WG Verification & ~~Applications~~

Overview of activities

Flora Gofa

WG 'Verification & ~~Applications~~' Activities

I. Software: MEC/FFV2, VAST, more….
II. Observation types: Crowd source data/PP-APOCS
III. Verification Caveats
IV. EPS Verification: PPCARMENS
V. Common Verification: Restructure Activity & CPlots
VI. Data driven Models Evaluation

MEC (H. Anlauf)

Felix Fundel
MEC/FFV2 Update

Humidities
- Changed coefficients for the calculation of saturation water vapour (from Magnus-Tetens/Murray (1967) to IFS/BUCK (1981), soon Hardy (1998))
- This effects the conversion of TD to RH from radiosonde and aircraft reports
- Also required an adaption in FFV2

Wind
- predictions are now taken from FF10M instead of the lowest model level
- Model level winds are only fallback if 10m wind is not available
- 10 min. average is preferred over instant wind if both are available
- Backward compatible with `&veri_obs`: `u10_use_mlevel = .true. ! default: .false.`

Global Radiation
- Both short names ASWDIR_S and ASWDIR_S_OS are accepted where ASWDIR_S_OS is preferred if both fields exist (might cause temporal inconsistencies)

AI Models
- Support for AI Models on reduced set of model levels (e.g. AICON)

Treatment of 10m Winds

Old:

- No use of winds above 100m station height in DA
- Therefore those winds had quality flag ‘rejected’ (state=7)
- For verification those winds were used, i.e. all data with state=7 had to be read but only wind data were used
- Own set of quality criteria for wind had to be defined

New:

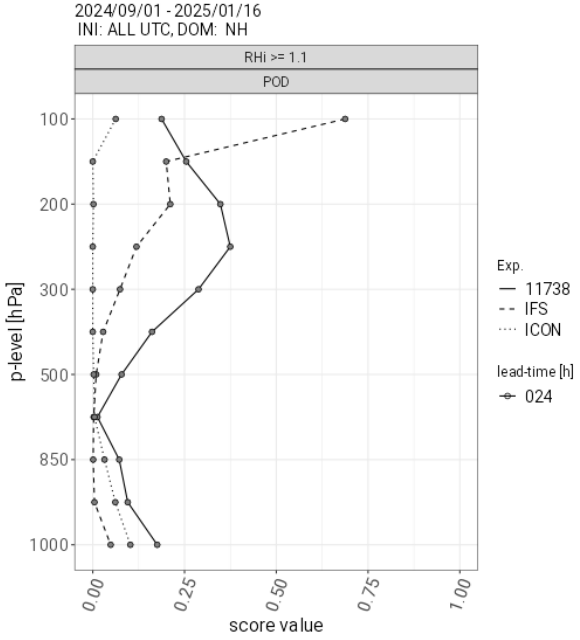
- ICON DA now uses all wind observations
- Verification can rely on DA quality flags now (as for other variables)
- Old behaviour can be forced by namelist entry (useRejectedWinds ‘TRUE’)

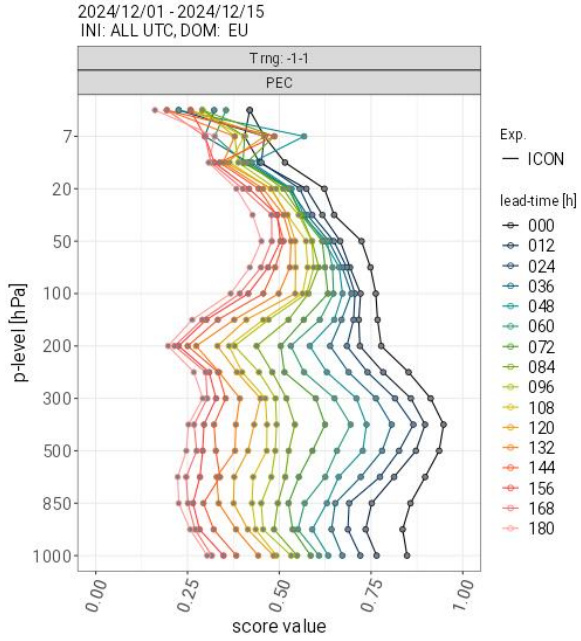
Still some optional filters for wind based on sso_stdh and, hardcoded, for too large wind differences between obs and model (>40m/s) exist.

Felix Fundel
MEC/FFV2 Update

Rel. Humidity over Ice RH_i & Categorical Verification of Vertical Profiles

- RH_i can now be calculated during verification and used in TEMP verification
- Continuous verification of RH_i is mostly not helpful
- Therefor a categorical verification of upper-air observation was implemented
- This allows to verify events, e.g. RH_i >110% (left, POD) or T+-1K (right, hit-rate)
- (Outcome of a project on a 2-moment ice microphysics for ICON)





FFV2

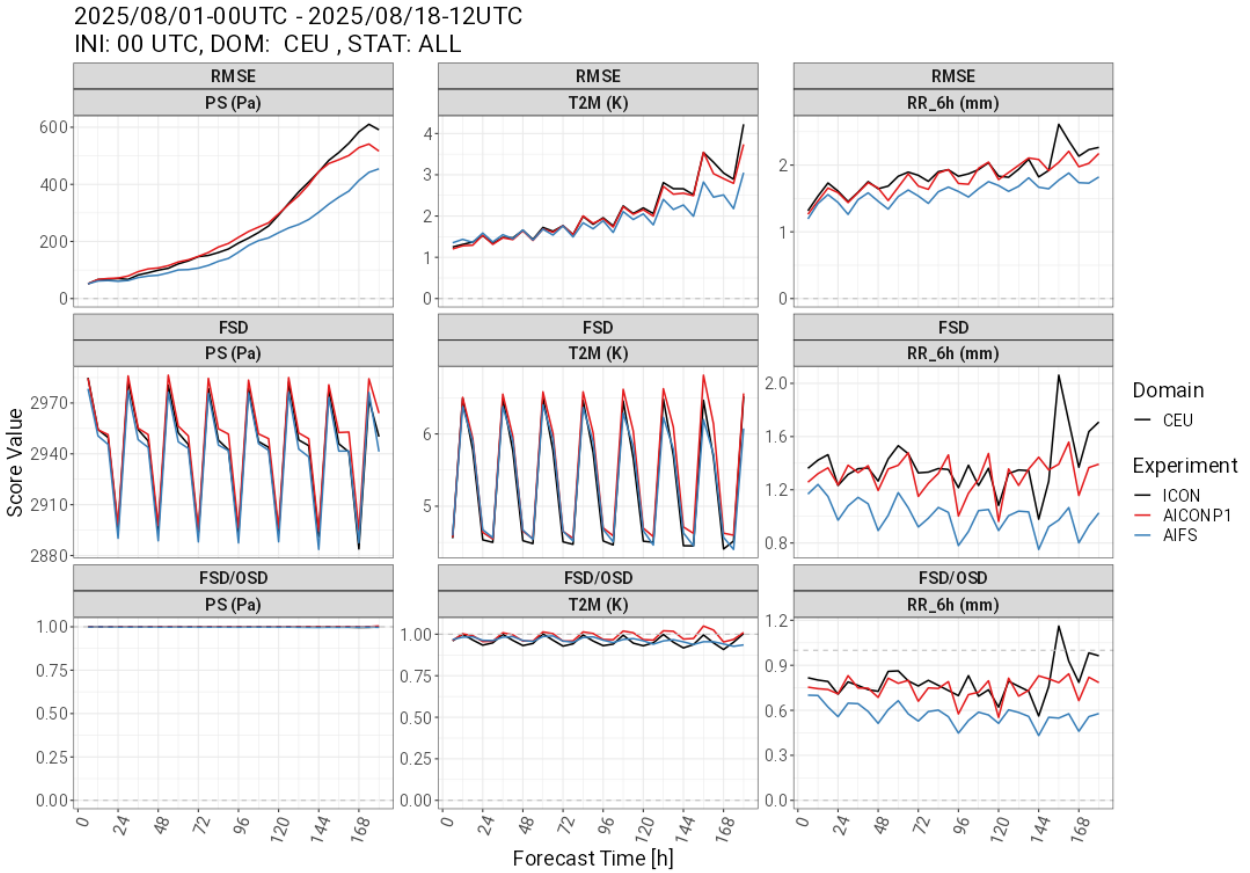
New Scores

Threshold Weighted CRPS

- Proper ensemble score with weight on (rare) events
- Test phase (include with namelist entry *TW 'TRUE'*)
- Unclear added value to e.g. Brier score

Activity

- Quantify forecast and observation activity (spatial variability)
- Smoother fields will be reflected by a reduced std. dev. of the forecasts
- Reduced activity can explain a reduction in RMSE
- Relevant for current AI models that show reduced RMSE due to smooth forecast fields
- Still, a more detailed view on activity is important, e.g. a scale dependent analysis using power-spectra



MEC, FFV2 and Shiny Apps on dkrz DACE gitlab

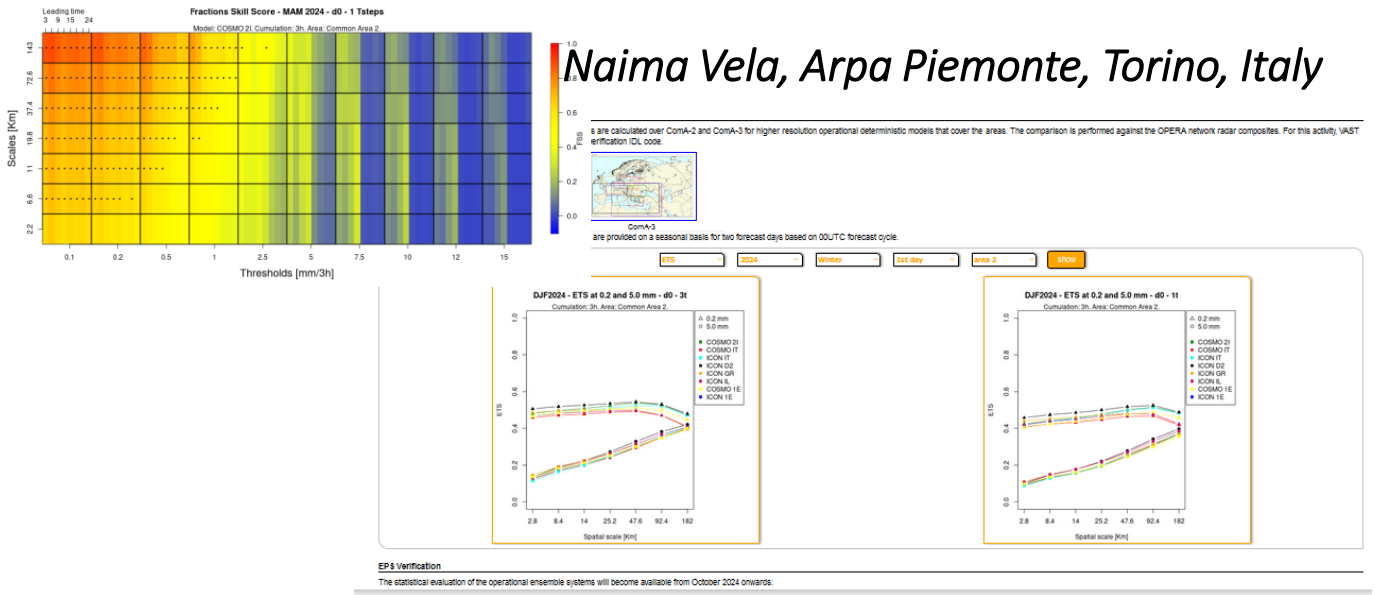
- Discontinuation of code management on my private gitlab
- https://gitlab.dkrz.de/dace/feedback_file_verification
- All COSMO partners should have access

spatial verification approaches

DET models

Various appoaches have been explored through COSMO projects (e.g. PPAWARE), object oriented (SAL, MODE), **Scale-separation (wavelets), neighborhood methods (e.g. FSS)**

For Common Plots: VAST 3.0 (Versus Additional Statistical Techniques) is a COSMO software package composed by Bash, Fortran 90 and R scripts to verify the dichotomic parameters of the ICON-LAM. Radar OPERA composite used for truth.
Non categorical scores are also included (FSS, BSS and ETSr). o



Naima Vela, Arpa Piemonte, Torino, Italy

EPS models

- Ensemble fields are spatially complex → classic scores fail.
- Need metrics that capture **spread, displacement, scale & object uncertainty**.
- Currently: no **common toolkit or agreed set of EPS metrics** in COSMO
- **Adopt tools**: METplus (MODE/MTD/FSS/CRA), R SpatialVx (SAL, ISS).
- **Define core metrics** for EPS probabilities/members.

Task to be included in new PP proposed by WG EPS

PP CARMENS

Project Objectives

- Each participating service to have **set up a web shiny platform where at least one season of scores based on local EPS system outputs, will have been calculated (MEC+FFV2)** and uploaded. - All the applications of the system, mec, ffv2, shiny currently have the ability required to fulfil the project.
- An EPS dedicated shiny platform on COSMO web that the stats (rdata) files from each EPS system plus COSMO-LEPS will be uploaded at least for a season following all the prerequisites of CP verification.
- After the completion of the PP this infrastructure will be used for COMMON PLOTS activity for following year.

PP-Leader Ioan-Stefan GABRIAN

with contributions from F. Fundel, F. Gofa, D. Boucouvala, T. Andreadis (HNMS), F. Batignani (CoMET), P. Khain, (IMS), A. Pauling, Pirmin Kaufmann (MCH), S. Dinicila (NMA) , Enrico Minguzzi(Arpae), Thomas

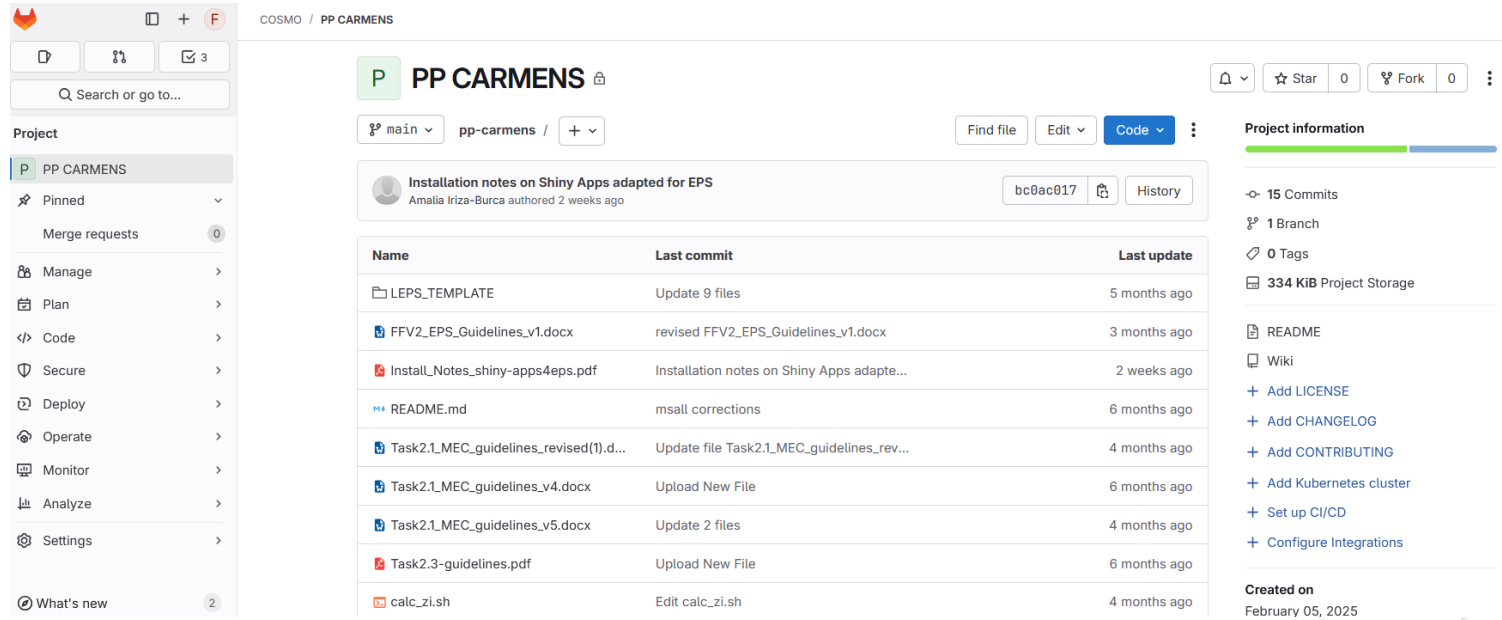
Task 2	MEC and Rfdbk system adaptations for EPS systems	Status	Start	End
2.1	MEC instructions for Feedback Files based on EPS model output	**Finalized: 05.2024**	2022-09	**2024-02**
	• IFS ENS cannot be processed by MEC			
2.2	MEC instructions of Rfdbk for ENS production and probabilistic scores	**Finalized: 07.2024**	2022-09	**2024-02**
2.3	EPS evaluation guidelines as part of Common Plot activity	**to be closed at a later date**	2022-09	**2024-02**
	• *Draft of the deliverable available*			
2.4	Scripts to produce verification scores for EPS forecasts	**Finalized: 07.2024**	2022-09	**2024-02**
	• Script examples/templates to produce the Scorefiles from model output			

