SPECIAL PROJECT PROGRESS REPORT

All the following mandatory information needs to be provided. The length should *reflect the complexity and duration* of the project.

Reporting year 2020

Project Title: **Testbed for the Evaluation of COSMO Model Versions**

Computer Project Account: **SPITRASP**

Principal Investigator(s): Amalia Iriza (NMA,Romania)[1]
Ines Cerenzia (Arpae-SIMC, Italy)[2]

Affiliation: National Meteorological Administration (NMA)[1]
Environmental Agency of Emilia-Romagna – Hydro-Meteo-Climate Service (Arpae-SIMC)[2]

Name of ECMWF scientist(s) collaborating to the project (if applicable) Umberto Modigliani and his staff,
Andrea Montani

Start date of the project: 2018

Expected end date: 2020

Computer resources allocated/used for the current year and the previous one

(if applicable)
Please answer for all project resources

		Previous year		Current year	
		Allocated	Used	Allocated	Used
High Performance Computing Facility	(units)	5 000 000	4 073 739.44 (81.5%)	5 000 000	54934.91 (1%)
Data storage capacity	(Gb)	1000	1600	1000	1560

Summary of project objectives (10 lines max)

The main objective of the **"Testbed for the Evaluation of COSMO Model Versions"** Special Project is to perform testing of new COSMO model versions prior to their official release using the software environment built on the ECMWF platform during previous SPITRASP projects (2013-2015, 2016-2018). This evaluation of new model versions carried out according to source code management procedures and using the Test Suite platform is taken into account before any operational implementation and release of an official model version. The NWP test suite currently represents a benchmark for rigorous testing of all new model features and allows the model developers to produce guidelines for the selection of a new operational implementation of the model. Several model versions and configurations have been installed and tested up to now in the framework of the SPITRASP special projects, while more are expected to be evaluated using this platform.

Summary of problems encountered (10 lines max)

No problems encountered.

Summary of plans for the continuation of the project (10 lines max)

Activities (including use of resources) to evaluate the new official version (6.00) of the COSMO model prior to its release are planned for this year.
Maintenance of the Test Suite.
Extending the Test Suite to evaluate ICON model official releases – migration to ICON-LAM: prepare an EcFlow suite on HPC, ICONTOOLS and ICON executable already available.

List of publications/reports from the project with complete references

M. Milelli: "Other WG6 Activities", The 21st COSMO General Meeting, Rome, Italy, 9 - 13 September 2019
F. GOFA: "Verification and Case Studies. overview of Activities", The 21st COSMO General Meeting, Rome, Italy, 9 - 13 September 2019
I. Cerenzia: "NWP Test Suite", The 21st COSMO General Meeting, Rome, Italy, 9 - 13 September 2019
A. Iriza-Burca: "NWP Test Suite Suggestions", The 21st COSMO General Meeting, Rome, Italy, 9 - 13 September 2019
I. Cerenzia, A. Iriza-Burca, M. Bogdan, F. Gofa, A. Iriza-Burca, F. Fundel, H. Reich (contributors): "Numerical Weather Prediction Meteorological Test Suite: COSMO 5.05 vs. 5.06", COSMO-Model Report, in preparation

Summary of results

The NWP test suite procedure was adopted by COSMO in order to perform carefully-controlled and rigorous testing, including the calculation of verification statistics, for any COSMO model test-version. Following the source code management procedure, this testing phase should offer the necessary information on the model forecasting performance, in order to determine whether the upgrade of a model test-version to a new release version is possible. For previous testing procedures, the VERSUS system has been used to perform verification. All activities were performed first during the frame of the COSMO Priority Task NWP Test Suite (2013-2015) and as part of special projects at ECMWF (2013-2015, 2016-2017).

Phases I & II: Model set-up & Model Configuration and Execution of Runs

The SPITRASP special project "Testbed for the Evaluation of COSMO Model Versions" approved for 2018-2020 ensures the ECMWF computer resources which were used for the aim of this task both for simulation and for archiving purposes. The platform has been in use for some time and was

employed for a large number of COSMO versions (up to 5.05), with an enriched and modified procedure for the past years. For previous tests, starting with version 5.03 of the model, all versions are implemented on the Cray HPC. Starting from version 5.04a (quasi 5.05) of the COSMO model, the 2.8km horizontal resolution of the model is also tested using the NWP Suite, in addition to the previously used setup at 7km. Starting from version 5.05, the single precision version of the model is also evaluated.

