Common Area Verification Activity with Rfdbk/MEC Application: MAM 2020 First Results

A. Iriza-Burcă[1], J. Linkowska[2] and F. Fundel[3]

1. National Meteorological Administration, Bucharest, Romania
2. Institute of Meteorology and Water Management National Research Institute, Warsaw, Poland
3. Deutscher Wetterdienst, Offenbach am Main, Germany

1 Introduction

The Common Plots (CP) Verification Activity aims to evaluate the performance of the various operational model implementations of COSMO over the same geographical areas, based on the same set of observations. For these activities, a common verification tool called VERSUS is used, ensuring the use of the same verification practices. This allows for a fair comparison of the results and an easy assessment of model performance.

The CARMA priority project (PP CARMA) aims to replace VERSUS with the MEC-Rfdbk software developed by DWD, as a Common Verification Software. This software uses the Model Equivalent Calculator (MEC, [1]) to produce Feedback Files [2], which are then used by R verification scripts based on the Rfdbk package [3] to produce verification scores. The MEC-Rfdbk system is used operationally at DWD for the verification of both the COSMO [4] and ICON [5] model chains.

The advantages of this verification system proposed to be adopted for CP preparation activity include the data pre-processing where all the observations are quality controlled by data assimilation, co-located observational and model data using the same observation operator to produce one combined file, allowing a fast and simple calculation of standard verification scores and online visualization of results.

The first results of a cross model verification (with the MEC-Rfdbk software) for the Common Plot activities for the 2020 spring season (MAM 2020) are presented below. For these results, a set of three COSMO model runs are considered: COSMO-D2 (DWD), COSMO-PL (IMGW) and COSMO-RO (NMA).

2 Brief Description of Data and Methodology

COSMO 00 UTC model runs are evaluated, with forecast step every 3 hours. The horizontal resolution is 7km for the COSMO-PL and COSMO-RO models and 2.2km for COSMO-D2. The integration time for COSMO-PL and COSMO-RO is 72 hours, while for COSMO-D2 27 hours of forecast are available. As a consequence, scores are computed every 3 hours for 27 hours forecast when comparing all three models. However, for a comparison only between COSMO-PL and COSMO-RO, scores could be computed for 72 hours of forecast.

The integration domains for COSMO-D2 and COSMO-PL (figure 1, top row) are the operational ones (also evaluated for the official Common Plots activities), while the COSMO-RO integration domain for the current evaluation (figure 1, bottom row) differs from the operational set-up of the model employed in NMA. This was done in order to cover the Common Areas, which are otherwise outside of the COSMO-RO domain. No other modifications are done compared to the operational set-up of the model.