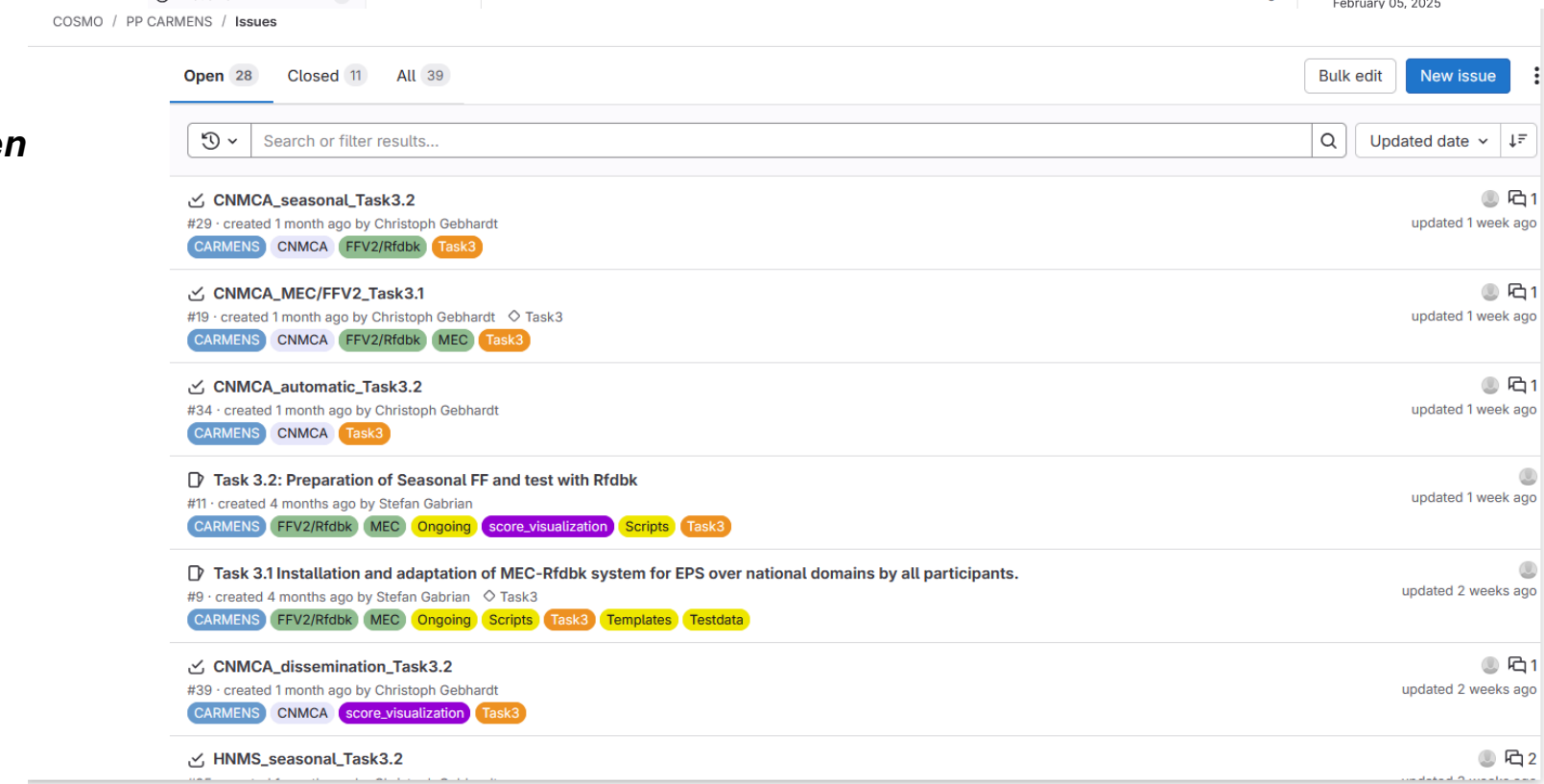
COSMO PP CARMENS Task 2 Deliverables available at:
https://www.cosmo-model.org/view/repository/wg5/PP-CARMENS/Deliverables

Task 3	Guidelines and templates for MEC/ FFV2/Shiny Apps for ENS – available on GitLab.dkrz repository DWD - everything done	End
3.1	IMS MEC/FFV2 and FF & scores produced (at least the test) Local Shiny adaptation - unknown at the moment	024-02
3.2	HNMS & NMA Workflow & Tests / Some issues with missing fields – solutions provided, TBD shortly	024-02
3.3	MCH Workflow & Tests / Scores for TEMP are produced with FFV2 and visualized with Shiny Tackling cumulated field issues in order to finish production for surface parameters	024-02
3.4	CNMCA & IMGW Adaptation of the scripts and namelist for MEC done / Work in progress	024-02
	ARPAE Activity just started	

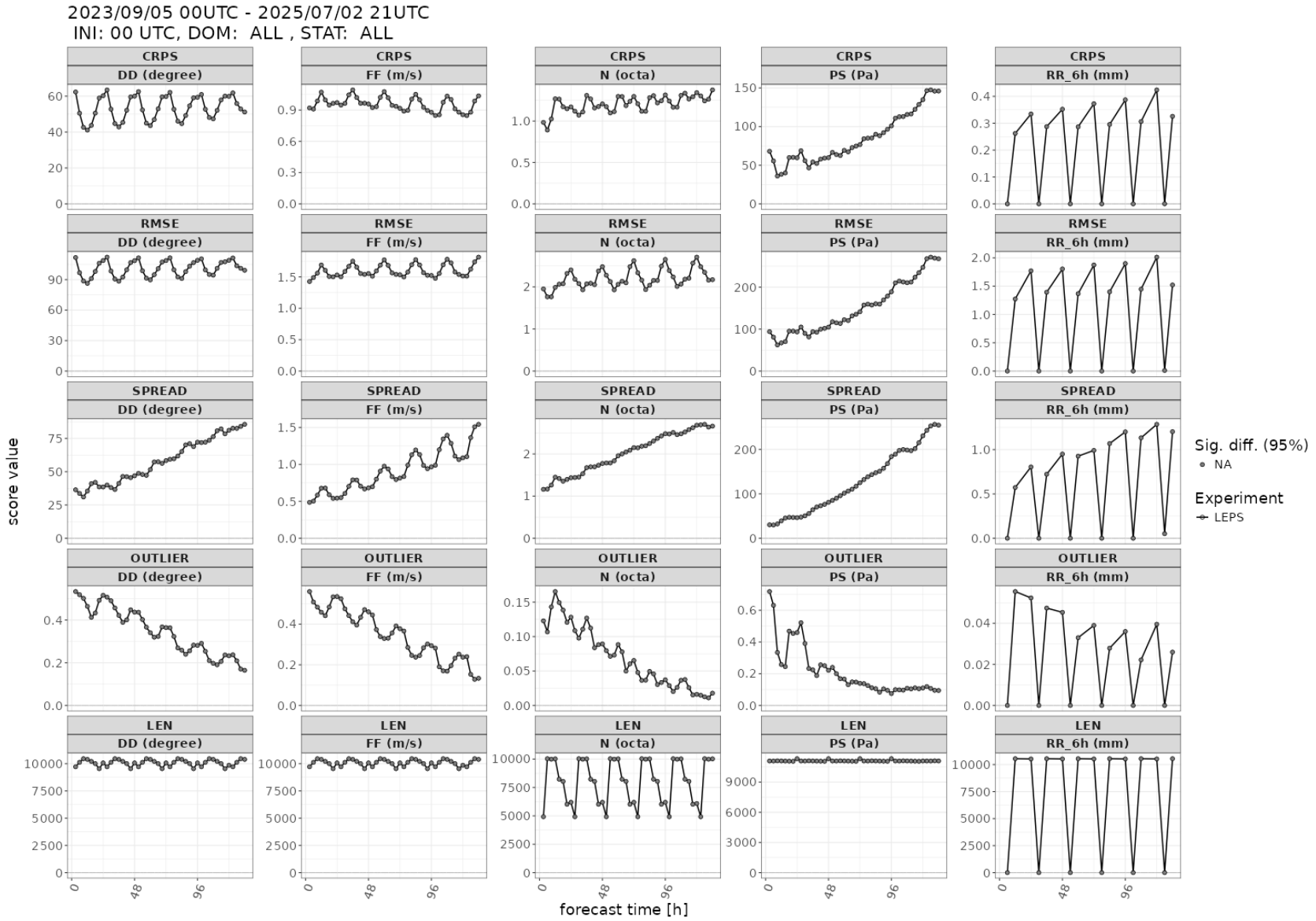
PPCARMENS Management and problem solving

https://gitlab.dkrz.de/cosmo/pp-carmen





S. Gabrian, NMA
PP-CARMENS



COSMO-LEPS

III. Verification Caveats

WMO-No. 8

Guide to Instruments and Methods of Observation (WMO-No. 8)

Work = Force · displacement

5.8.2 Peak gusts and standard deviation

Before specifying the appropriate response characteristics of wind-measuring systems, it is necessary to define the gust duration as required by the application. Wind extremes are mainly used for warning purposes and for the climatology of extreme loads on buildings, constructions and aircraft. It is important to realize that the shortest gusts have neither the time nor the horizontal extent to exert their full damaging effect on large constructions. WMO (1987) concludes that a gust duration of about 3 s accommodates most potential users. Gusts that persist for about 3 s correspond to a "wind run" (duration multiplied by the average wind speed) of the order of 50 to 100 m in strong wind conditions. This is sufficient to engulf structures of ordinary suburban/urban size and to expose them to the full load of a potentially damaging gust.

——, 1987: *The Measurement of Gustiness at Routine Wind Stations: A Review* (A.C.M. Beljaars). Instruments and Observing Methods Report No. 31. Geneva.

However …

Country	NWS	Gust Definition
Germany	DWD	3 s (to be confirmed)
France	MeteoFrance	3 s (to be confirmed)
Austria	GeoSphere	2 s
Italy	Meteo Alto Adige	3 s (except mountain peaks: 1 s)
Italy	ARPA Lombardia, ARPA Piemonte	5 s
Italy	Centro Funzionale Valle d'Aosta	1 s
Switzerland	MeteoSwiss	1 s

Measured Frequencies of Occurrence

All Swiss Stations; Winter (DJF) 2023/24, Summer (JJA) 2024, Winter 2024/25

Score	Season	3s Gusts (WMO)	1s Gusts (MCH)	Relative Change 3s → 1s
Frequency of medium gusts ≥ 12.5 m/s (≈ 25 kn)	Wi 23/24	9.9%	11.0%	11% increase
	Su 2024	2.90%	3.50%	21% increase
	Wi 24/25	7.6%	8.4%	11% increase
Frequency of strong gusts ≥ 20 m/s (≈ 70 km/h)	Wi 23/24	1.88%	2.27%	21% increase
	Su 2024	0.227%	0.295%	30% increase
	Wi 24/25	1.36%	1.65%	21% increase
Frequency of very strong gusts ≥ 25 m/s (= 90 km/h)	Wi 23/24	0.612%	0.769%	26% increase
	Su 2024	0.058%	0.072%	24% increase
	Wi 24/25	0.367%	0.493%	34% increase

Pirmin.Kaufmann@MeteoSwiss.ch

Freq. Strong Gusts

Diurnal Cycle

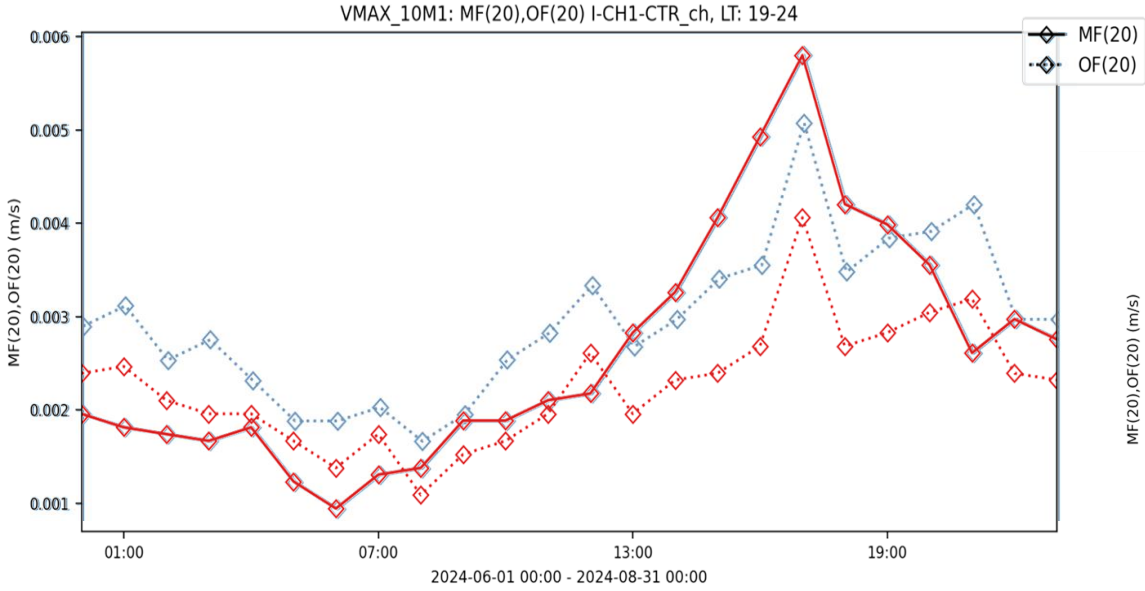
Gusts > 20 m/s (≈ 70 km/h)

