period considered: 2006-2010 (ERA5, 2010-2016). Reference dataset is EOBS.

The best skill-score is achieved by members 1 and 2, where the post-processing is based on the higher-density station network. Ensemble members 3, 4, and 8 display the worse skill-scores, because of the post-processing based on the lower-density station network. Member 7 falls in between those two clusters: for spatial scales larger than 50Km is more similar to members 1 and 2, while for the finer spatial scales it is similar to members 3, 4 and 8. Member 7 makes use of ALADIN 5.5Km background (instead of a downscaling from HARMONIE at 11Km) and the post-processing is based on the higher-density station network.

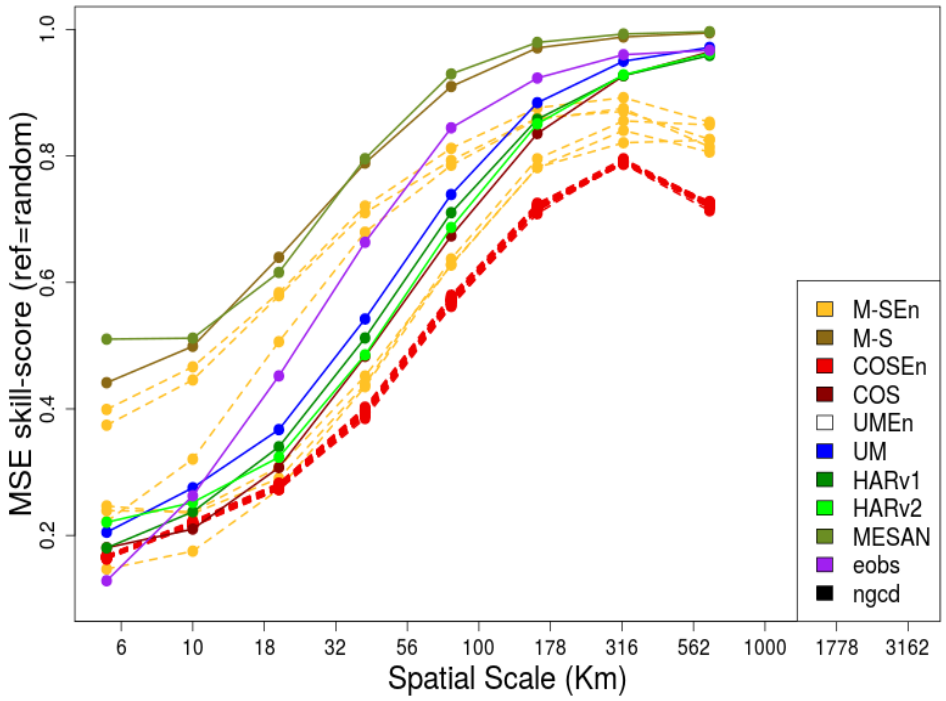


**Figure 4.3.2.19:** Scale-separation MSE skill-score, the datasets have been rescaled over the 5Km ETRS-LAEA coordinate system. Time period considered: 2006-2010 (ERA5, 2010-2016). Reference dataset is NGCD.

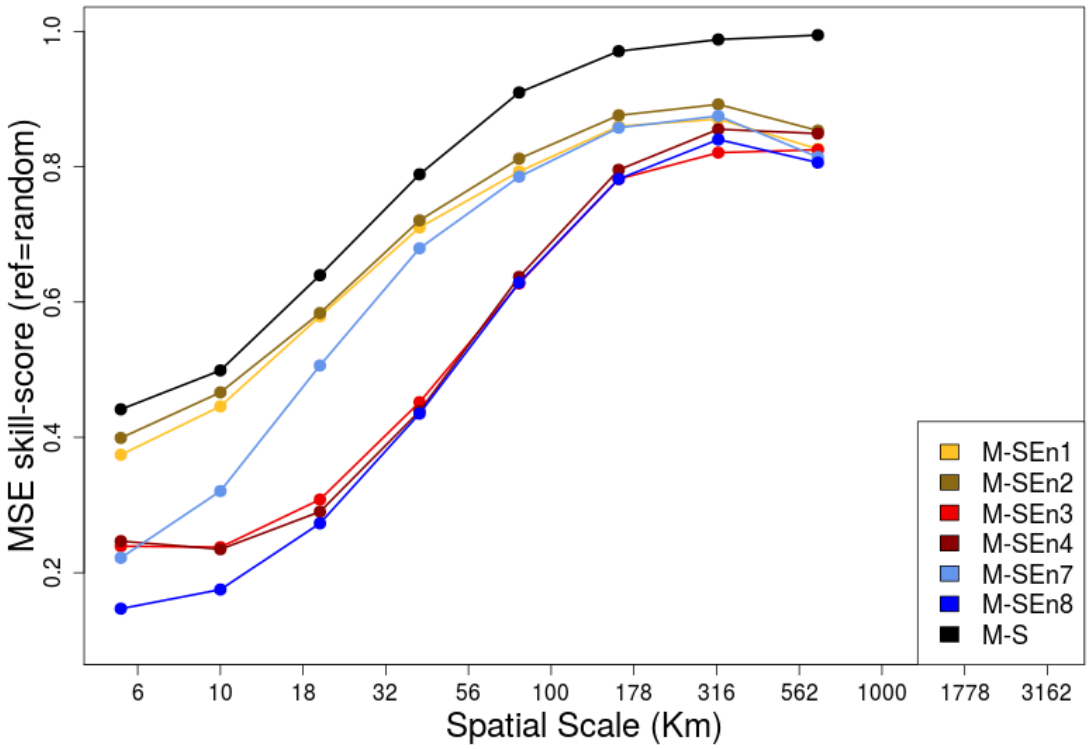


**Figure 4.3.2.20:** Scale-separation MSE skill-score for MESCAN-SURFEX, the datasets have been rescaled over the 5Km ETRS-LAEA coordinate system. Time period considered: 2006-2010 (ERA5, 2010-2016). Reference dataset is NGCD.

## Fennoscandia – Main outcomes

Regional reanalyses:

- Added value compared to global reanalyses, especially in complex terrain along the Norwegian coast. In general, regional reanalyses represent precipitation fields with a spatial structure more similar to the observational gridded datasets than global reanalyses.
- Tend to overestimate precipitation amounts and frequency, especially in complex terrain (i.e. for the highest quantiles). However, the UERRA reanalyses provides satisfactory results (especially for the wet-day frequency) when compared with the hindcast dataset NORA10, which is currently used at MET Norway.
- Regional reanalyses simulate realistic precipitation features at high-resolution spatial scales. Regional reanalyses often shows much more detailed precipitation structures than observational gridded datasets in region of low station density.
- HARMONIE v2 show the best performance of all regional reanalyses. In particular, it is the only reanalysis that can represent not only the precipitation hot-spot in western Norway but also the dry area of Lapland in the north.

- COSMO-ENS provides satisfactory results both on the representation of precipitation and of its uncertainty, as shown by the Brier skill-score.

