Quality Assurance in HIRLAM-C 2017: Status and Outlook

EWGLAM/SRNWP Meeting 2017

ECMWF

Bent Hansen Sass

Quality Assurance in HIRLAM-C

OUTLINE

- Main tools used in routine verification and validation
- Benchmarking of HIRLAM RCR centre results and HIRLAM ensembles against ECMWF ensembles
- Forecast Quality Assessment of HARMONIE-AROME by HIRLAM duty-forecasters
- Why do we need high resolution models to predict extremes ? A recent example
- Brief summary of the evolution process from NWP verified in `fixed points´ to NWP verified for `scales in space and time´ : Why is this transition important ?
- Outlook: : How should a future verification strategy look like ?

Main tools used in routine verification and validation

Documentation via www.hirlam.org

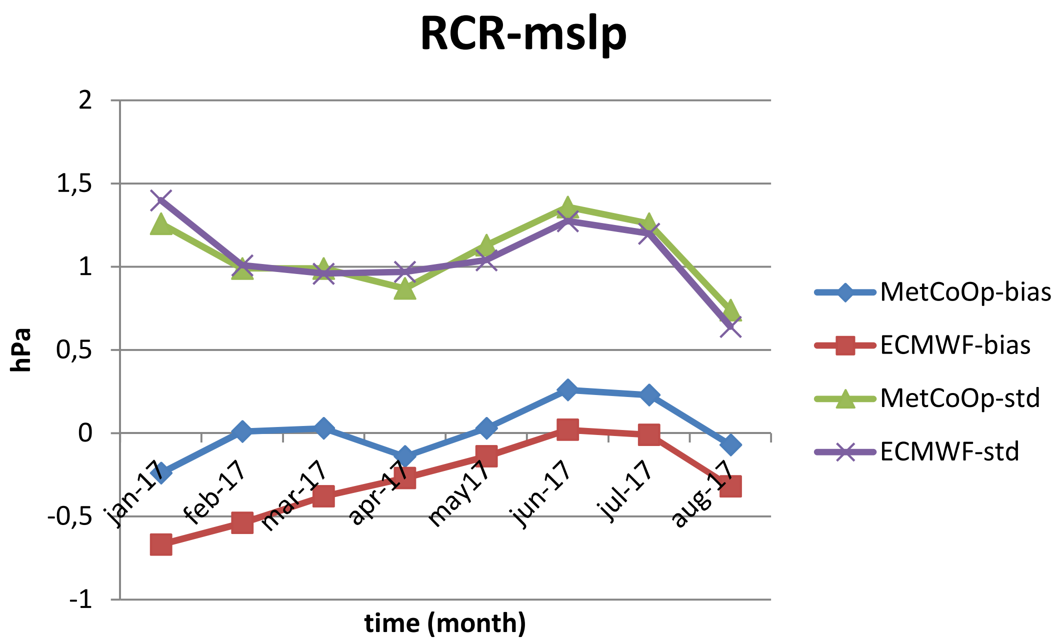
- **MONITOR**
  (point verification with many parameters and options for comparing models)

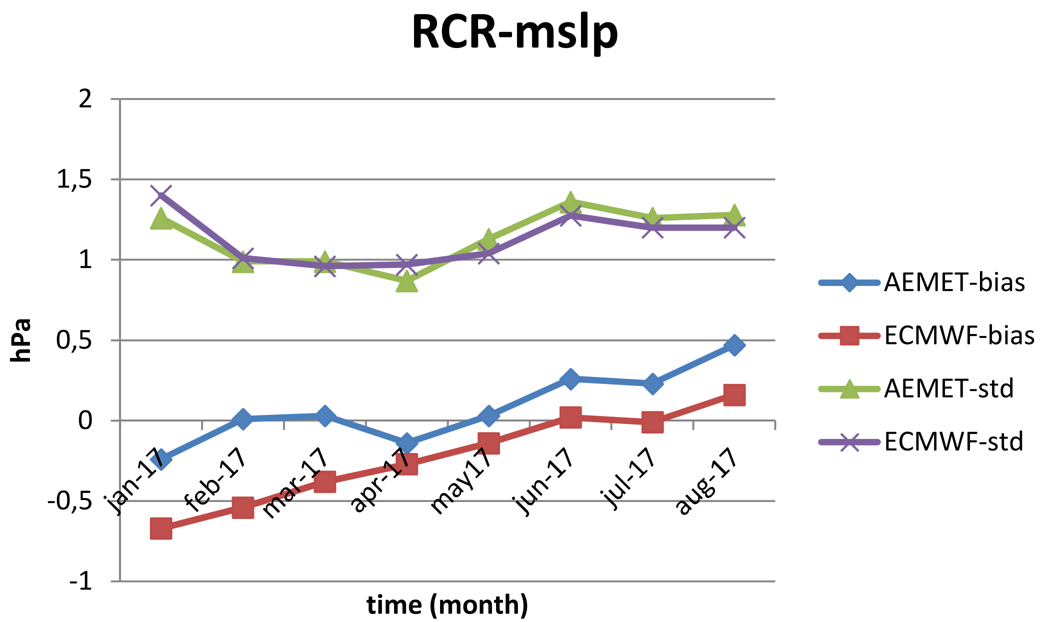
- **HARP**
  (**H**IRLAM **A**LADIN **R** verification **P**ackage) :
  point verification based probabilistic verification of ensembles +
  beta-release of spatial verification, e.g. containing
  FSS (Fractions skill score) , Roberts , N.M., and Lean, H.W., 2008
  SAL (`Structure, Amplitued and Location' score), Wernli et al. 2008

- **OBSMON**: Observation monitoring (needed for data-assimilation)

- **National verification scores** (supplementary input) ,
  e.g. special computations of FSS, SAL and
  SWS (Significant Weather Score) , Sass and Yang 2012

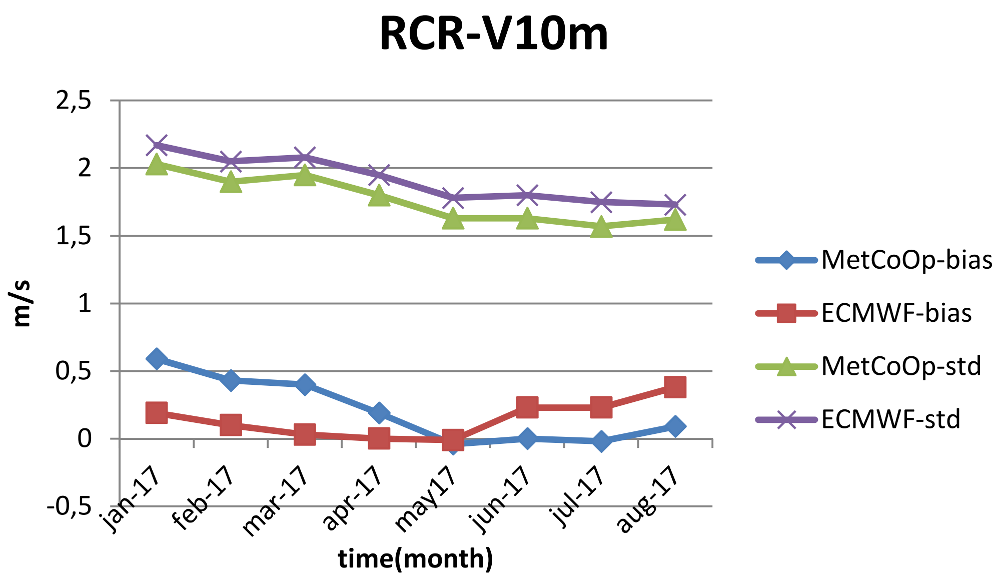
Benchmarking of HIRLAM RCR centre results against ECMWF.
RCR Centres in 2017: AEMET and MetCoOp
(all stations , valid at +24 h)

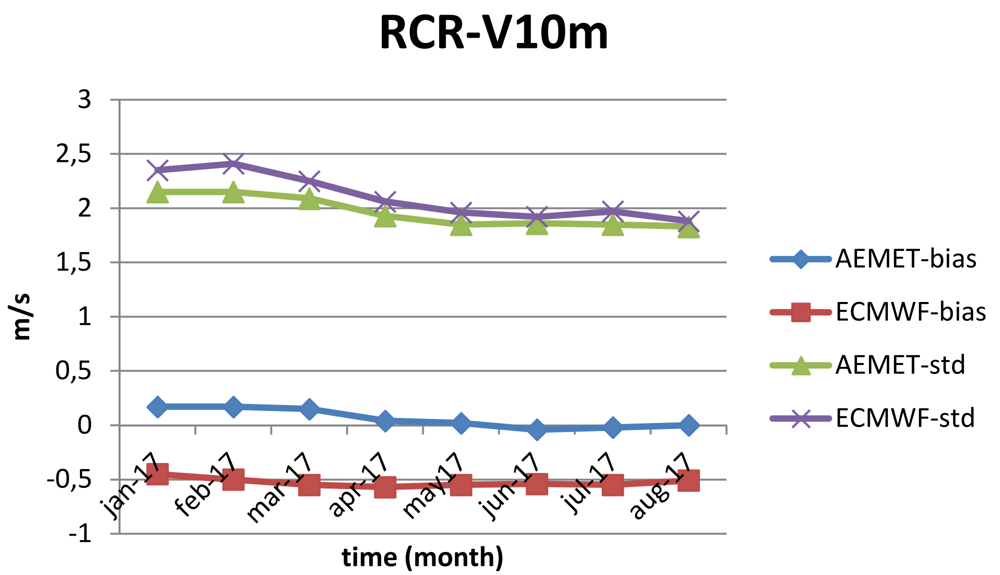




EWGLAM/SRNWP Meeting 2017
ECMWF
Bent Hansen Sass

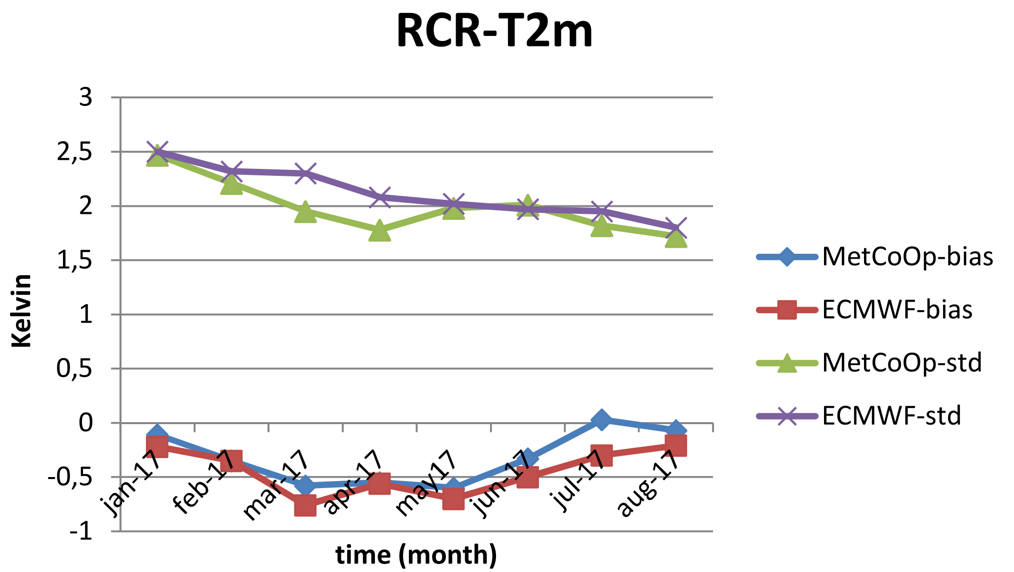
Benchmarking of HIRLAM RCR centre results against ECMWF
RCR Centres in 2017: AEMET and MetCoOP
(all stations , valid at +24 h)