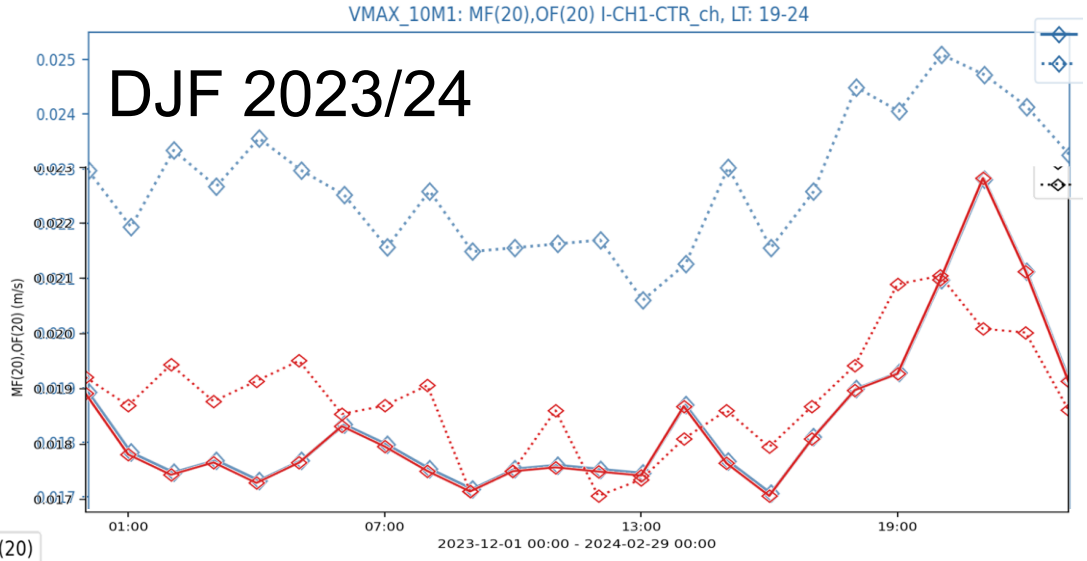
◇···◇ Obs: 1s gusts

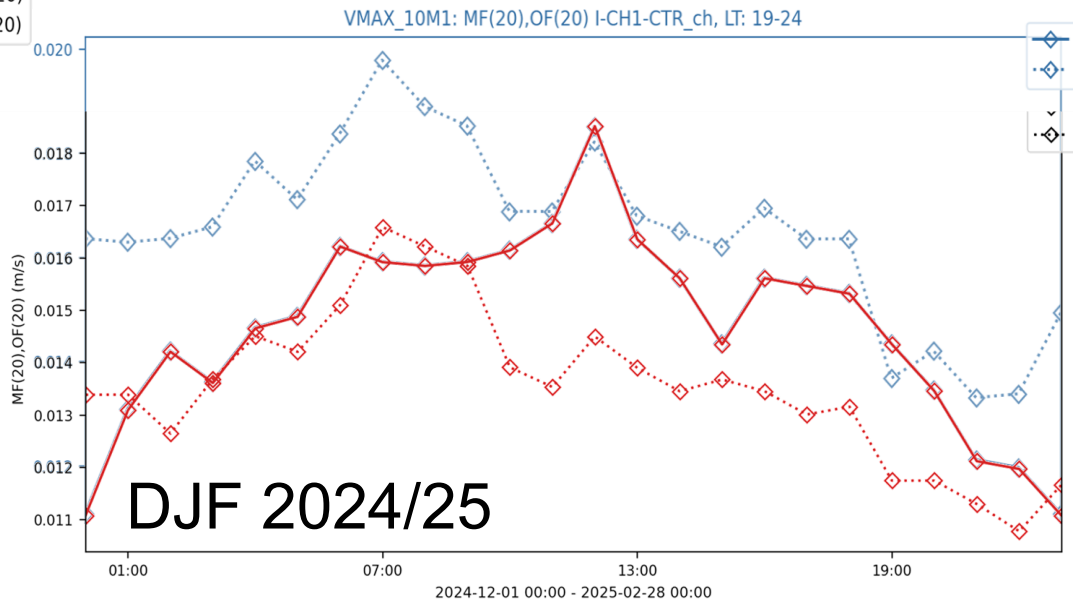
◇—◇ ICON-CH1-EPS

◇···◇ Obs: 3s gusts

JJA 2024





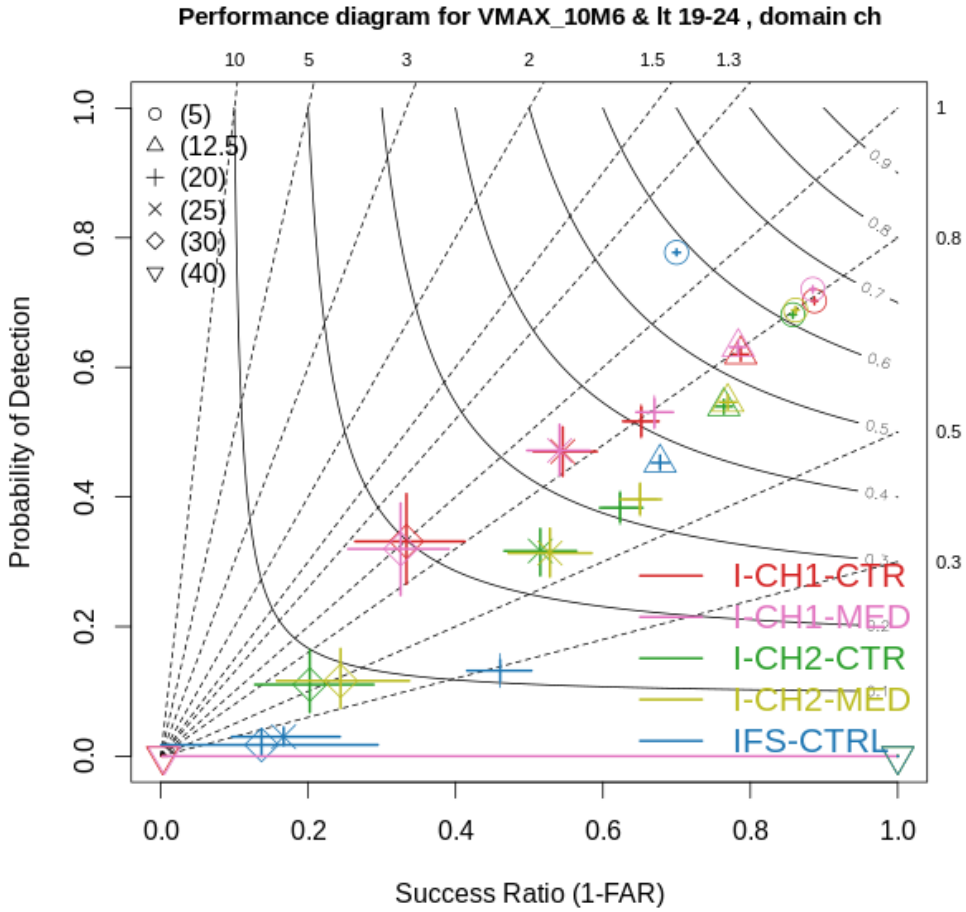


Pirmin.Kaufmann@MeteoSwiss.ch

Model Performance 1 s Gusts 6 h Max.

Winter 2024/25

- Weak gusts:
  (seemingly) underestimated by ICON models
  (overestimated by IFS)
- Very strong gusts:
  about ok in ICON-CH1
  underestimated in ICON-CH2
  (missing in IFS)



Pirmin.Kaufmann@MeteoSwiss.ch

Change in ICON-CH1-EPS Scores

Score	Sea.	Change 1s → 3s Gusts	Relative Change
Frequency Bias ≥12.5 m/s	DJF JJA DJF	0.87 → 0.96 0.97 → 1.17 0.88 → 0.98	Decrease of underestimation (9% less) Overestimation appears (17%) Underestimation vanishes (10% less)
Frequency Bias ≥20 m/s	DJF JJA DJF	0.81 → 0.98 0.89 → 1.16 0.89 → 1.08	Decrease of underestimation (17% less) Change from under- to overestimation Change from under- to overestimation
Frequency Bias ≥25 m/s	DJF JJA DJF	0.90 → 1.13 1.06 → 1.34 1.03 → 1.38	Change from under- to overestimation Increase of overestimation (28% more) Increase of overestimation (35% more)
Eq. Threat Score ≥12.5 m/s	DJF JJA DJF	0.44 → 0.44 0.25 → 0.24 0.44 → 0.44	No change (similar for other thresholds and for Threat Score)

Pirmin.Kaufmann@MeteoSwiss.ch

II. Observation types: Crowd source data/PP-APOCS

Common Plots Activity Restructure

The activity concerns the calculation and representation of verification results using the operational ICON-LAM implementations in each service, both det and eps with MEC-FFV2 platform

Reasoning:

- Common Plot verification activity has steadily expanded, both in terms of scope and complexity.
- Integration of deterministic (limited area and global), ensemble models and additional statistical requirements has increased the operational workload
- Important to reformulate and reassign roles to ensure the sustainable continuation and quality of the CP activity.

Structure & Responsibilities

1. MEC-FFV2 Seasonal Management (S, Gabrian, F. Fundel, S. Denicila

A. Align FF files from various models and run FFV2. Upload correctly named Rdata files (nomenclature description file available) on shiny COSMO web pages for all appropriate Common Areas and for different statistical outputs (continuous, categorical, by station, upper air, etc.), **for det and eps models**

B. Update the system with new FFV2 statistical package versions

Incorporate new features in CP activity: e.g. 6h precipitation in all timesteps, additional conditional verification tests, new scores, etc.

2. Observation retrieval and adaptation (M.Bogdan, F. Gofa)

Retrieve observations (SYNOP, TEMP) from MARS convert to netcdf, upload on common server. Retrieve hdf5 precipitation files from OPERA. Convert and Upload on common server.

3. Preparation of FF for "Common models"- IFS (F. Fundel)

IFS: Retrieve models output with required fields from det global model. Run MEC and create FFs for all seasons, upload them on common server

4. Preparation of FF for "Common models"- ICON-Gb (F. Fundel)

ICON: Retrieve models output with required fields. Run MEC and produce FFs for all seasons, upload them on common server

Structure & Responsibilities

5. Preparation FF/Rdata files for COSMO/ICON-LEPS (E. Minguzzi, F,Fundel?)
COSMO/ICON-LEPS: Get full member model output (ECMWF hpc) and run MEC for FF productiion. Run FFV2 for seasonal scores. Transfer to COSMO shiny server

6. Spatial Verification, VAST (N. Vela)
Collect and adapt input GRIB1/2 Precipitation files for **VAST2.0**
Run VAST for FSS/POD/FAR/FBI/TS scores for a number of spatial windows
Prepare seasonal/annual summary (for COSMO-GM presentation and newsletter section).
Update and revise VAST software, update manual

7A. Seasonal scores: Det models: ***M. Bogdan (NMA), F. Fundel (DWD), D.Boucouvala (HNMS), N. Zaccariello and F. Batignani (CNMCA), A. Surowiecki (IMGW-PIB), A. Pauling (MCH), P. Khain (IMS), T. Gastaldo (ARPAE).***
Produce FF (MEC) for the complete simulation domain for deterministic models and cntr EPS
Upload them on *common server*.. Extract GRIB precipitation files and upload them common FTP server. Provide information on operational model version/changes of fcst data. Present verification analysis over national domains during Model Errors annual workshop and GM

7B. Seasonal scores: EPS models ***F. Fundel (DWD), P. Khain (IMS), A. Pauling (MCH),.....***
Produce FF (MEC) for the complete simulation domain for EPS limited area models.
Run FFV2 for ensembles.. Upload Rdata files on COSMO shiny server.. Provide information on operational model version/changes of fcst data

Structure & Responsibilities

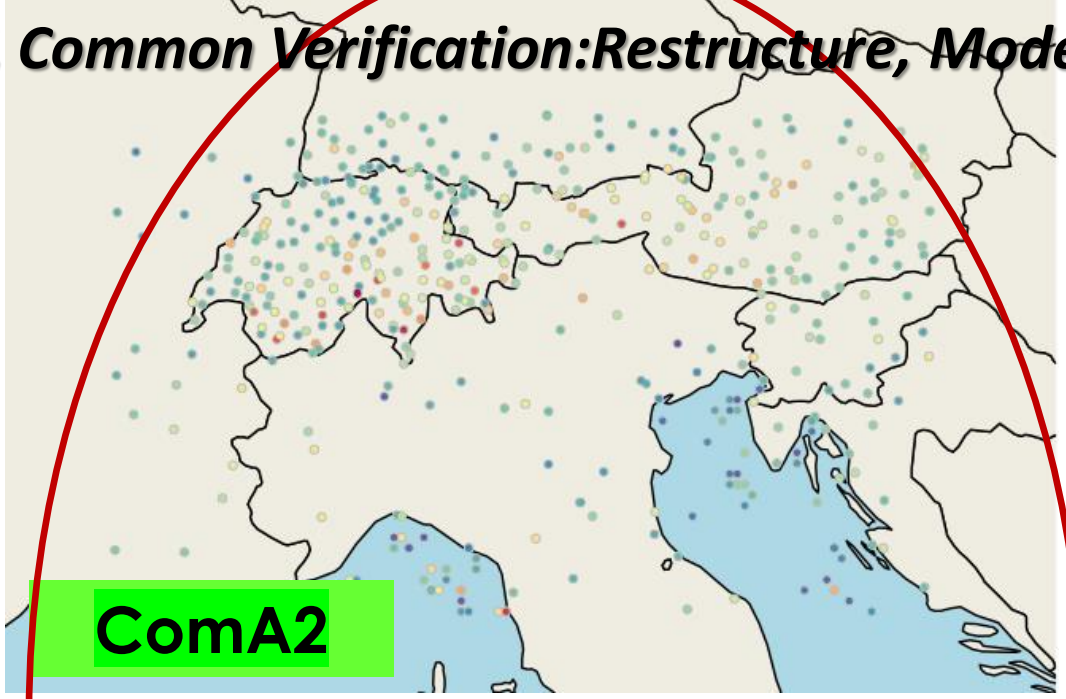
8. Reporting (WC V/A chairs)

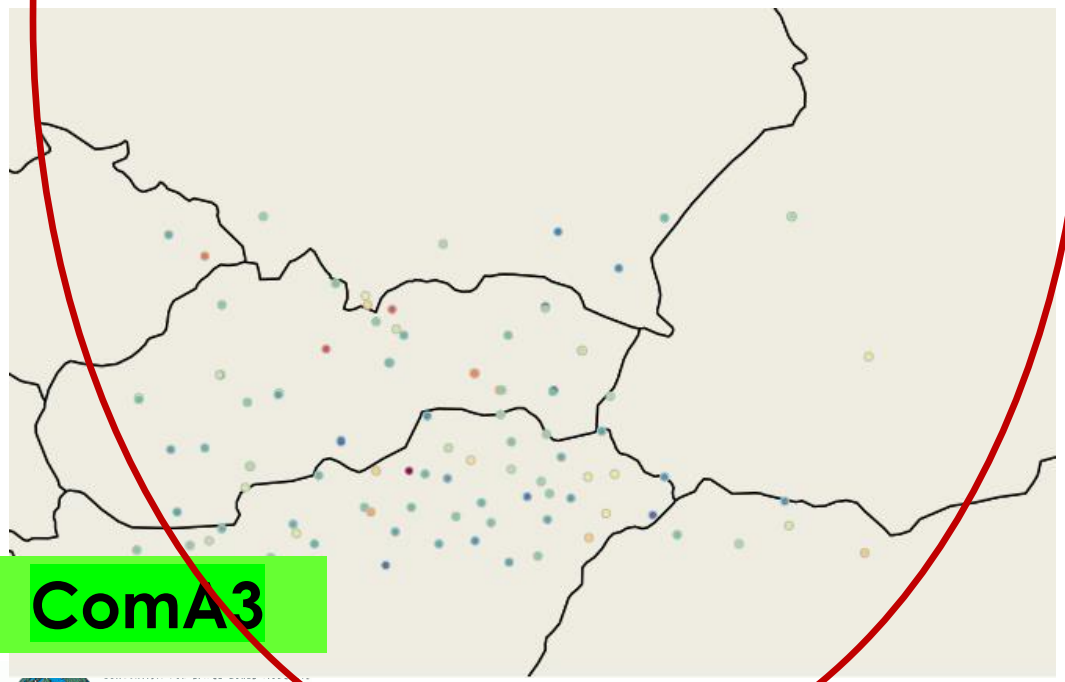
Preparation of annual guidelines for CP activity. Analysis and presentation of main verification outcome. Presentation during GM. Organization of annual WG meeting on Model errors. Newsletter contribution with annual highlights of CP activity.

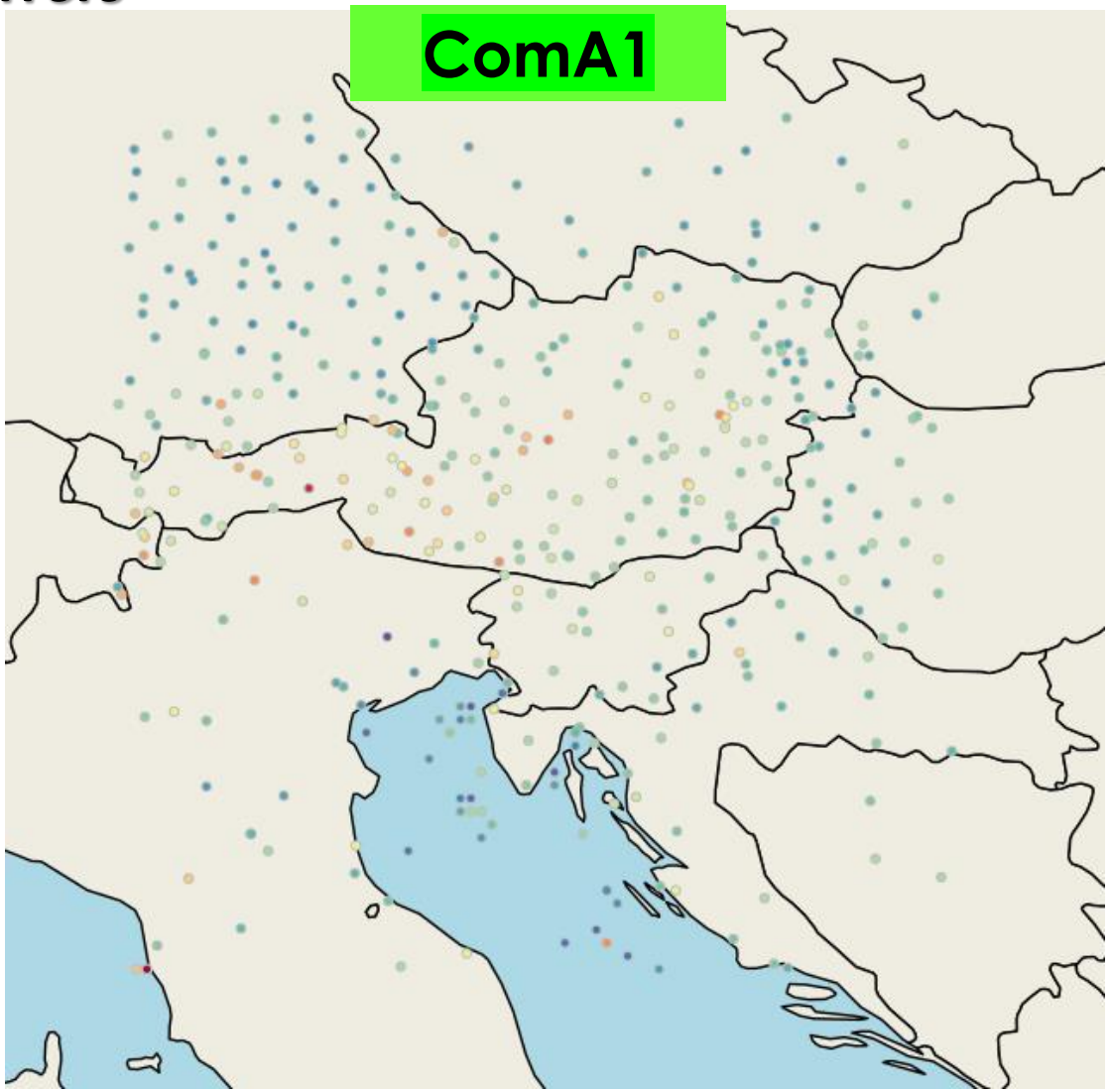
Total FTEs for Common Plots Activity for 2025/2026: XXX FTEs

