The Challenge of Verifying the AROME-RUC Precipitation Forecasts

***Phillip Scheffknecht**, Florian Meier, Christoph Wittmann*

ZAMG
Zentralanstalt für Meteorologie und Geodynamik

Overview

Motivation and Methodology
- Basic Challenge and Idea
- The Tool: **Panelification**
- Scores and Simplifications

Example Output
- Deeper Look at an Event
- Results for the Summer 2019

Discussion & Outlook

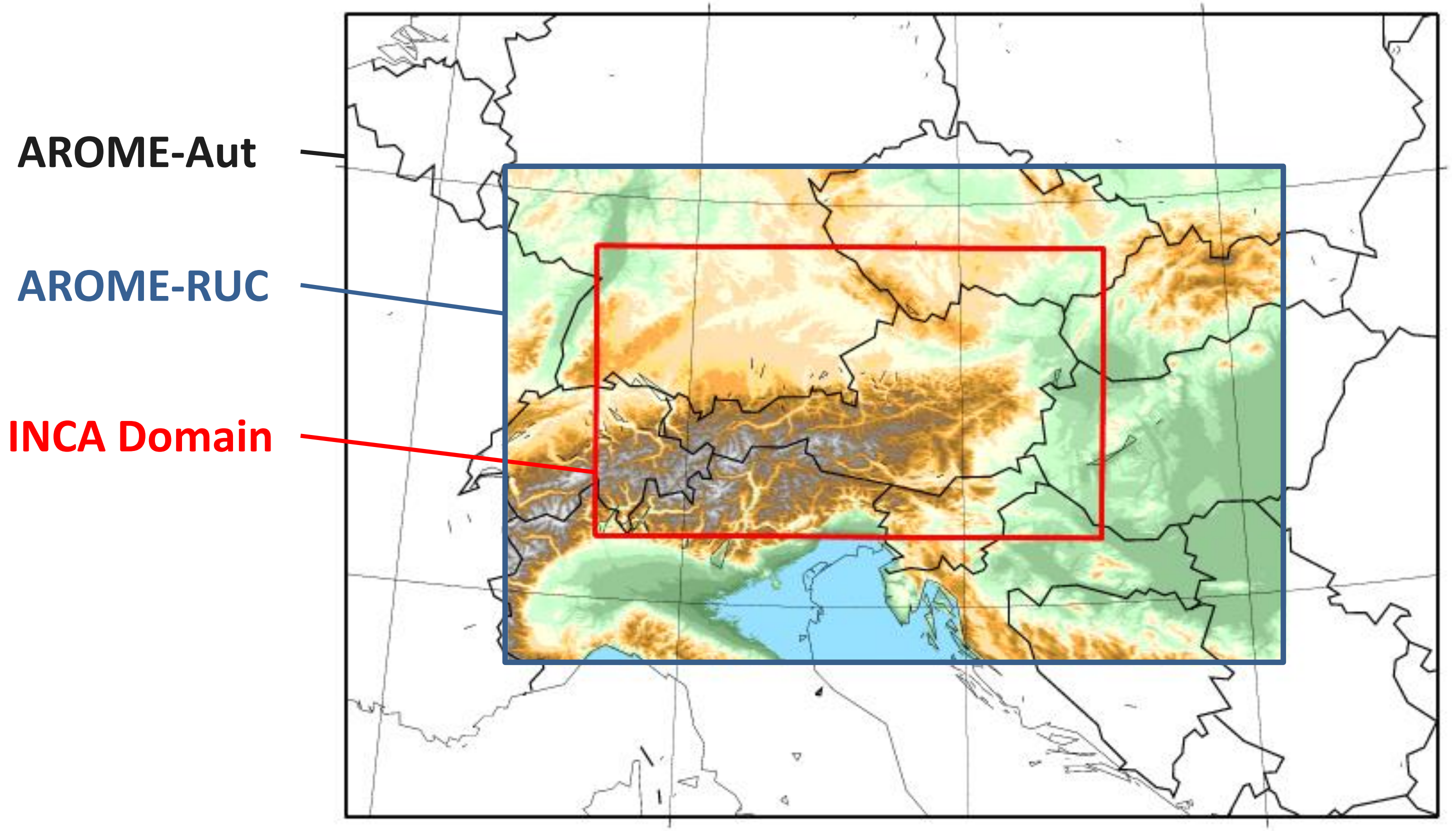
Overview

- **Motivation and Methodology**
  - Basic Challenge and Idea
  - The Tool: **Panelification**
  - Scores and Simplifications
- **Example Output**
  - Deeper Look at an Event
  - Results for the Summer 2019
- **Discussion & Outlook**

The Austrian AROME-RUC Nowcasting System

- Idea: fill gap between classical nowcasting systems and short range NWP
- Hourly forecasts up to 12h with hourly 3D-Var and 25 min cutoff time available within 1h
- 900x576x90 Grid Points at 1.2km horizontal grid spacing
- LBC from AROME-AUT, hourly OI soil assimilation

AROME–Nowcasting Domain & Topography



Original Motivation – Quickly Determine **Relative Model Performance**

- **Originally:** A possibility to **quickly examine** the results of past forecasts, compare the runs and models, give an **overview over performance**
- Based upon archived model forecasts
- **Challenges:**
  - High Resolution spatial data -> classic metrics like MAE, Bias, perform poorly in some instances
  - Lots of Data to process
  - Lots of Models to compare
  - Calculate verification and present it in a way that allows quick judgment by experts

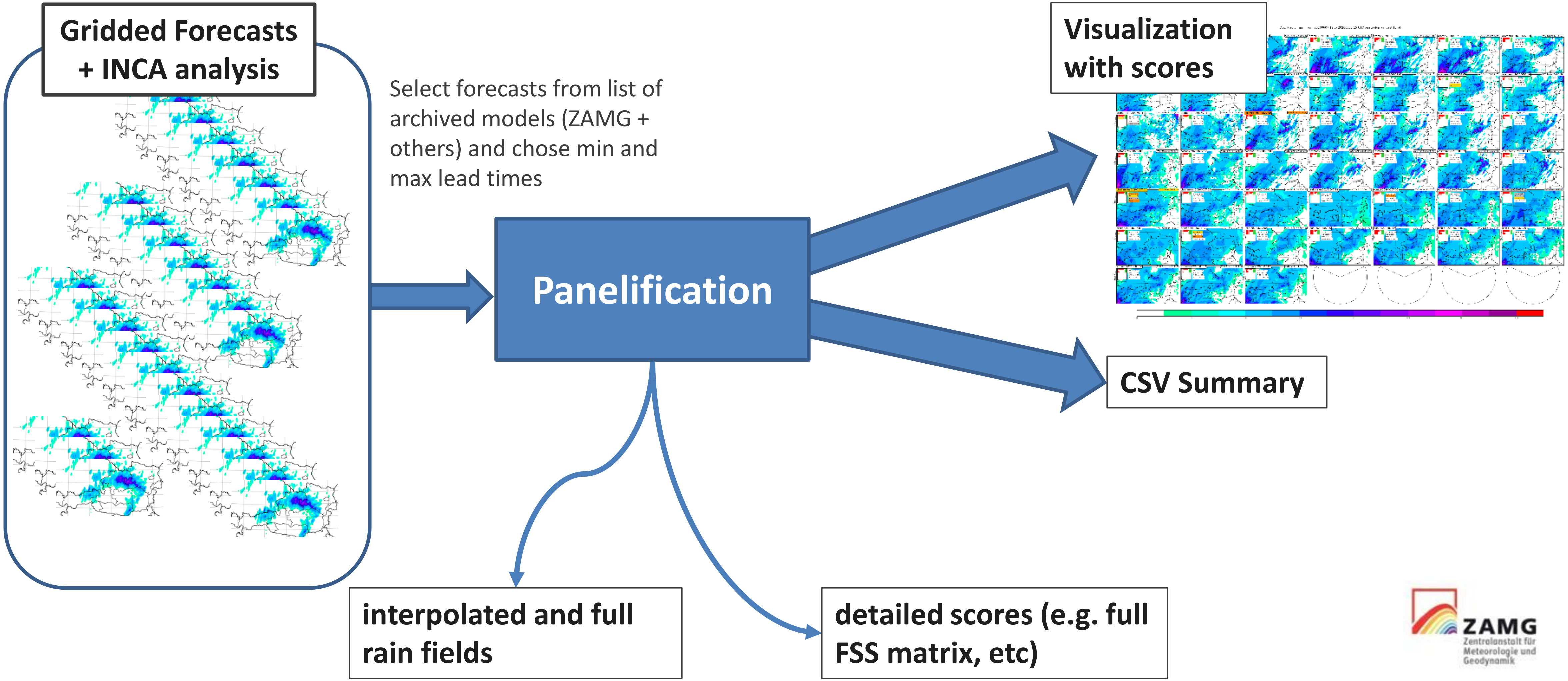
Ranking Forecasts – Experimenting with Further Simplification

- **But which forecast is the best?
  Is this or that model better?
  Is the new configuration better?
  Which configuration works best for our case study?**

- The answer is almost always an ***it depends*…**
  - Try to come up with a measure for ranking the simulations, **knowing that this is losing information!**
  - Make sure that the resulting ranking corresponds to what an expert would deem a **good forecast.**

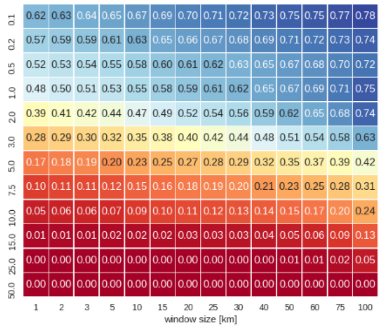
The Resulting Tool: **Panelification**

- It started as a small visualization and verification python programme, but grew slowly