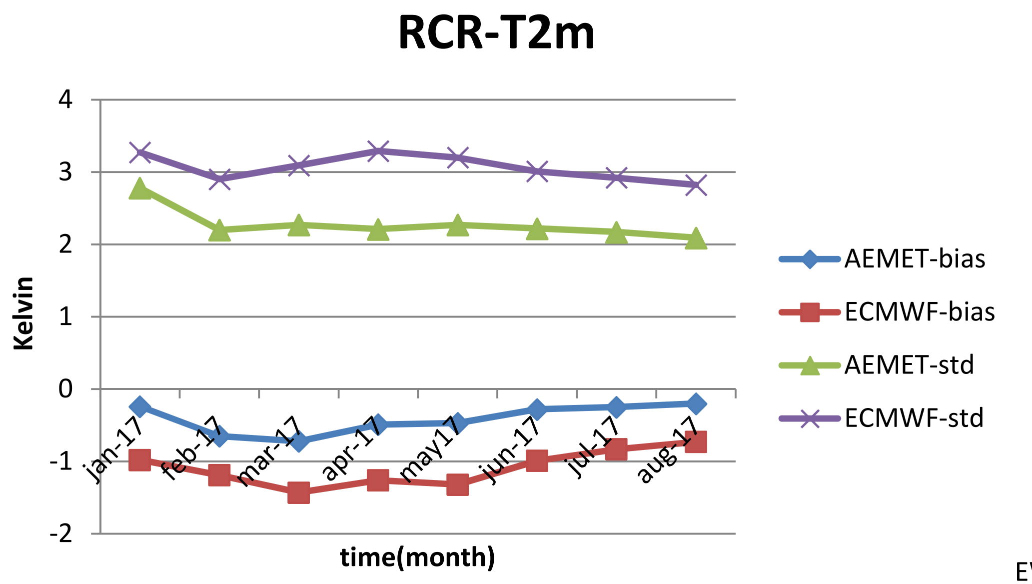




EWGLAM/SRNWP Meeting 2017
ECMWF
Bent Hansen Sass

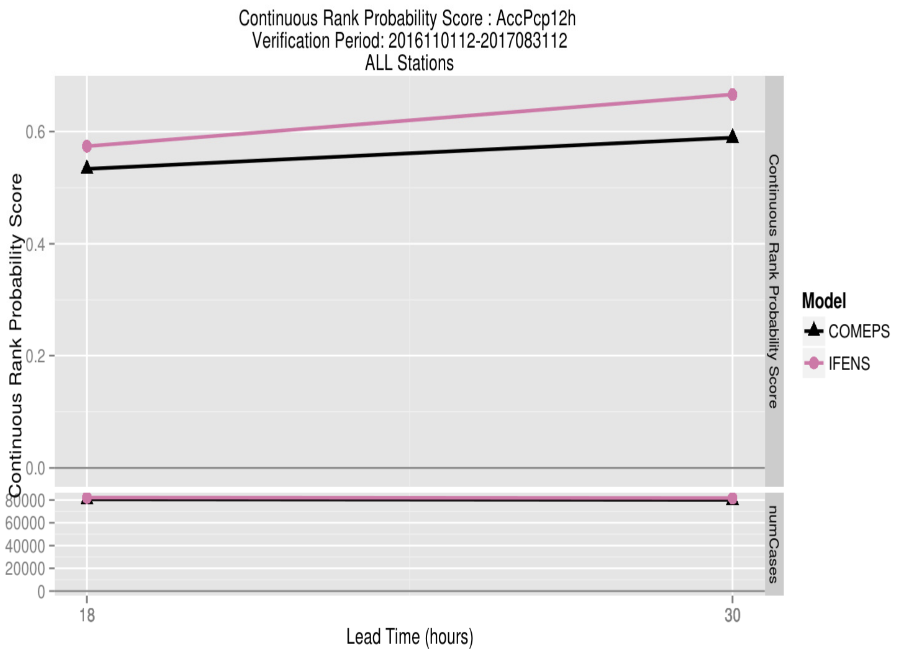
Benchmarking of HIRLAM RCR centre results against ECMWF
RCR Centres in 2017: AEMET and MetCoOP
(all stations , valid at +24 h)

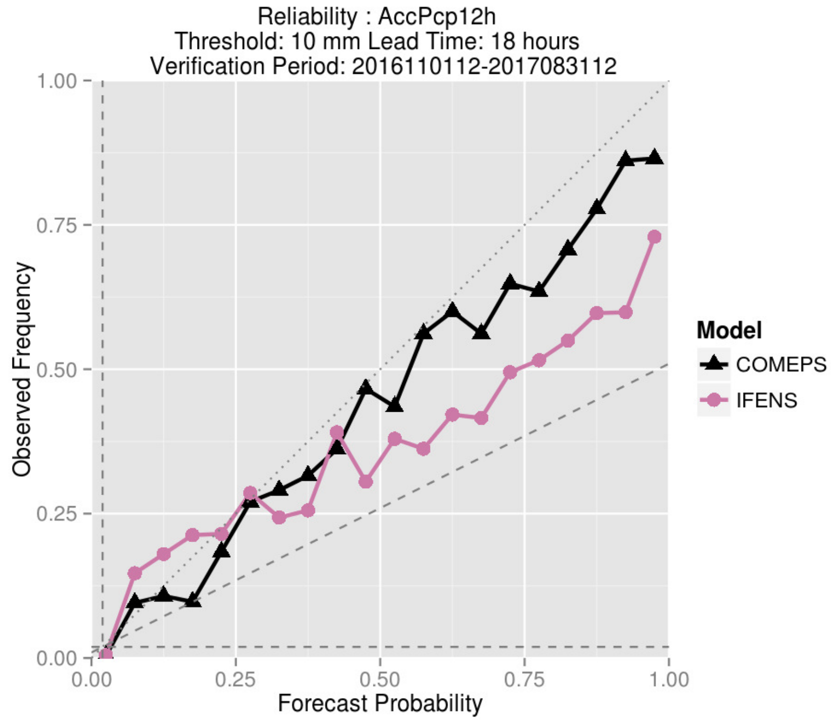




EWGLAM/SRNWP Meeting 2017
ECMWF
Bent Hansen Sass

HARP ensemble probabilistic verification
DMI COMEPS benchmarked against IFS-ENS



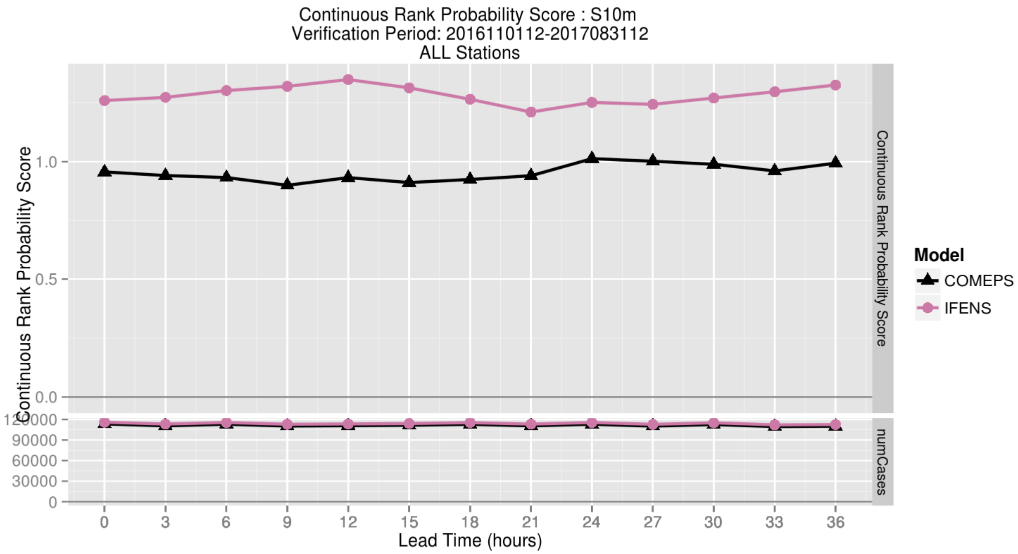


Precipitation continuous rank probaility score

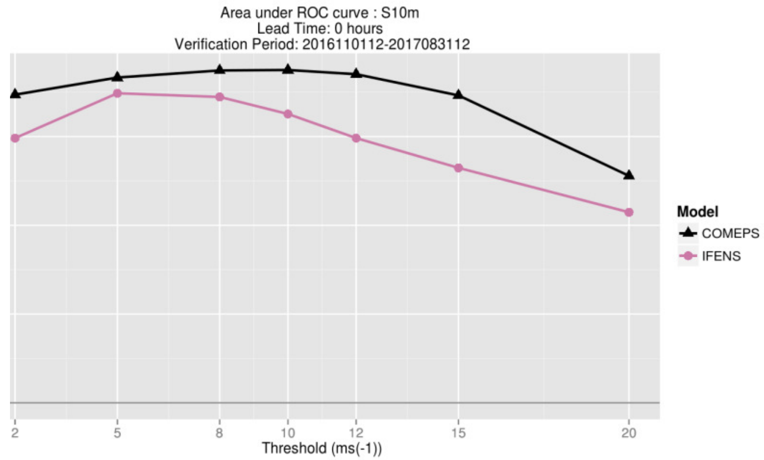
Precip. 10 mm threshold, reliability curves

EWGLAM/SRNWP Meeting 2017
ECMWF
Bent Hansen Sass

HARP ensemble probabilistic verification
DMI COMEPS benchmarked against IFS-ENS



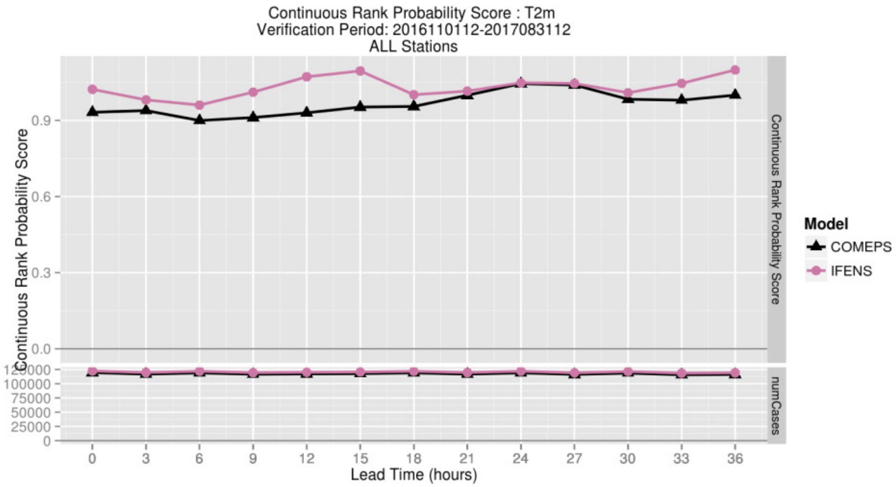
V10M Continuous rank probability score



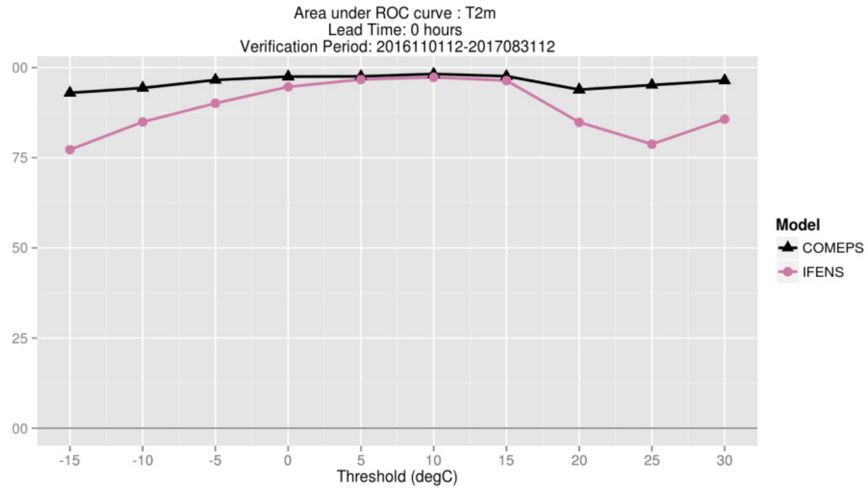
V10m : Area under ROC curve

EWGLAM/SRNWP Meeting 2017
ECMWF
Bent Hansen Sass

HARP ensemble probabilistic verification
DMI COMEPS benchmarked against IFS-ENS



T2M Continuous rank probability score



T2M Area under ROC curve

Forecast Quality Assessment of HARMONIE-AROME by duty-forecasters in HIRLAM institutes

A TABLE of the quality assessment of HARMONIE-AROME by users is intended to reflect NWP quality according to the needs of the forecasters when using HARMONIE for `high quality work´ in the meteorological services. Focus is on forecasts up to 24 hours.