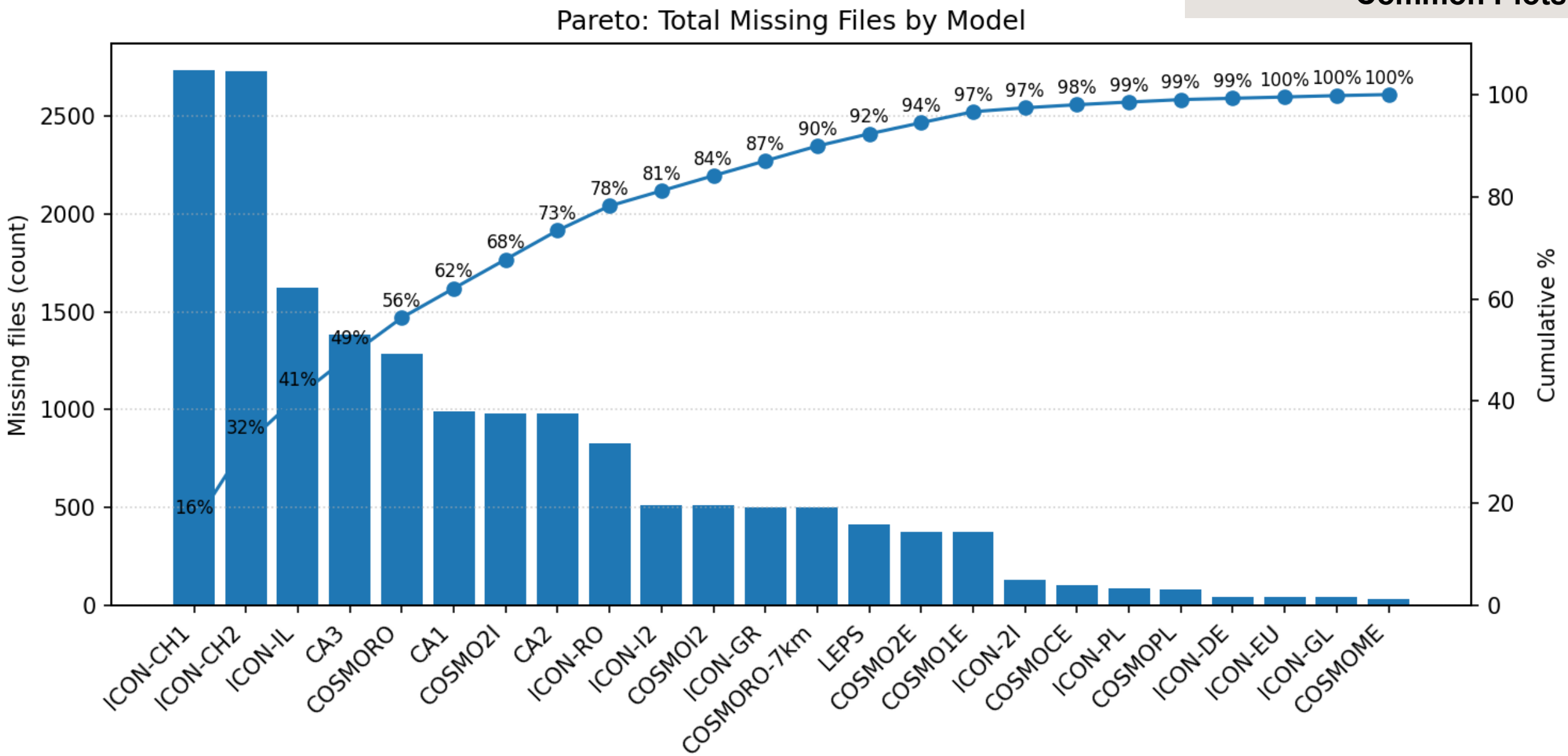
Currently: 0.91FTEs

V. Common Verification:Restructure, Model Errors





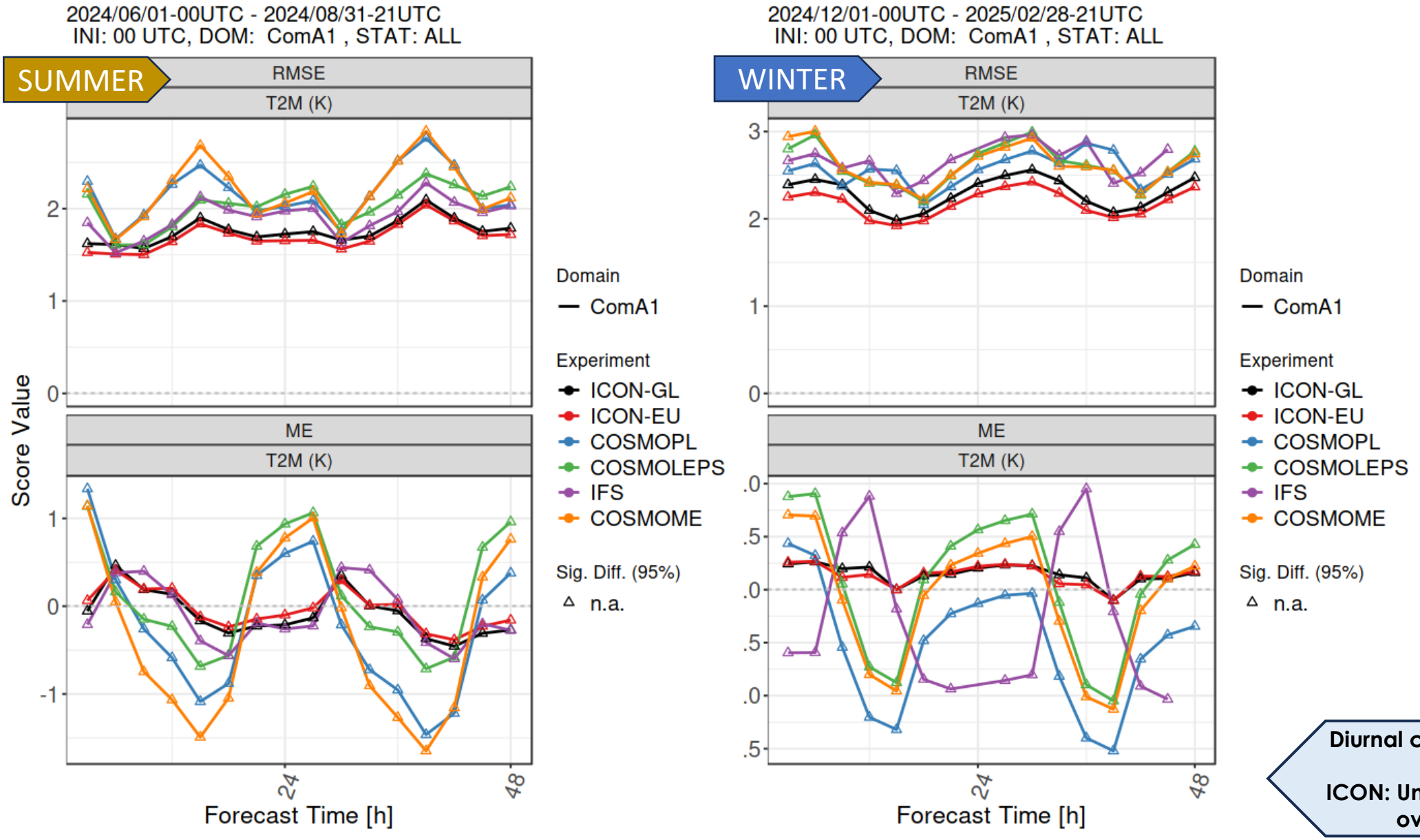



Pareto: Total Missing Files by Model
Missing files (count)
Cumulative %
0
500
1000
1500
2000
2500
0
20
40
60
80
100
16%
32%
41%
49%
56%
62%
68%
73%
78%
81%
84%
87%
90%
92%
94%
97%
97%
98%
99%
99%
99%
100%
100%
100%
ICON-CH1
ICON-CH2
ICON-IL
CA3
COSMORO
CA1
COSMO2I
CA2
ICON-RO
ICON-I2
COSMOI2
ICON-GR
COSMORO-7km
LEPS
COSMO2E
COSMO1E
ICON-2I
COSMOCE
ICON-PL
COSMOPL
ICON-DE
ICON-EU
ICON-GL
COSMOME

Driving Models

ComA1

2mT

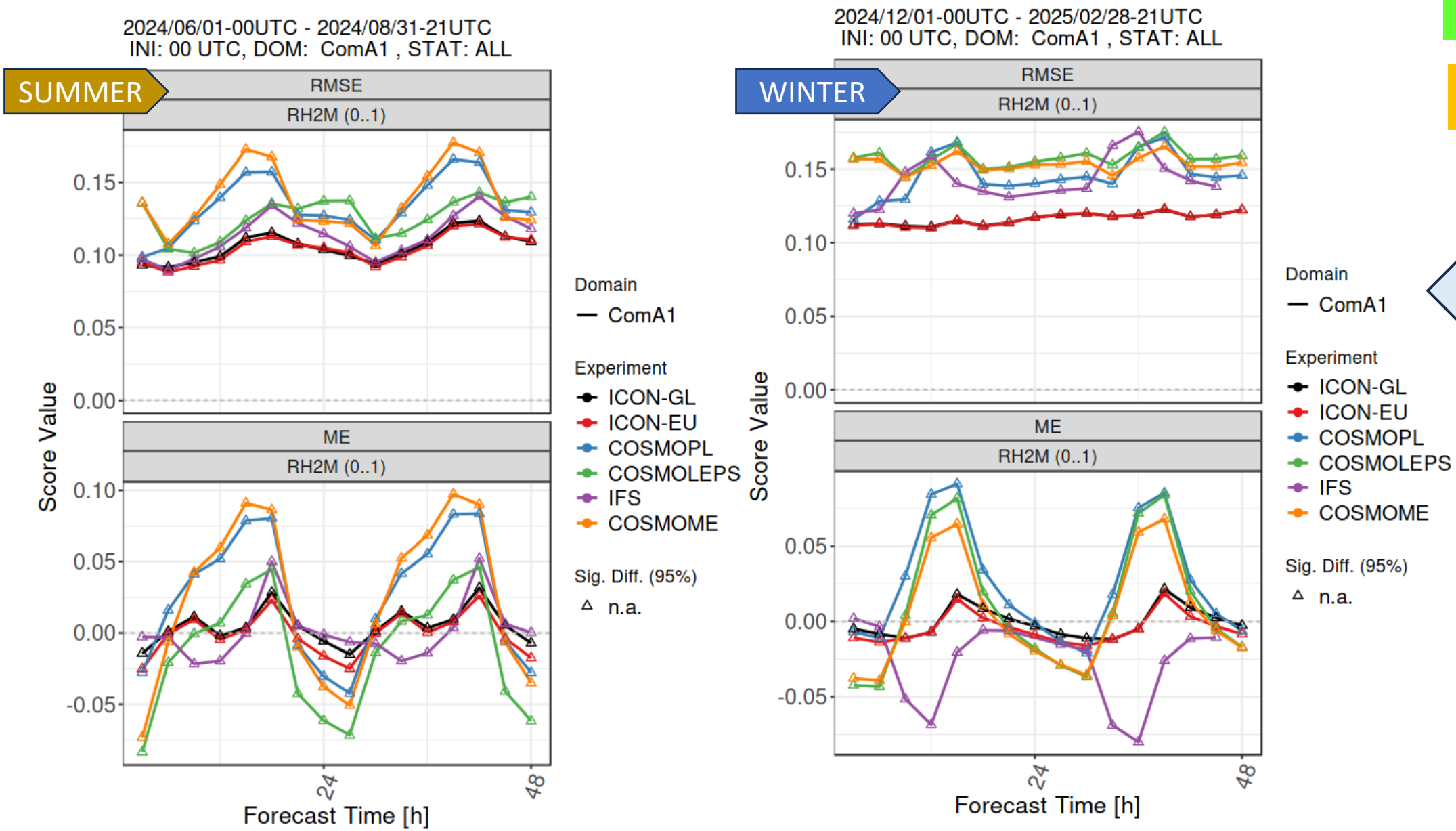


Diurnal cycle stronger in COSMO+IFS
Opposite phase
ICON: Underestimation @night mainly, overestimation @midday