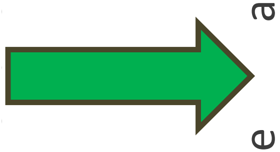
Simplification of the FSS Display

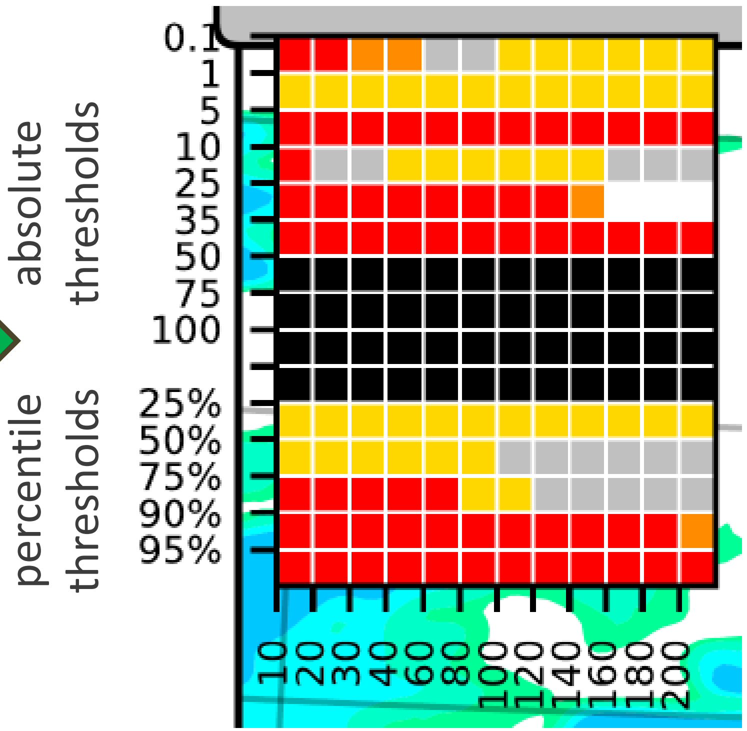
- Reduce information clutter by removing the numbers
- Focus on fast visual aid for comparison



	1	2	3	5	10	15	20	25	30	40	50	60	75	100
0.1	0.62	0.63	0.64	0.65	0.67	0.69	0.70	0.71	0.72	0.73	0.75	0.75	0.77	0.78
0.2	0.57	0.59	0.59	0.61	0.63	0.65	0.66	0.67	0.68	0.69	0.71	0.72	0.73	0.74
0.5	0.52	0.53	0.54	0.55	0.58	0.60	0.61	0.62	0.63	0.65	0.67	0.68	0.70	0.72
1.0	0.48	0.50	0.51	0.53	0.55	0.58	0.59	0.61	0.62	0.65	0.67	0.69	0.71	0.75
2.0	0.39	0.41	0.42	0.44	0.47	0.49	0.52	0.54	0.56	0.59	0.62	0.65	0.68	0.74
3.0	0.28	0.29	0.30	0.32	0.35	0.38	0.40	0.42	0.44	0.48	0.51	0.54	0.58	0.63
5.0	0.17	0.18	0.19	0.20	0.23	0.25	0.27	0.28	0.29	0.32	0.35	0.37	0.39	0.42
7.5	0.10	0.11	0.11	0.12	0.15	0.16	0.18	0.19	0.20	0.21	0.23	0.25	0.28	0.31
10.0	0.05	0.06	0.06	0.07	0.09	0.10	0.11	0.12	0.13	0.14	0.15	0.17	0.20	0.24
15.0	0.01	0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.05	0.06	0.09	0.13
25.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.02	0.05
50.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

window size [km]



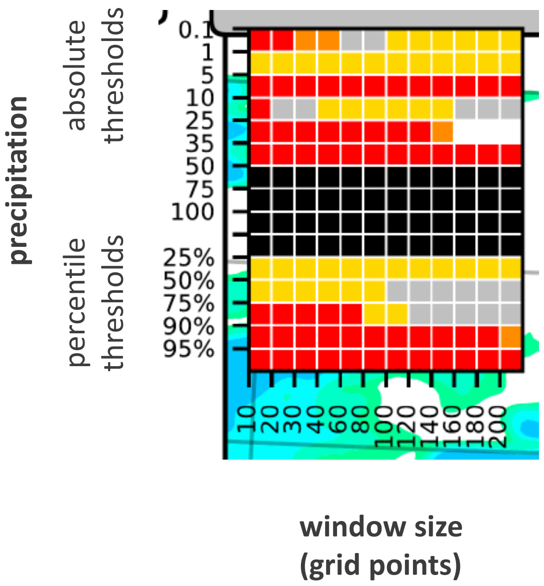


window size (grid points)

green	1.0 (perfect)
gold	best of all simulations
silver	second place
bronze	third place
white	above usefulness threshold
red	below usefulness threshold
black	NaN (not observed)

Experimental Forecast Ranking

- **Simple:** simply rank forecasts according to a single metric, e.g. MAE
- **Simple, but…:** combine these ranks into a single rank, mixes different metrics
- **Experimental:** Condense the information contained in the FSS into a single number and rank the forecasts accordingly.



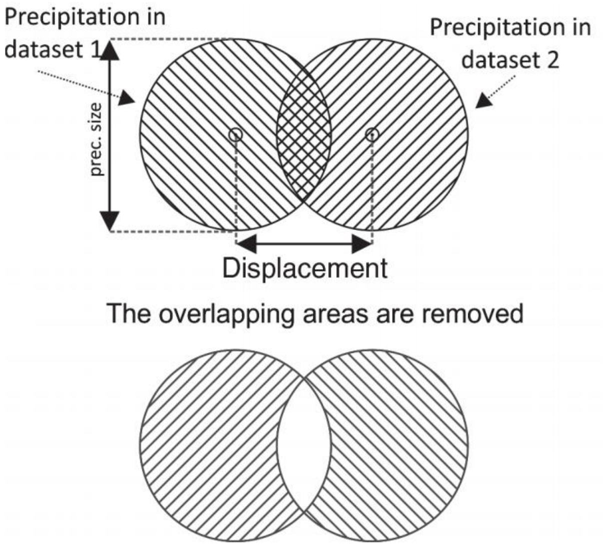
Test Score:

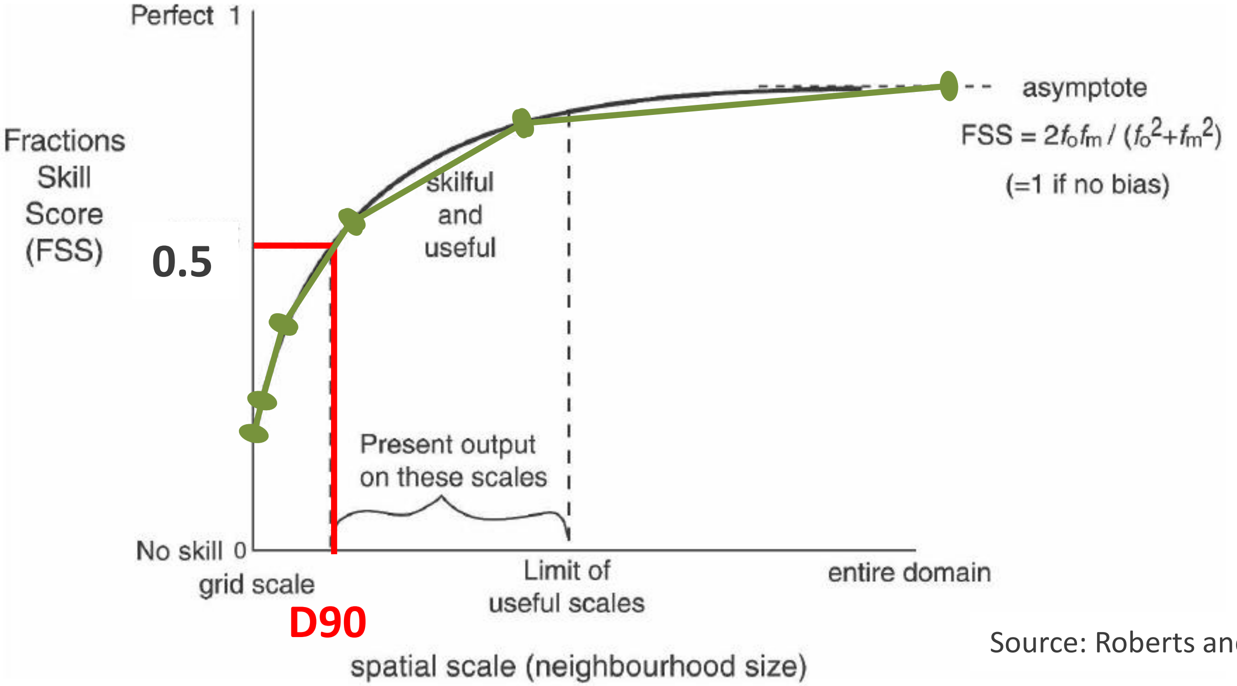
$$FSS\ Rank\ Score = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{1}{R_{T_n,W_m}}$$

One over *R* Summed over all thresholds and window sizes

D90: Displacement of the 90th precipitation percentile

- Use 90th Percentile -> removes bias
- D90 is defined as the window size at which the FSS exceeds 0.5, the threshold for a skillful and useful forecast





Approximation:

1. Remove Overlap
2. Calculate FSS for 1, 2, 4, 8, … 2k windows
3. Stop when FSS > 0.5
4. Linearly interpolate to 0.5

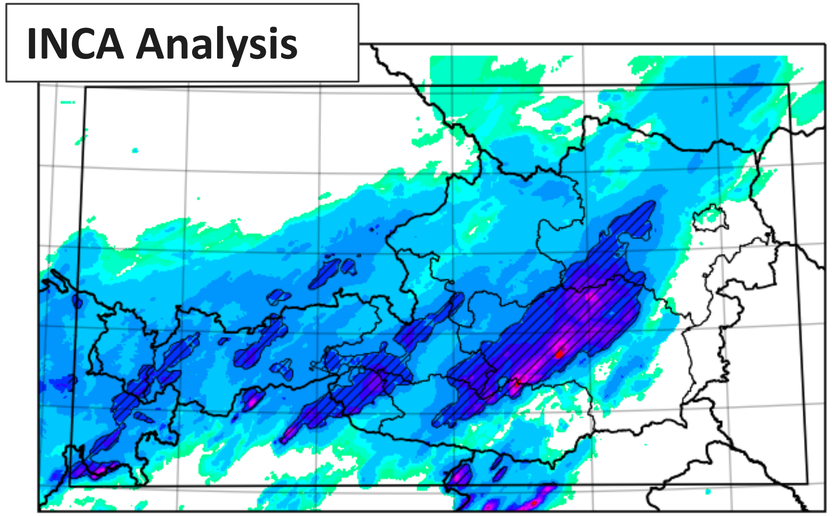
Source: Roberts and Lean (2007), **Skok and Roberts (2018)**