The results of the TABLE apply to user assessments in 2017, mainly from March – May . The numbers written to the boxes of individual score categories are the number of independent assessments from forecasters for this category.

Definition of score numbers

`0´means forecasts have **extremely poor quality** on average , e.g. predicted changes often poorer than a forecast predicting no changes (persistence)

`1´ means that forecasts have **poor quality** on average , e.g. predicted changes are often incorrect with rather large errors.

`2´ means forecasts have **fair quality** on average, but possibly having **high variability of quality**.

`3´ means forecasts have **good quality** on average, the majority of forecasts have good predictive value.

`4´ means forecasts have **extremely good quality**, e.g. with excellent predictive value. It is possible to assign decimal numbers, e.g. 3.5 means a quality assessment between 3 and 4.

Forecast Quality Assessment of HARMONIE-AROME by duty-forecasters in HIRLAM institutes

HARMONIE Quality Assessment by forecastsers

Parameter	Score = [0 : 1.0 [	Score = [1.0 : 2.0[	Score = [2.0 : 3.0[	Score = [3.0 : 4.0]	average
mslp			1	10	3,4
v10m			1	11	3.2
t2m		1	4	6	2,6
rh2m		1	4	1	2.5
fog		2	6	1	1.9
cld			6	6	2.6
prrn		2	4	5	2,4
prfr		1	4	3	2.4
ceiling			1		2.0
cape			1		2,0
pp1				1	4.0
pp2				1	4.0

Parameter definitions:

mslp: mean sea level pressure

v10m: 10 metre wind

t2m: 2m temperature,

rh2m: relative humidity at 2 metres**;**

fog: prediction of fog,

cld : total cloud cover;

‘**ceiling**: prediction of celiling;

`cape´: convective available potential energy;

pp1: postprocessed `lightning’ product**;**

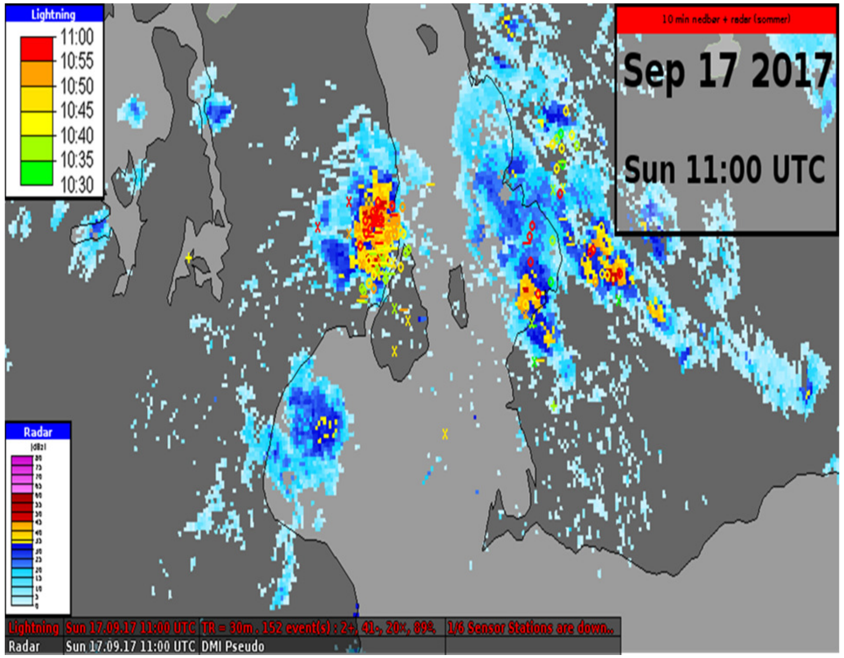
pp2: postprocessed radar reflectivity product.

Why do we need high resolution models to predict extremes ?

A recent forecast example

- EWGLAM/SRNWP Meeting 2017
- ECMWF
- Bent Hansen Sass

Sept 17, 2017, Copenhagen cloudburst

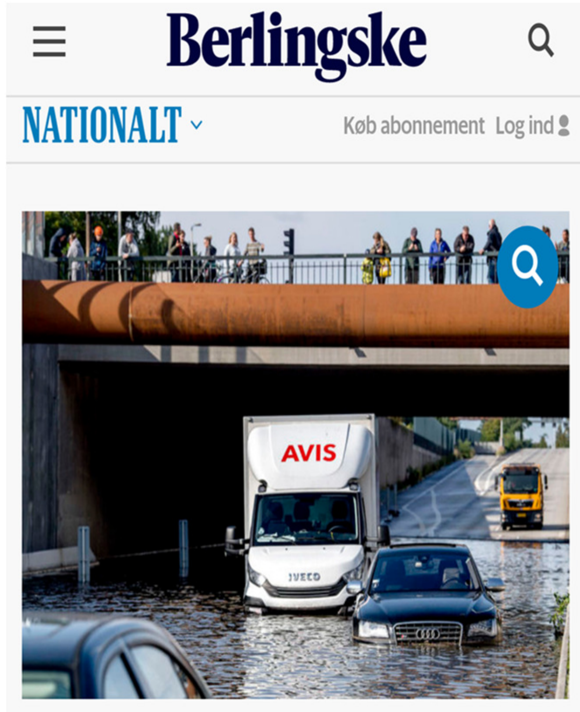


DMI regional forecast issued in the morning was correct for general weather (" scattered showers with risk of hail and thunder ")

But small scales in space and time are not described in detail !

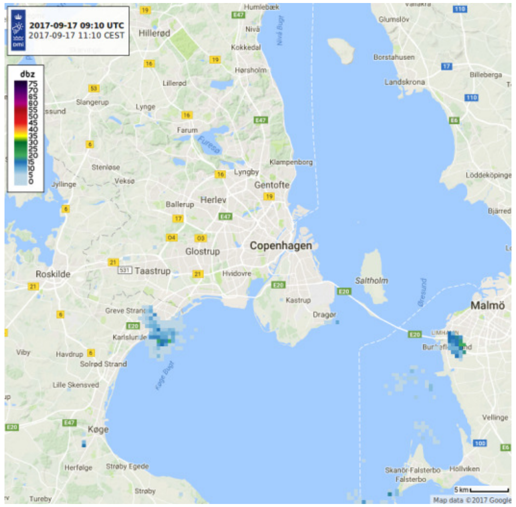
Løbet blev aflyst:
Københavns halvmaraton ramt af voldsomt uvejr

Copenhagen half-Maraton was canceled

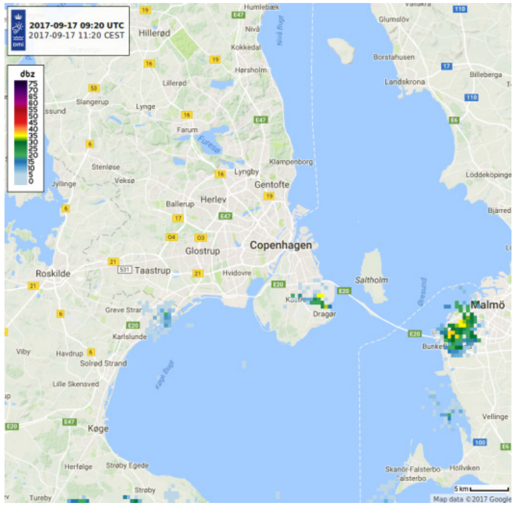


Lyngbyvej floded (once again …)