Driving Models

ComA1

RH

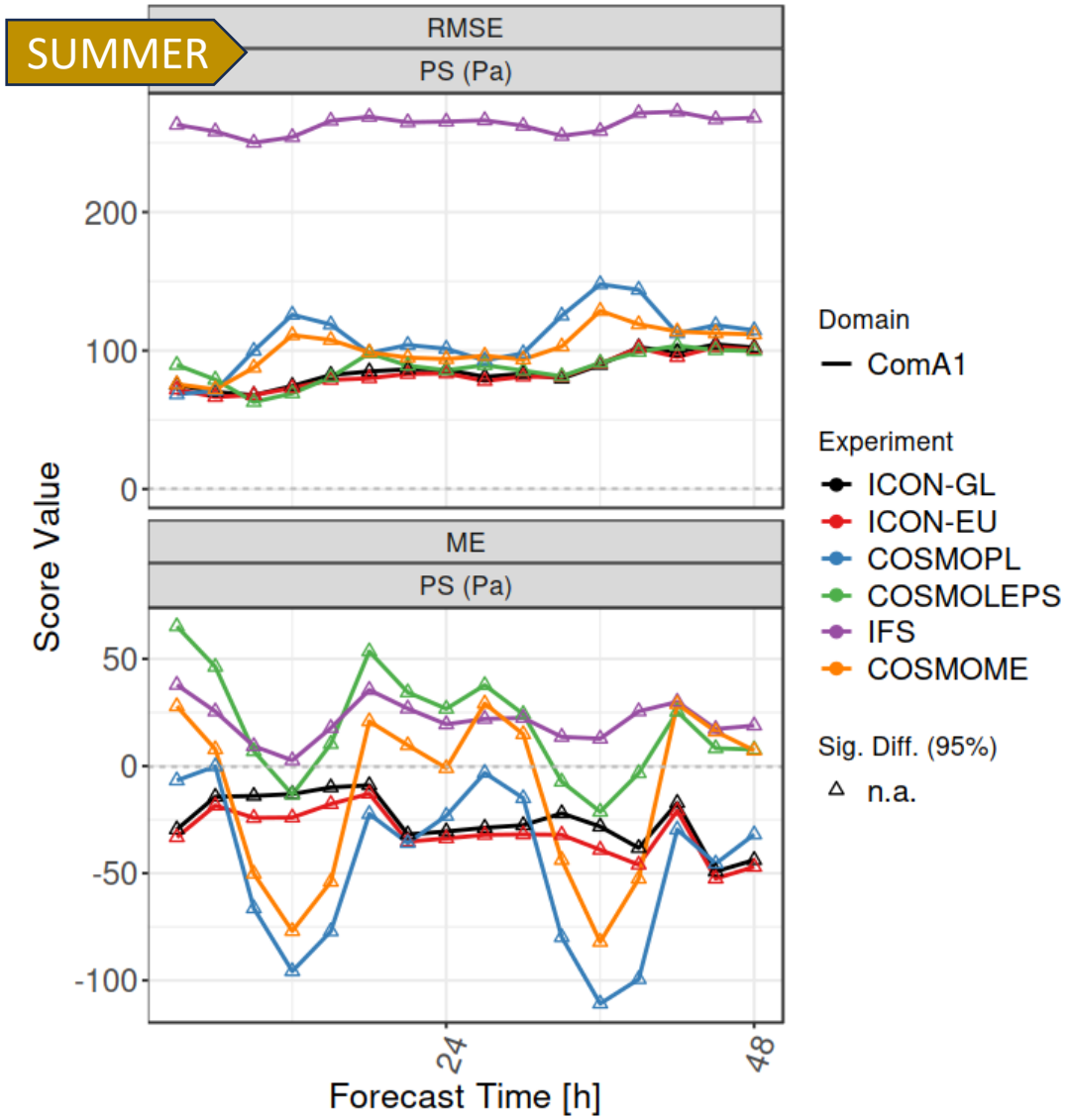


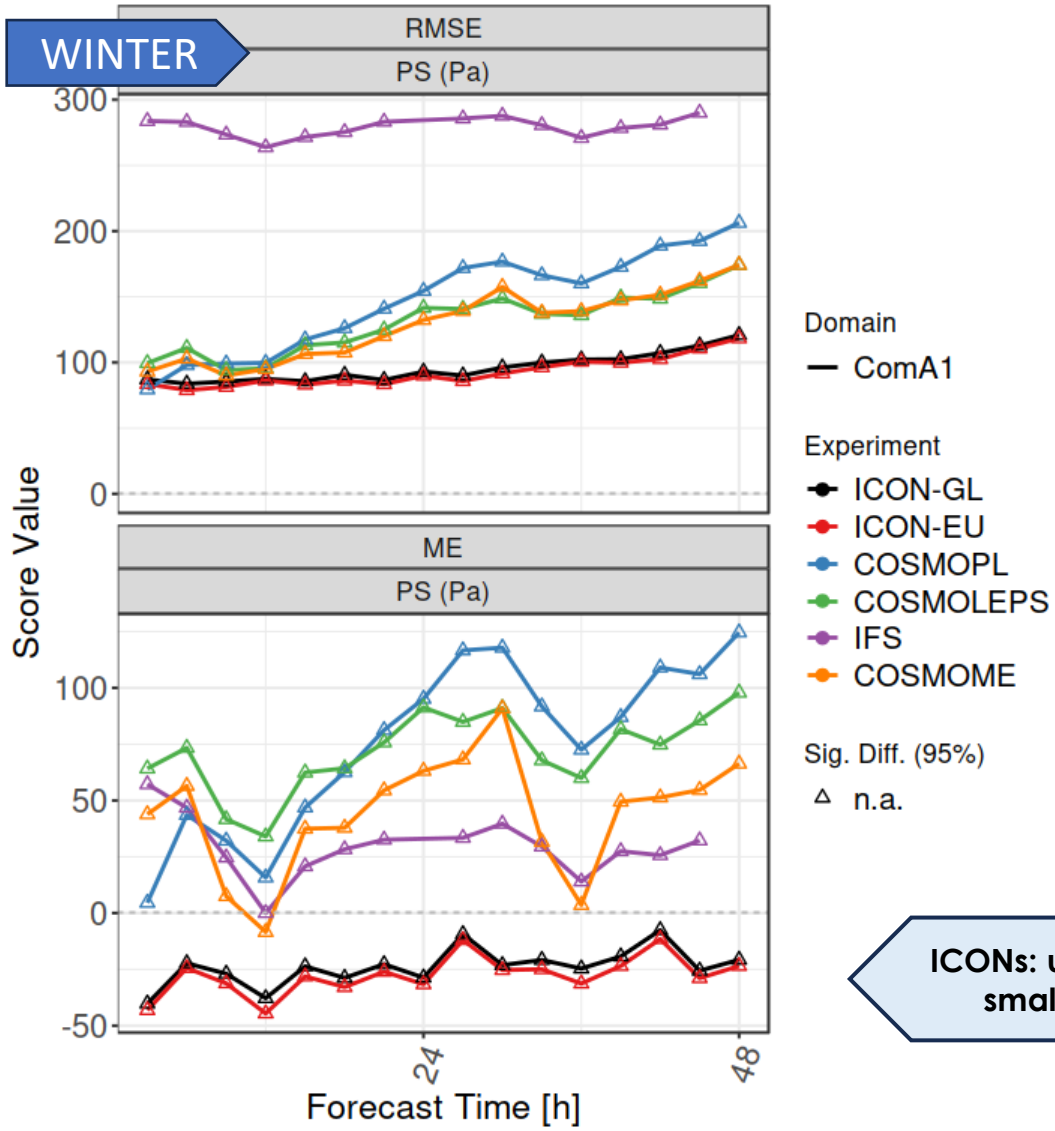
ICON+ IFS in general drier

Driving Models

ComA1

Pa



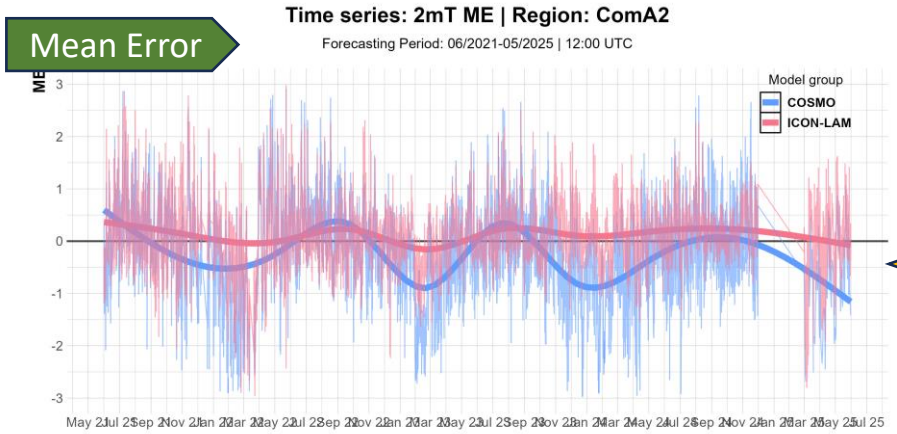


ICONs: underestimation winter, smaller increasing trend

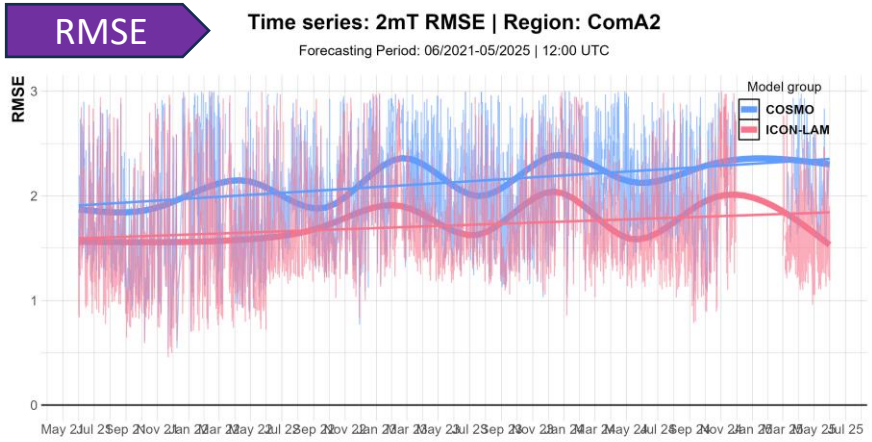
2mT @12+24UTC: HRES COSMO/ICON ComA2. 2021-2025

Factorial timeseries linked to ME and RMSE, thicker lines represent the smoothed average. Used: geom_smooth() that adds a regression line to a plot, and it uses a loess smooth when there are fewer than 1000 observations, and a GAM when there are more.

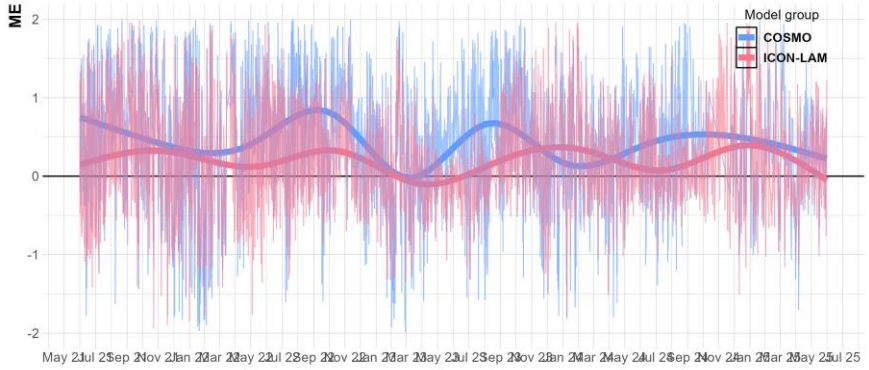
Mean Error



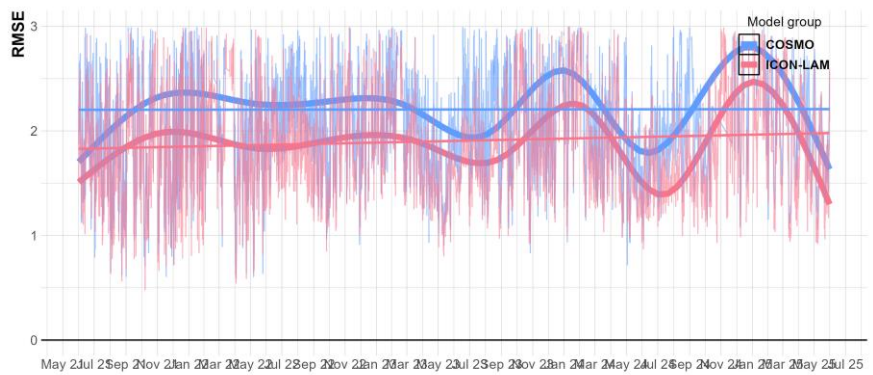
RMSE



Mean Error



RMSE



2 m T

SYS: RMSE diurnal cycle present in all models. Large errors in run initialization. Reduced RMSE/Bias DC with ICON

SYS : Larger errors at night and early morning in winter and at midday in the summer.

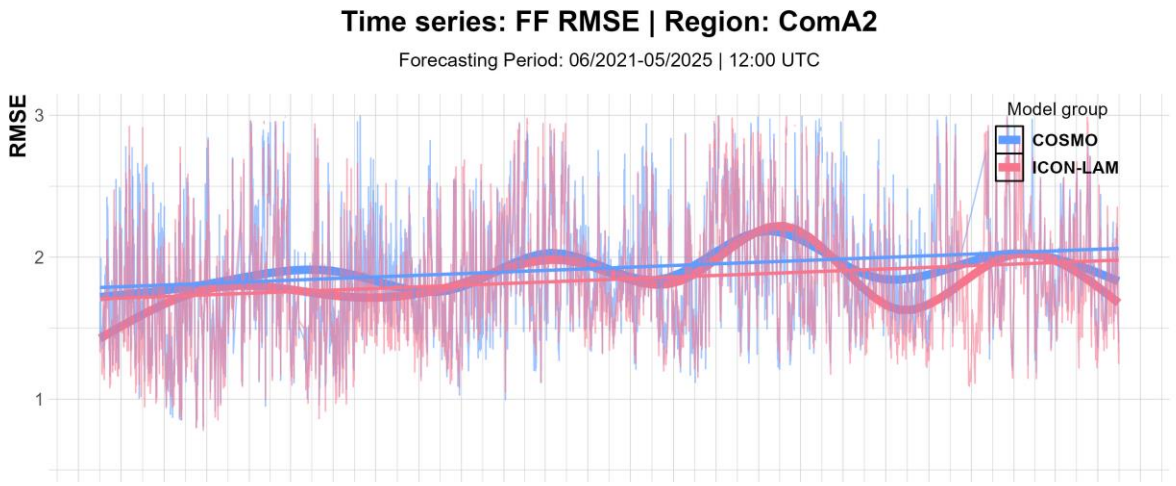
SYS : Underestimation during summer warm hours. Overestimation during summer at night

SYS : ICON tendency to underestimate mainly in the winter night hours

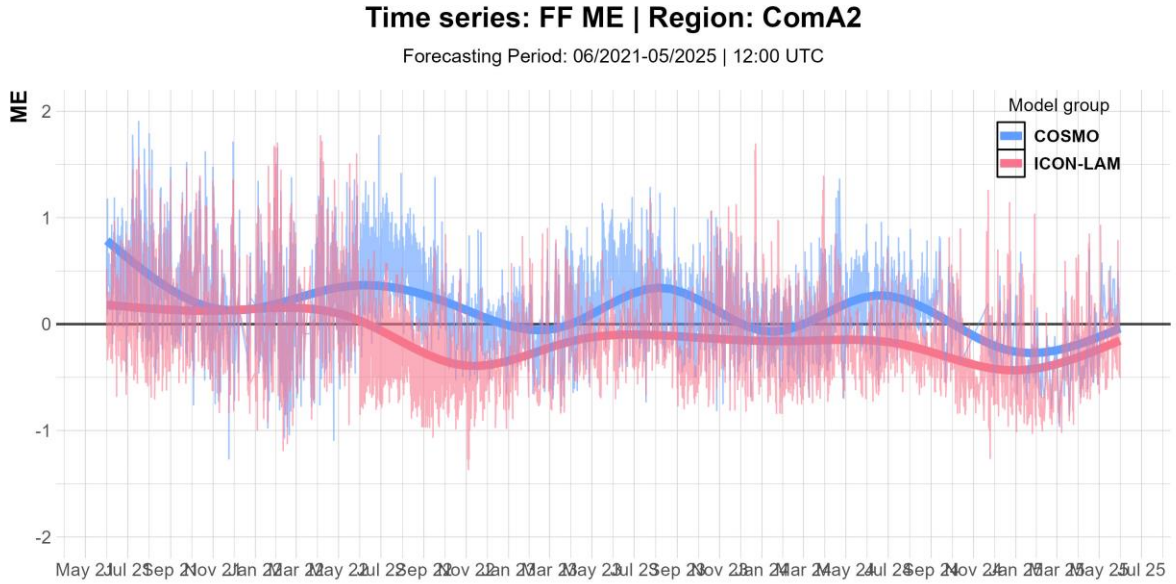
Possible Model Error attributes: issues in the surface energy budget, representation of land-atmosphere interactions, errors in cloud cover or radiation schemes, incorrect representation of soil moisture or boundary Layer parameterization.

WinsSp @12UTC: HRES COSMO/ICON ComA2, 2021-2025

Mean Error



RMSE



WindSp

HighRes LAM: No differences in RMSE variability among models. Underestimation mainly from ICON-LAMs

SYS: Larger error cycle in summer

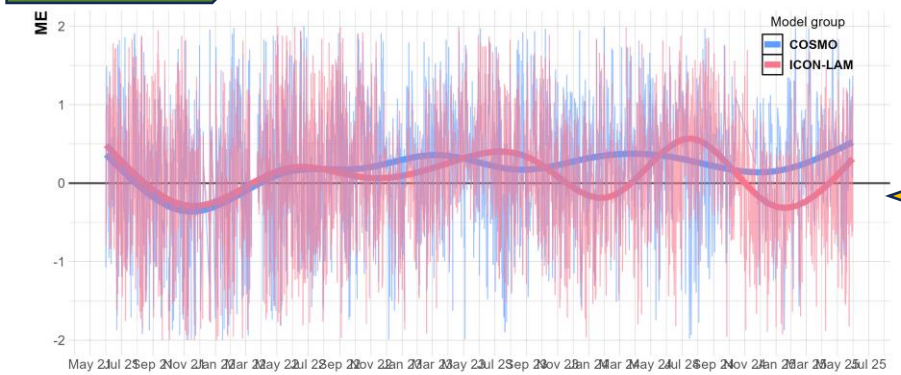
ModDep: negative bias in all seasons with max during nighttime mainly in winter

Possible Model Error attributes: errors in the momentum fluxes or frictional effects in the boundary layer, inaccuracies in the pressure field, particularly in the simulation of high and low-pressure systems (wind direction)

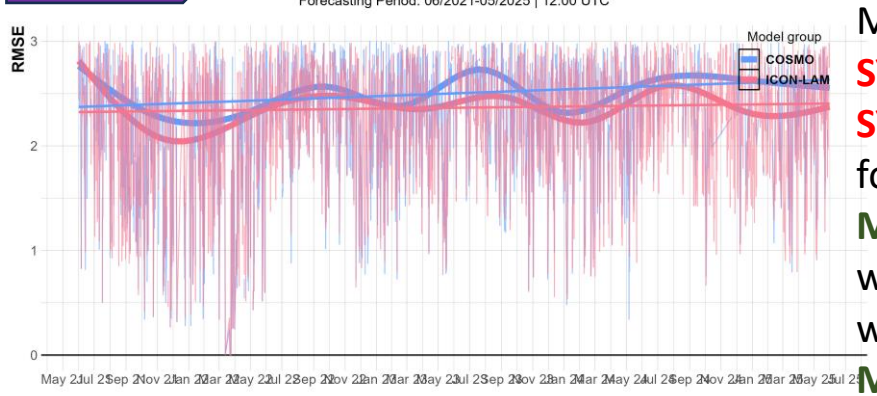
TCC @12+24UTC: HRES COSMO/ICON ComA2, 2021-2025

Factorial timeseries linked to ME and RMSE, thicker lines represent the smoothed average. Used: geom_smooth() that adds a regression line to a plot, and it uses a loess smooth when there are fewer than 1000 observations, and a GAM when there are more.

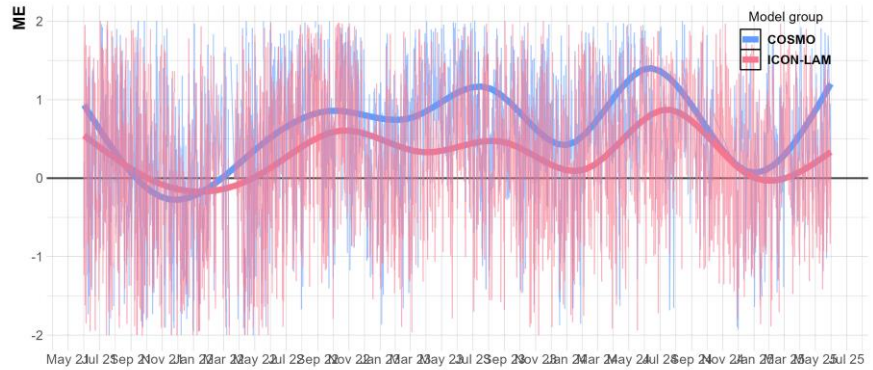
Mean Error



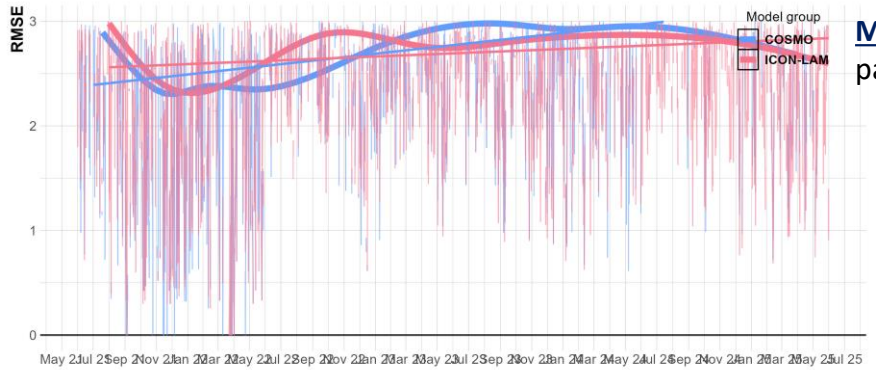
RMSE



Mean Error



RMSE



TCC

SYS : Diurnal cycle of both ME/RMSE for TCC

SYS : **La**rger errors in summer

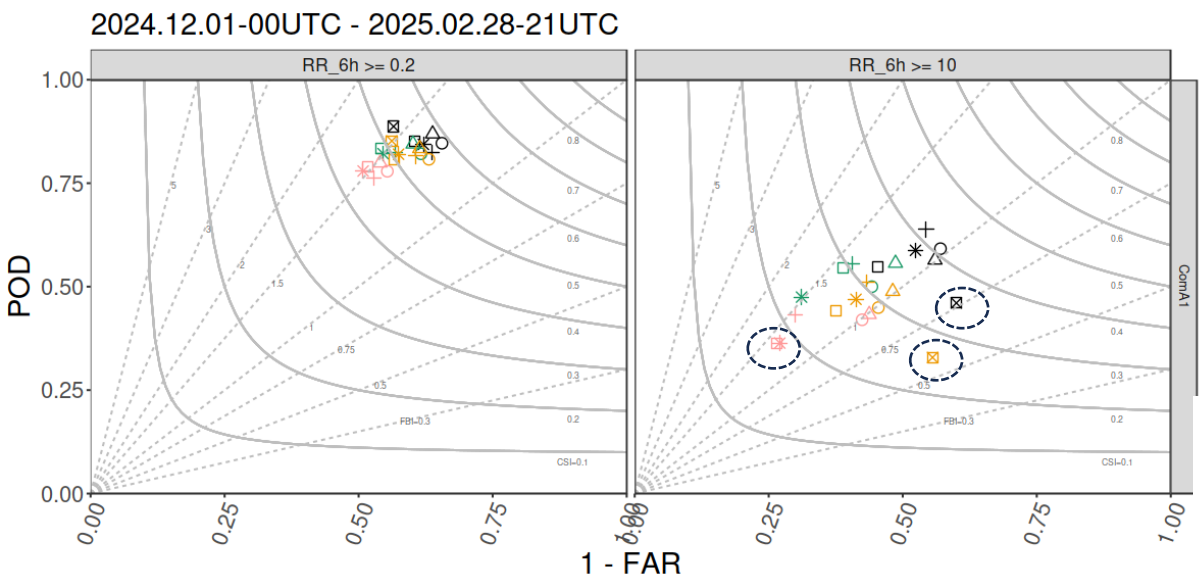
SYS : Higher errors during nighttime for all models (overestimation)

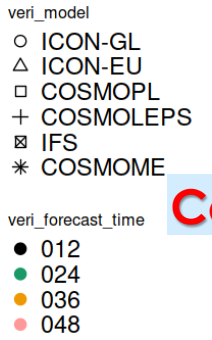
ModDp: Higher underestimation in warm hours of the day with ICON in winter

ModDp: Ambiguous performance in RMSE (higher errors at night) with a tendency to overestimation during warm months night hours.