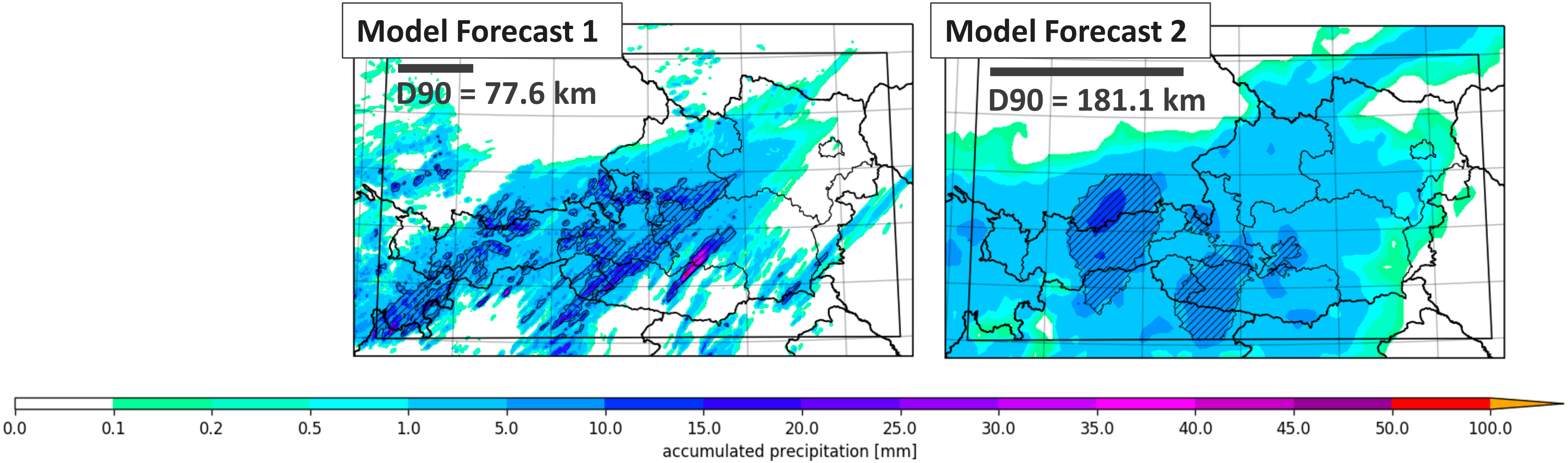
ZAMG Zentralanstalt für Meteorologie und Geodynamik

D90: Displacement of the 90th precipitation percentile

- Example plot that shows the 90% of the grid points with the most intense precipitation.

- This shows the **before** eliminating the overlap





ZAMG Zentralanstalt für Meteorologie und Geodynamik

Overview

Motivation and Methodology
- Basic Challenge and Idea
- The Tool: **Panelification**
- Scores and Simplifications

➡ **Example Output**
- Deeper Look at an Event
- Results for the Summer 2019

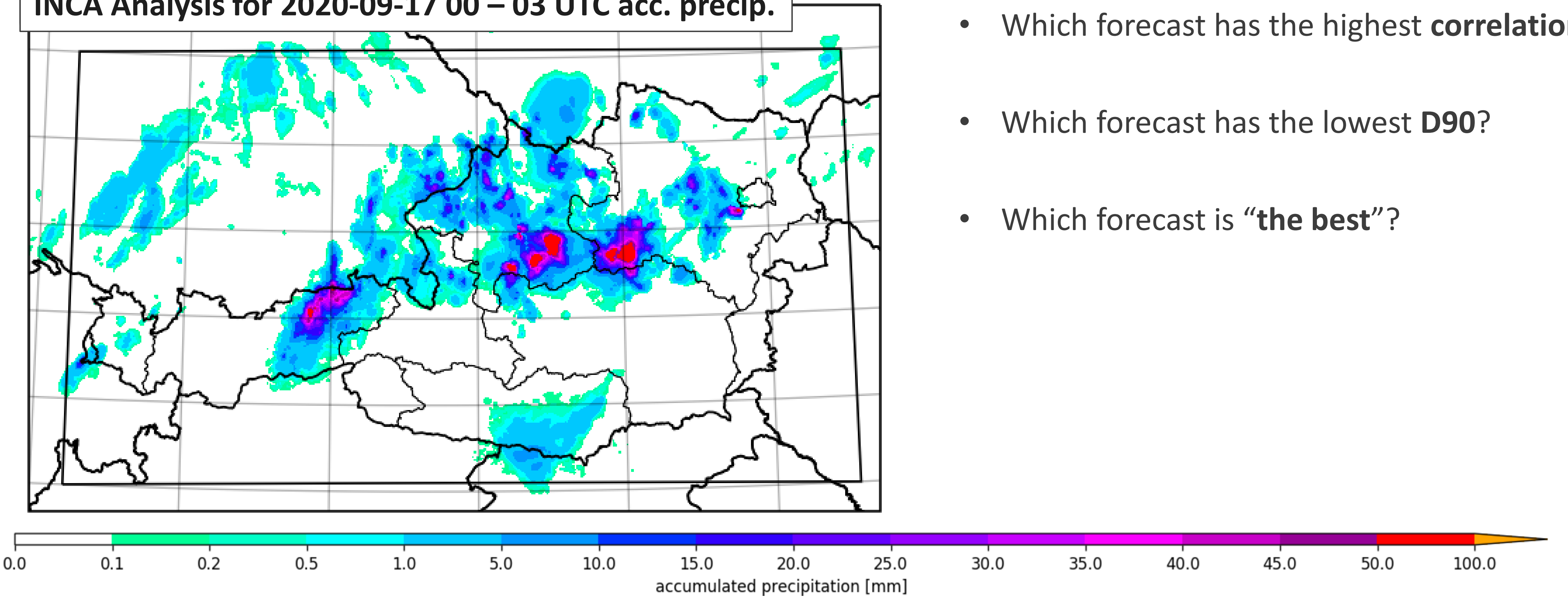
Discussion & Outlook

Detailed (Mini-)Example for a Single Nocturnal Convective Forecast

- Nocturnal Convection observed on 17 September 2020

- Different metrics will respond differently to these features

INCA Analysis for 2020-09-17 00 – 03 UTC acc. precip.



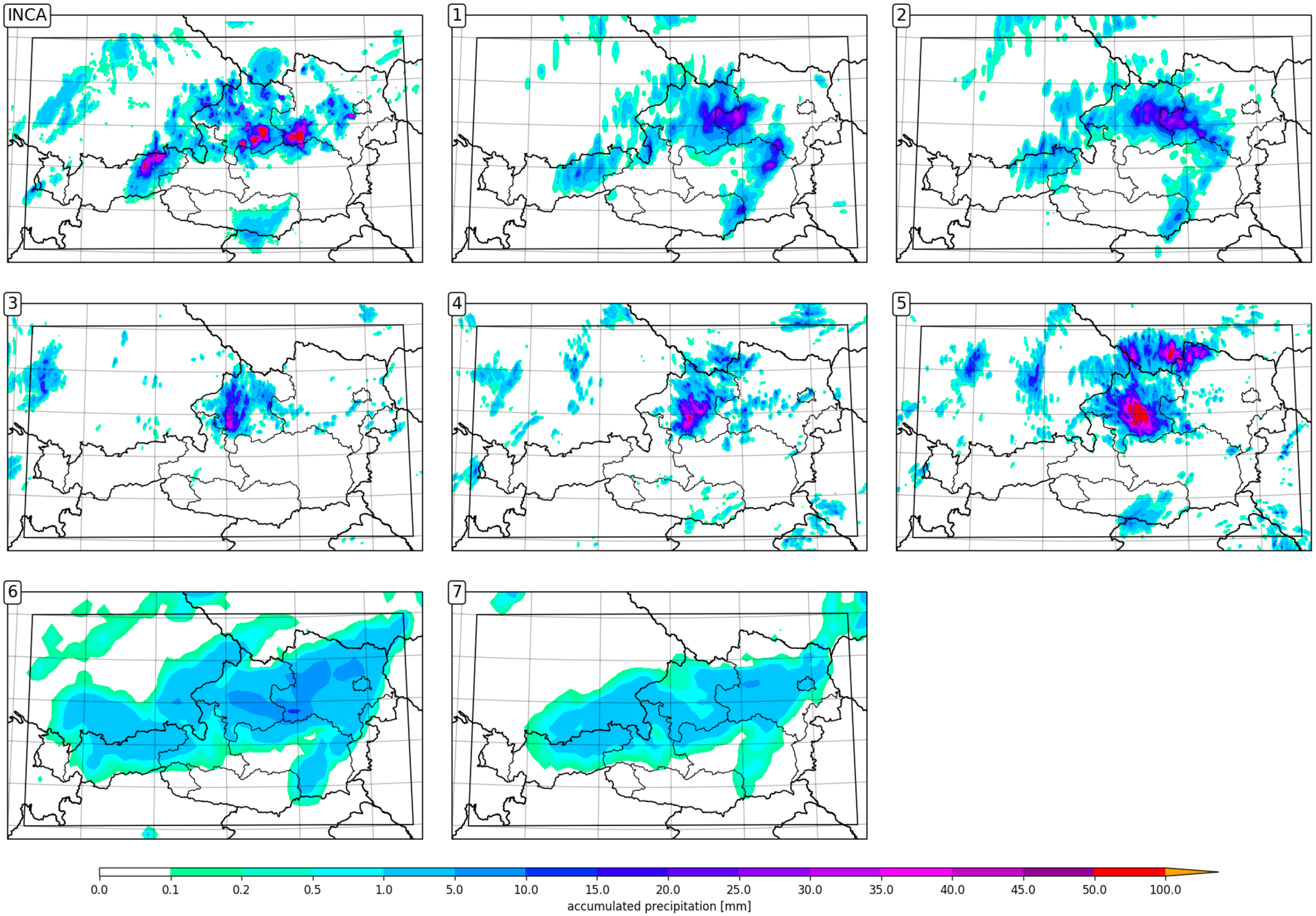
- Which forecast has the lowest **MAE**?

- Which forecast has the highest **correlation**?