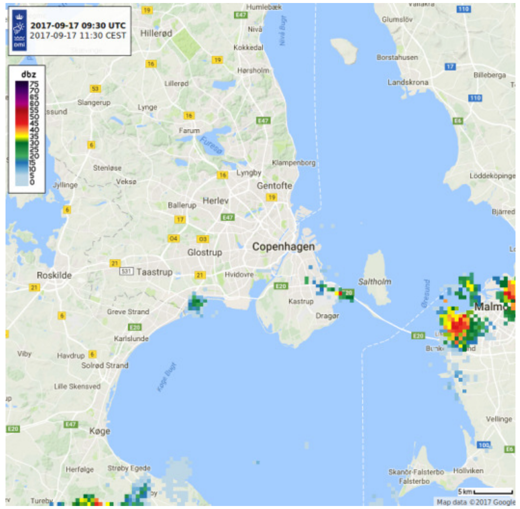
Copenhagen Cloud Burst CASE 2017-09-17



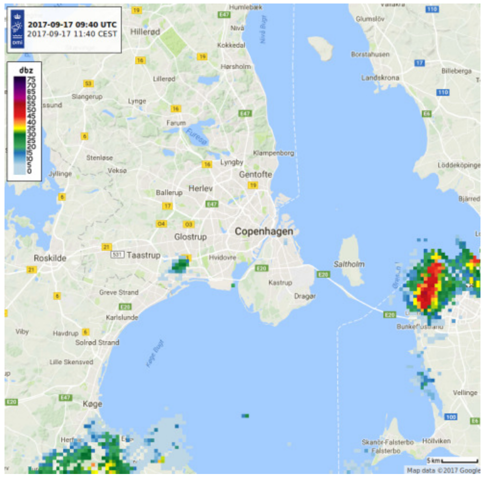
Radar image : 9:10 UTC



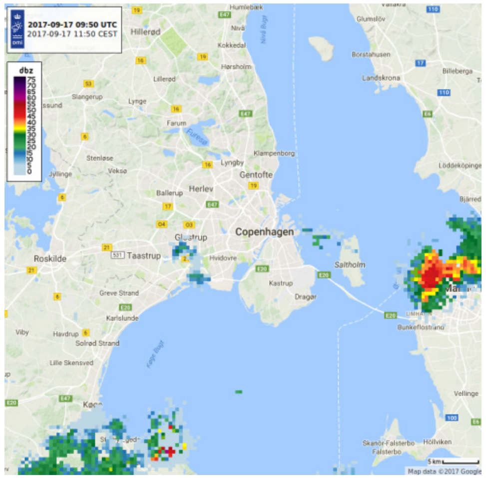
Radar image : 9:20 UTC



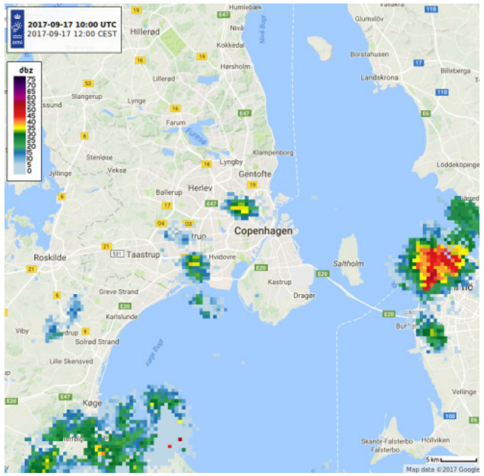
Radar image : 9:30 UTC



Radar image : 9:40 UTC

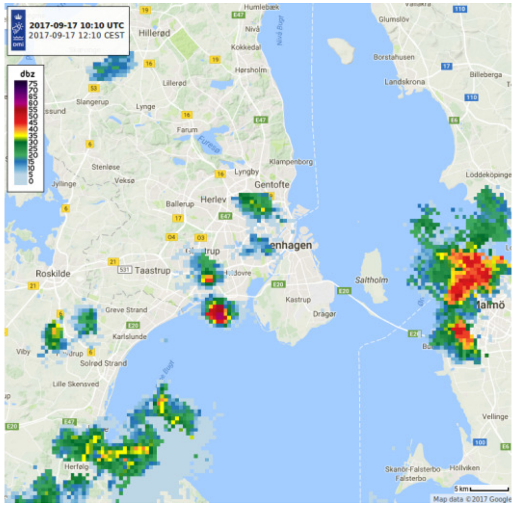


Radar image : 9:50 UTC

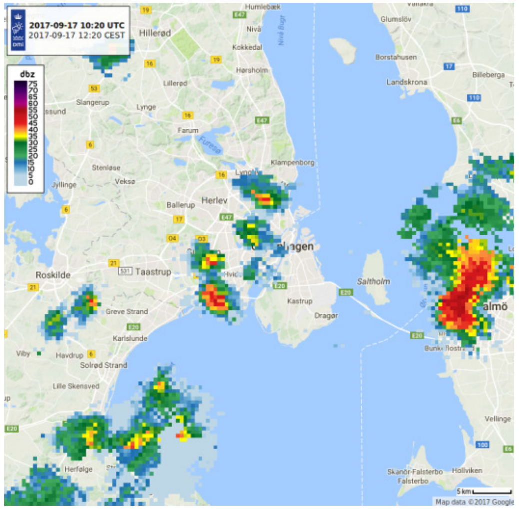


Radar image : 10:00 UTC

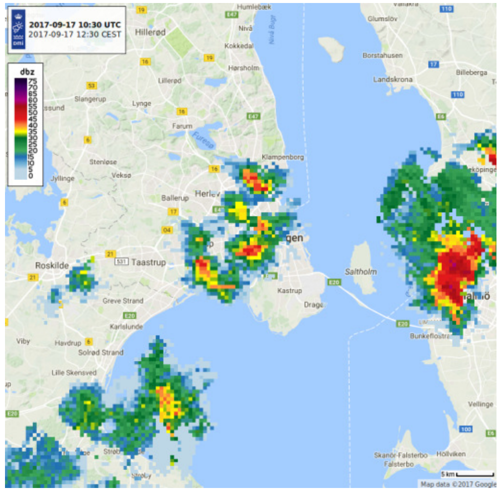
Copenhagen Cloud Burst CASE 2017-09-17



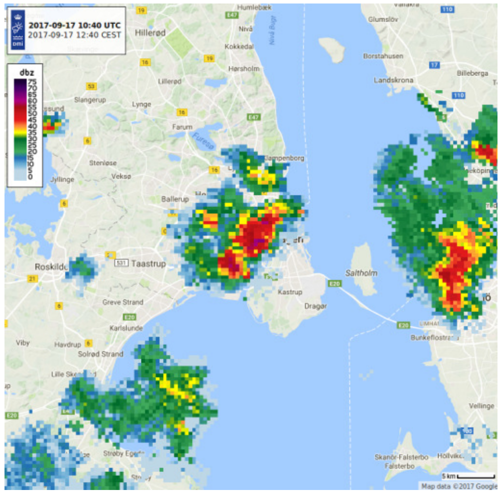
Radar image : 10:10 UTC



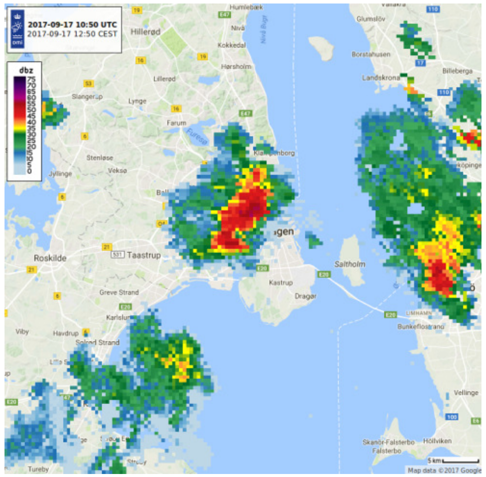
Radar image : 10:20 UTC



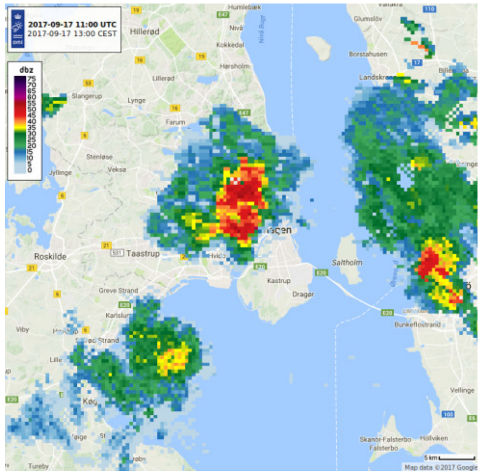
Radar image : 10:30 UTC