Downscaling datasets:

- Additional value compared to regional reanalyses, especially in regions with dense station network.
- The local density of the station network is the most important factor in determining the quality of the post-processed precipitation fields.
- Results are more similar to the observational gridded datasets than for regional reanalysis with regard to both precipitation amounts and frequency of wet days.
- The spatial structures of the precipitation field are similar to the observational gridded datasets, though the downscaling datasets reach a very high detail of the precipitation pattern even in complex terrain.
- MESCAN-SURFEX products provide a more detailed information than MESAN.

General comments:

- The biggest differences from the reference and the lowest Brier skill score are found in complex topography, for higher precipitation amounts and in areas characterized by a sparse station network.
- Annual cycle is mostly well reproduced in all datasets.
- The spatial distribution of annual accumulated precipitation and the 95% quantile of daily precipitation are well reproduced by all datasets.
- The time interval 1986-1990 has also been considered in the evaluation (see Supplementary materials) and the results were similar to the ones obtained for 2006-2010.
- User should be aware of the effective resolution of datasets, which for datasets as E-OBS and the reference NGCD is coarser than the grid resolution. Besides, because of the significant differences in the station network, the observational gridded datasets in the northern part of Fennoscandia are affected by larger uncertainties.

# 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.

**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.

**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].

**Value of method**

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

## 5.2 Examples of application

The comparison of global radiation, using CM SAF satellite observation data and reanalysis data from COSMO-REA 6 and HARMONIE is shown in deliverable D3.5 [Borsche et al., 2016]

# 6. Method E: Ensemble based methods

## 6.1 Method description

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.
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, 2008], 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.

## 6.2 Examples of application

Method E was investigated for precipitation, using gridded datasets for evaluation. The results are presented in section 4. Method E is also used for the verification of wind speed. Reference data are station measurements and thus the results are presented in section 3.

# 7. Conclusion

The evaluation of reanalysis datasets, produced during the UERRA project, includes the investigation of different parameters (wind speed, temperature, precipitation and radiation, as well as climate indices) on various spatial and temporal scales. The comparison with further datasets, like global reanalyses, makes it possible to assess the quality of the UERRA datasets not only in absolute terms but also to determine the added value of the higher reanalysis resolutions. Five various evaluation methods are used, which were determined during the UERRA workshop (D3.1). The application of these methods was demonstrated in D3.5, using preliminary data sets. It was realized that only the combination of multiple methods can characterize the complexity of reanalysis systems. In combination, the methods are shown to be appropriate tools for reanalyses comparisons, and can be applied for user friendly estimates of regional reanalyses uncertainty.

**Method A** (Use of observation feedback statistics) is applicable to assess the fit between the reanalysis and assimilated observations. The comparison of background and observation provides a tool for data quality control as well as monitoring of the data assimilation system. The advantage is that observation operators are already applied to the model fields, so that the comparison is optimal in the sense that like-is-compared-with-like. It is harder to interpret when different reanalysis systems are compared with each other (as they differ in their observation operators, assimilation methods and quality control, all of which have an influence on the feedback statistics. Method A has been applied with the UKMO reanalysis.

**Method B** (Comparison against station observations) allows comparison with data, regardless whether they are assimilated or not. It ignores the observation operator. It may be the most user friendly practice. The main issues are strong location dependencies, which have to be pointed out carefully to the users to avoid wrong expectations in areas not investigated but potentially of interest to the users.

This is demonstrated in section 3, where wind speed of the various regional reanalyses of UERRA is validated with station data over Germany. For each reanalysis system one can find a station, where one regional reanalysis outperform the others. Averaged over all station locations, the regional reanalyses show significant better correlations than ERA-Interim. The regional systems do not vary significantly, though HARMONIE tends to have somewhat smaller correlation. For higher elevated stations all reanalyses lose correlation, and the bias increases, due to higher discrepancies between modelled and real topography. Considering only stations beneath 500m height COSMO-REA6, COSMO-REA12 and HARMONIE are nearly unbiased, whereas UM and MESCAN overestimates the wind speed of more than 0.2 m/s, especially in the northern part of Germany. In general, all reanalysis systems overestimate low wind speeds and underestimate high wind speeds, which could be explained by the spatial resolution of the reanalyses. Thus the bias strongly depends on station location, wind speed and model system. The analysis of further scores and skill scores document good performance of all model systems, also for extreme events. However, it is much more useful to work with or investigate percentiles, rather than absolute values. In the latter case the results can become worse, due to strong local biases. In addition to the analysis of 10m wind speed, the use of level wind speed up to 100m height, shows good results for the regional systems COSMO-REA12, HARMONIE, MESCAN and UM as well. For tower locations over sea they demonstrate significant better correlations than the global systems ERA-Interim and ERA20C. The UERRA reanalyses reproduce the annual cycle of wind speed for all heights.

**Method C** (Comparison against gridded station observations) provides the opportunity to assess a model on various spatial scales. The application with respect to climate indices and precipitation show varying results of practicability. Climate indices, based on daily minimum and maximum temperature were computed for UM and HARMONIE and compared to the gridded data set E-OBS. Differences between reanalysis and E-OBS temperature show

strong local variance, which is mainly caused by the inhomogeneous density of observations, used for E-OBS. Moreover, high differences between model and E-OBS are reached in regions with pronounced topographic features, due to the use of different topography maps in E-OBS and the reanalyses. Hence, the assessment of reanalysis skill is difficult, since differences between reanalysis and E-OBS can often be related to E-OBS characteristics rather than to problems in the reanalysis systems. The annual cycle of various investigated regions of daily minimum and maximum temperature is reproduced well by the regional systems. The differences towards E-OBS are largest in the coldest season. Histograms of daily temperature show remarkably similarity between reanalyses and E-OBS. However there are exceptional cases as well. At the Iberian Peninsula, HARMONIE shows a shift of +2°C, which underlines again the strong local differences of model behaviour. The investigation of climate indices, (frost days, tropical nights, ice days and summer days), results in partially high differences of more than 40 days per year. While HARMONIE tends to overestimate the four indices on large scales, UM shows a more balanced picture for frost days and summer days. Due to the strong local variations and discrepancies towards E-OBS the reanalyses seem not to be applicable for the computation of climate indices. This could be based on the fact, that absolute thresholds are used for climate indices, which are not suitable for the reanalysis datasets, due to strong local biases and possibly their dependence on the nominal and actual spatial resolution. For precipitation the outcomes show stronger variations between the model systems. In general all models are able to reproduce the spatial precipitation patterns of reference gridded data sets. An added value of regional reanalyses compared to global reanalyses is identified, especially in complex terrain. However, the reanalyses tend to overestimate precipitation amounts and frequencies. While COSMO-REA6 shows the best performance for the Alpine region, HARMONIE v2 shows the best performance for Fennoscandia. The downscaling experiment MESCAN in particular, shows closer fit to the observations than the regional reanalyses at locations with high observation density, which is not surprising, as the observations are used in MESCAN. The annual cycle is well reproduced in all data sets for the investigated locations. In general the model fitness decreases with higher precipitation amounts, more complex terrain, lower catchment size and station density.