Model Error attributes: shortcomings in cloud parameterization, radiation errors

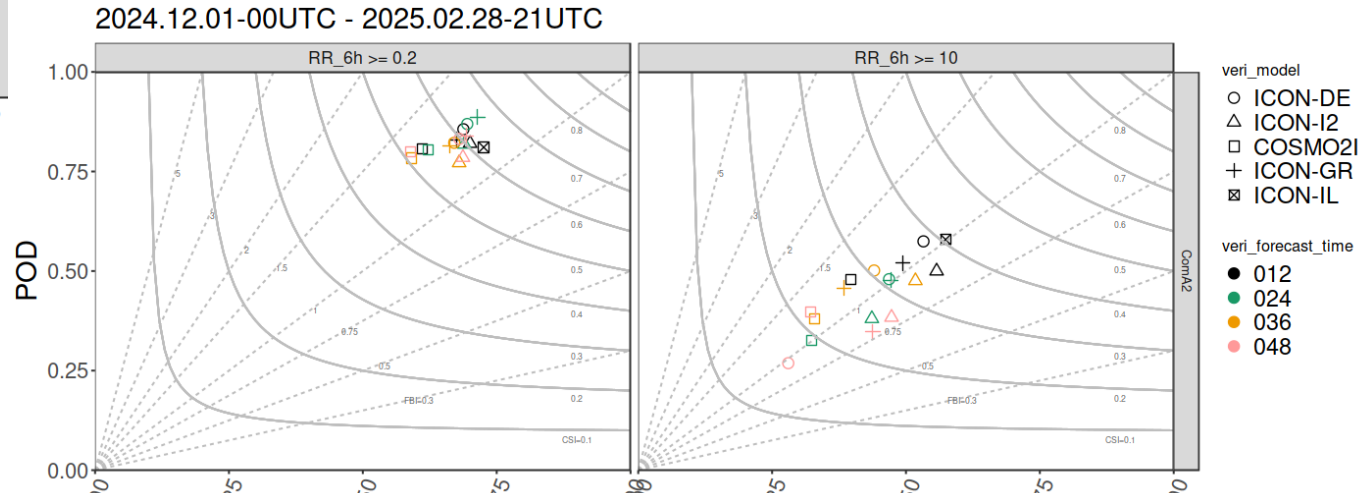
DJF2025



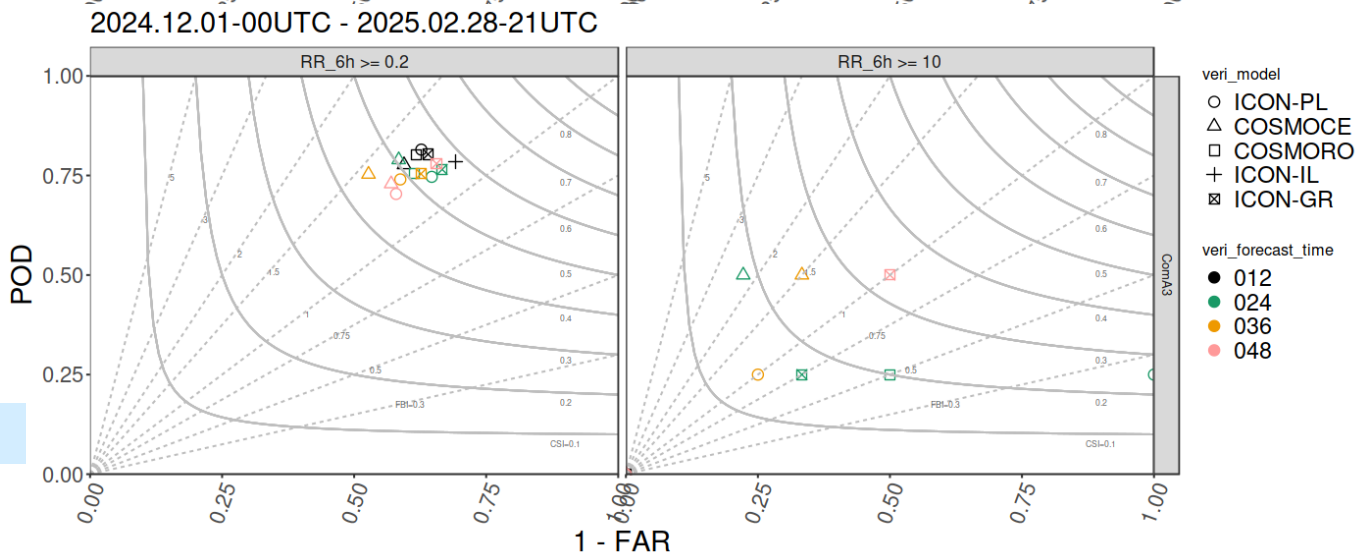


ComA1

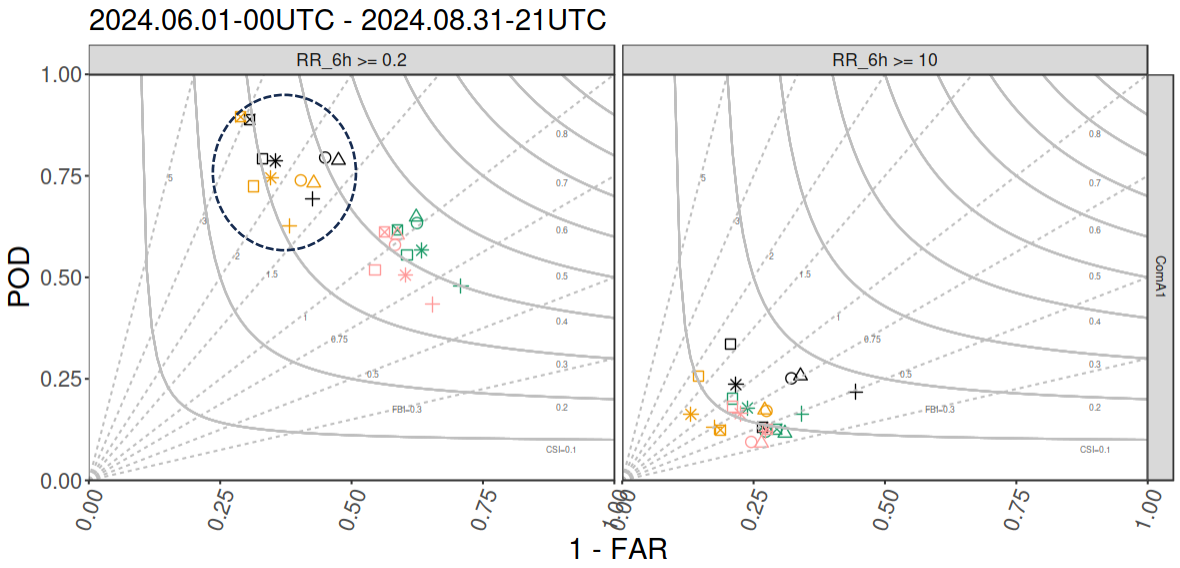
ComA2

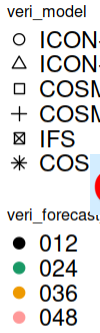


ComA3

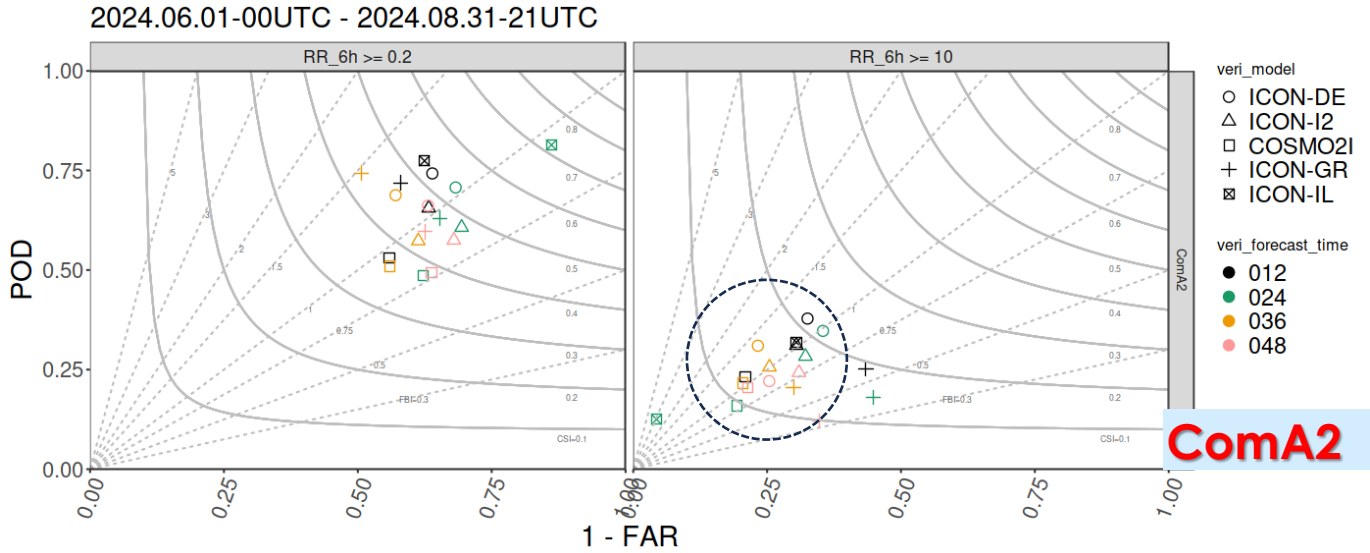


JJA2024

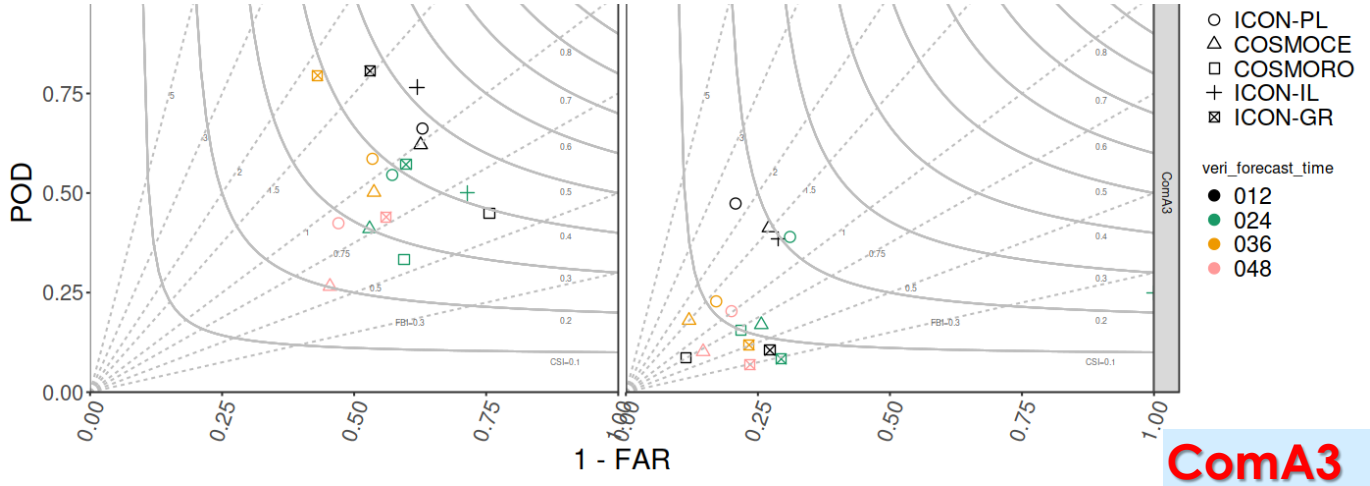




ComA1



ComA2



ComA3

Precipitation

Clearer differences in performance in higher precipitation amounts

MOD: ETS, FAR are higher for ICON but with a tendency to underestimate small preci amounts

Possible Model Error attributes: convective parameterization, small-scale convective or orographic precipitation processes (moisture transport, cloud formation, etc,)

Discussion on MODEL ERRORs

2st Meeting (Videoconference) 24.01.2024
Based on Common Area and National Domain verification
Standard and Conditional Verification, Fuzzy on precipitation

Focus:

- Reporting of systematic errors of ICON-LAMs (dependence on: season, hour, geographical location, weather, other parameters)
- Tuning on systematic model errors
- Precipitation headache updates

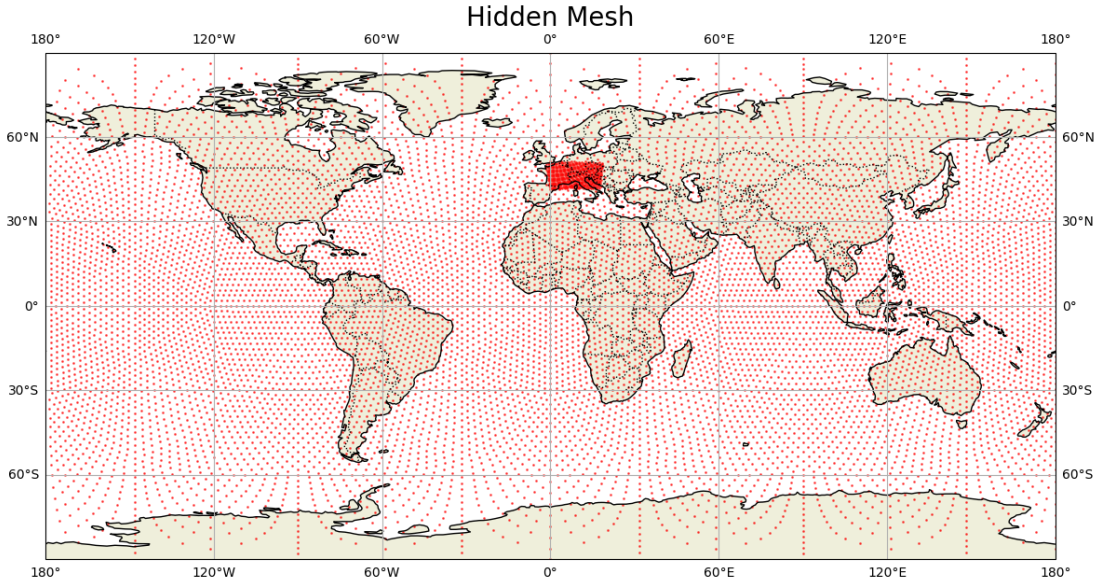
Addressing Systematic Errors
Model Tuning/Calibration
Improved Parameterizations