Radar image : 10:40 UTC

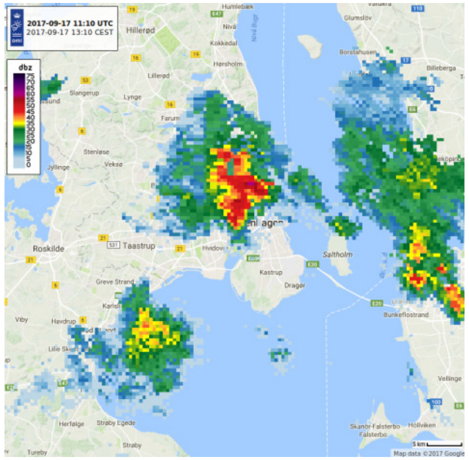


Radar image : 10:50 UTC

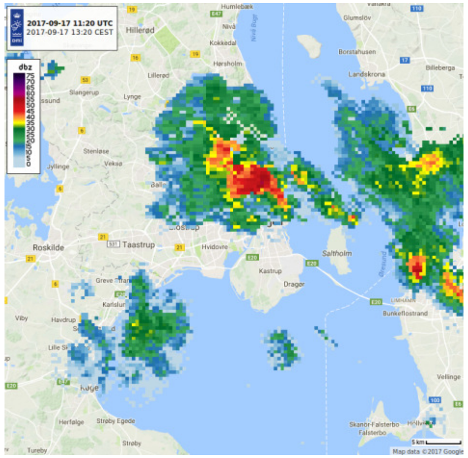


Radar image : 11:00 UTC

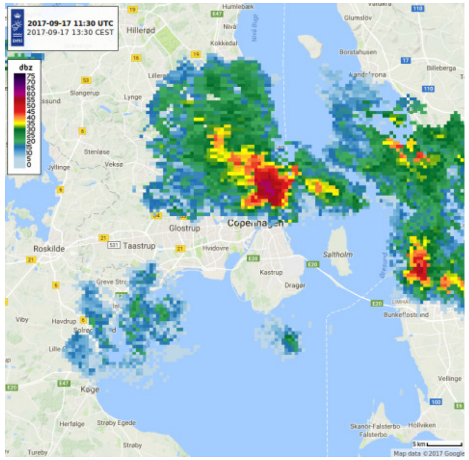
Copenhagen Cloud Burst CASE 2017-09-17



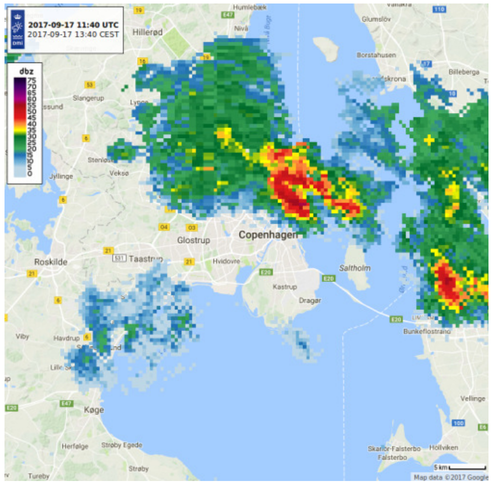
Radar image : 11:10 UTC



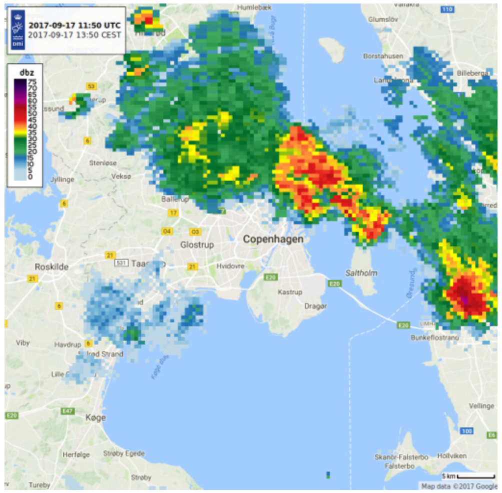
Radar image : 11:20 UTC



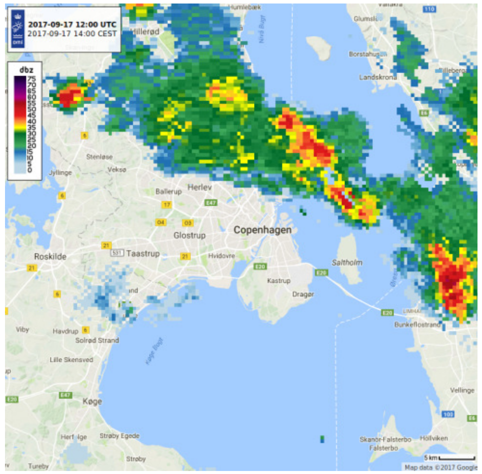
Radar image : 11:30 UTC



Radar image : 11:40 UTC

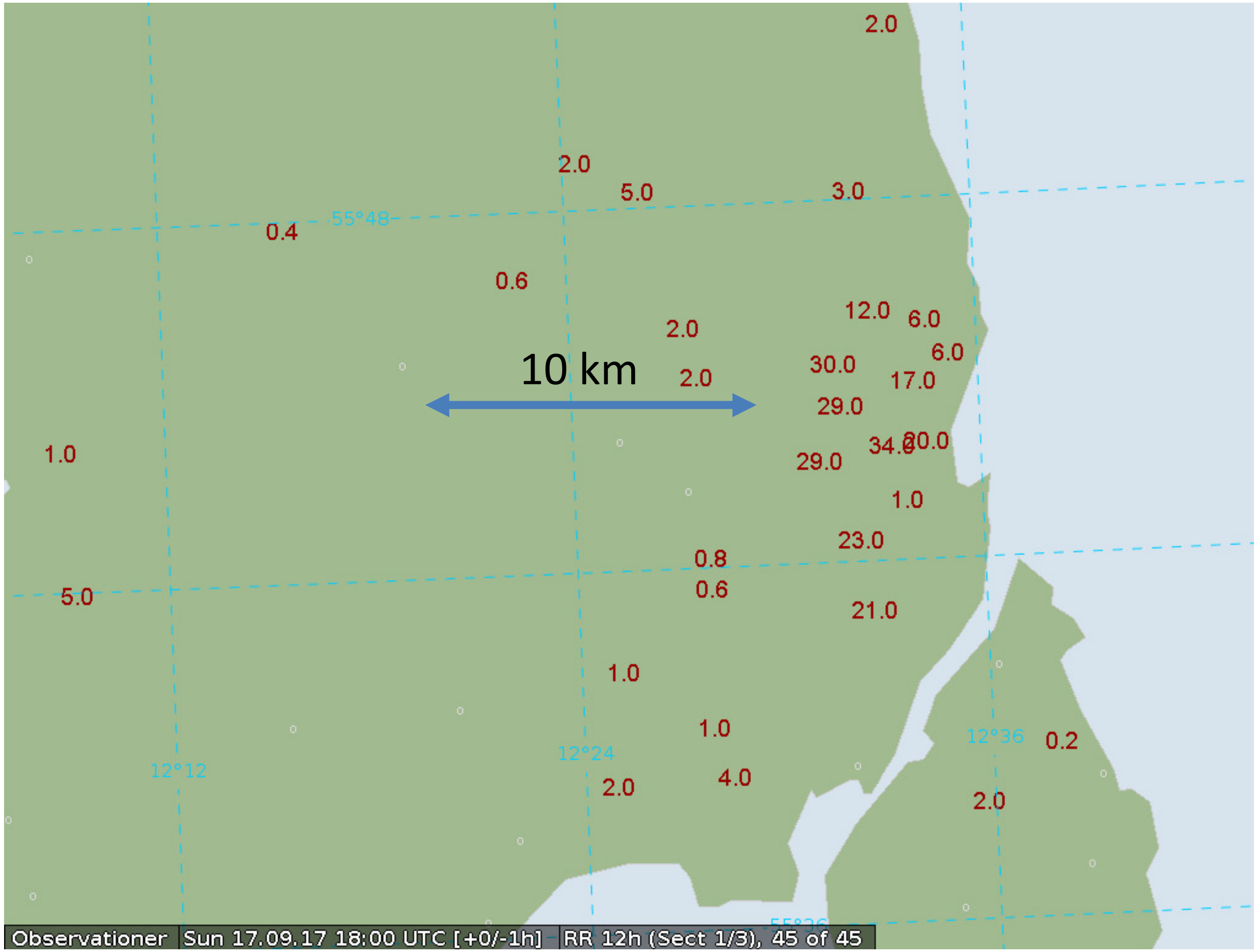


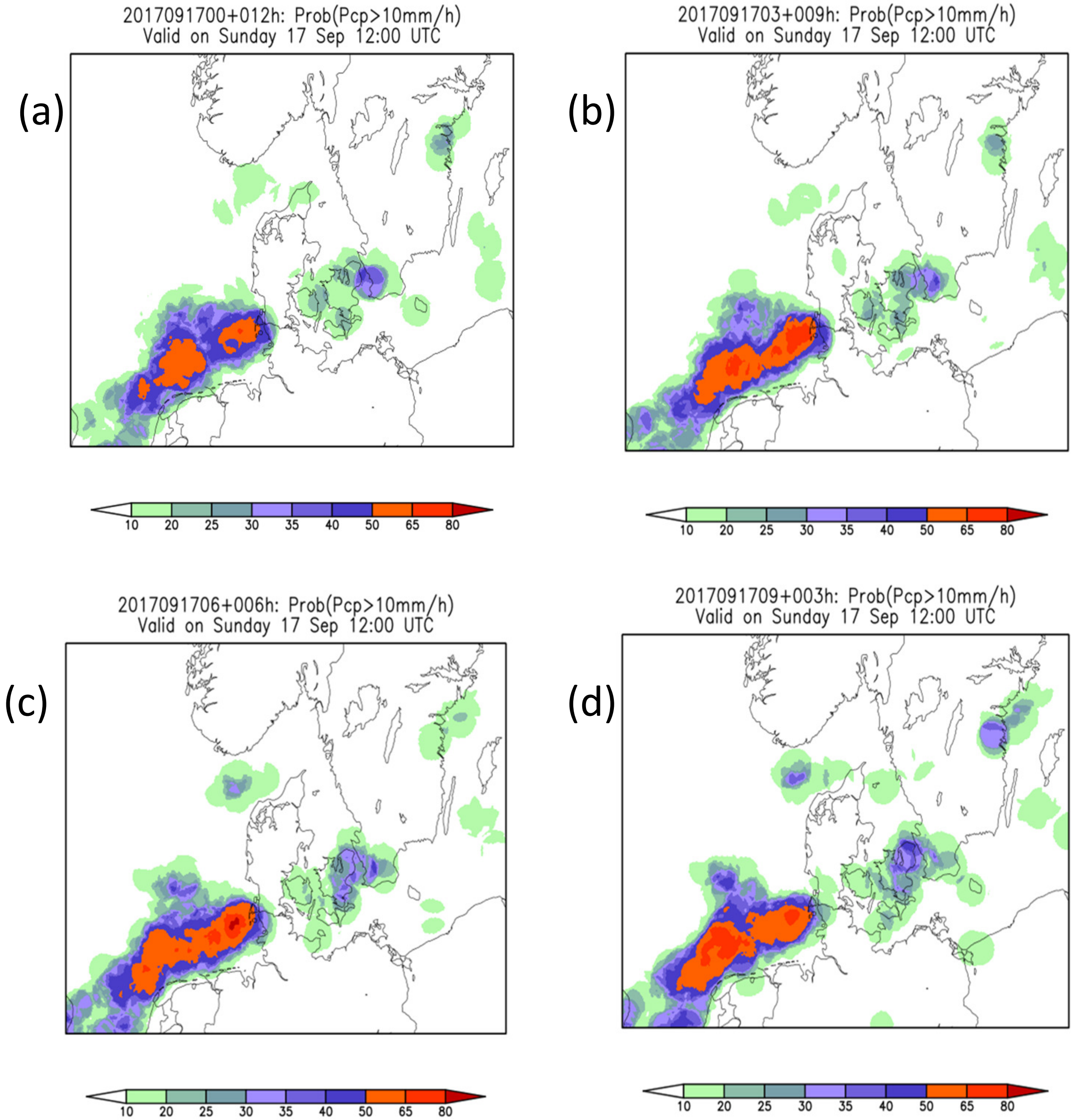
Radar image : 11:50 UTC



Radar image : 12:00 UTC

Spatial variation of accumulated precipitation (6 UTC- 18 UTC) in the area of Copenhagen 17 September 2017