**Method D** (Comparison against satellite data) assessed radiation data of reanalysis systems. Regional reanalyses offer spatial datasets of various radiation components, which are mainly important for the energy sector. For global radiation COSMO-REA6 and HARMONIE show a good overall agreement with satellite data from CM SAF on a yearly scale. COSMO-REA6 has in general a negative bias around 10 percent, while HARMONIE results are more heterogeneous. Over land the deviations are generally approximately below 5 %, except at some locations at the Mediterranean coast. Over the Mediterranean Sea HARMONIE has a negative bias and over the North West Atlantic a positive bias. Moreover, both models underestimate high radiation and COSMO-REA6 overestimates low radiations as well. The annual cycle is well reproduced for both model systems and daily correlations are higher than 0.97 for various spatial areas.

**Method E** is important for uncertainty estimations of UERRA ensembles. Typical techniques are rank histograms, CRPS, Brier score, reliability diagrams and ROC curves. They were used in deliverable D2.14 as well, considering a short time period for the summer and the winter season, exemplarily. The additional investigations of ensemble data sets shown here enlarge the examined time period. For temperature and wind speed the outcomes show a strong underdispersion of UM and COSMO-REA12 ensemble spread.

Considering wind speed, the spread in summer is slightly higher than in winter. The analysis of the Brier score for wind speed identifies advantages for COSMO-REA12 over the UM ensemble, due to a lack of reliability in the British model system. This could be caused by less spatial resolution, which is three times higher for COSMO-REA12 than for UM. For precipitation COSMO-REA12 shows satisfactory results concerning the uncertainty, i.e., the spread can serve as a proxy for the precipitation uncertainty.

All reanalysis data sets, produced during the UERRA project, were validated in this deliverable, assessing various time periods and model variables. The model systems show overall a good performance. However, it was shown in method B and C that it is important to carefully consider the used reference observation datasets, when evaluating the model fitness. It is recommended to use more methods and different comparison data sets, to get a more universal picture of reanalysis fitness. The added value of regional reanalyses was demonstrated with several examples. The application of reanalysis ensembles for uncertainty estimation was shown and discussed for the single-model UERRA data sets, and the multi-model UERRA ensemble.

# 8. References

Bach, L.., Schraff, C., Keller, J. D., Hense, A.: Towards a probabilistic regional reanalysis for Europe: evaluation from experiments, Tellus, 68, doi:10.4302/tellusa.v68.32209, 2016

Bazile, R., Abida, R., Szczypta, C., Verelle, A., Soci, C., and Le Moigne, P.: Ensemble surface MESCAN analysis. Deliverable D3.5 of project: 607193 UERRA. Available at http://uerra.eu/project-overview/all-deliverables.html

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.

Borsche, M. et al.: Preliminary report of assessment of regional reanalyses - first results. Deliverable D3.5 of project: 607193 UERRA. Available at http://www.uerra.eu/publications/deliverable-reports.html

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., Ross, G. and Stephenson, D.B.: A new intensity-scale approach for the verification of spatial precipitation forecasts. Meteorological Applications, 11(2), 141-154, 2004

Casati, B.: New developments of the intensity-scale technique within the Spatial Verification Methods Intercomparison Project. Weather and Forecasting, 25(1), 113-143, 2010.

B. Casati, A. Glazer, J. Milbrandt, V.Fortin. An intensity-scale skill score to assess the added value of enhanced resolution. Poster at the 15th EMS Annual Meeting & 12th European Conference on Applications of Meteorology (ECAM), 07–11 September 2015, Sofia, Bulgaria

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

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.

Frei, C., and F. Isotta, 2017: From single estimates to ensembles: A probabilistic spatial analysis for the Alpine region. (in preparation).

Gisnås, K., Etzelmüller, B., Lussana, C., Hjort, J., Sannel, A.B.K., Isaksen, K., Westermann, S., Kuhry, P., Christiansen, H.H., Frampton, A. and Åkerman, J.: Permafrost Map for Norway, Sweden and Finland. *Permafrost and Periglacial Processes*, *28*(2), pp.359-378, 2017

Gupta, H. V., and Kling, H. and Yilmaz, K. K. and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. of Hydrology, 377, 80-91, doi: 10.1016/j.jhydrol.2009.08.003, 2009

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. (Atmospheres) 113:D20119}, doi:10.1029/2008JD010201, 2008

Isotta, F. A., Frei, C., Weilguni, V., Tadic, M. P., Lassegues, 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., Vertacnik, G.,: The climate of daily precipitation in the Alps: development and analysis of a high-resolution grid dataset from pan-Alpine rain-gauge data. INTERNATIONAL JOURNAL OF CLIMATOLOGY, 34, 1657-1675, 10.1002/joc.3794, 2014

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

Jermey, P. et al.: RA uncertainty evaluation. Deliverable D2.14 of project: 607193 UERRA. Available at http://www.uerra.eu/publications/deliverable-reports.html

Jermey, P. et al.: Ensemble variational DA diagnostics. Deliverable D2.3 of project: 607193 UERRA. Available at http://www.uerra.eu/publications/deliverable-reports.html

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

Klein Tank, A. M. G. et al.: Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment. Intern. J. Climatol. 22:1441-1453. Data and metadata available at http://www.ecad.eu, 2002

Lussana, C., Saloranta, T., Skaugen, T., Magnusson, J., Tveito, O. E., and Andersen, J.: Evaluation of seNorge2, a conventional climatological datasets for snow-and hydrological modeling in Norway. Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2017-64. Manuscript under review for journal Earth Syst. Sci. Data, 2017

Murphy, A. H..: A new vector partition of the probability score. J. of applied meteorology, 12, 595-600, doi: 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2, 1973

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

Grimit, E. P. and Mass, C. F.: Measuring the Ensemble Spread–Error Relationship with a Probabilistic Approach: Stochastic Ensemble Results. AMS Journal, 135, 203-221, doi: 10.1175/MWR3262.1, 2007

Reistad, M., Breivik, Ø., Haakenstad, H., Aarnes, O.J., Furevik, B.R. and Bidlot, J.R.: A high-resolution hindcast of wind and waves for the North Sea, the Norwegian Sea, and the Barents Sea. Journal of Geophysical Research: Oceans, 116(C5), 2011

Soci, C., Bazile, E., Besson, F. and Landelius, T.: High-resolution precipitation re-analysis system for climatological purposes. Tellus A: Dynamic Meteorology and Oceanography, *68*(1), p.29879, 2016

# 9. Supplementary Materials

## 9.1 Verification scores based on the contingency table

For dichotomous events the verification of forecasts is based on the 2x2 contingency table, see Table 9.1. The beneath scores are used in section 3.

| | Observed YES | Observed NO |
|---|---|---|
| Forecast YES | Hits (a) | False alarms (b) |
| Forecast NO | Misses (c) | Correct negatives (d) |