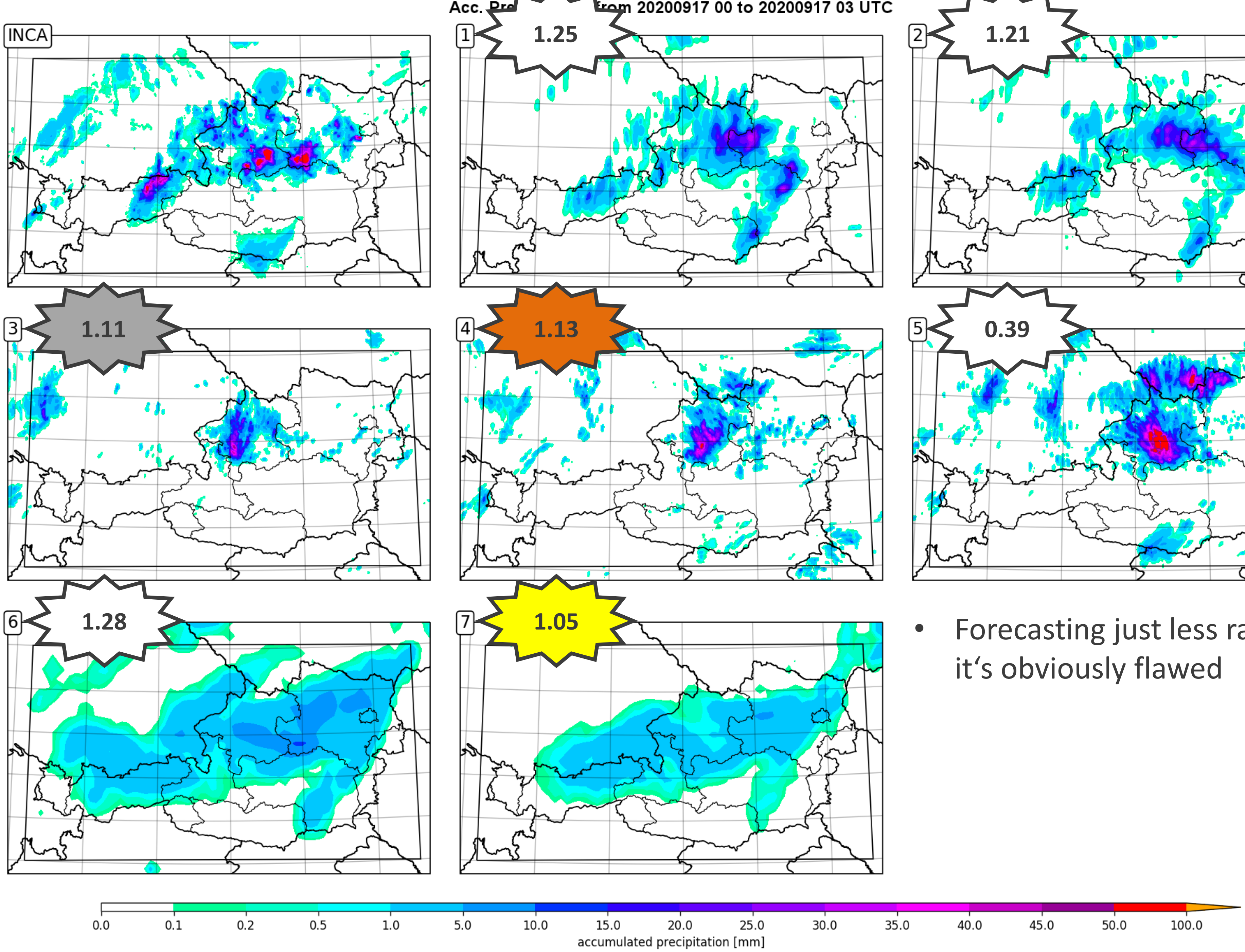
- Which forecast has the lowest **D90**?

- Which forecast is "**the best**"?