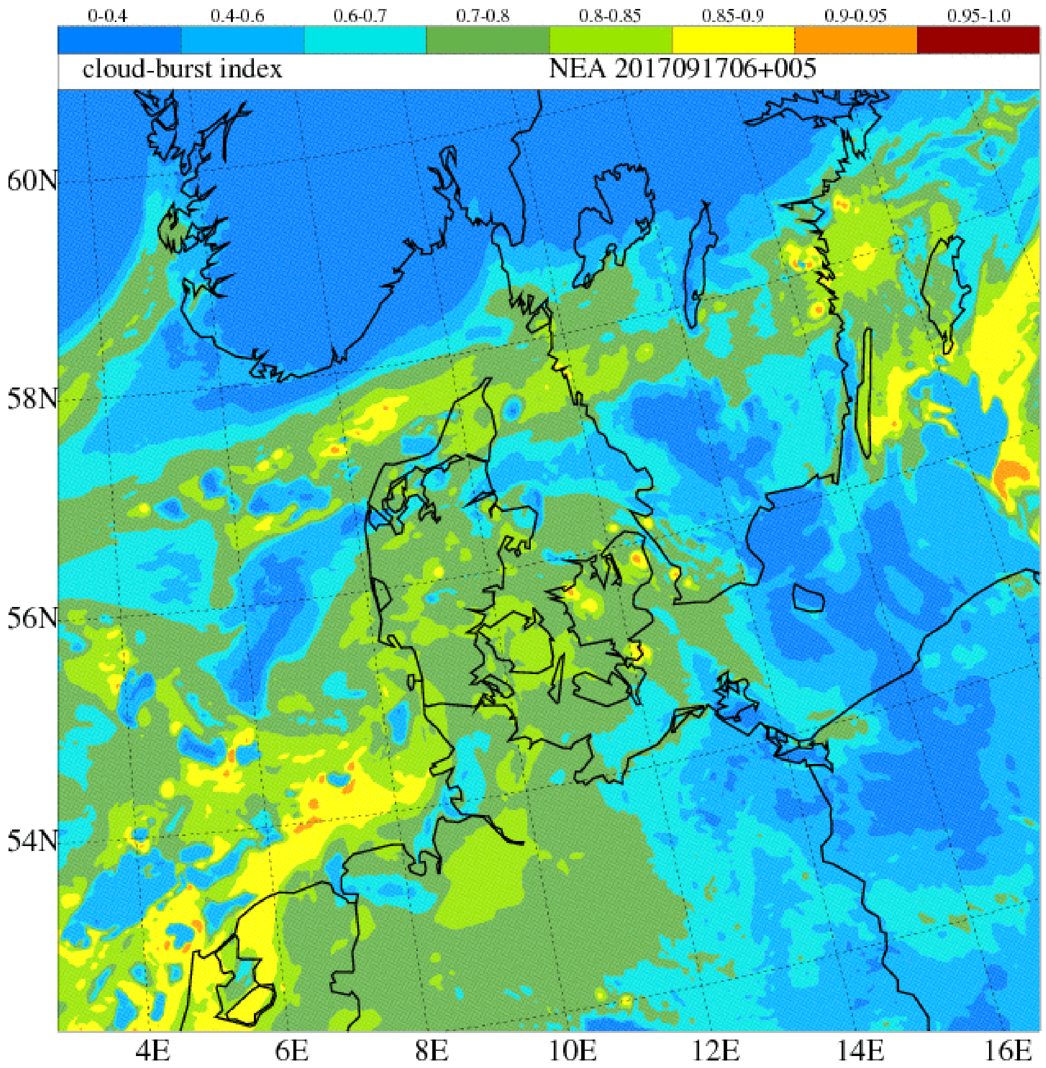


COMEPS `probabilities´

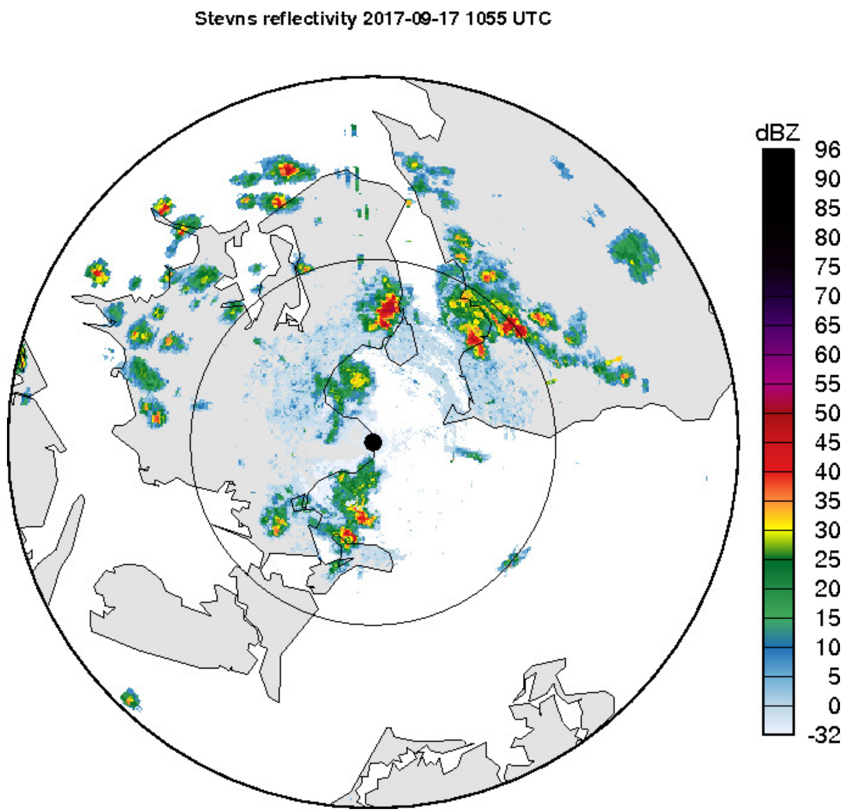
Pcp > 10mm/hour

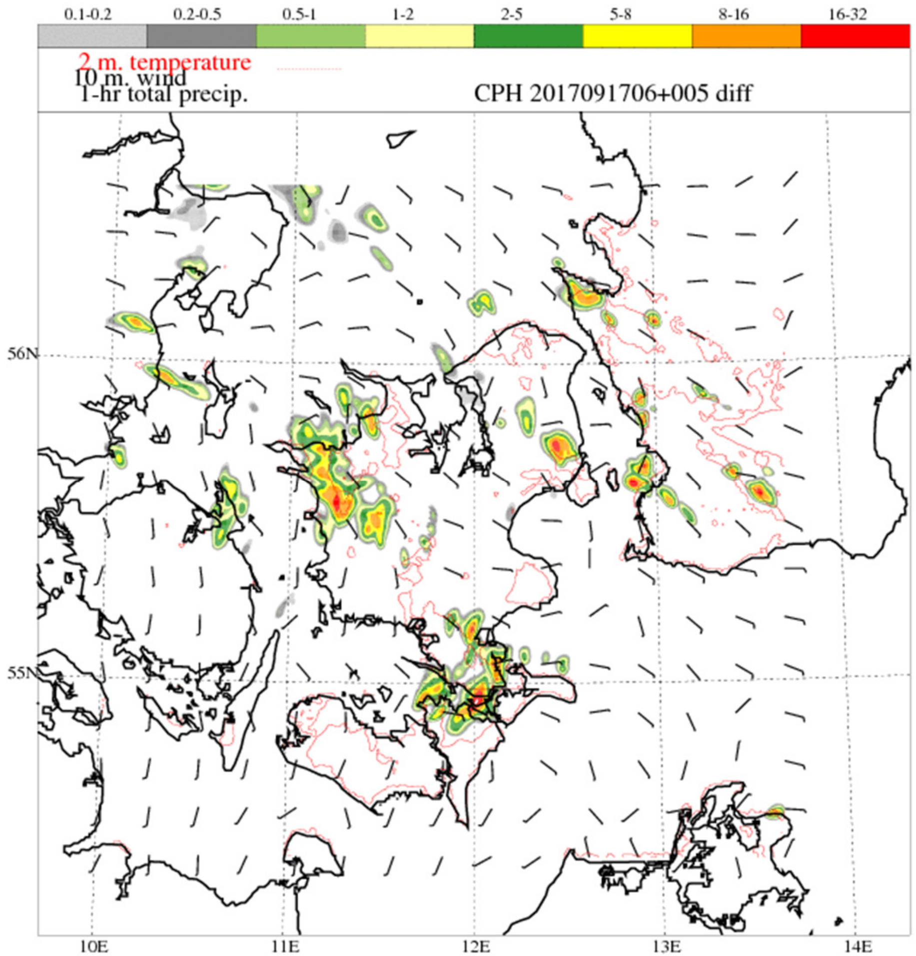
valid at 12 UTC 17/9 2017

Predictions starting
at 00UTC (Fig.a)
at 03UTC (Fig.b)
at 06UTC (Fig.c)
at 09UTC (Fig.d)



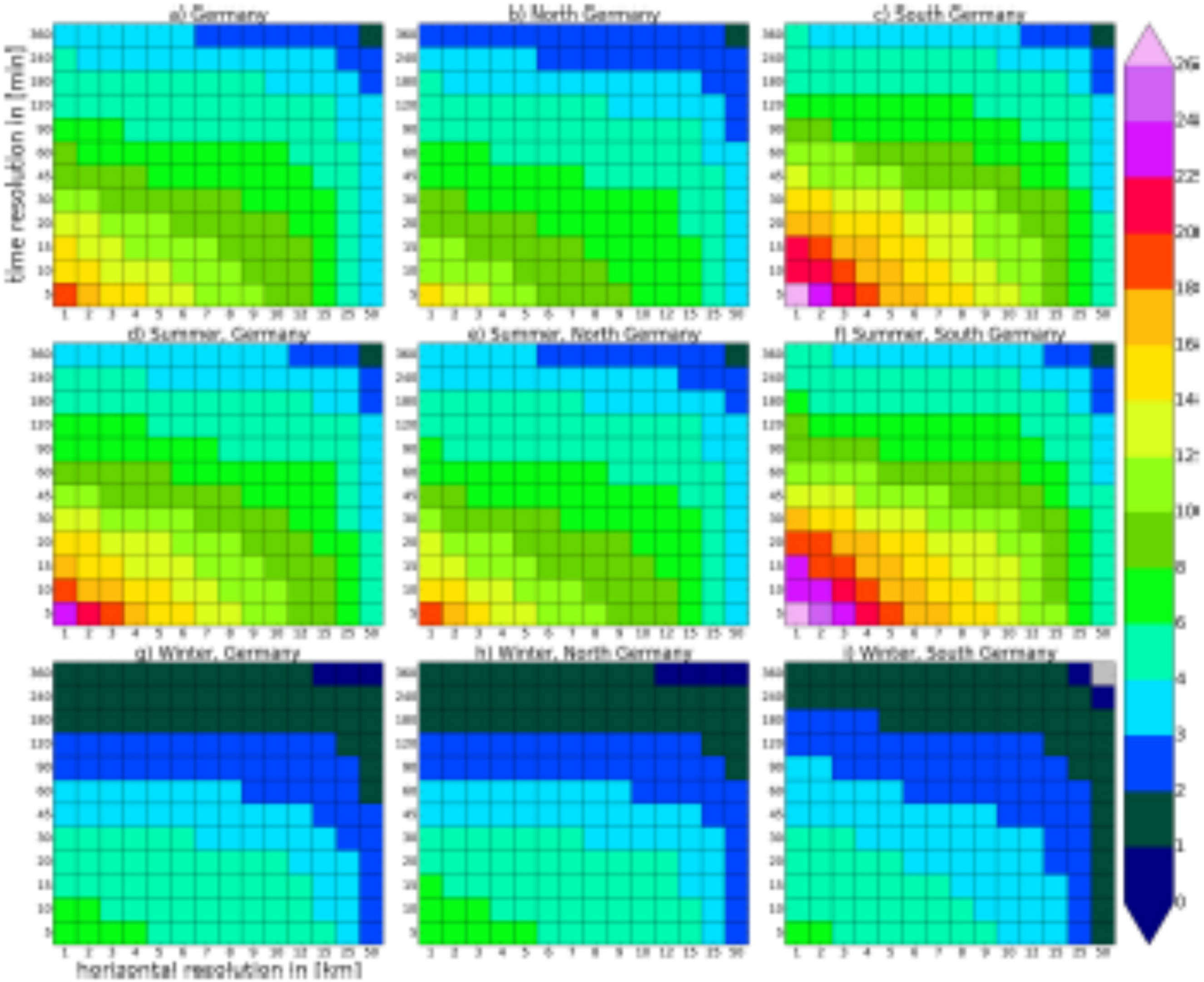
A new diagnostic cloud burst index in DMI did indicate high probability right north of Copenhagen at +5 hours (correct) . At +6 hours the index went to very low value again (correct).





Promising forecast with HARMONIE-AROME 500m model (right) to be compared with radar picture (left)

PRECIPITATION from RADAR: Large rainfall intensities and small time scales fit together



From Eggert et al., May 2015: Atmos. Chem. Phys. 15, 5957-5971

99th percentile of convective extreme precipitation intensities (mm/hour) as a function of resolution in space and time, deduced from German radar systems.

Disinction between Entire Germany (first column) Northern Germany (second column) and Southern Germany (third column) .

Also a distinction is made between ”entire year” , ”summer ” and ”winter”

Vertical axes show time resolution , horizontal axes shows spatial resolution

We see , e.g. second row, third plot , that large precipitation intensities are well correlated with small time scales (deduced pependency down to 1 km scale

Brief summary of the evolution process from

NWP verified in `fixed points´ to

NWP verified for `scales in space and time´ :

Why is this transition important ?

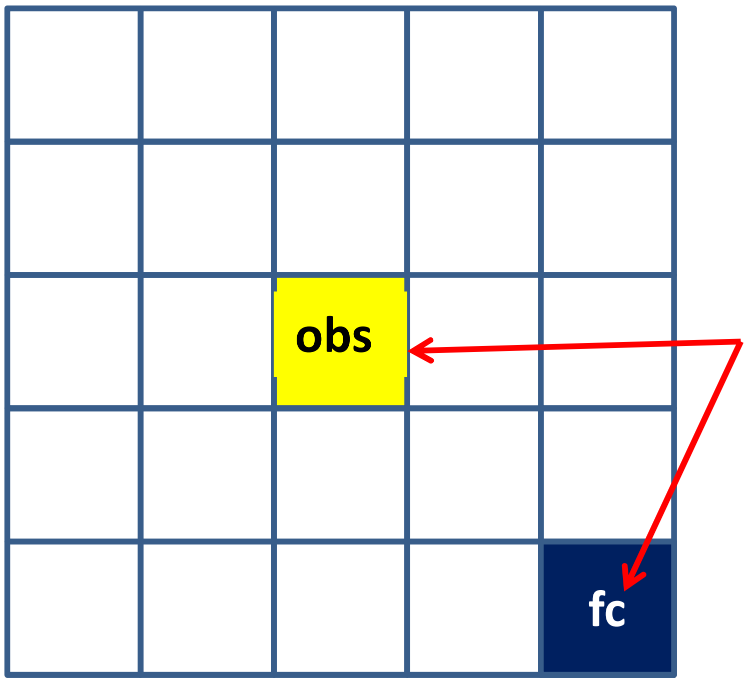
From `point verifications´ to `spatial verification methods´

Diagnosing predictable spatial- and time scales

” SPATIAL WINDOW” matters
especially when predicting extremes

SUGGESTION:

For a given threshold to be forecasted look for and `optimal´ upscaling distance to be used. This may be determined on the basis of verification using different upscaling.



BASIC CHALLENGE :

Theory says that there is not predictability on GRID SCALE.