**Table 9.1**: Contingency table of dichotomous forecast

**Hit rate (probability of detection):** $POD = \frac{a}{a+c}$

**False alarm ratio:** $FAR = \frac{b}{a+b}$

**False alarm rate:** $F = POFD = \frac{b}{b+d}$

**Threat score (critical success index):** $TS = CSI = \frac{a}{a+c+b}$

**Accuracy (fraction correct):** $FC = \frac{a+d}{total}$

**Log odds ratio:** $logOR = log(\frac{a*d}{b*c})$

**Extreme dependency score:** $EDS = \frac{ln(p)-ln(POD)}{ln(p)+ln(POD)}$ with $p = baserate = \frac{a+c}{total}$

**Extremal dependence index:** $EDI = \frac{\ln(F)-\ln(POD)}{\ln(F)+\ln(POD)}$

**Symmetric extremal dependence index:** $SEDI = \frac{\ln(F)-\ln(POD)+\ln(1-POD)-\ln(1-F)}{\ln(F)+\ln(POD)+\ln(1-POD)+\ln(1-F)}$

**Equitable Threat score (Gilbert skill score):**
$GSS = ETS = \frac{a-hits(random)}{a+c+b-hits(random)}$

$$hits(random) = \frac{(a+c)(a+b)}{total}$$

**Heidke skill score:** $HSS = \frac{(a+d) - expected\ correct\ random}{total - expected\ correct\ random}$

$$expected\ correct\ random = \frac{1}{total}[(a+c)(a+b) + (d+c)(d+b)]$$

**True skill stats (Peirce's skill score, Hanssen and Kuipers discriminant):**

$$TSS = PSS = HK = \frac{a}{a+c} - \frac{b}{b+d}$$

## 9.2 Histograms of daily minimum and maximum temperature

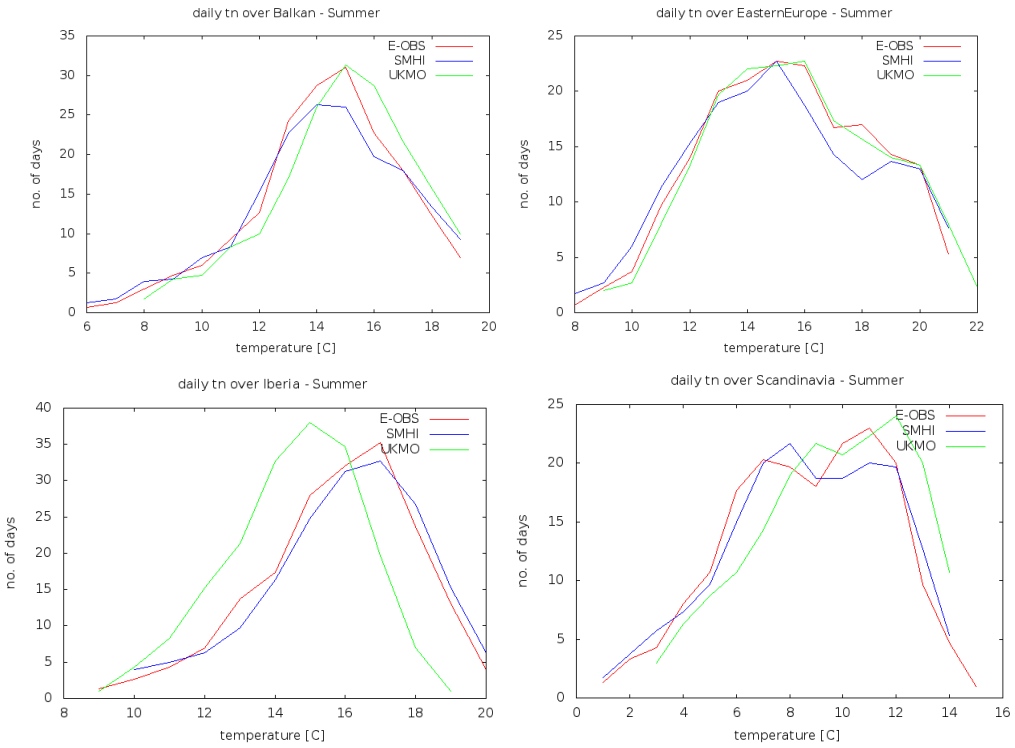


**Figure 9.1**: Histograms of daily minimum temperature during summer over selected areas. Red colours denote the E-OBS data; blue and green denote the SMHI and UKMO reanalysis respectively.

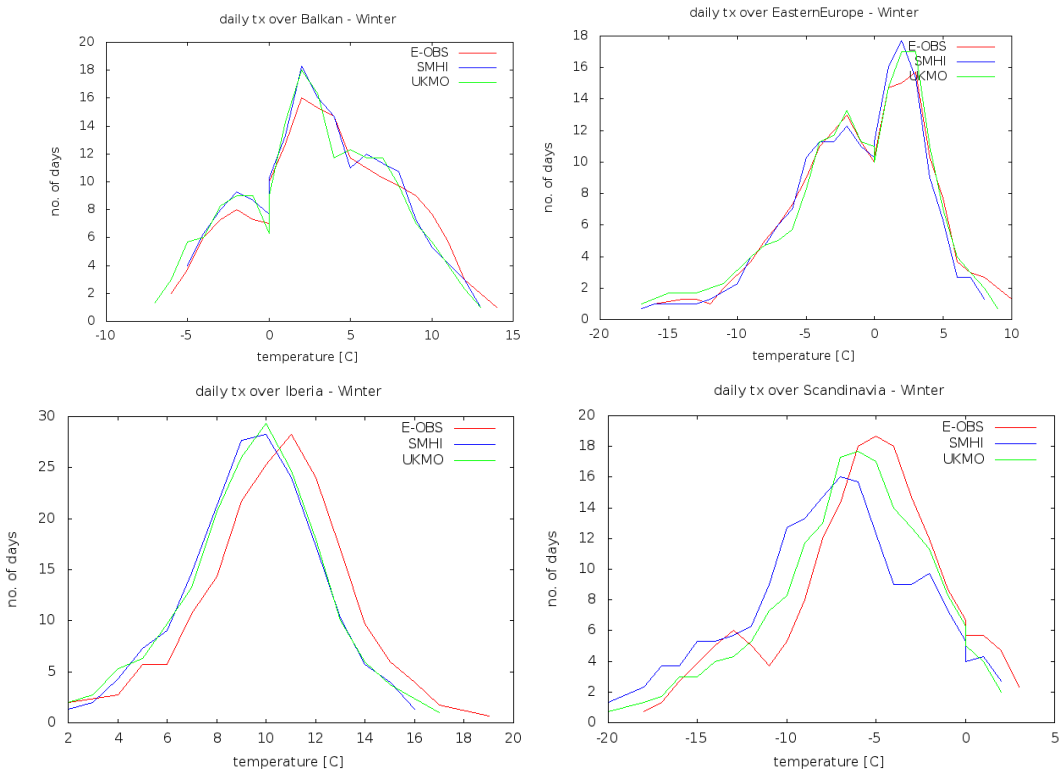


**Figure 9.2**: Histograms of daily maximum temperature during winter over selected areas. Red colours denote the E-OBS data; blue and green denote the SMHI and UKMO reanalysis respectively.