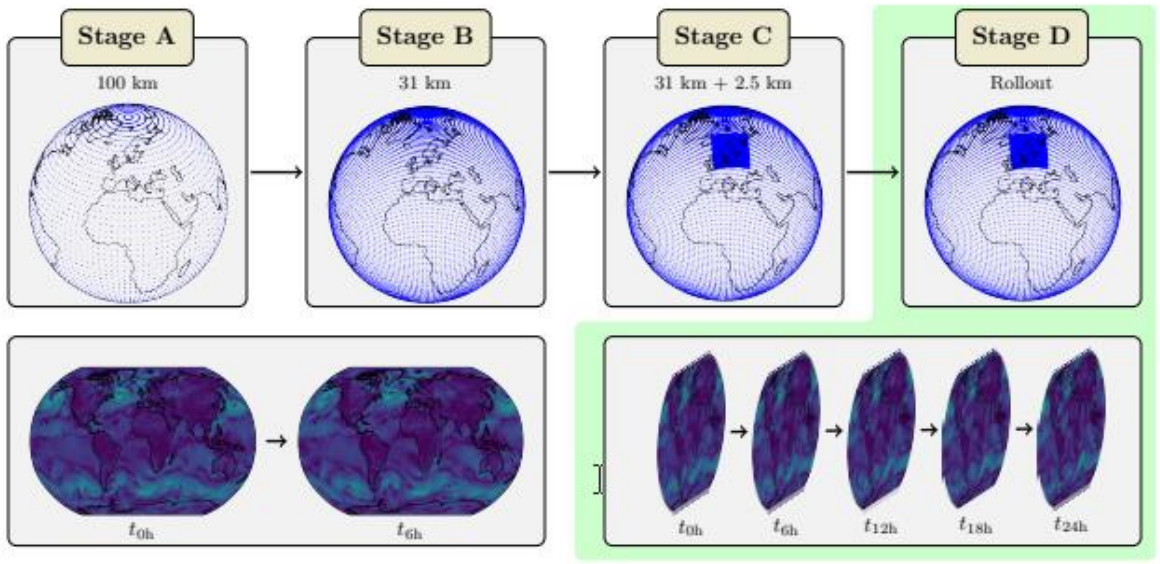
Newsletter section on Models Performance

A. Pauling, et al.
Comparison of NWP and ML verification results

Stretched grid approach (Nipen et al. 2024)

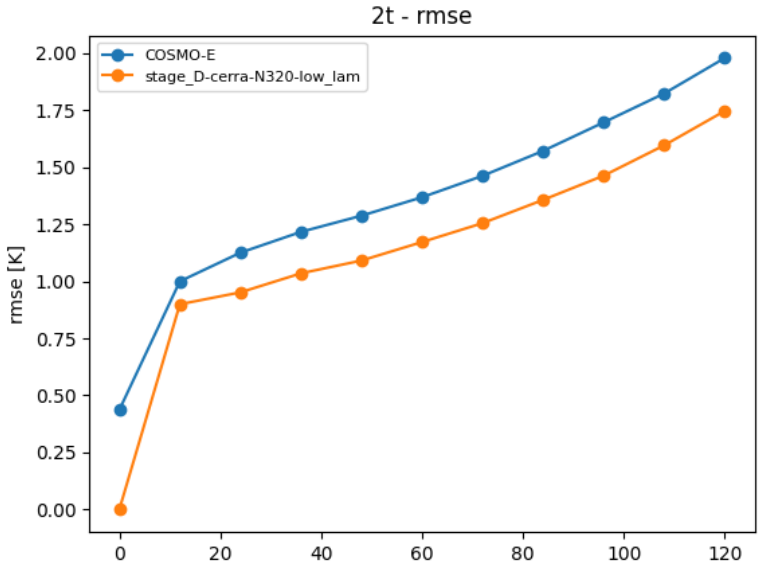
- Combine global and regional dataset in a single multi-resolution grid
- Transfer-learning using four-stage training procedure

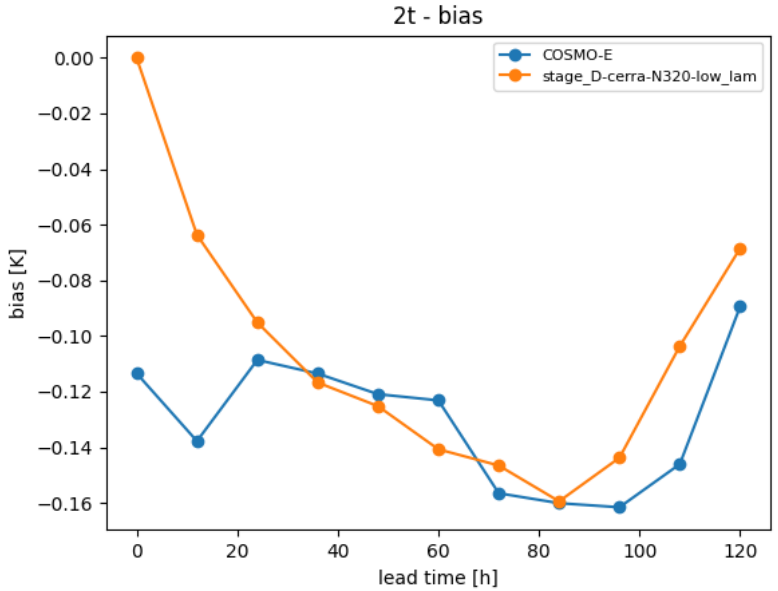


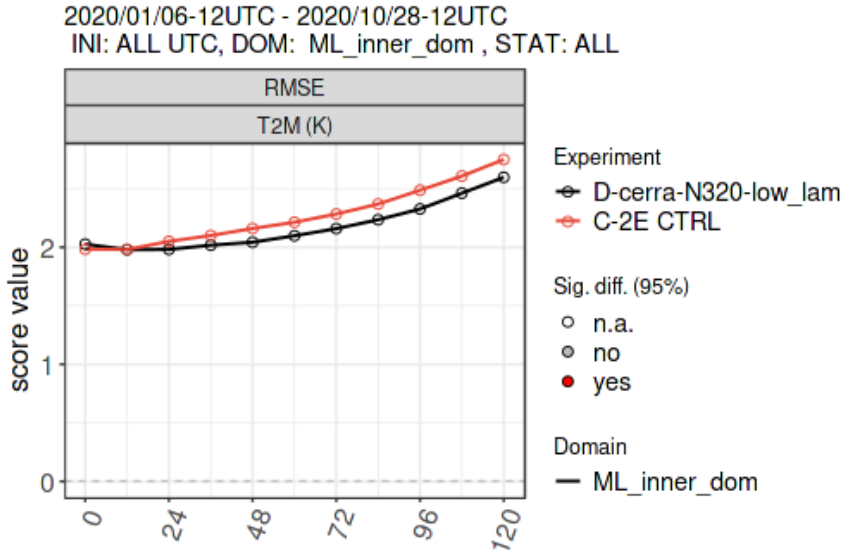


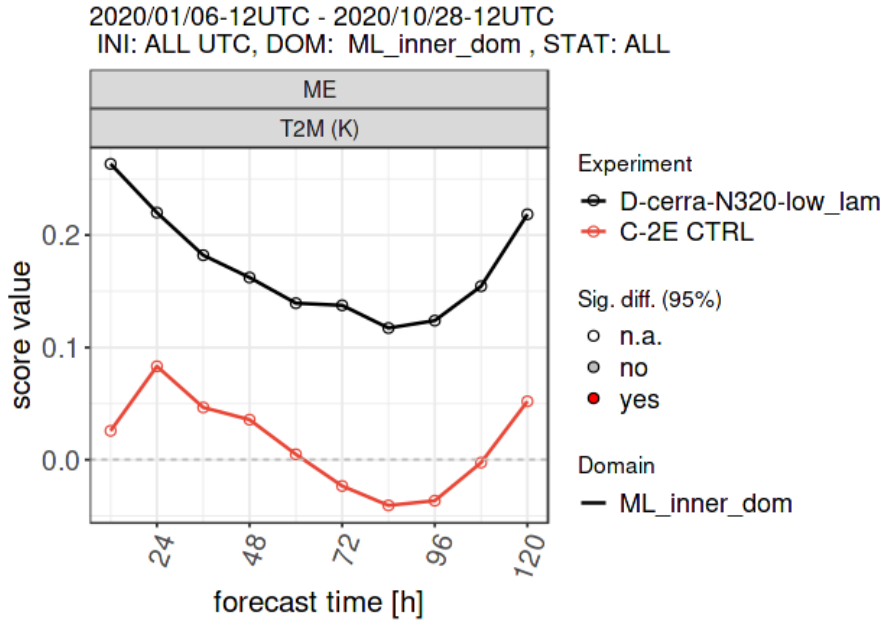
NWP/ML comparison: T_2M

A. Pauling, et al. Comparison of NWP and ML verification results



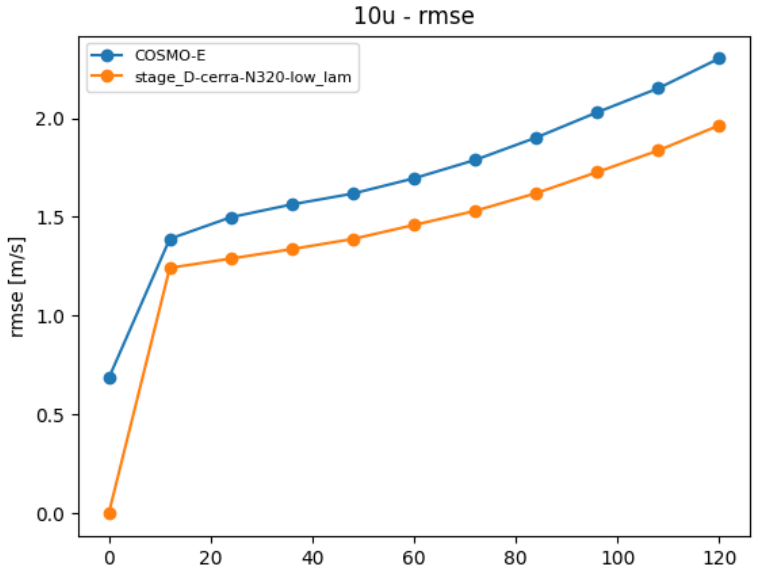


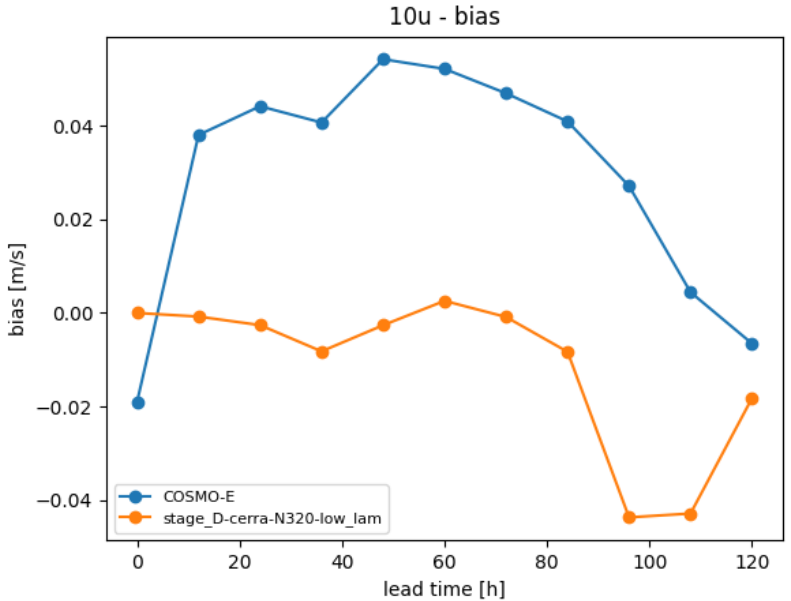


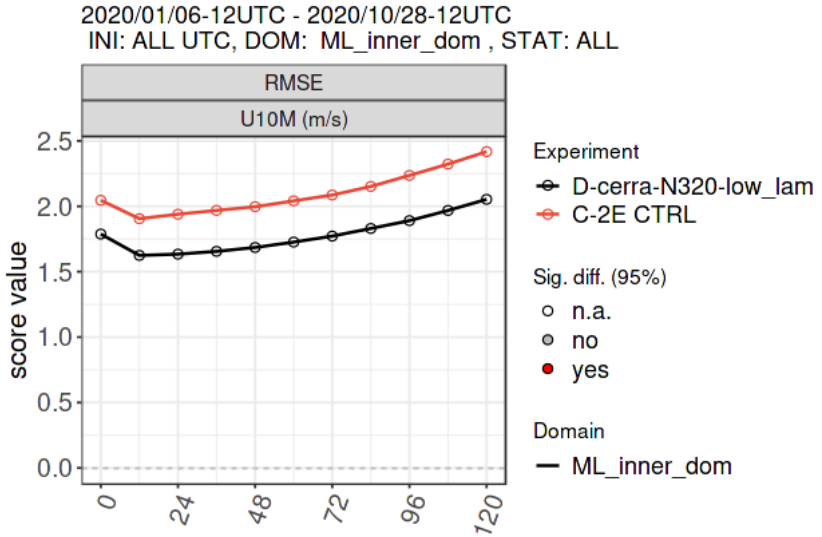


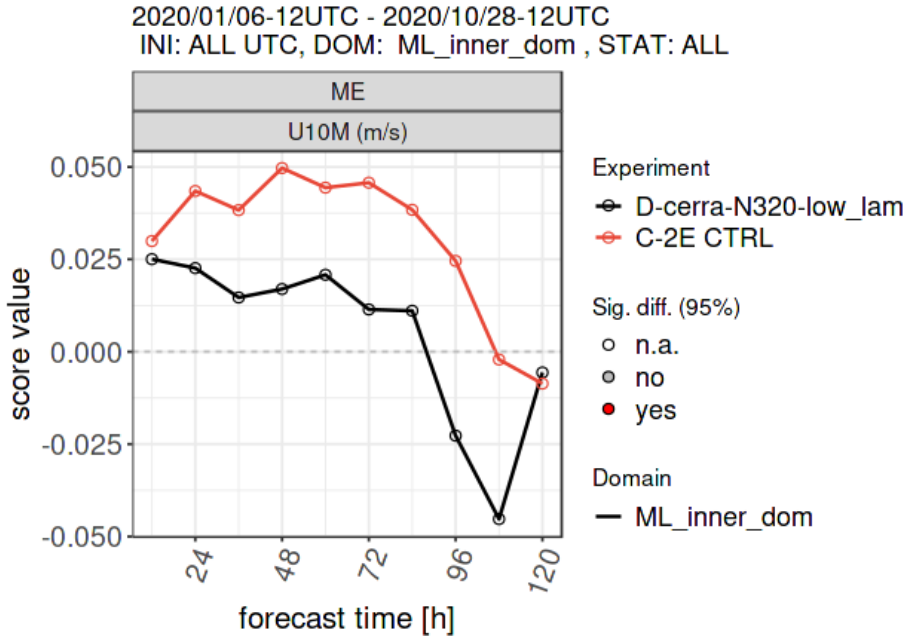
NWP/ML comparison: U

A. Pauling, et al. Comparison of NWP and ML verification results





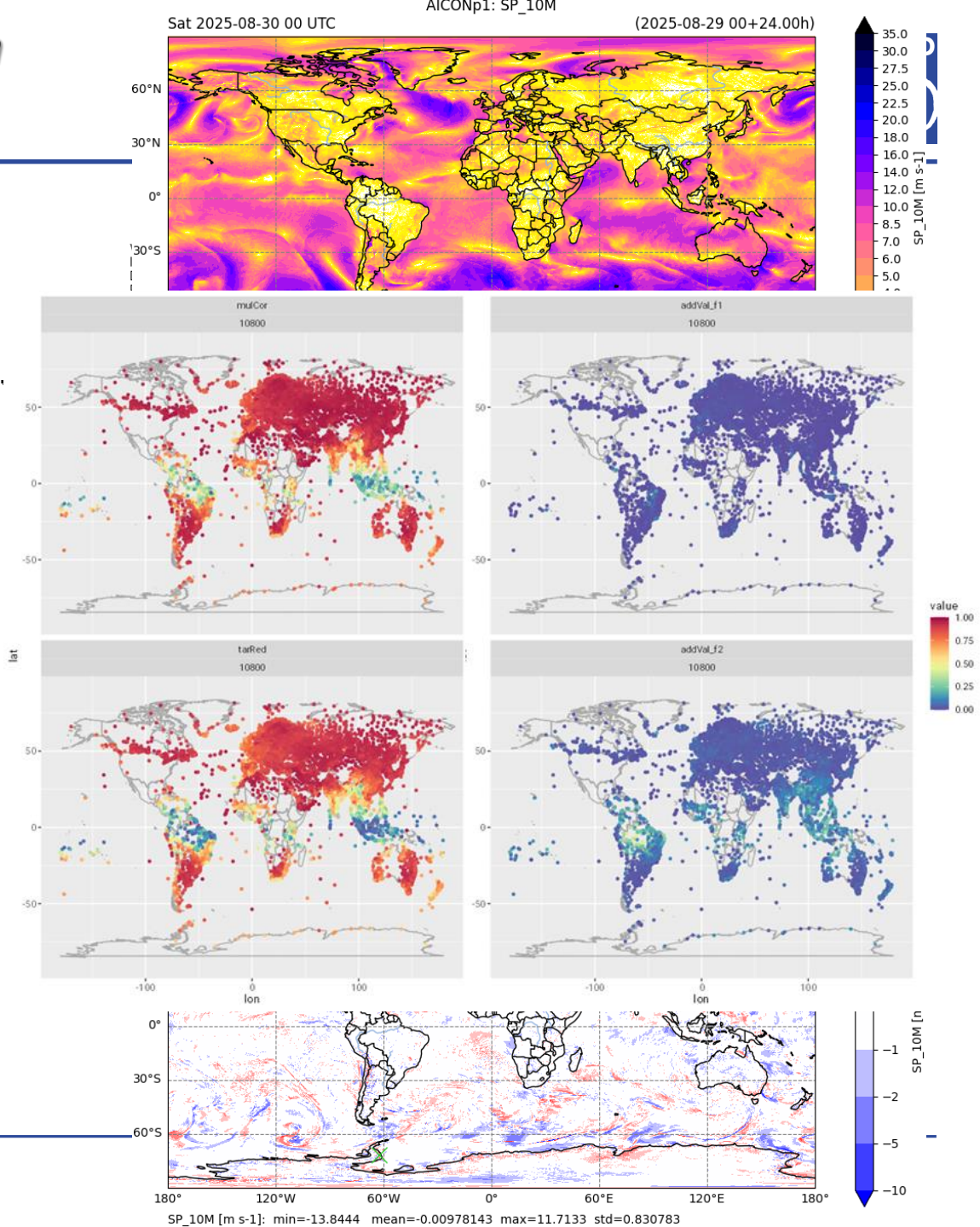




VI. Data driven Models Evaluation

AICON

- AICON will become internally operational this week
  - More intense evaluation also by forecasters
  - Limited output FF10M, T2M, TD2M, RH2M, TOT_PREC and U, V, RH, T, Z on 13 model levels (close to standard pressure levels)
  - Some artefacts like unphysical values or unrealistic precipitati at some grid cells need to be addressed
- AICON verification
  - Standard scores look promising
  - Additional methods and metrics are under development
  - Scale dependent activity (spatial variance) (Britta Seegebrech
  - Simple, observation based activity estimate in FFV2 (Felix Fundel)
  - Additionally explained variance after Glowienka-Hense et al. (2020) (Sabrina Wahl)
- Upcoming developments
  - AICON-LAM