During this year (starting from version 5.06), the forecast mode that was applied to the test simulations until now is being replaced with the hindcast mode. This change is aimed at reducing the computational costs and time for testing. Starting from version 5.05 of the model (previously tested), evaluations are performed in both double and single precision model versions for the 7 km horizontal resolution setup and only in double precision for the 2.8 km horizontal resolution configuration**.** For version **5.06 (**test**)**, the double precision hindcast configurations (both 7 km and 2.8 km horizontal resolution) were tested against version 5.05_1 (benchmark/operational) of the model (5.05 with a bug fix). The single precision (SP) setup for v5.06 was tested against the double precision (DP) one, with single precision runs for version 5.05_1 also available.

The directory structure and the archiving procedures for COSMO-5.05_1 and COSMO-5.06 (hindcast mode) model versions follow the ones used for the previous implementations. Int2lm version 2.05 was used for the interpolation of initial and lateral boundary conditions

After completion of the simulations, model outputs were processed together with the corresponding observations using the MEC (Model Equivalent Calculator) software, aimed at producing the necessary Feedback files. Rfdbk (DWD developed) software that utilizes R libraries was used to process Feedback Files in order to produce verification scores. The VERSUS verification software is no longer used for new model version evaluation. Model output obtained from the experiments is locally stored in the ECFS system. All the necessary software (MEC, Rfdbk) used for NWP Test suite purposes are also implemented on ecgate.

Initial conditions are provided by ECMWF HRES analysis, whereas lateral boundary conditions are introduced with a 3 hourly frequency and they include the ECMWF HRES analyses (at hours 00, 06, 12 and 18UTC) and short cut off analyses (at hours 03, 09, 15 and 21 UTC). Soil is initialized from ICON-EU, then runs free (for both model resolutions). Model output was available as 1 daily hindcast file for each day of the month, containing 1 forecast each 3 hours (one step), with lead-times from 0 to 21 (in total 62 daily files). The domain characteristics are presented in Table 1 and figure 1.

Table 1: Main features of the models used in the NWP Test Suite

	ECMWF HRES	**COSMO 7p0**	**COSMO 2p8**
Grid points (nx x ny)	901 x 501	661 x 471	1587 x 1147
Model levels	137	40	50
Resolution (dx x dy)	0.1 x 0.1	0.0625 x 0.0625	0.025 x 0.025
Forecast range (h)	72	72	48

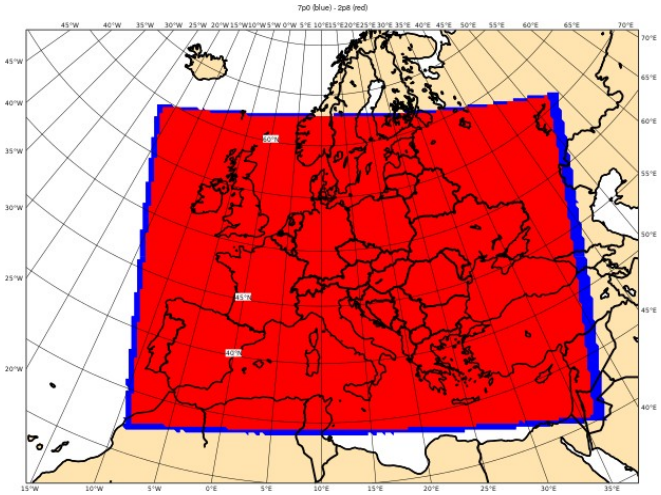


Fig. 1 Integration domain for the COSMO model at 7 km of horizontal resolution (blue) below the domain for 2.8 km of horizontal resolution (red).

The ECMWF Special Project SPITRASP has an allocation of 5.000.000/year (2018-2020). The costs of the test suite in hindcast mode for the configurations v5.05_1 and v5.06 are presented in Table 2 (double and single precision for 7 km, double precision only for 2.8 km).