## 9.3 Additional material to precipitation analysis over Fennoscandia

### Summary statistics on the 0.25° regular grid

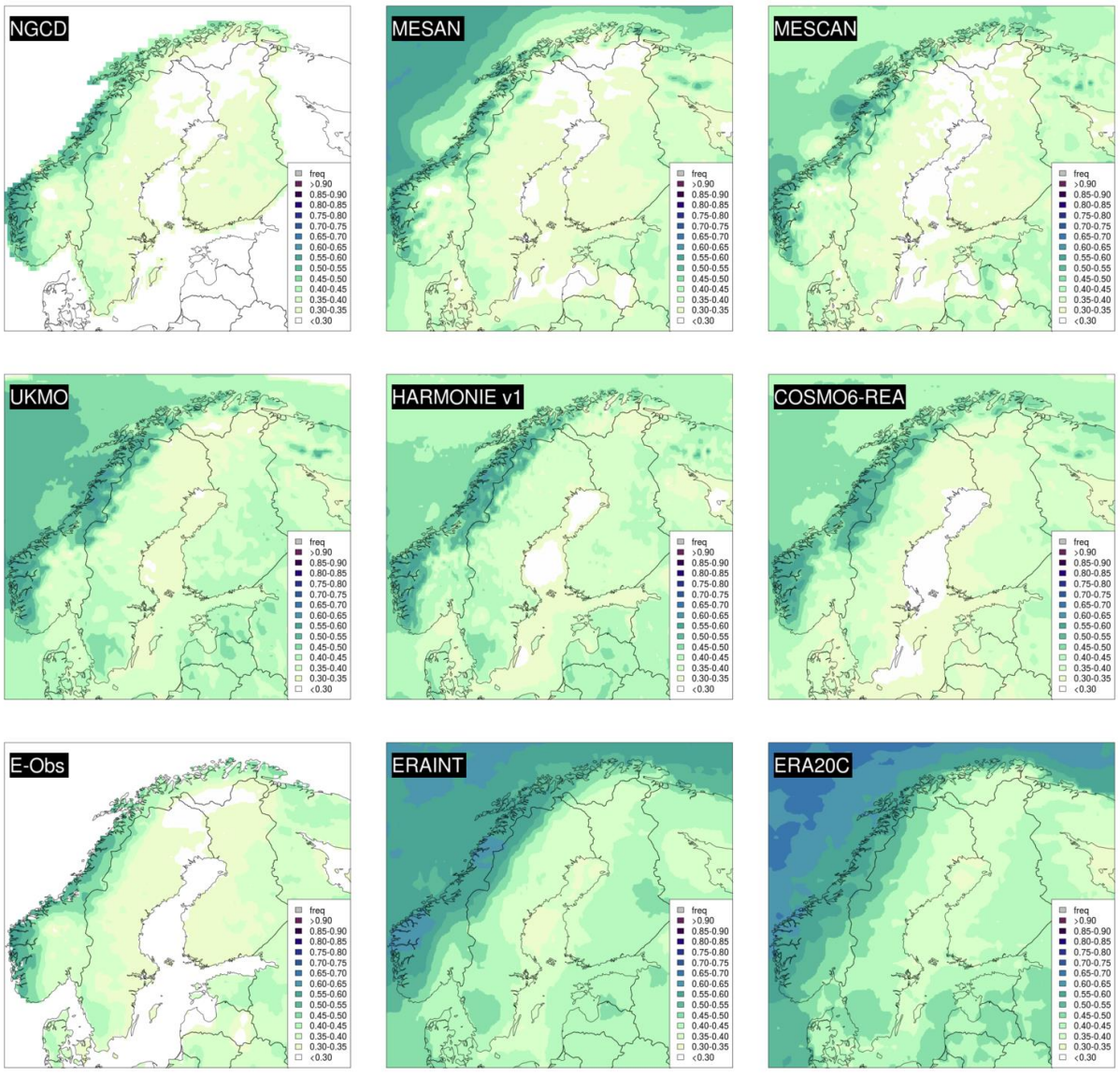


**Figure 9.3**: Annual frequency of wet days (≥ 1mm/d, fraction, 2006-2010). Rescaled to 0.25° regular grid. Reference: NGCD.