Normally 6 or more grid moints are associated with the smallest scales that can be handled by the model

Forecasting `obs´ correctly on gridscale is not likely to happen, but operating on predictable scales gives better chance

Common NWP verification practices and evolution trends: From `point verifications´ to `spatial verification methods´

- **Most users of NWP models need to know the risk of extremes. Traditional model verification computing BIAS an RMS in points fail to produce fair verifications for extremes due to the `double penalty'**

- **First a high resolution model capable of producing extremes is penalized for not verifying an extreme on the spot (point) where it is observed**

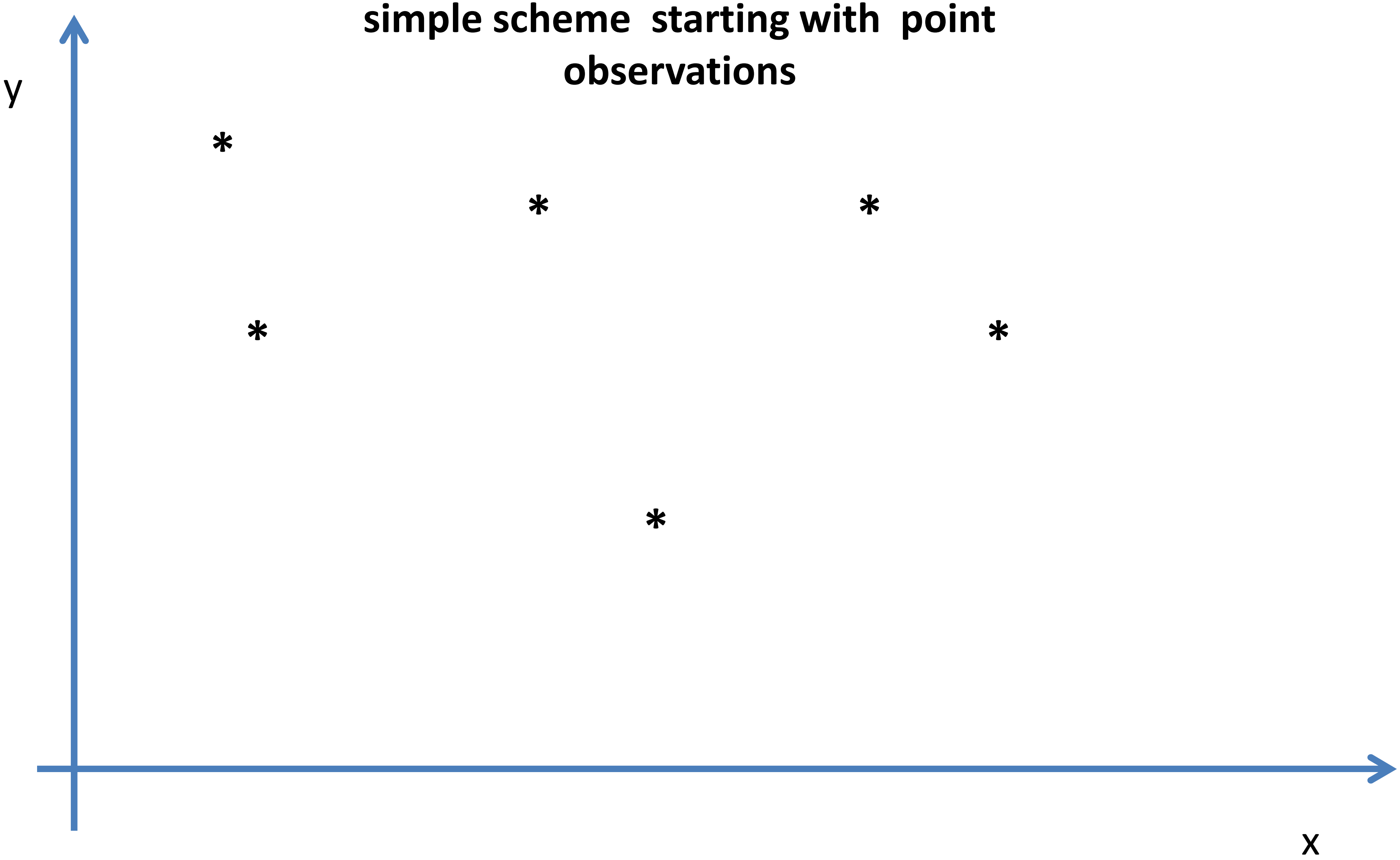
- **Secondly, it might be penalized for predicting the extreme at a location where it is not observed.**

- **Also predictions you would associate with `hedging´ will often have low RMS (predictions avoiding large errors but without forecasting risk of extremes well) :**

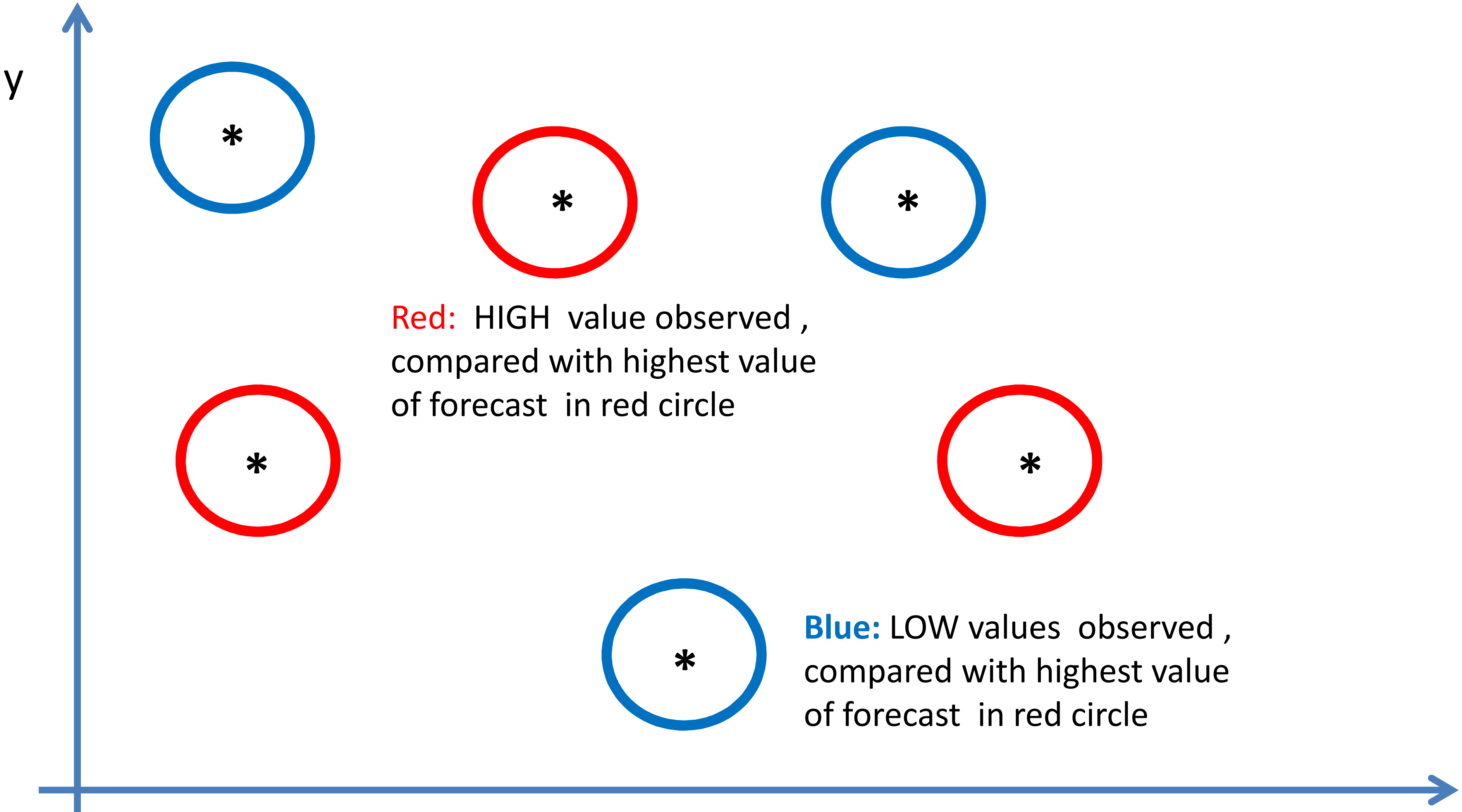
- **HENCE there is a need for SPATIAL verification methods which have been in focus in recent years.**

EXAMPLE:

Transition to spatial verification: Example of simple scheme starting with point observations



Transition to spatial verification: Signifcant Weather Score (SWS), Spatial tolerance included in the prediction , the dimention Dof the circles , [ref. 1]



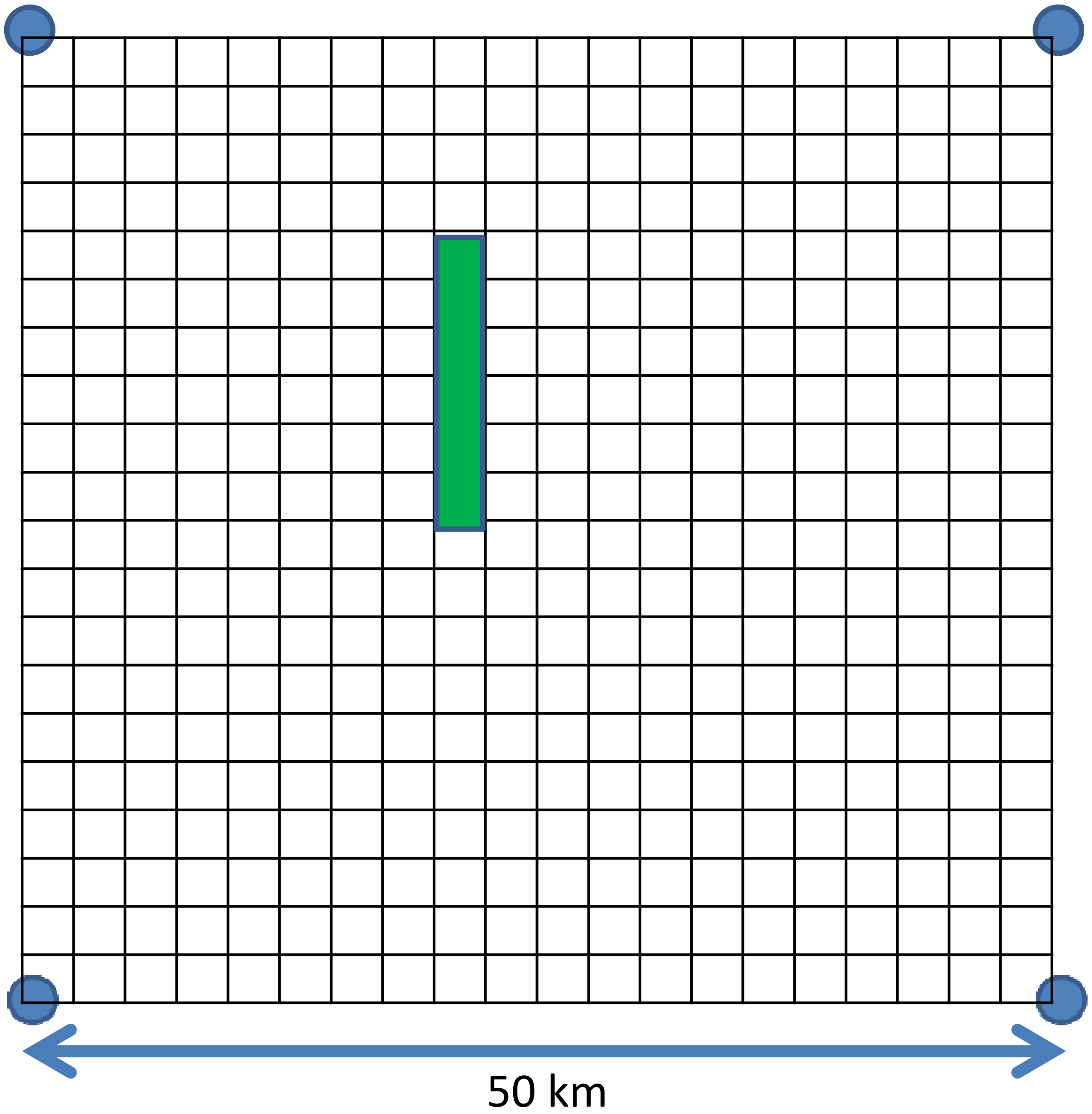
SWS=(1+ Σ Fmeso) / (1 + Σ Fref) , k= 1,N

where the Fmeso and Fref measure the success of the prediction with mesoscale prediction

NB: In the early days of EWGLAM a common station verification package was made. Characteristics of the ratio between model grid size and the distance between observations have changed over time !