The hindcast mode costs in terms of BU and simulation time are about one third and a half of the forecast mode, for both horizontal resolutions of the COSMO model. Differences are mainly due to reduced time range of hindcast simulations (24h long) compared to the 72h and 48h forecast range set for COSMO in forecast mode at 7km and 2.8km respectively.

Table 2 Cost of the suite for configurations v5.05_1 and v5.06 (on Cray).

INT2LM from IFS to COSMO-7km	
HRES → 7p0 ~ 17 BU, ~ 1.5min,	
EC_total_tasks=72, EC_nodes=1	
COSMO-5.05_1	***COSMO-5.06***
7p0_DP ~ 890 BU, ~ 5 min	**7p0_DP** ~ 800 BU, ~ 5min
EC_total_tasks=720, EC_nodes=20	EC_total_tasks=720, EC_nodes=20
7p0_DP→2p8_DP ~ 115 BU, ~ 6 min	**7p0_DP→2p8_DP** ~ 121 BU, ~ 6 min
EC_total_tasks=72, EC_nodes=2	EC_total_tasks=72, EC_nodes=2
2p8_DP ~ 13.535 BU, ~ 53 min	**2p8_DP** ~ 12.708 BU, ~ 50 min
EC_total_tasks=972, EC_nodes=27	EC_total_tasks=972, EC_nodes=27
7p0_SP ~ 712 BU, ~ 4 min	**7p0_SP** ~ 761 BU, ~ 4 min
EC_total_tasks=720, EC_nodes=20	EC_total_tasks=720, EC_nodes=20

To evaluate version 5.06 (test version) against version 5.05_1 (operational) as benchmark, tests were performed for runs at 7.0 km, 40 model levels, in hindcast mode, double precision and single precision and for runs at 2.8 km, 50 model levels, in hindcast mode, double precision

Since the tests were performed in hindcast mode, some post-processing of the model output files was necessary in order to process the available information. The grib1 files were remapped by changing the timeRangeIndicator, while the corresponding files for 00 UTC were used twice: as analysis for one day and as forecast from the previous day with step 24.

Some problems with precipitation, which did not arise in previous tests of model versions, were encountered, due to the switch from forecast to hindcast mode. In hindcast mode, total precipitation (and other parameters such as wind gust) are not coded as accumulated fields. Moreover, model

output format is grib1, which can only encode accumulated fields with stepRange up to 255 hours (~10 days), while the experiments are performed for two 30-days periods.

The solution for the ensuing issues was keeping hourly cumulated grib1 files, but splitting the period into 10-days file batches, due to the limitations of up to 255 hours (~10 days) cumulation interval in this format. Total precipitation messages were regribbed in order to be defined as accumulated fields from the beginning of the month to the ongoing hour, as if the simulation were continuous from the beginning of the month. Additionally, in order to process the final/first dates from each 10-days period, hindcast files containing model estimates for the previous date (step 240 from the previous 10-days run) were also regridded as step 00 for the current 10-days run.

This approach provided a working solution for evaluating the performance of the model in estimating precipitation quantities and will be further used to test the next COSMO version prior to its official release.

Phase III: Model Output Verification

Grid-to-point comparisons was employed to compare gridded surface and upper-air model data to point observations, taking into account around 3600 selected NWP suite stations situated in an area covering **-25/24/65/65 (W/S/E/N)**. Suspect observation values (forecast-observation greater than a specific limit are excluded) and included in the verification test to eliminate errors connected to observations. The verification modules for testing v5.06 were the following:

- **surface continuous parameters** 2m temperature (T2M), 2m dew point (TD2m), wind speed (FF), total cloud cover (N), surface pressure (PS): ME, RMSE, SD, R^2, TCC (tendency correlation), LEN (number of observations), OMEAN, FMEAN (observed and forecast mean);
- **precipitation verification** (6h, 12h) for selected thresholds (greater than 0.2, 0.4, 0.6, 0.8, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 25, 30): ETS, FBI, POD, FAR, LEN etc.;
- **upper air verification (TEMP based)** – Temperature (T), dew point (TD), relative humidity (RH), wind speed (FF) and wind direction (DD) for selected pressure levels (250., 500., 700., 850., 925., 1000.): ME, MAE, RMSE, SD, etc.