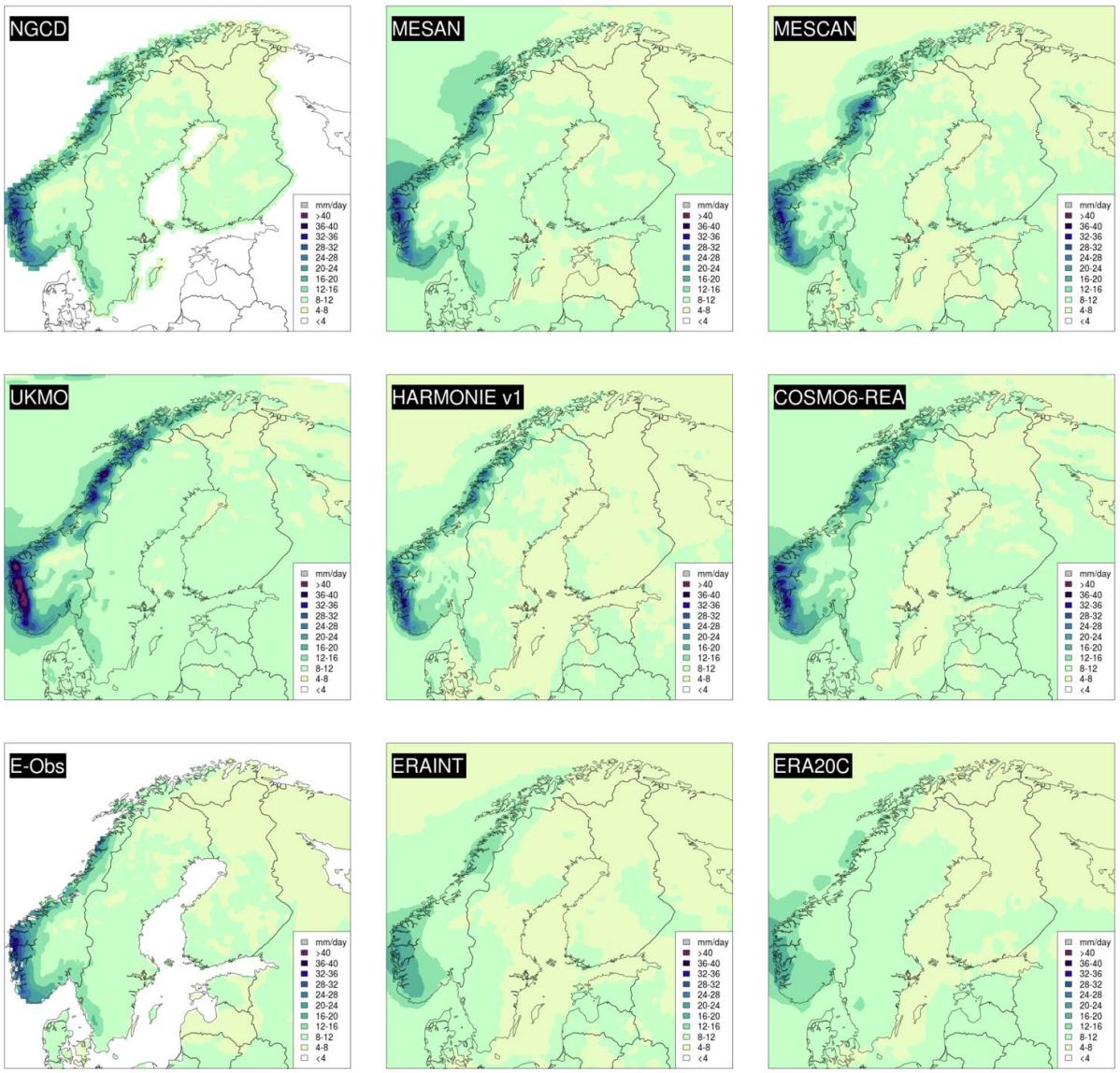


**Figure 9.4**: 95% quantile of daily precipitation (mm/d, 2005-2008). Rescaled to 0.25° regular grid. Reference: NGCD.

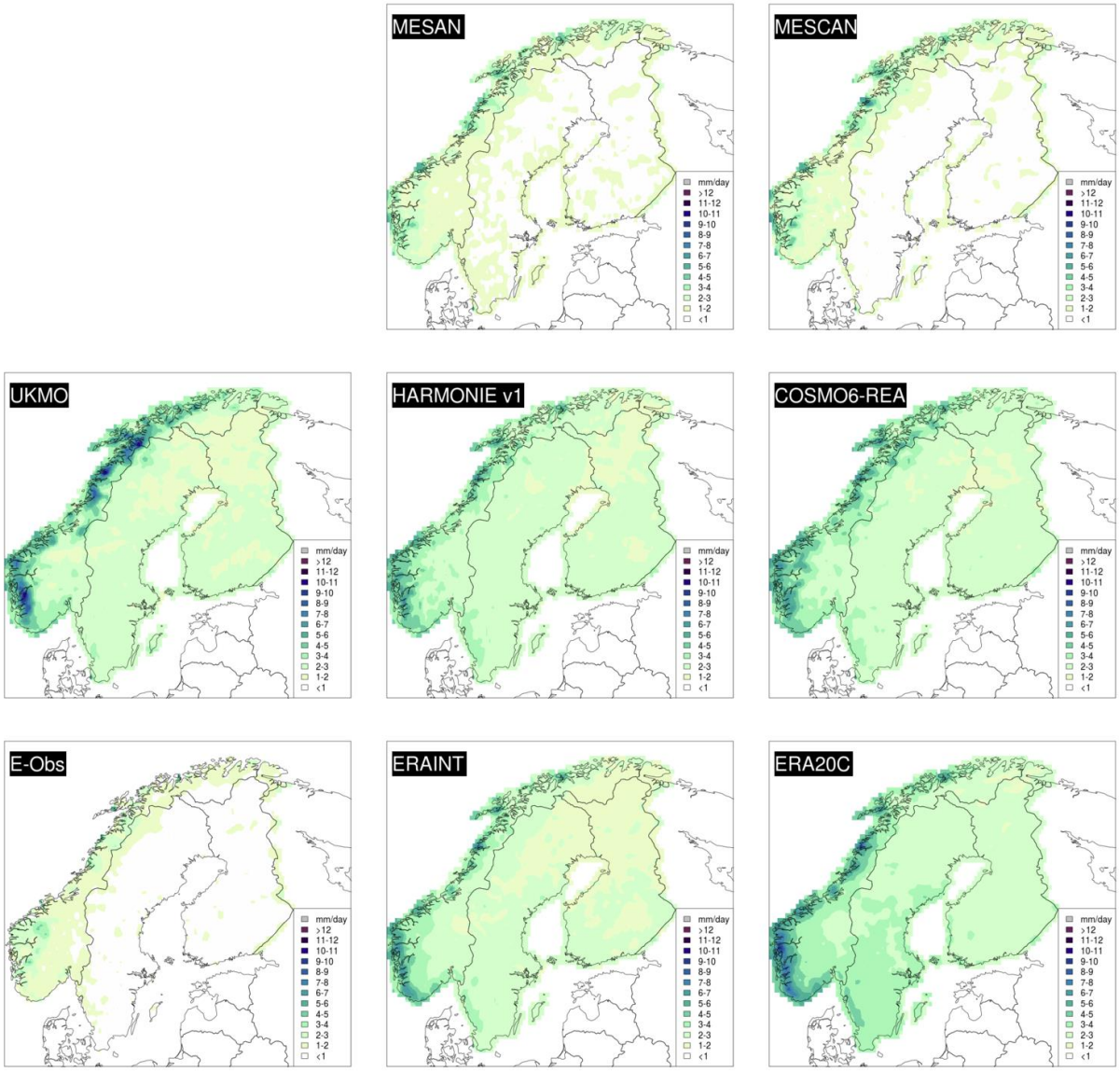


**Figure 9.5:** root Mean Square Error of daily precipitation (mm/d, 2006-2010). Rescaled to 0.25° regular grid. Reference: NGCD.