Verification of data driven Forecast Models
Challenges & Alignment with NWP Practices

COSMO

- **Reference Datasets Matter**

NWP: usually against *observations* (radiosondes, SYNOP, satellite)
AI/ML: often against *reanalyses (area)*
Need: establish *common reference frameworks*

- **Metrics**

RMSE, ACC, CRPS often look "better" for AI (smoothed fields)
Need: introduce measures to quantify forecast and observation activity, spatial variability (variance, power spectra)

AI/ML Models

ICON

- **Extremes and Physical realism**

NWP can capture extremes but scores worse on average.
Need: Verify tails with threshold scores, metrics sensitive to extremes, spatial structures & rare events

Verification is not only about performance scores, it is the scientific check that ensures reliability, transparency, and trust in model forecasts

WG5 activities contributors

F. Fundel, N. Vela, S. Gabrian, A. Iriza Burca,
S. Dinicila, M.S.Tesini, F. Batignani, D. Boucouvala, P. Kaufmann, M. Grzelczyk,
A. Surowiecki *P. Khain, M. Bogdan, F.Gofa,….*

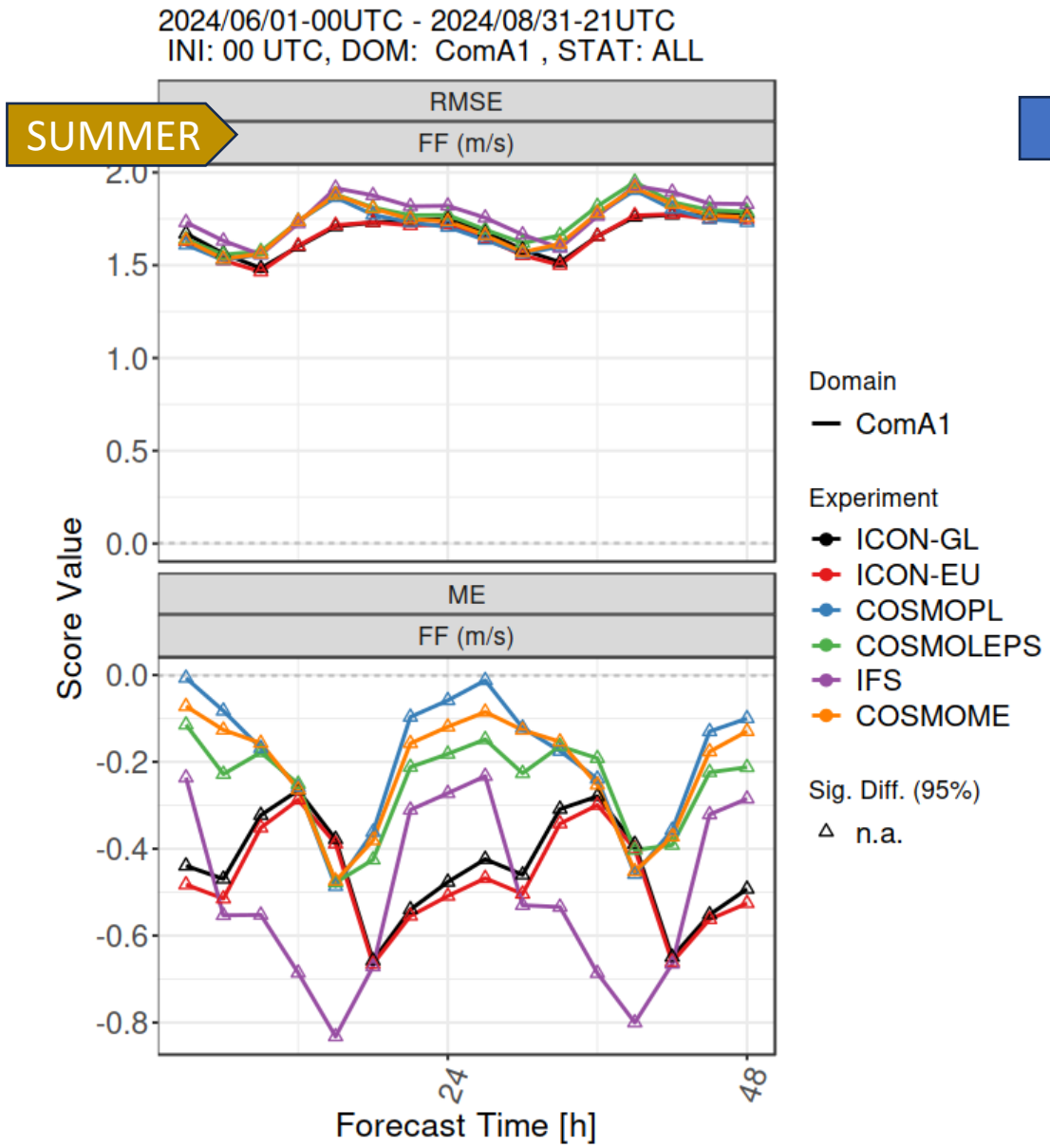

Thank you all

Driving Models

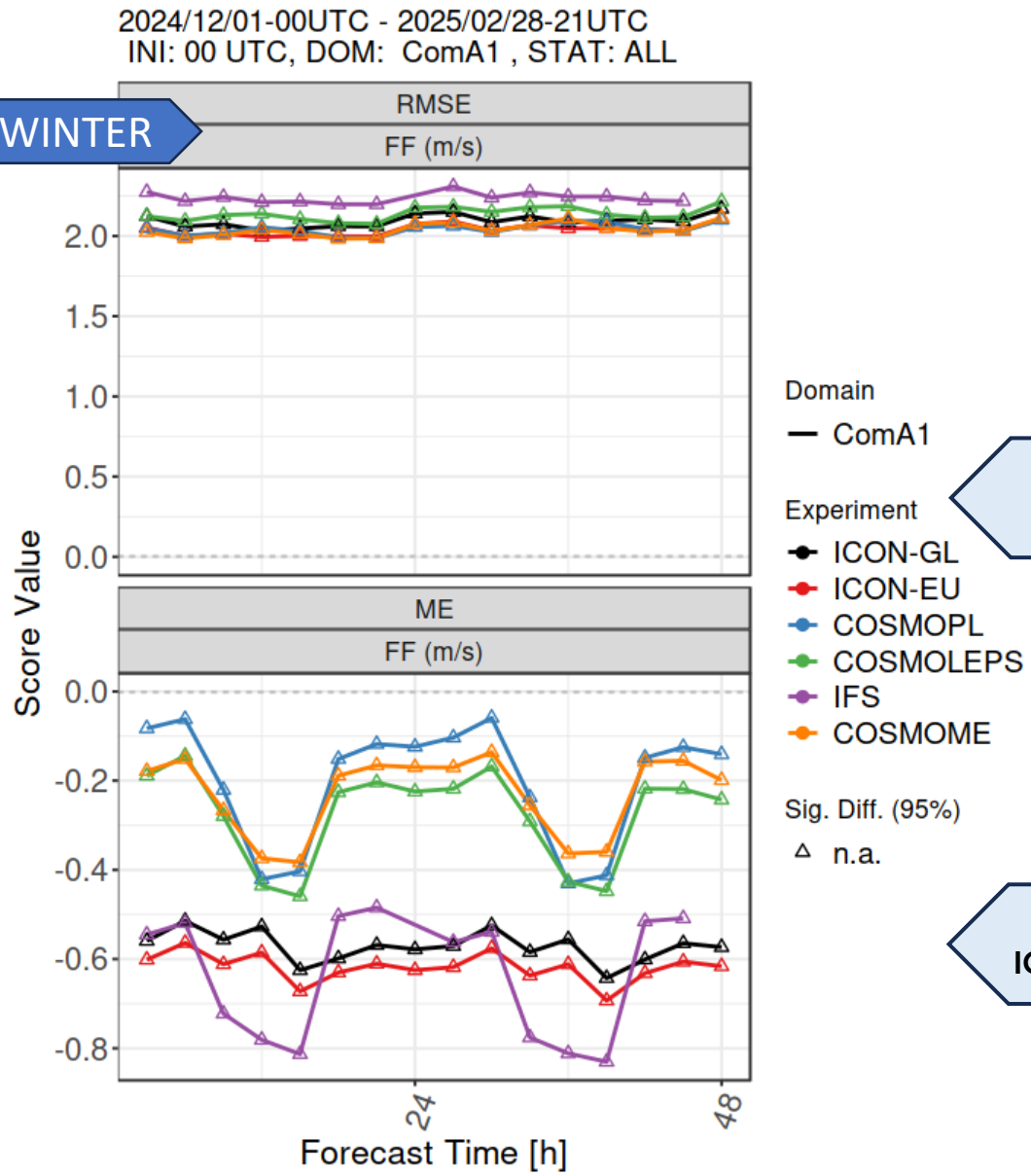
ComA1

WindSp

SUMMER



WINTER



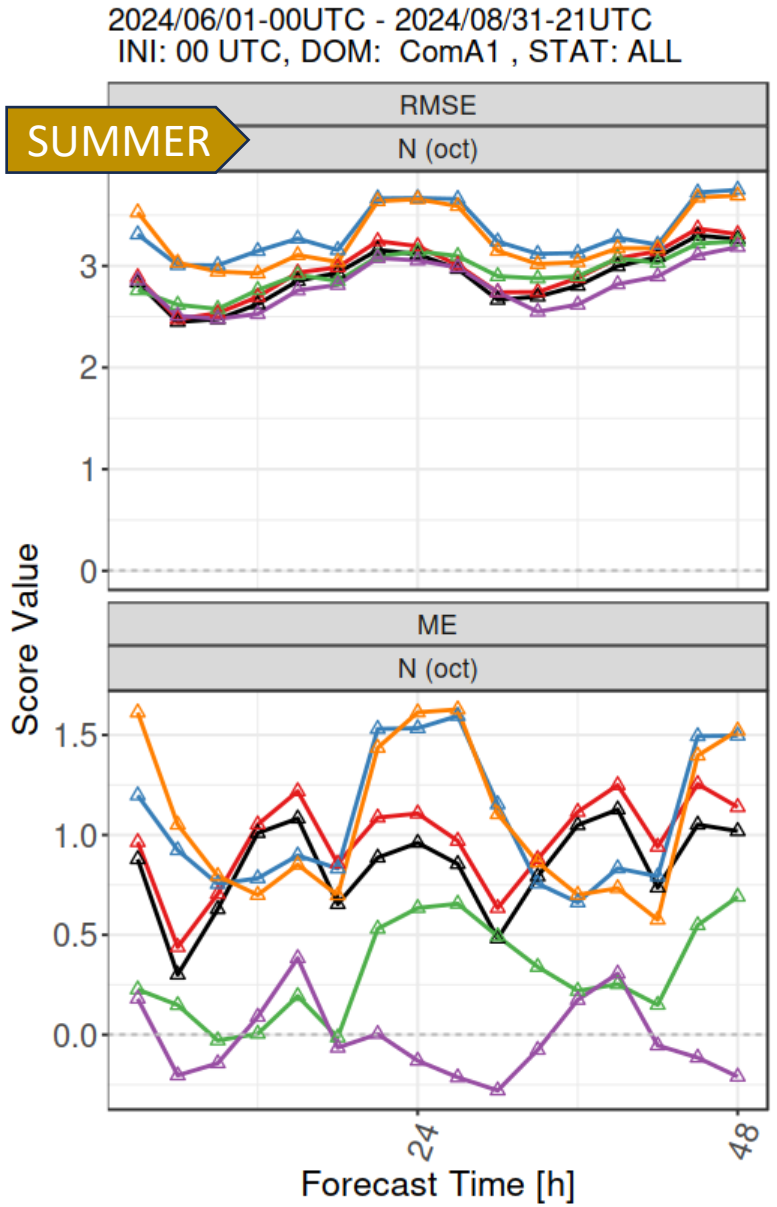
IFS (purple) larger errors

Underestimation ICON(black/red)+IFS

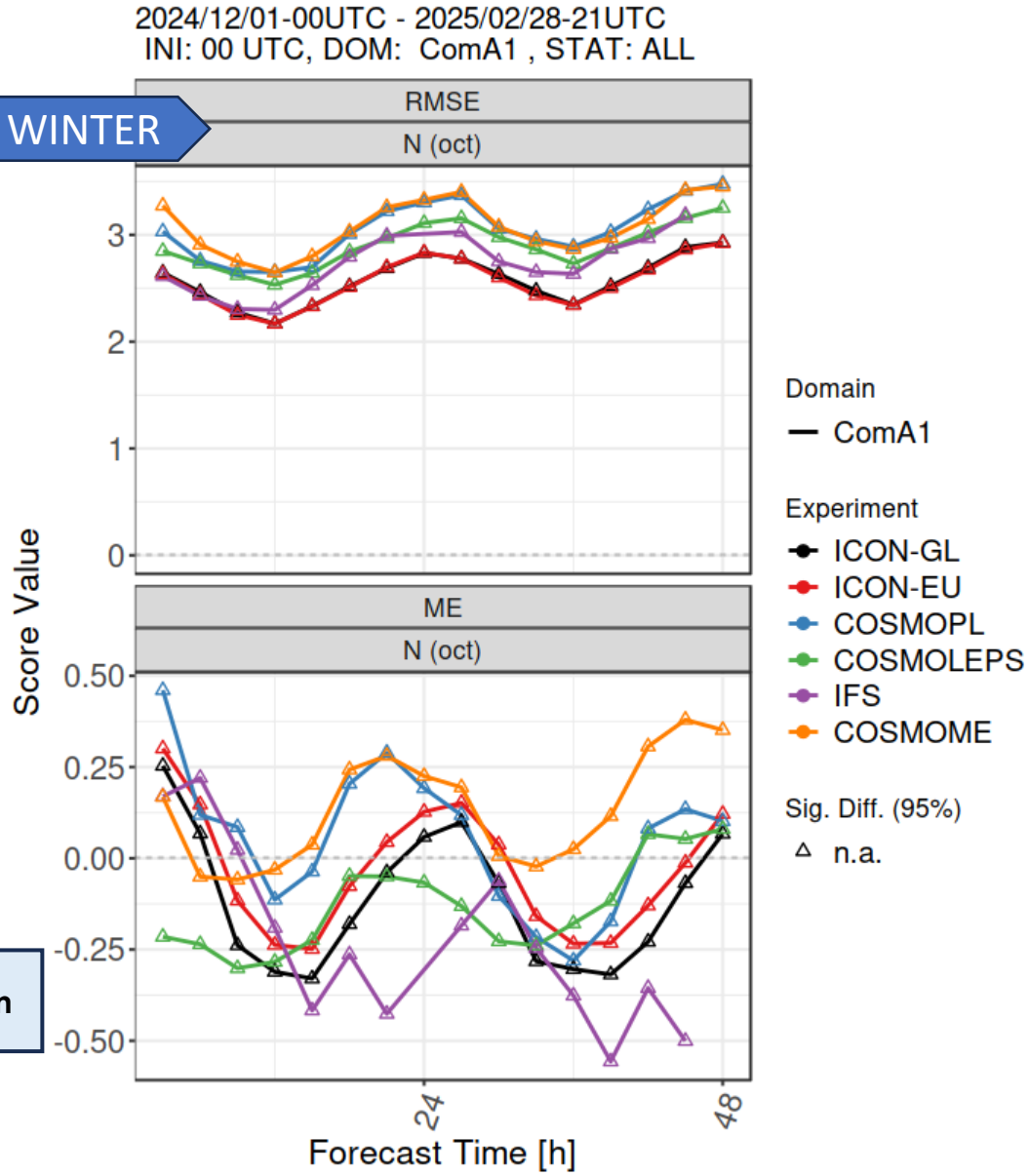
Driving Models

ComA1

TCC



ICON stonget overestimation in summer



NWP/ML comparison: TD_2M

A. Pauling, et al. Comparison of NWP and ML verification results