The verification results presented here (figures 2 – 4) are a summary of the derived statistics, with the complete overview of all the statistical analysis available on the COSMO web-site. As a side note, on the figures, COSMO v5.05_1 (v5.05 with a bug fix) is denoted in short as v5.05.

For the **double precision runs (figure 2)**, 2m temperature differences are insignificant for both seasons between the two model versions and the two resolutions. Both model versions for 7km, exhibit an underestimation during the warm hours of the day in the summer and during all hours in the winter month. The summer underestimation is slightly reduced for the 5.06 version, while the daily cycle for ME during summer can be detected in the 24h statistical results.
Both COSMO 5.05_1 and COSMO 5.06 underestimate the values forecasted for 2m dew point temperature for both seasons, while an increased underestimation is shown for the month of July when model 5.06 is used and this is apparent in both model resolutions.

For PS, both model versions exhibit an overestimation during the winter month and underestimation in the summer, which is slightly reduced with the use of v5.06 (7km), while also RMSE reduction for this season is observed.

With respect to 10 meter wind speed, behavior of NWP test results exhibit almost identical results for both seasons and resolutions for both model versions, mainly underestimating observed values with the 7km resolution model and overestimating with the 2.8km one.

Total Could Cover is a parameter that exhibits no change in the results between 5.05 and 5.06

versions for both resolutions.

The scores for the upper air parameters (e.g. in figure 4) in general also show similar behaviour for both models.
RH exhibits small differences between the two model versions with the 7km resolution and the 2.8km one. For the winter period, there is small reduction in the overestimation of RH during the night in the lower atmosphere and lower underestimation in the layer 700-400mb for the 7km version. During the warm hours there is a smaller reduction in underestimation. During summer, the new version performs slightly better, with reduced overestimation of RH from the surface up to almost 500mb, with no difference between the two model versions. Overall, there is a small positive impact of version 5.06.

Temperature comparison for the two model versions gave insignificant differences in most cases. Only during winter ME graphs exhibit a very small increased underestimation of error in levels lower than 500mb and the coarser resolutions. With the 2.8km implementation during summer, there is a small increase in the overestimation close to the surface.

The outcome from FF performance comparison is similar, with no strong trend in the impact on the performance with the 7km version model. ME exhibits a steady behaviour with both resolutions in the winter, with overestimation that with the newer version in slightly reduced in the upper atmosphere. For the summer period, there is no difference with the coarse resolution models, while for the 2.8km resolution there is an increase in the overestimation almost for all levels during summer, while there is no difference in the performance during winter.

For the **single precision runs**, the verification was performed for the 5.06 version of the COSMO model at 7 km resolution, against the double precision run of the same version. The performance between the two precision schemes is almost identical for all surface parameters examined. For PS, a very small increase during summer for all hours is shown in the ME and RMSE values when the SP scheme is adopted.

Small changes in the performance are noticed between the two precision schemes (v5.06) with regards to upper air parameters. For RH, small changes are shown in the middle atmosphere, with no specific tendency in the error. For T, the differences between the two precision schemes for 5.06 are shown in the upper atmospheric levels, with a very small reduction during summer. The more noticeable differences are exhibited with FF, where there are small differences during the hours of the day for both months, more obvious during the winter with respect to the ME values in upper atmosphere.

It should be mentioned however, that the difference in the comparison between the two precision schemes for model version 5.06 are almost in all cases smaller than the differences between model versions of DP mode.

For the forecast of **total precipitation** (12h accumulation, figure 3), the statistics of the two versions of the model are quite similar. Small deterioration of FBI score is exhibited for high thresholds during summer and only in the 7km resolution model implementation while the opposite was shown in the 2.8km one. In general, underestimation of precipitation amounts for all thresholds, higher FAR and lower POD with increasing threshold. No significant resolution dependence of the performance of the new version with respect to the 5.05_1 one is noticed. With respect to the DP versus SP comparison, again the results are almost identical for July, with some differences in all scores in the higher threshold category.