## 1986-1990 vs 2006-2010: Mean annual precipitation

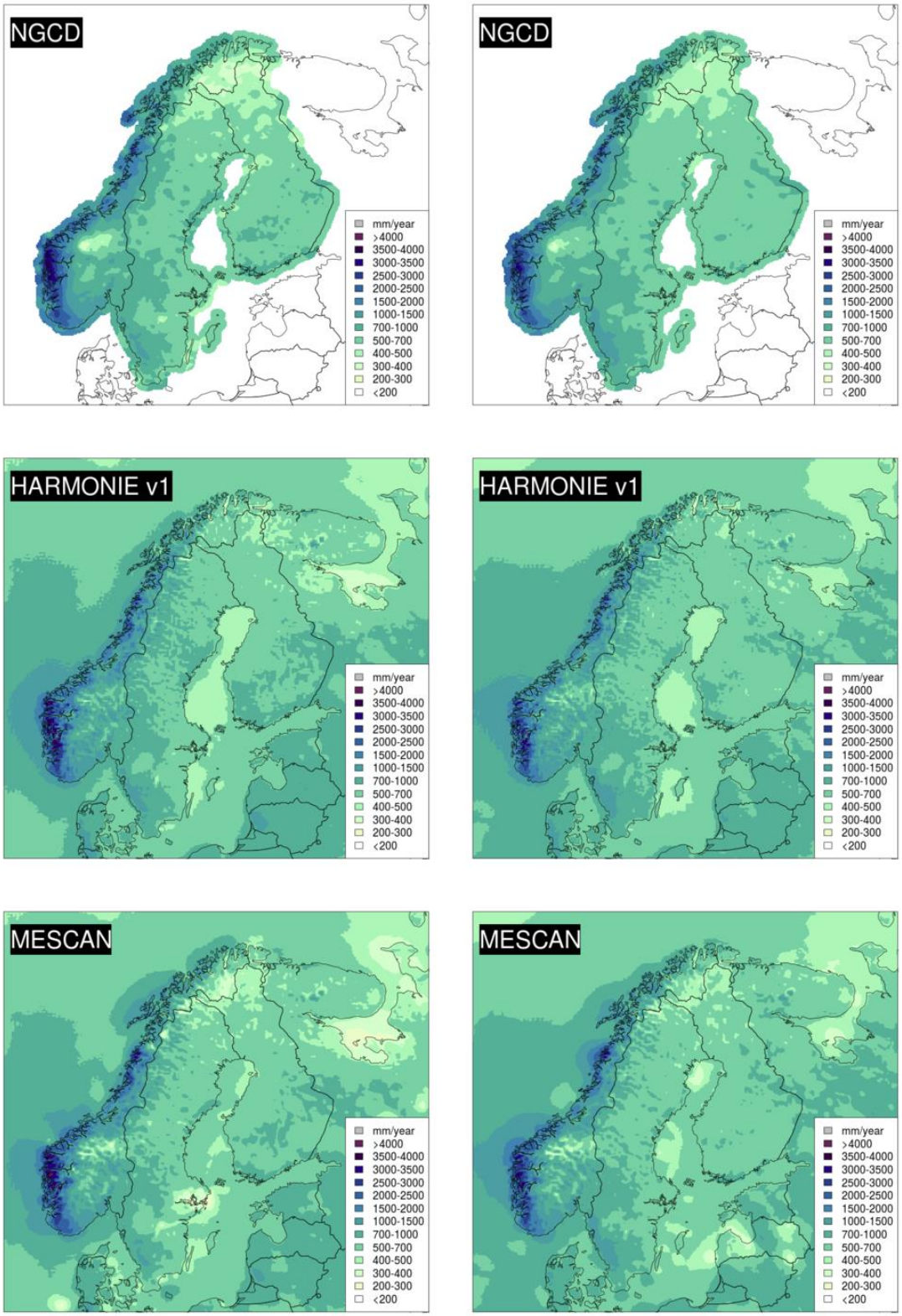


**Figure 9.6**: Mean annual precipitation (mm per year). Left panels: 1986-1990. Right panels: 2006-2010. Rescaled to 5km ETRS-LAEA coordinate system. Reference (top panels): NGCD.

## MESCAN vs HARMONIE: Mean annual 95% percentile

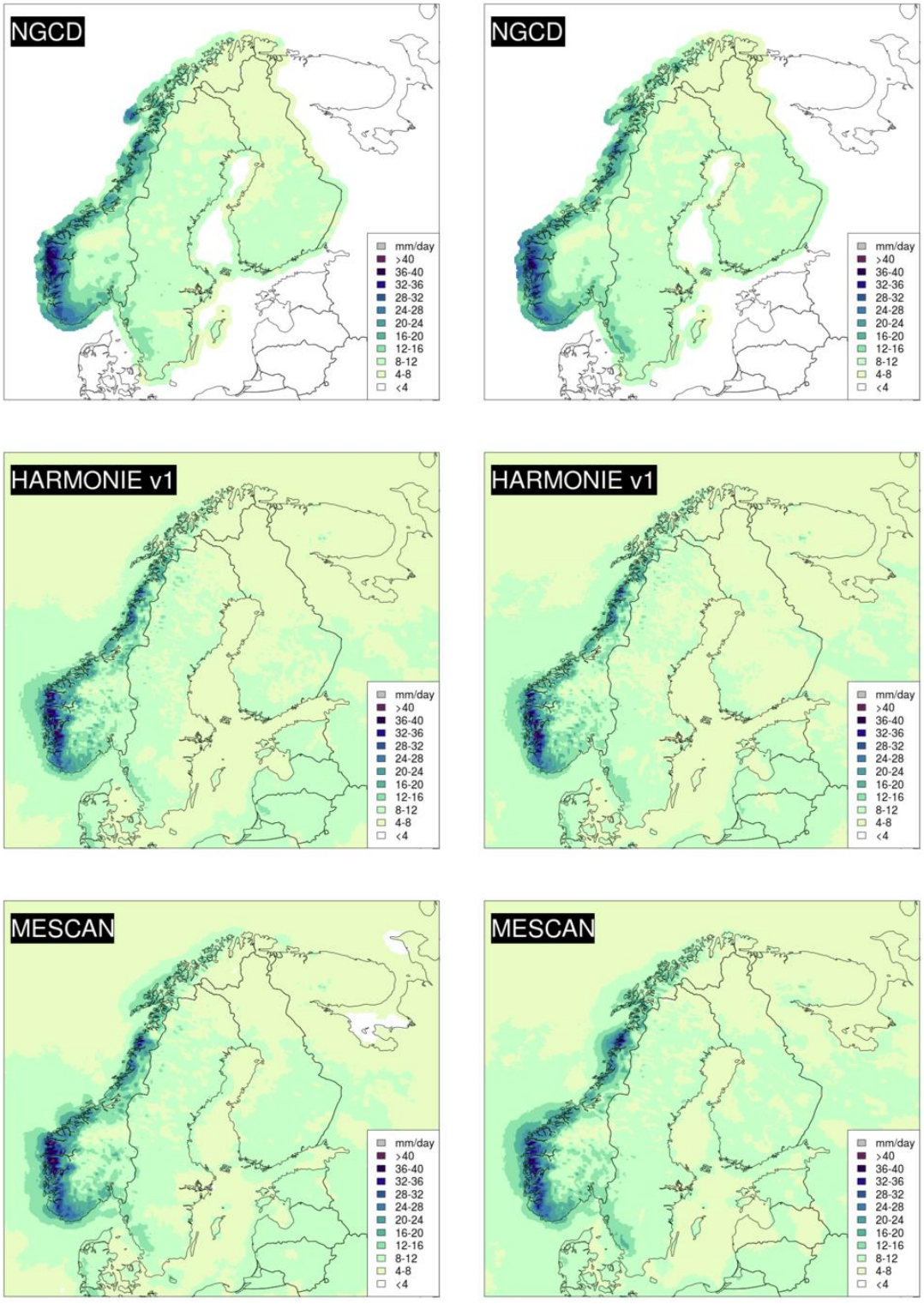


**Figure 9.7**: Mean annual 95% quantile of daily precipitation (mm per day). Left panels: 1986-1990. Right panels: 2006-2010. Rescaled to 5km ETRS-LAEA coordinate system. Reference (top panels): NGCD.