Acc. Precip. [mm] from 20200917 00 to 20200917 03 UTC

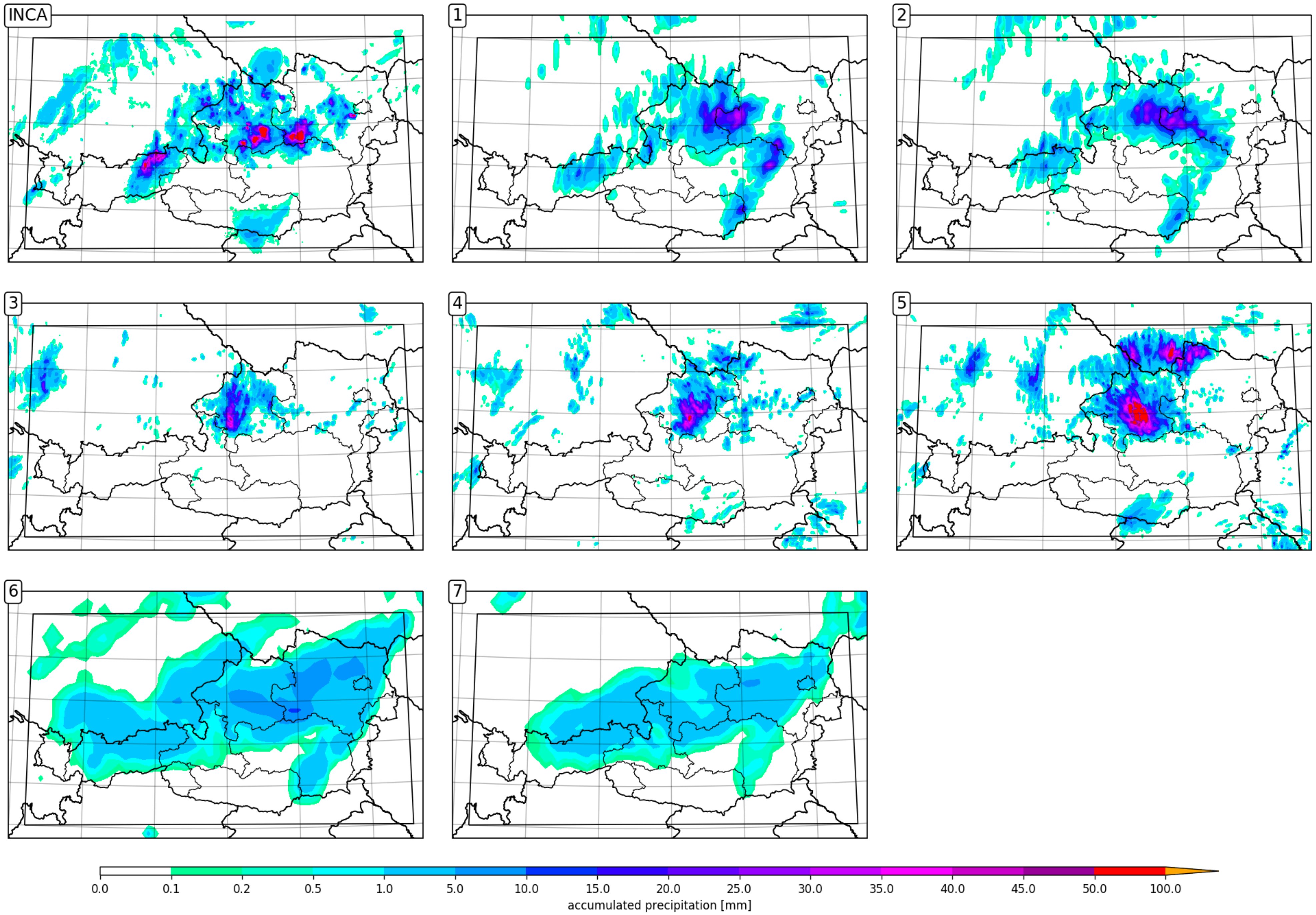


ZAMG
Zentralanstalt für Meteorologie und Geodynamik

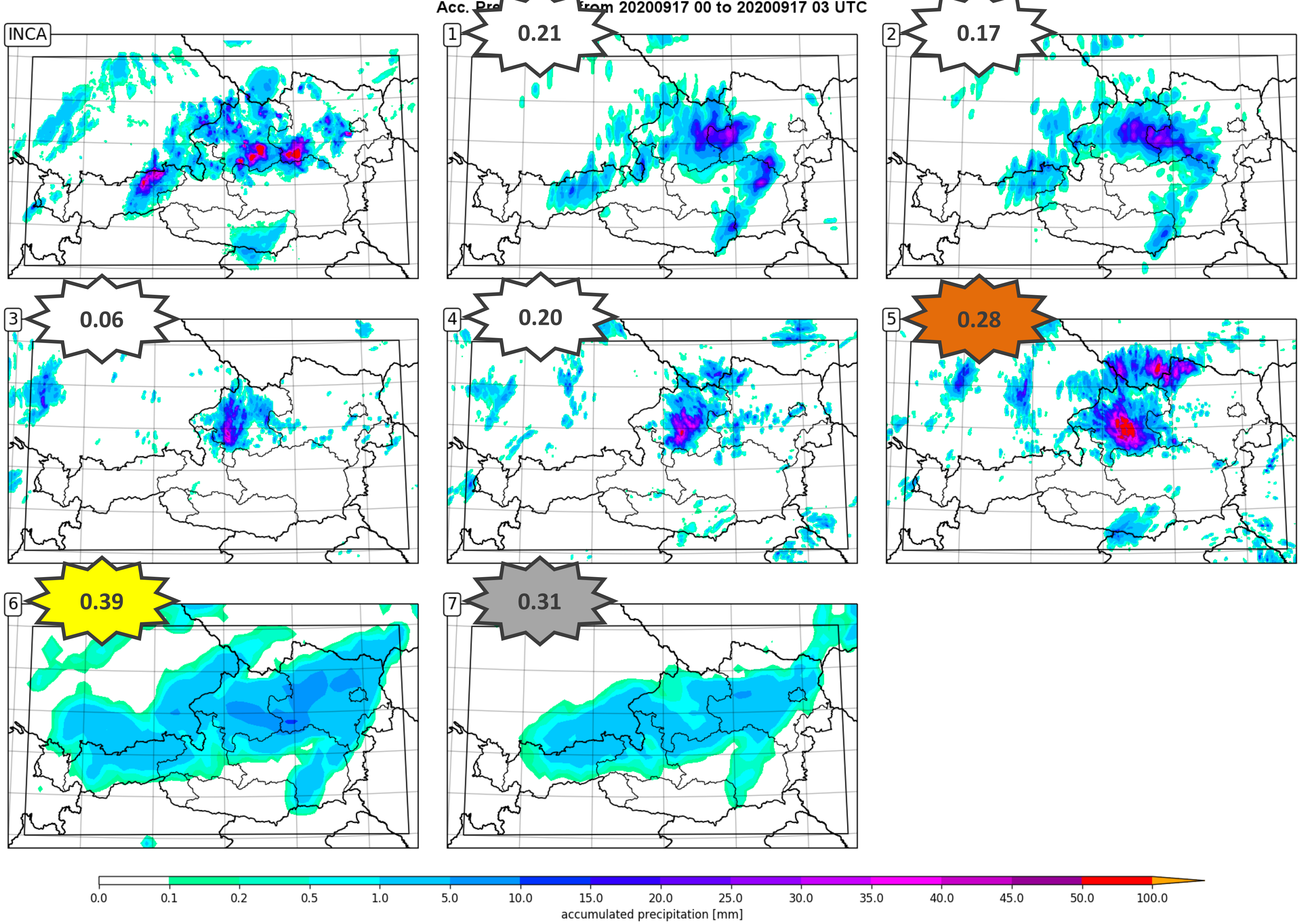


- Forecasting just less rain means less error, even if it‘s obviously flawed

Acc. Precip. [mm] from 20200917 00 to 20200917 03 UTC