During winter, the results are almost identical for both model versions at 7km resolution, with slightly more significant differences in FAR and FBI for the 20mm threshold, while for the 2.8km resolution more slightly larger differences can be noticed only in FAR, for the same threshold. Differences are insignificant between the DP and SP versions of v5.06.

For the forecast of precipitation (6h accumulation, figure 3), the statistics of the two versions of the model are, again, quite similar, with slightly more visible differences for the FBI score, especially in the higher threshold category for both periods and resolutions. With regards to the DP versus SP comparison, results are almost identical also for December, with some differences in FBI for the higher threshold categories.

Due to the post processing of precipitation grib files, the statistical indices cannot be compared to the optimum values they could acquire that will reveal the true performance of the model, but instead this comparison and conclusions are focusing on the relative performance of the two model versions.

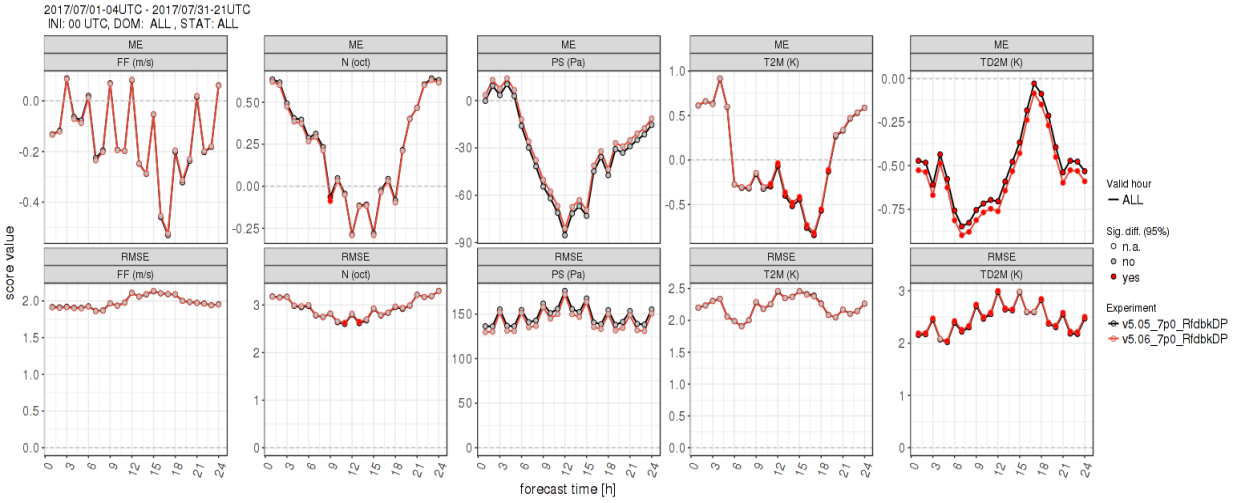
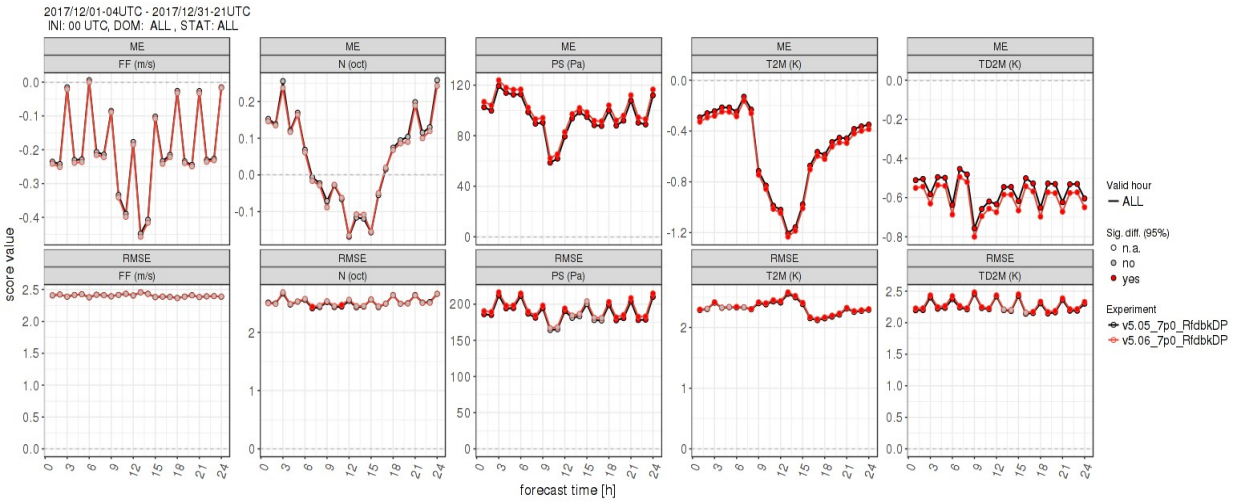