## MESCAN vs HARMONIE: Mean annual frequency of wet days

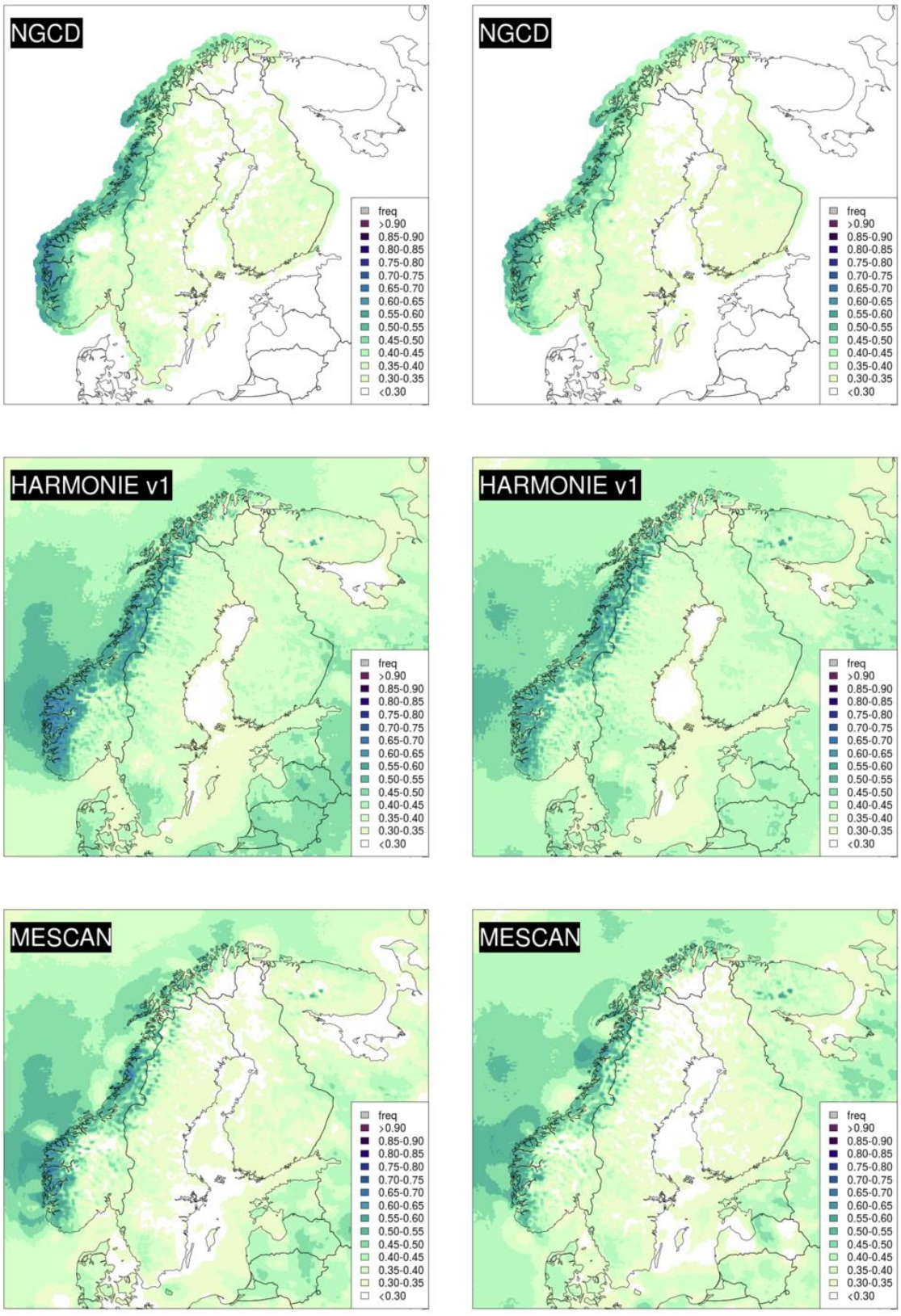


**Figure 9.8**: Mean annual frequency of wet days. Left panels: 1986-1990. Right panels: 2006-2010. Rescaled to 5km ETRS-LAEA coordinate system. Reference (top panels): NGCD.

## MESCAN vs HARMONIE: Root-Mean-Square Error

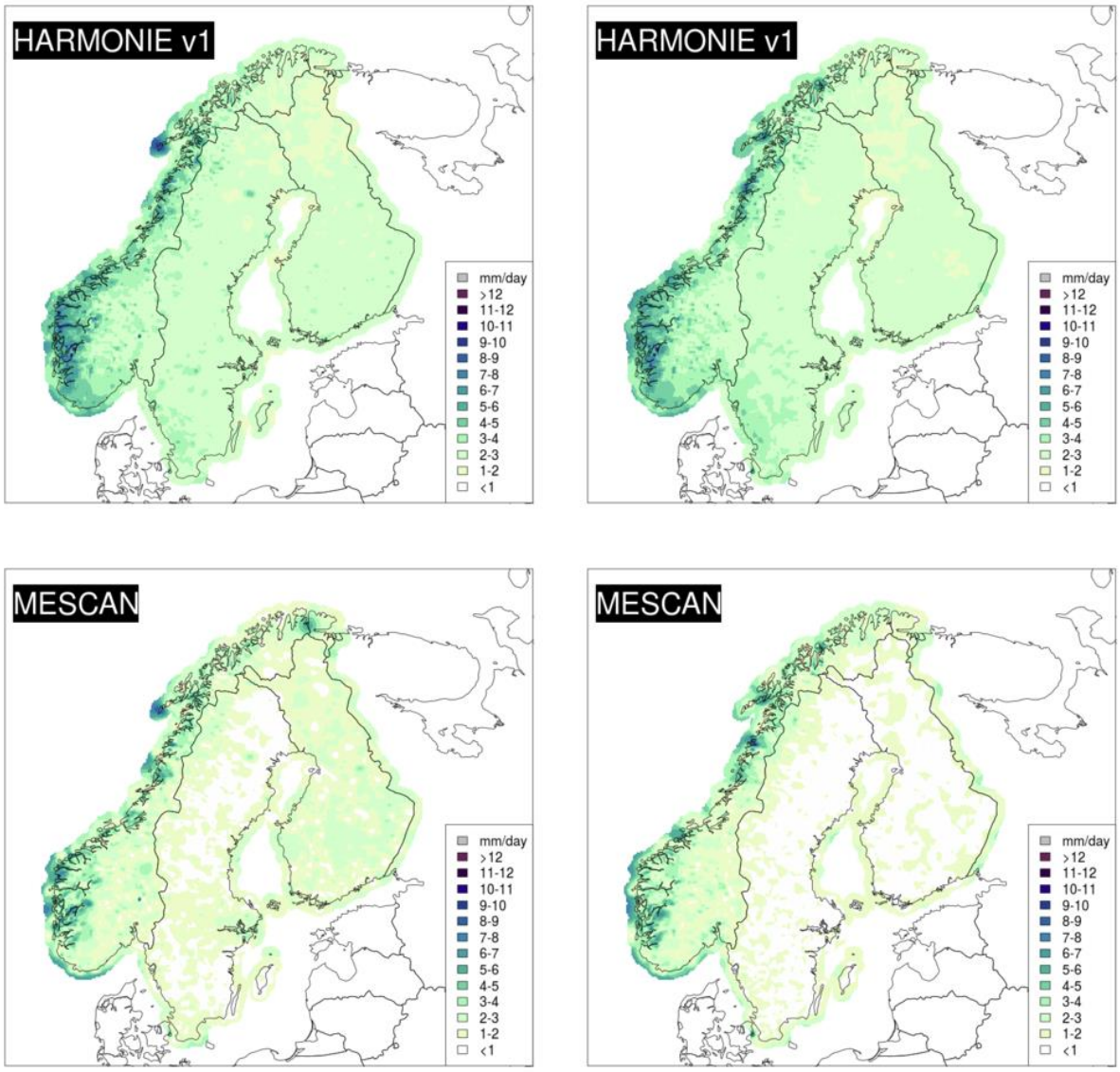


**Figure 9.9**: Root-Mean-Square Error. Left panels: 1986-1990. Right panels: 2006-2010. Rescaled to 5km ETRS-LAEA coordinate system. Reference: NGCD.