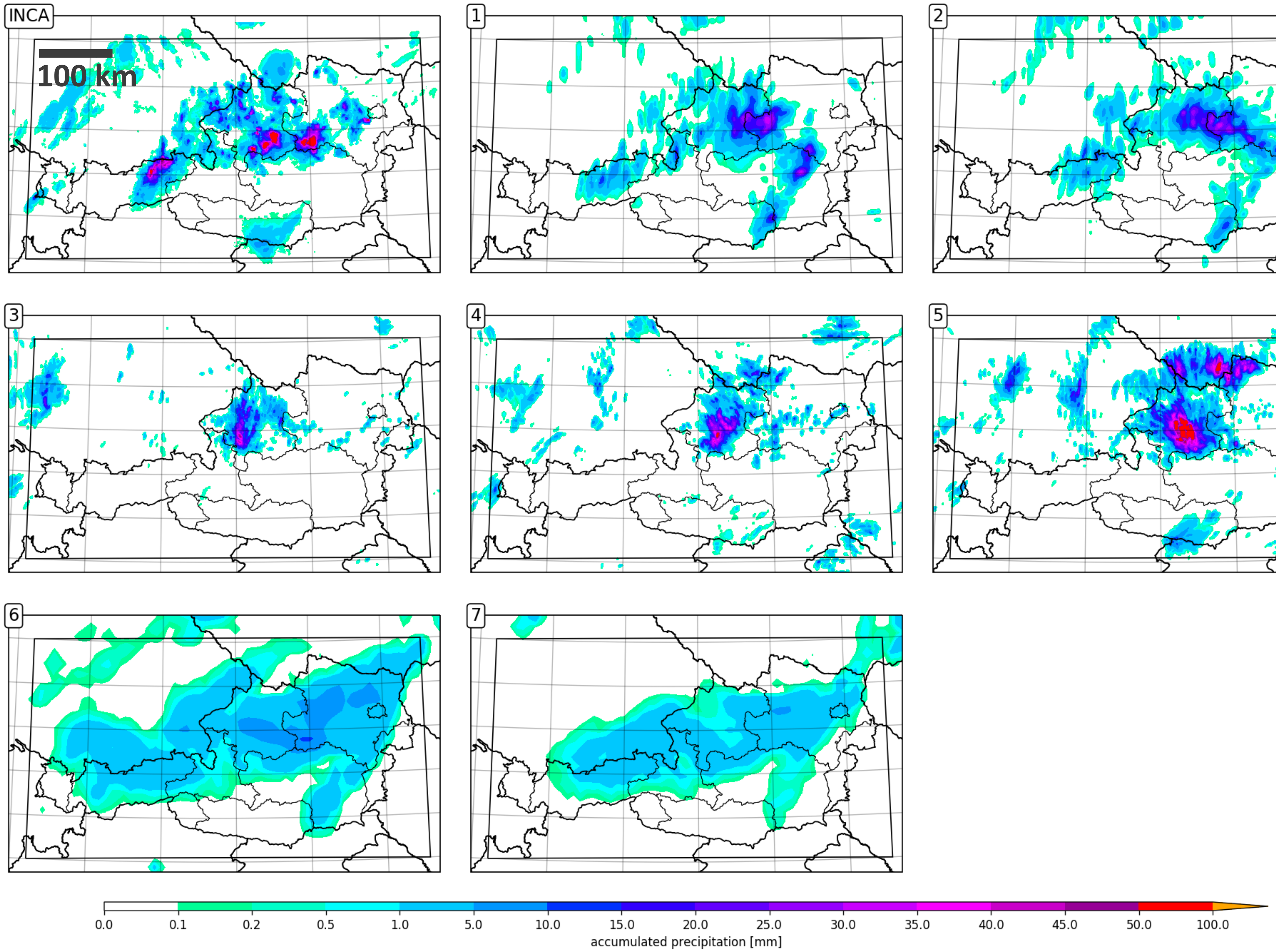


ZAMG
Zentralanstalt für Meteorologie und Geodynamik

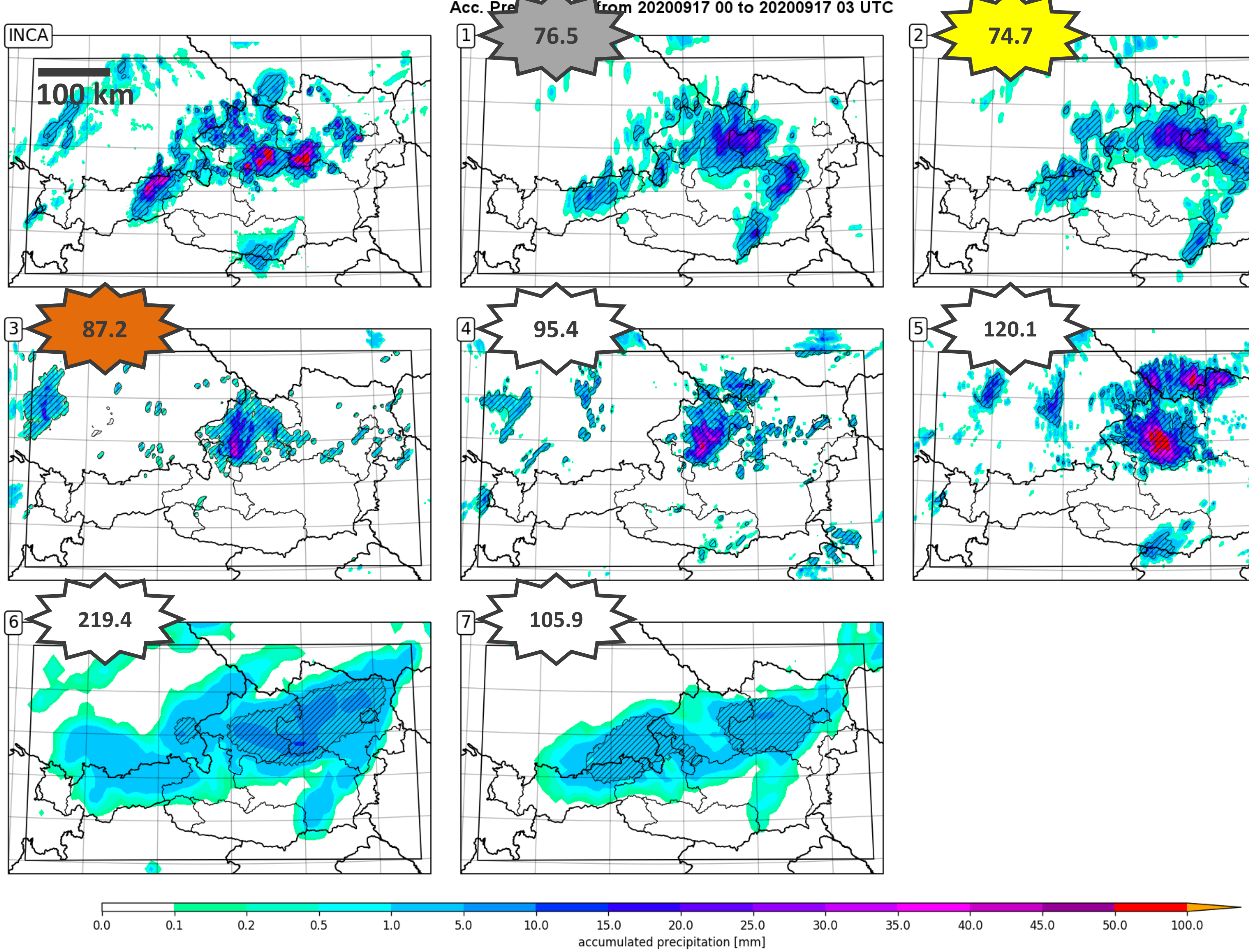


ZAMG
Zentralanstalt für Meteorologie und Geodynamik

Acc. Precip. [mm] from 20200917 00 to 20200917 03 UTC



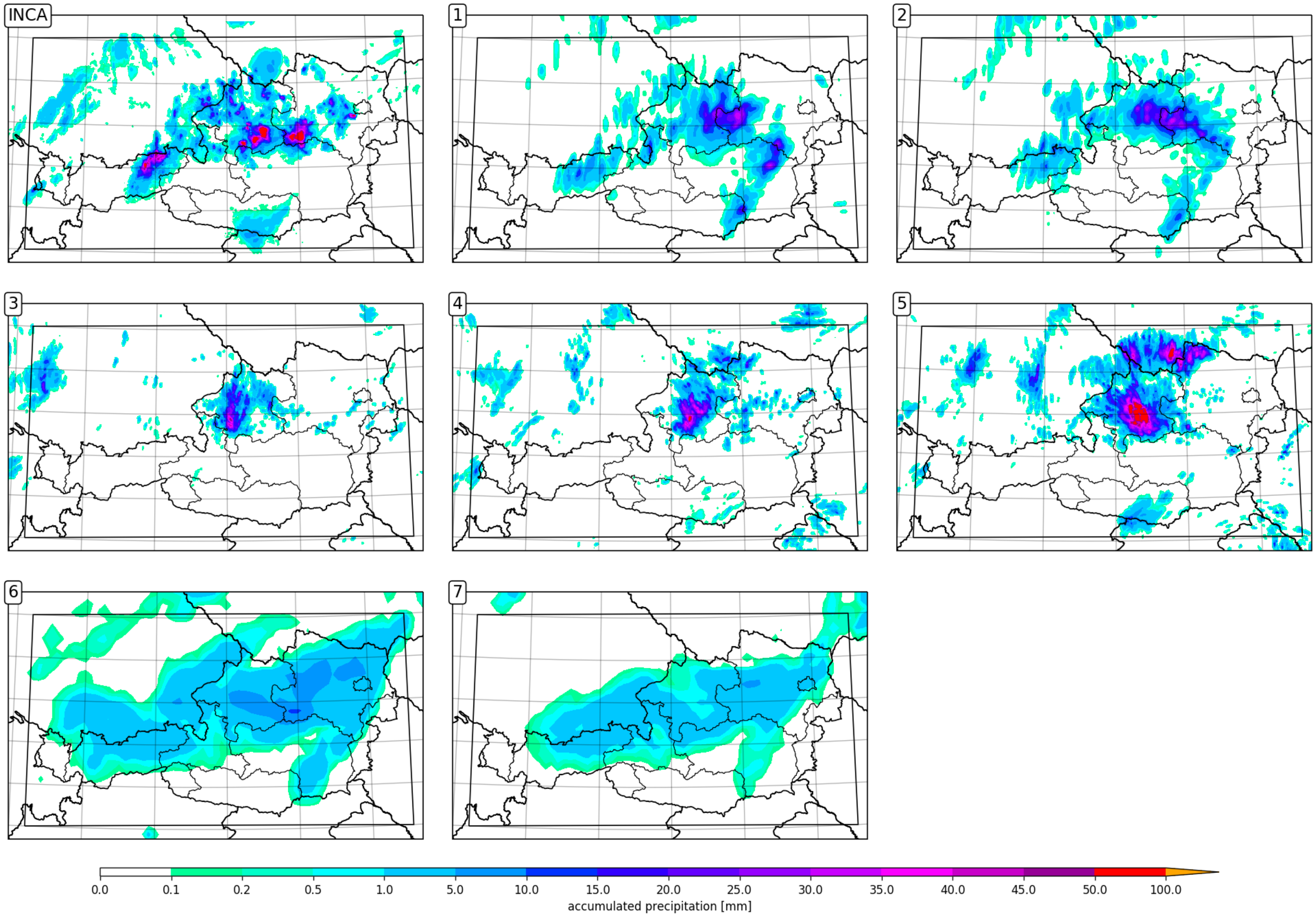
ZAMG
Zentralanstalt für Meteorologie und Geodynamik