Fig. 2 COSMO-7km DP Continuous parameters verification results - COSMO 5.06 (red) and COSMO 5.05_1 (black) mean error (ME, first row) and root mean square error (RMSE, second row) for: July 2017 (top), Dec 2017 (bottom). Red/gray filled dots indicate a significant/insignificant (95% level) difference of scores between the 2 model versions. Parameters (from left to right): 10 m wind speed, surface pressure, 2 m temperature and 2 m dew point temperature.



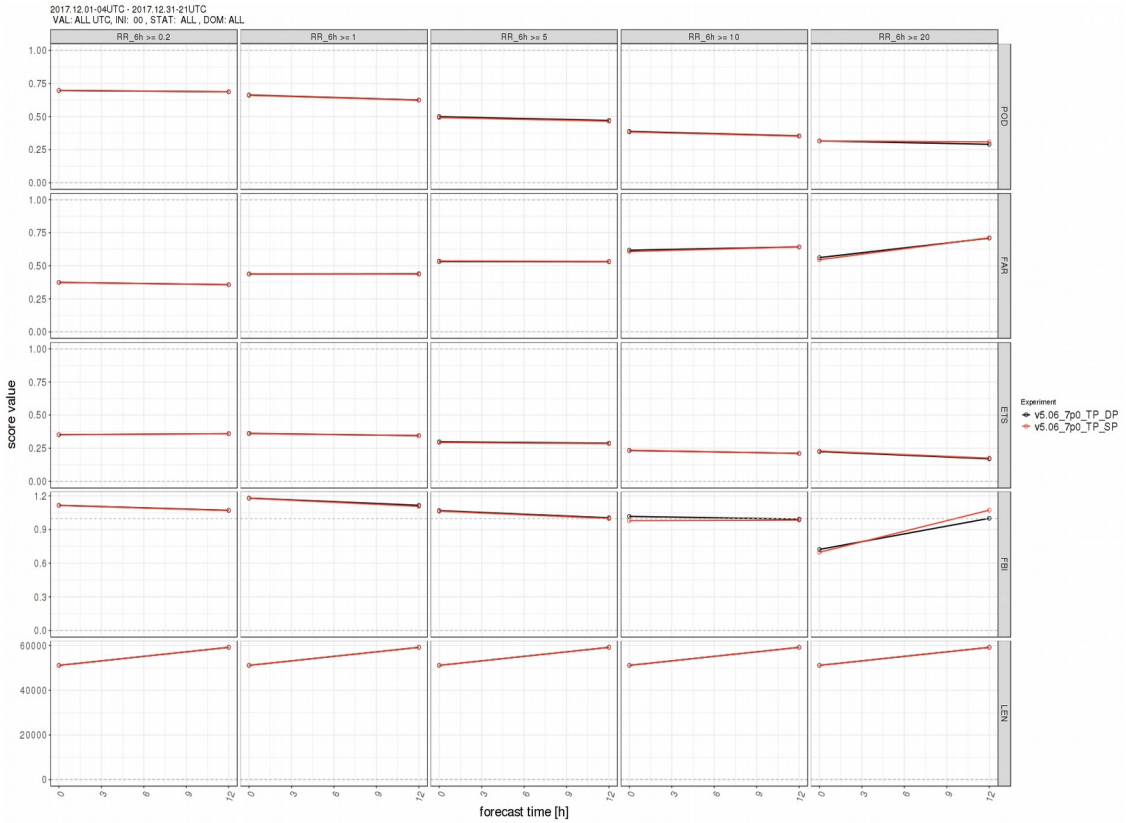
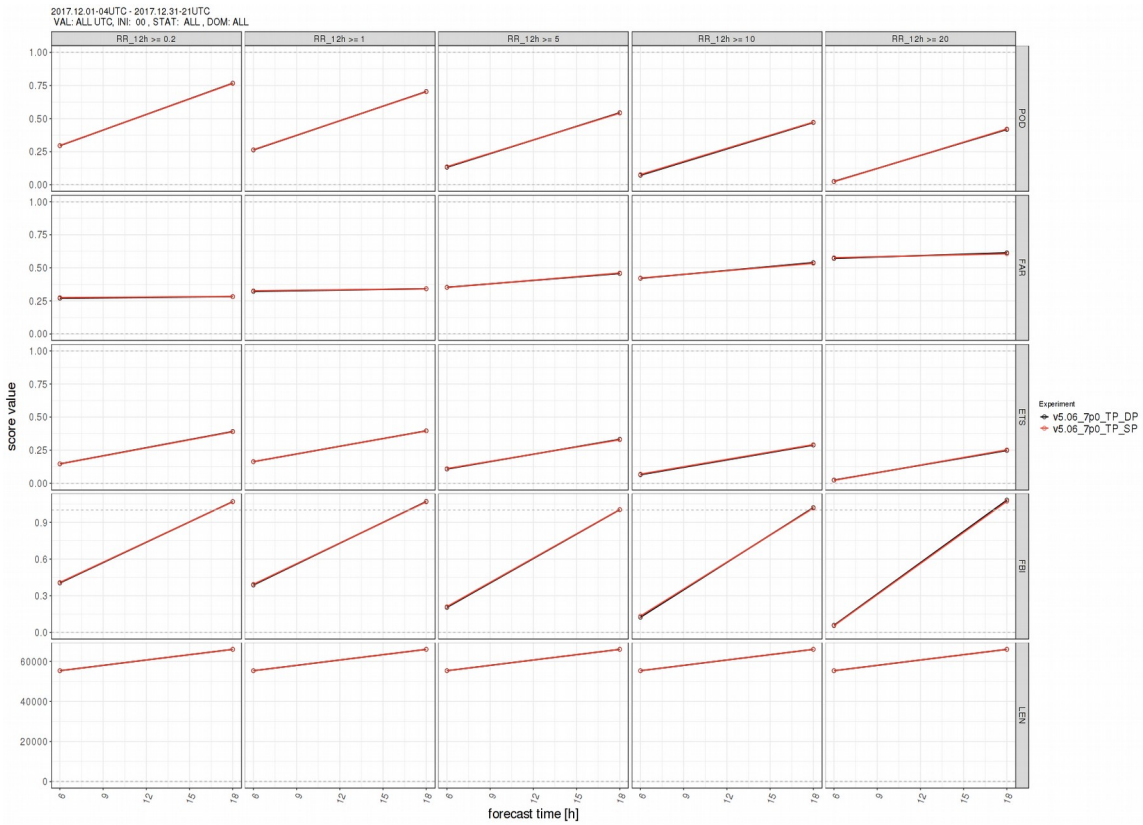


Fig. 3 COSMO-7km DP-SP precipitation verification results (00UTC run) for December 2017 - COSMO 5.06 double precision (black) and single precision (red): POD, FAR, ETS, FBI and LEN (top to bottom) for thresholds 0.2, 1, 5,10, 20mm (left to right), 6 hours cumulation interval (top), 12 hours cumulation interval (bottom)