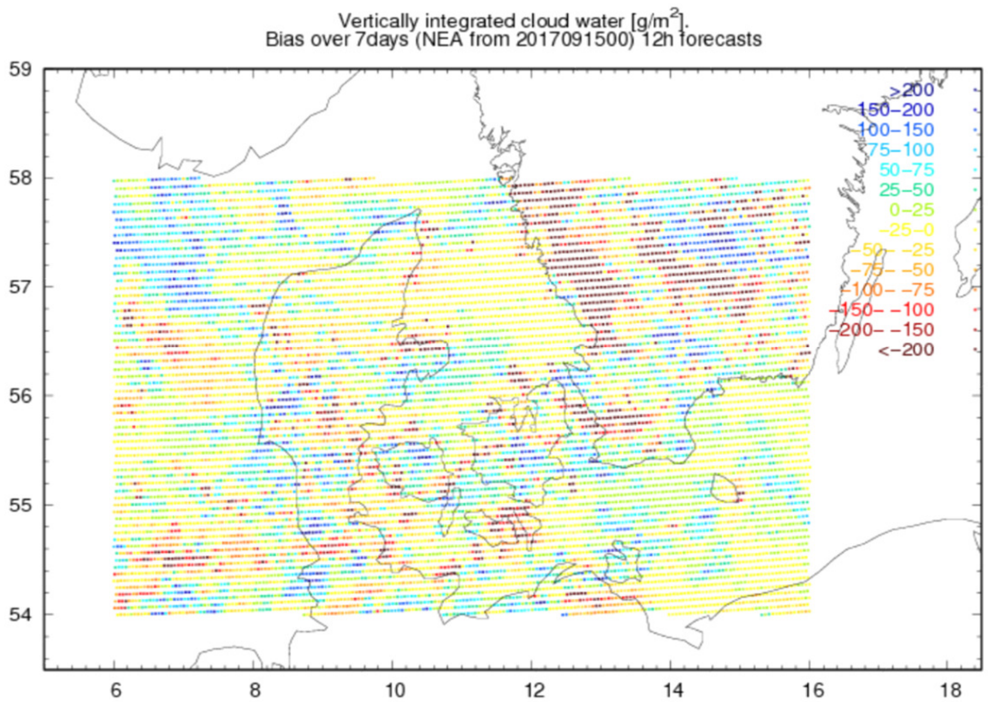
1985: **Typical grid size ~ 50 km ~**
Distance between observations **1 grid box inside 4 obs**

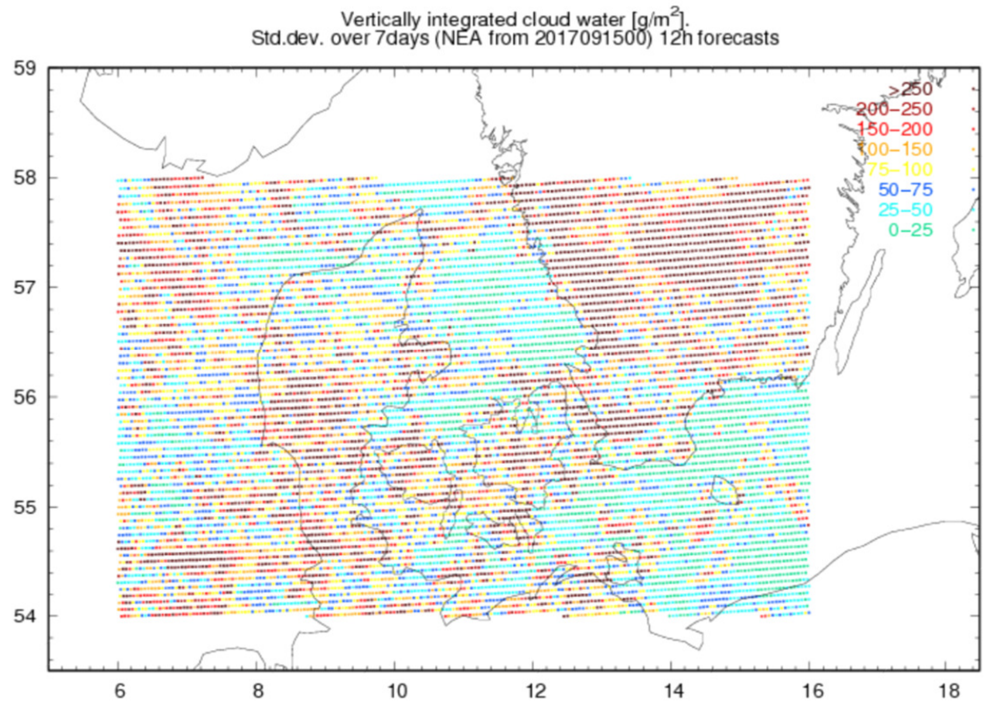
2015: **Typicall grid size ~ 2.5 km ~**
5 % distance between observations: **400 grid boxes inside 4 obs**


50 km

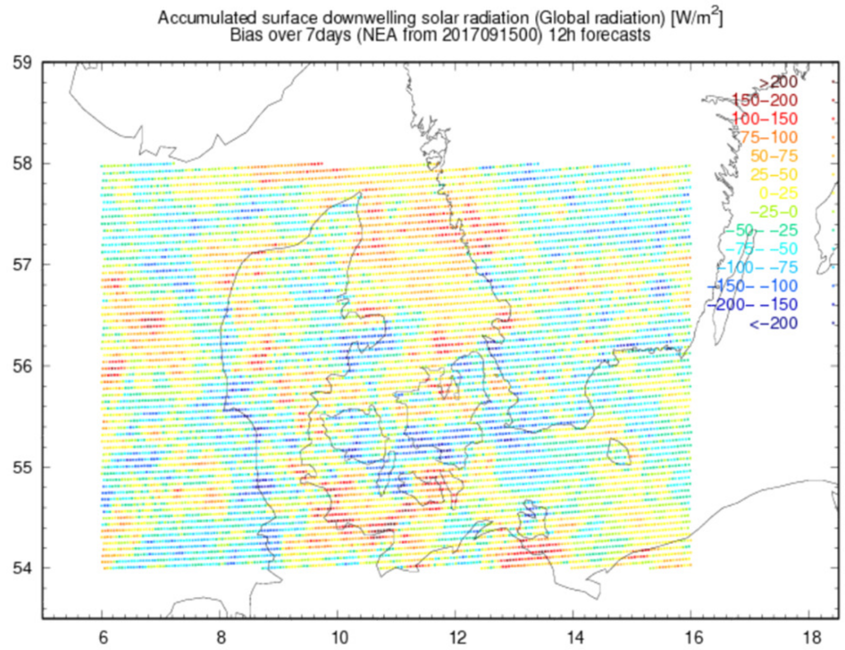
As a consequence between much smaller mesh size today compared with traditional surface obs distance :
USE suitable routine sateillite products at a comparable resolution to model resolution:

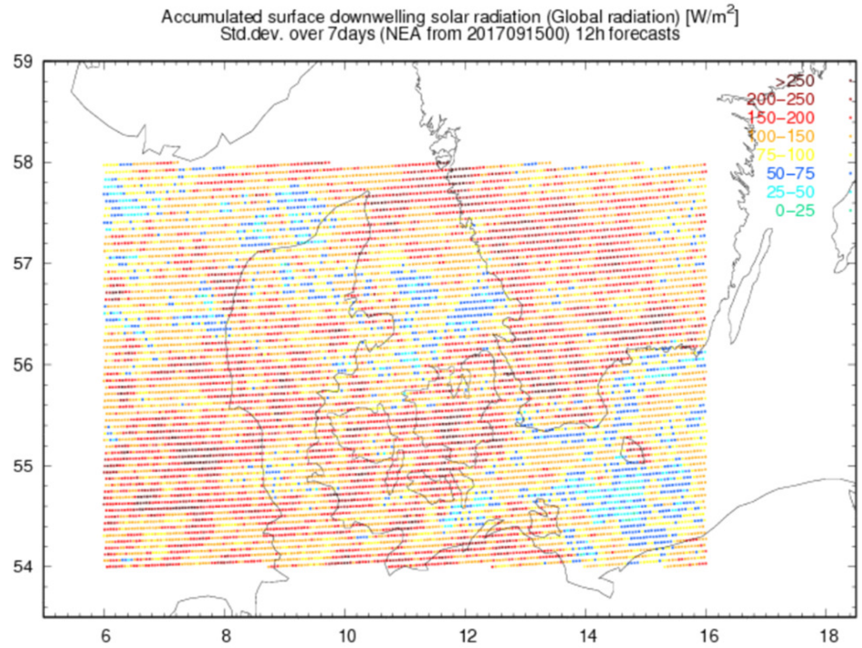
EXAMPLE 1: Maps of bias (left) and standard deviation (right) of HARMONIE-AROME vertically integrated cloud water compared with KNMI CLIMATE SAF deduced values (averaged over one week : 15/9-22/9 2017)





EXAMPLE 2: Maps of bias (left) and standard deviation (right) of HARMONIE-AROME downwelling global solar radiation compared with KNMI CLIMATE SAF deduced values (averaged over one week : 15/9-22/9 2017)





How should a future verification strategy look like ?

a) Case studies of significant/extreme cases

WHY ? Significant or extreme cases are the most relevant to society/users, forecasters must have confidence in model's ability to forecast extremes

DEMANDS: Several cases and model aspects to be dealt with

CHALLENGES: Procedure must be worked out for efficient launch and verification of (all) considered cases in order to be manageable and not creating bottle-neck in model developments

b) Run verification for long periods (months) for different seasons

WHY ? Statistical robustness is needed

DEMANDS: Efficient setup for execution and verification.

CHALLENGES: Improved mode `scalability´ desirable for fast execution.

How should a future verification strategy look like ?

c) Feedback from users

WHY ? **Users must feel that products from the model system has high quality and are reliable. Sometimes users identify issues not easily identified in standard verification**

DEMANDS: **Procedures for communication should be well defined. Quality scores and communication practice should be useful to both users and developers**

CHALLENGES: **Special and new useful postprocessed products might be needed and developed in a collaboration between developers and users.**

References :

(1) Eggert, B., Berg, P., Haerter, P.O. , Jacob, D. and Moseley, C. , 2015: Atmos. Chem. Phys. 15, 5957-5971

(2) Roberts , N.M., and Lean, H.W., 2008:
Scale –selective verification of rainfall accumulations from high-resolution forecasts of convective events . Mon. Wea. Rev., 136, 78-97

[3] Sass, B.H., and Yang, X, 2012: A verification score for high resolution NWP: Idealized and pre-operational tests , HIRLAM Tech. Rep. no. 69, 28pp [available from www.hirlam.org]

[4] Wernli, H. , Paulat, M., Hage, M. and Frei, C., 2008:
SAL- A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts , Mon. Wea. Rev. , 136, 4470 – 4487