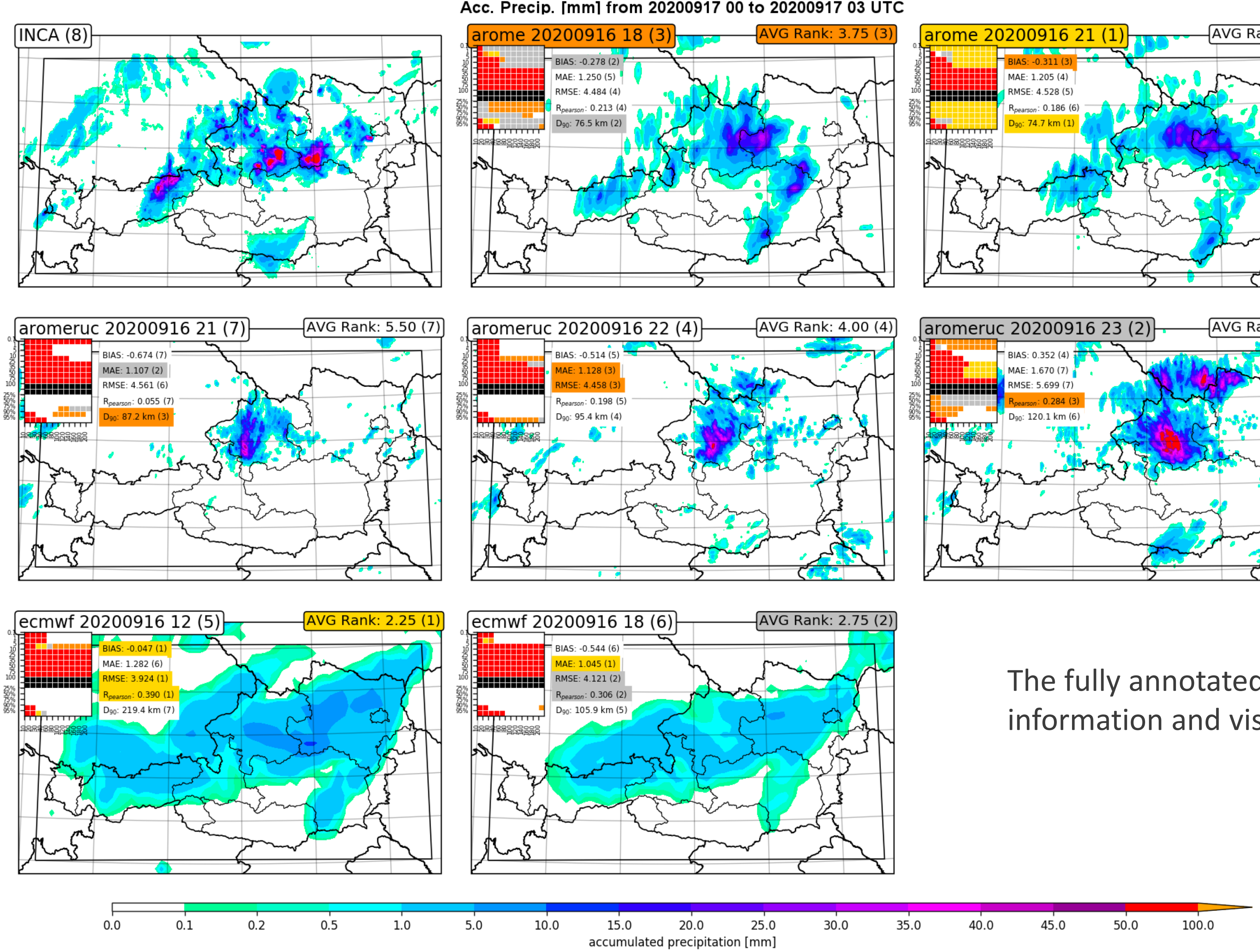
ZAMG

Zentralanstalt für Meteorologie und Geodynamik

Acc. Precip. [mm] from 20200917 00 to 20200917 03 UTC

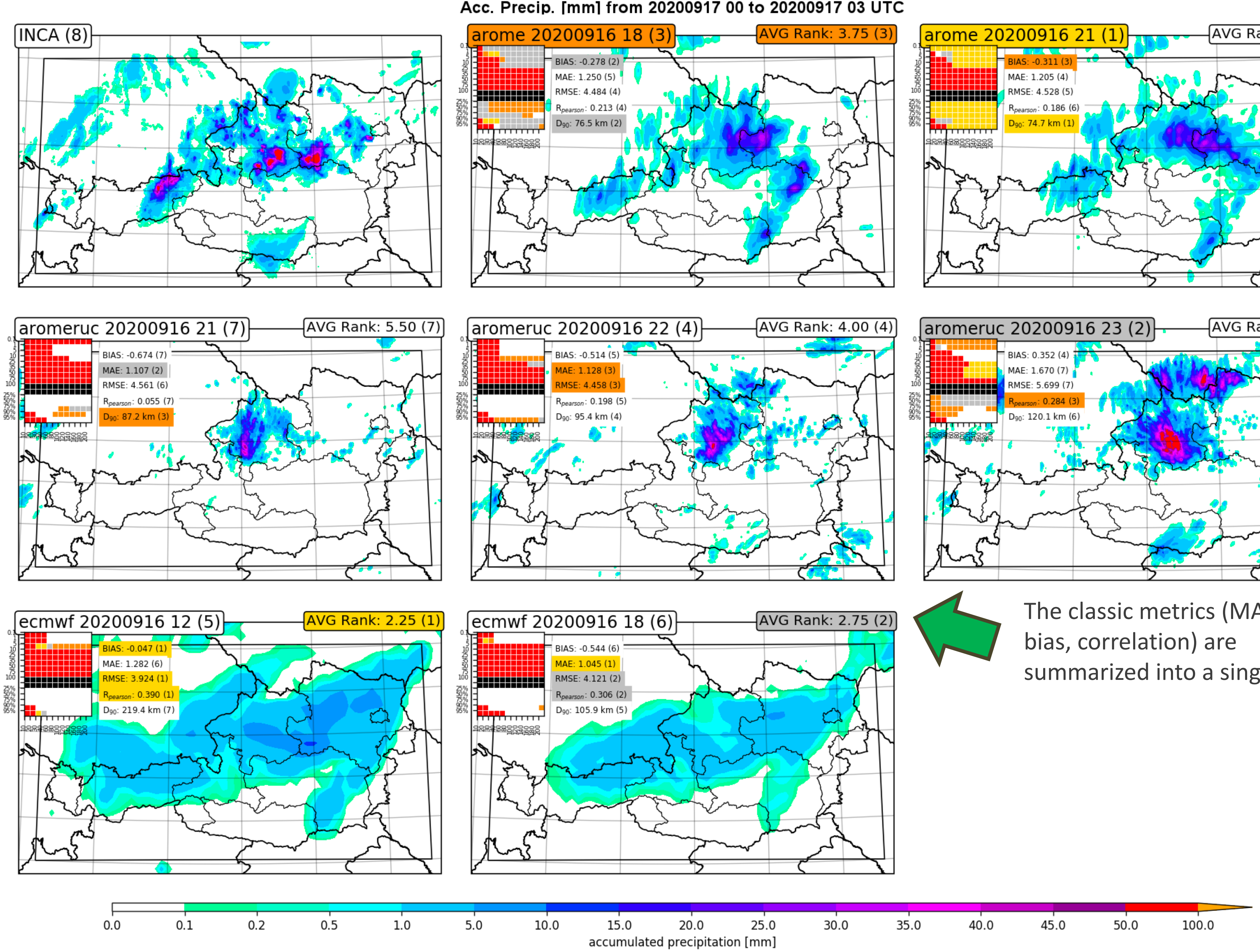


ZAMG Zentralanstalt für Meteorologie und Geodynamik

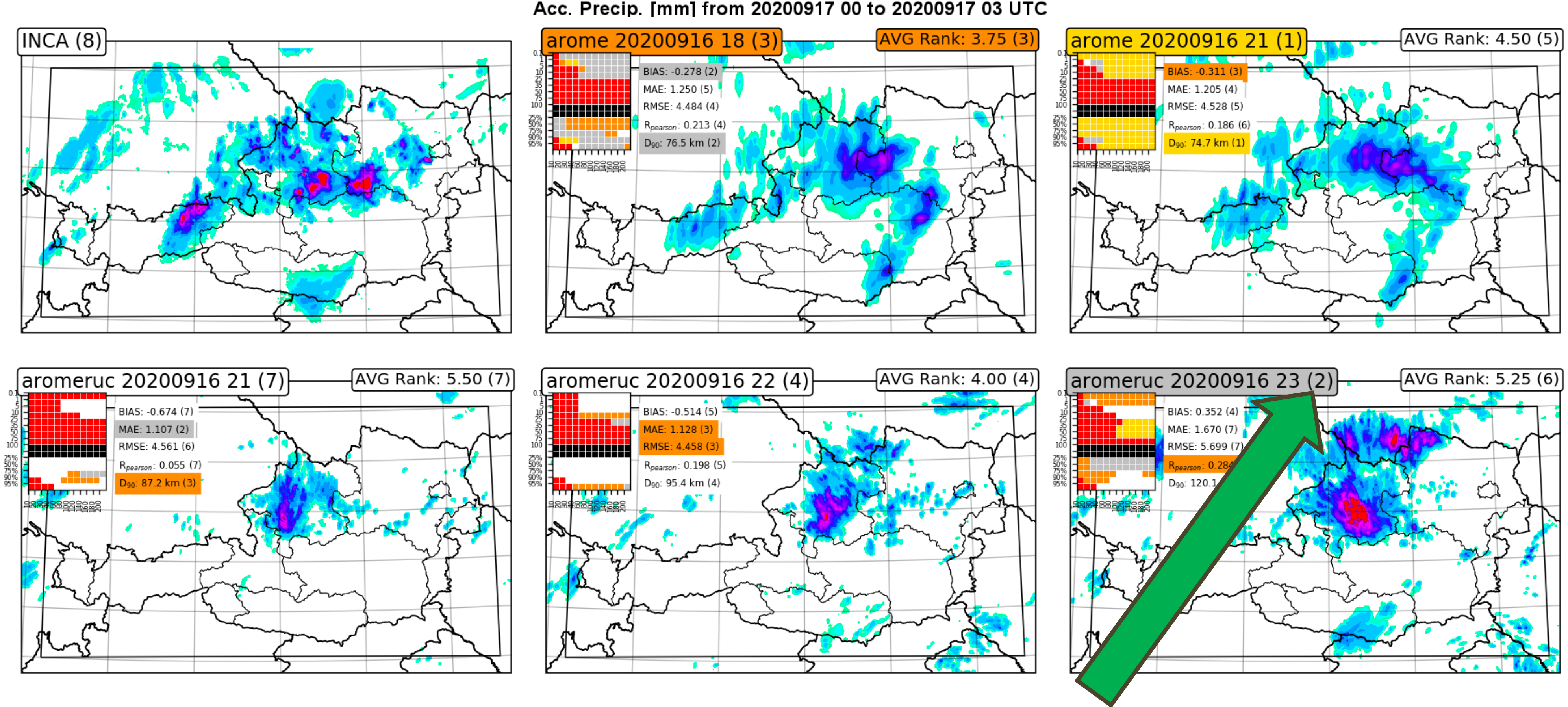


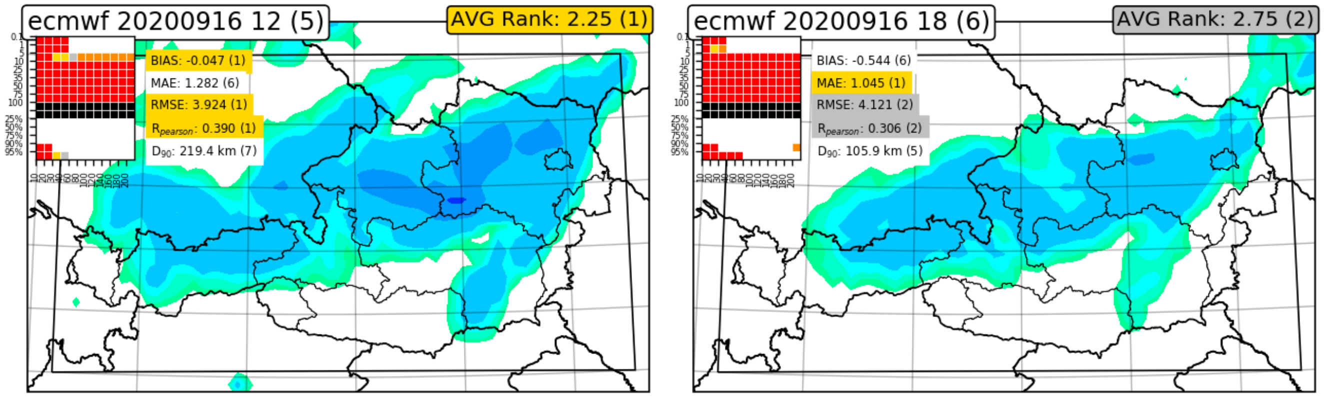
The fully annotated panels contain lots of information and visually aid comparison

ZAMG Zentralanstalt für Meteorologie und Geodynamik



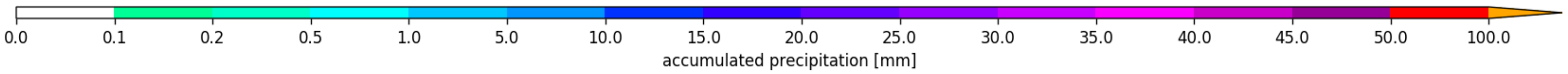
The classic metrics (MAE, RMSE, bias, correlation) are summarized into a single rank





The entire FSS-Matrix is used to calculate a single score, the forecasts are then ranked accordingly.

More detailed information enters this ranking.



ZAMG Zentralanstalt für Meteorologie und Geodynamik

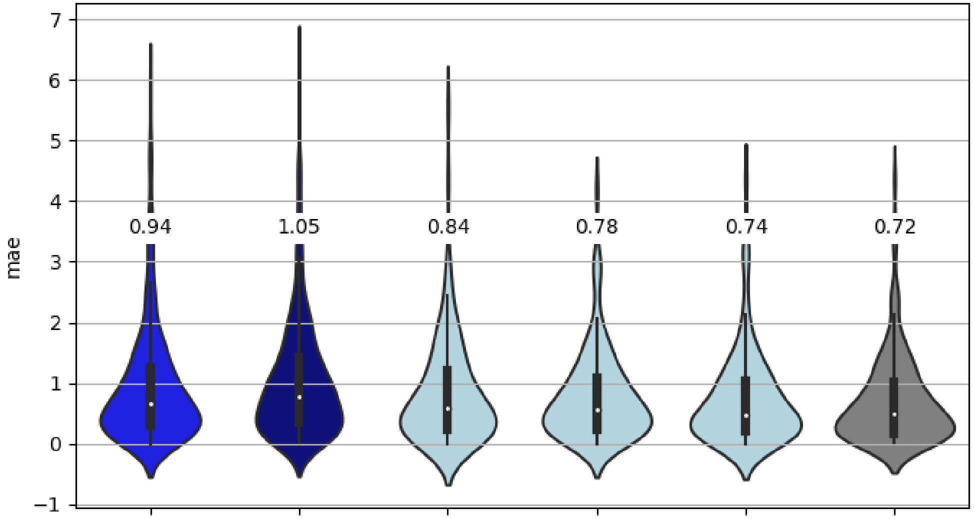
Scores for Summer 2020 (May – August)

- Comparing the hourly forecasts from **15 to 19 UTC (4 hours) for May – August 2020**

- Lead times vary for the different models, to emulate what is available for a nowcast after noon of each day
  - **AROME-Aut** 6 and 9 UTC
  - **AROME-RUC** 9 to 13 UTC
  - **CLAEF** 06 UTC
  - **ECMWF** 00 and 06 UTC

- Archived scores are evaluated (work in progress)

Scores for Summer 2020 (May – August)