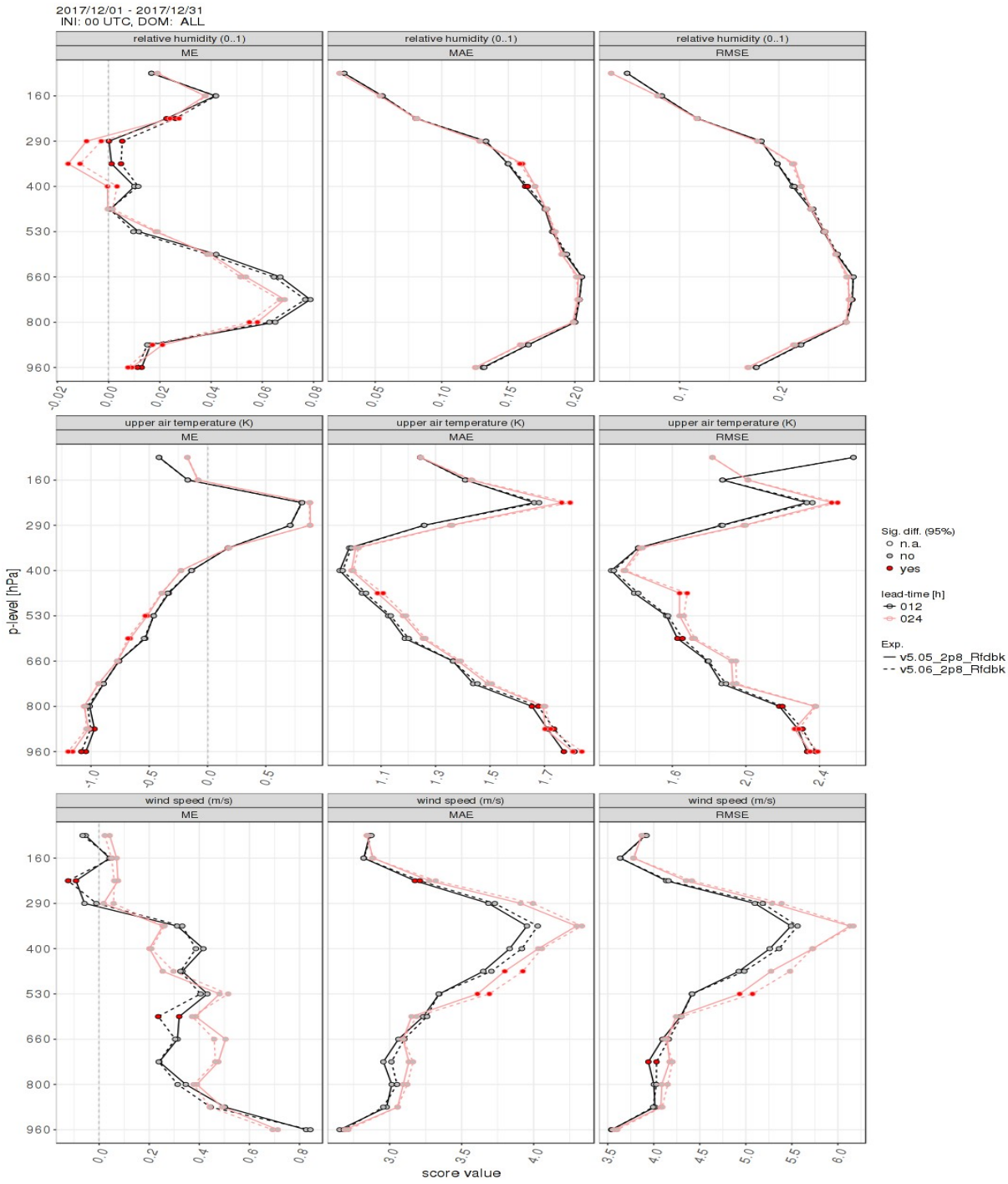


Fig. 4 COSMO-2.8km DP Upper Air parameters verification results - COSMO 5.06 (dashed line) and COSMO 5.05_1 (continuous line) mean error (ME, left), mean absolute error (MAE, center) and root mean square error (RMSE, right) for December 2017, +12 hours (black) and +24 hours (red). Parameters (top to bottom): relative humidity, temperature and wind speed.

Phase IV: Additional steps

Activities (including use of resources) to test the new official version (6.00) of the COSMO model prior to its release which is anticipated in the second part of the year.
Maintenance of the Test Suite.
Extending the Test Suite to evaluate ICON model official releases (migration to ICON-LAM: prepare an EcFlow suite on HPC, ICONTOOLS and ICON executable already available).