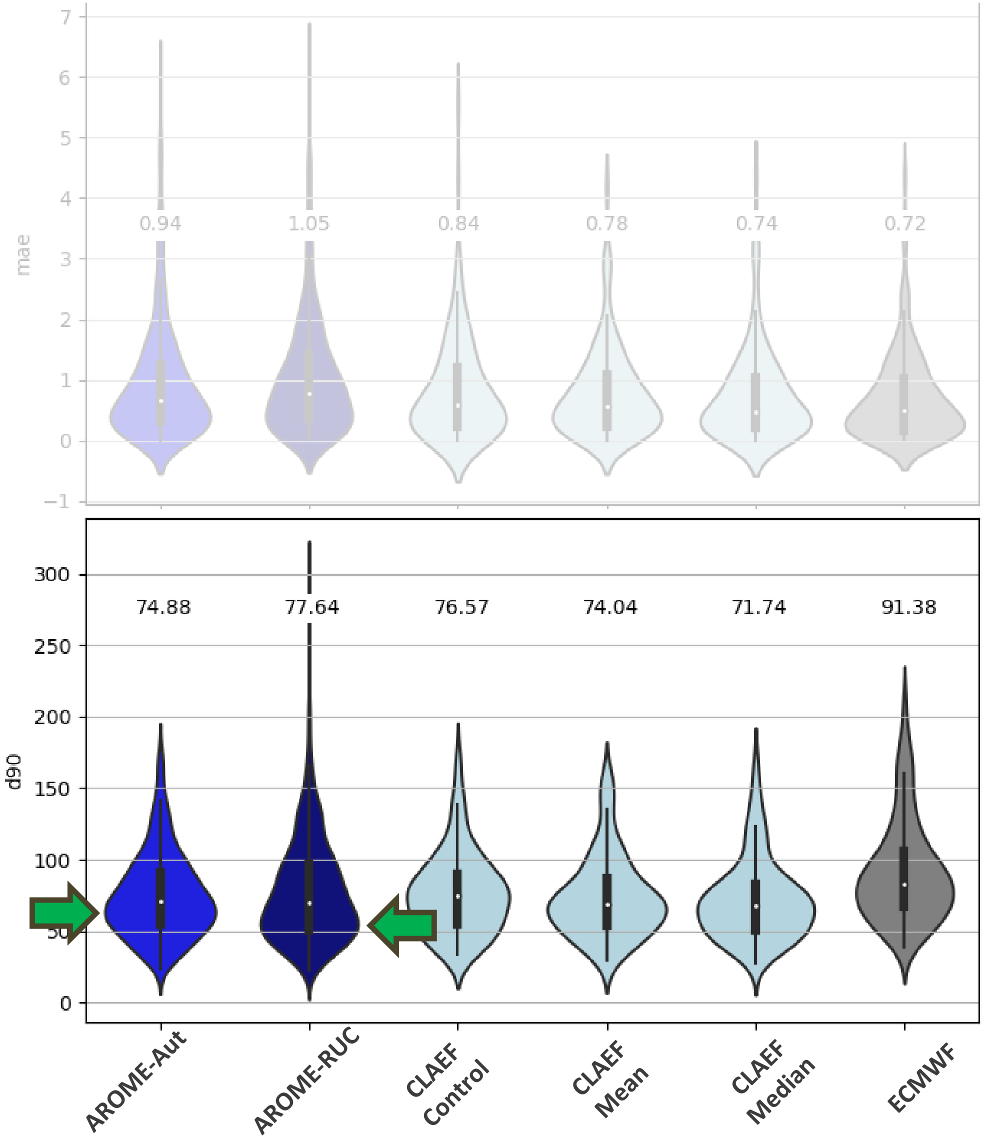


AROME-Aut
AROME-RUC
CLAEF Control
CLAEF Mean
CLAEF Median
ECMWF

MAE

- High resolution deterministic AROME-Aut, AROME-RUC, and CLAEF-Control show the highest MAE
- The ensemble mean and median and the global model perform best

Scores for Summer 2020 (May – August)

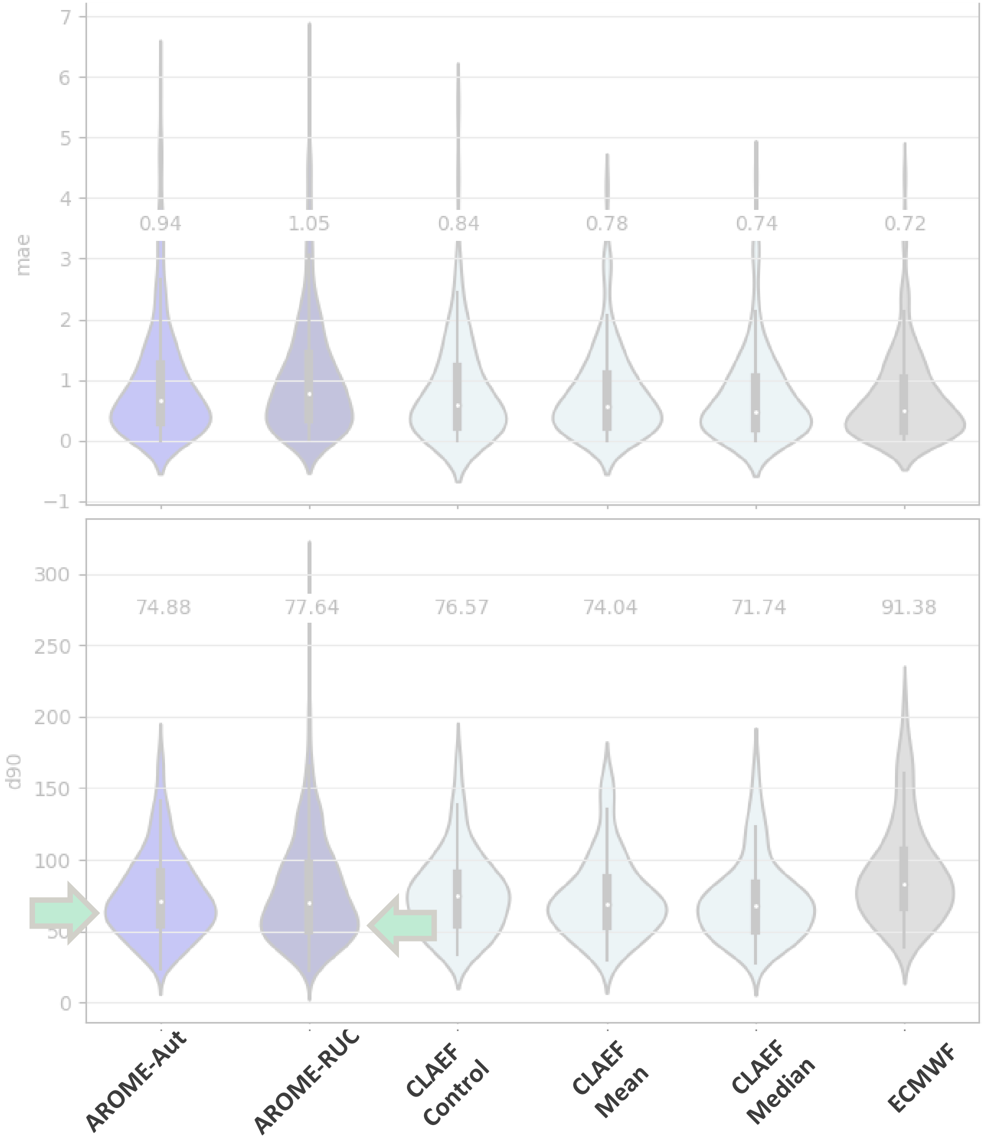


D90

- While AROME-RUC has the highest average, its median is slightly lower
- The convection permitting ensemble CLAEF outperforms both deterministic AROME versions

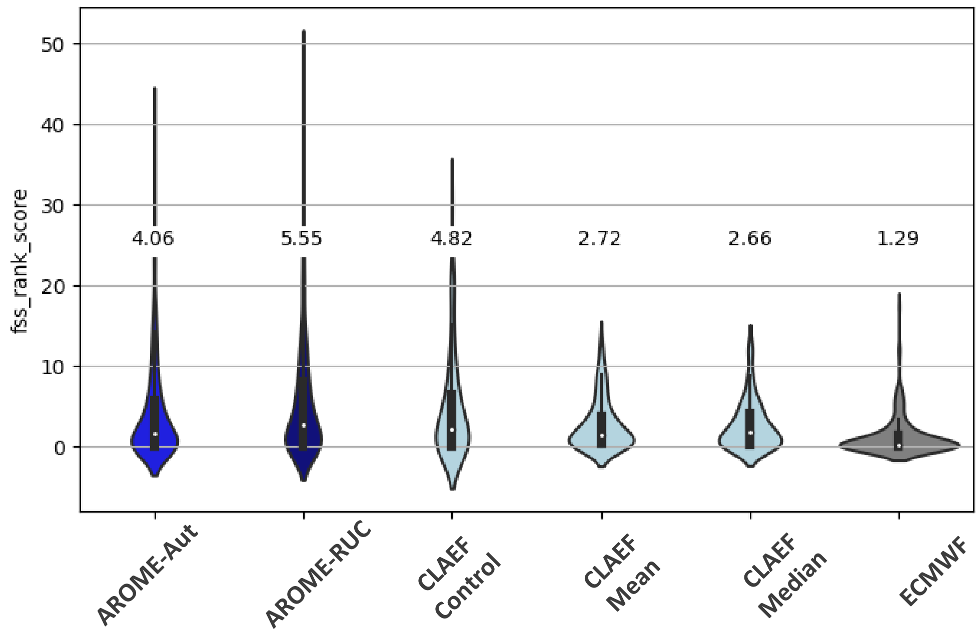
AROME-Aut AROME-RUC CLAEF Control CLAEF Mean CLAEF Median ECMWF

Scores for Summer 2020 (May – August)



FSS Rank Score

- The high resolution deterministic models are best at scoring high FSS values



Scores for Summer 2020 (May – August)

- AROME-RUC shows relatively **high variability** in ist results, AROME-Aut and CLAEF are more consistent

- The CLAEF mean and median and ECMWF IFS produce **low errors due to smoother fields** with less extreme rain, but have problems producing sufficient intensity and localization

- AROME-RUC performs slightly better when taking **spatial scales** and **higher intensities** into account using the Fraction Skill Score

Overview

Motivation and Methodology

- Basic Challenge and Idea
- The Tool: **Panelification**
- Scores and Simplifications

Example Output

- Deeper Look at an Event
- Results for the Summer 2019

⇨ Discussion & Outlook

Discussion

- As a side project, Panelification is growing slowly but steadily
- The ranking is, as of yet, **experimental**.
- Some of the FSS-derived scores might ultimately prove to be of little use
- **Best use as of now:** give modellers a tool to **quickly check on an interesting event**, allowing them to chose which **models, lead times, geographical areas and periods** to verify and, if desired, save some of the data for closer examination

Where do we go from here?

- **Currently:**
  - Panelification **runs daily** with a selection of **forecasts that is available to the forecasters** at ZAMG to compare the ones available **in practice**
  - Used for evaluating case studies

- **Planned:**
  - Deriving single values from the FSS matrices to compare forecasts
  - Test them by comparing the resulting rankings with rankings done by experts

Where do we go from here?

- **Currently:**
  - Panelification **runs daily** with a selection of **forecasts that is available to the forecasters** at ZAMG to compare the ones available **in practice**
  - Used for evaluating case studies

- **Planned:**
  - Deriving single values from the FSS matrices to compare forecasts
  - Test them by comparing the resulting rankings with rankings done by experts

Thank you for your attention!