doi:10.5676/dwd_pub/nwv/cosmo-nl_20_04

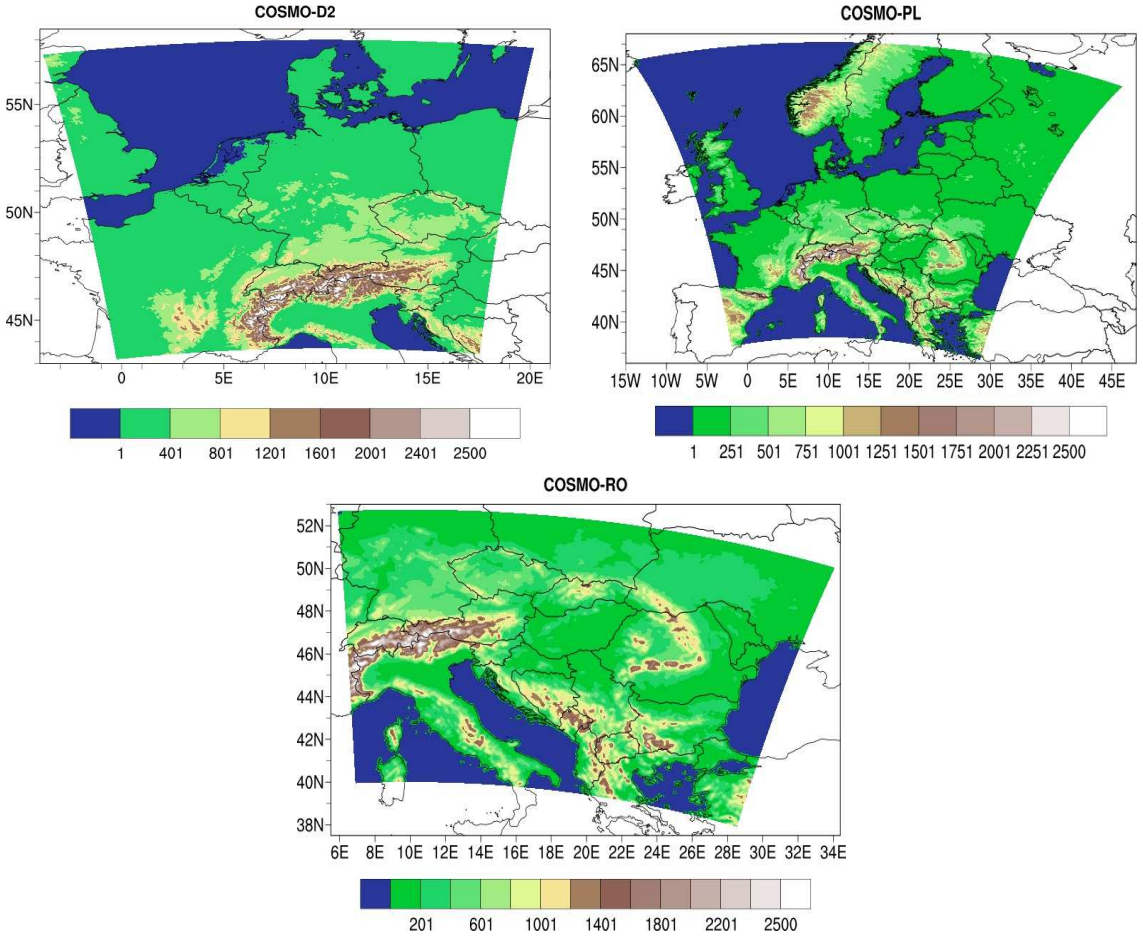

Figure 1: Integration domains for COSMO-D2 (top, left), COSMO-PL (top, right), COSMO-RO-7km (bottom) used for verification.

An overview of the model configurations can be seen in Table 1, while a more detailed description can be found on the COSMO web-site [6].

	res.	ie_tot	je_tot	ke_tot	ICLBC	DA	ant.
COSMO-D2	0.02	651	716	65	ICON-EU	KENDA,LHN	27h
COSMO-PL	0.0625	415	460	40	DAC/ICON	nudging	72h
COSMO-RO	0.0625	201	177	40	ICON	nudging	72h

Table 1: Overview of the model configurations.

As described in [7], COSMO models are generally verified over the same areas, with the same set of stations. Observations are retrieved in BUFR format from the MARS database and converted to netcdf format using the bufr2netcdf software [8]. The Rfdbk system allows to align the data, so that only those available for all evaluated experiments are used.

The two common areas for which the scores are computed are presented in figure 2 and include parts of Northern Italy, Austria, Slovenia, Croatia, Germany, Bosnia and Herzegovina, Hungary, Slovakia, and the Czech Republic.

For Common Area 1 (CA1, figure 2, left), 96 stations are selected, with altitudes between 0m and 753 meters. Although slightly more restricted geographically, Common Area 2 (CA2, figure 2, right) includes a larger number of stations than CA1. The stations in CA2 also cover a greater range of altitudes compared to CA1 (maximum altitude 3114m in CA2 compared to 753m in CA1). Table 2 offers an overview of station distribution for the two areas.

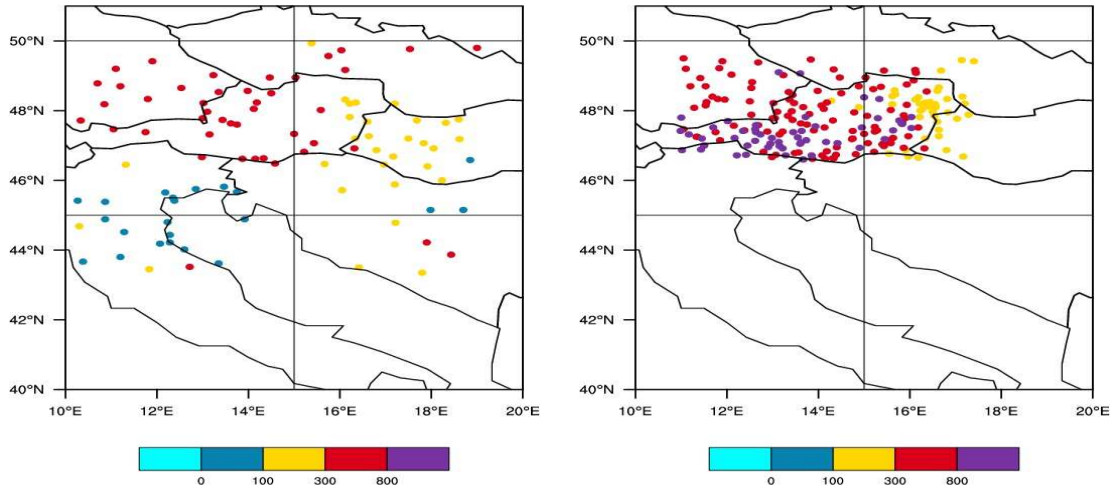

Figure 2: Location and altitude of synoptic stations in CA1 (left) and CA2 (right) used for the verification.

Common Area	Number of stations					Alt.	
	Total	alt≤100m	100m<alt≤300m	300m<alt≤800m	alt<800m	min	max
CA1	96	23	30	43	0	0m	753m
CA2	205	2	41	107	55	0m	3114m

Table 2: Overview of stations distribution for the 2 Common Areas, depending on altitude.

The results presented in this paper are computed taking into account all stations of interest. However, results are also available stratified by station altitude.

Standard Verification (seasonal) was performed for the following continuous parameters: 2 meter temperature – T2M (deg K), 2 meter dew point temperature - TD2M (deg K), surface pressure - PS (Pa), total cloud cover - N (octa), 10 meter wind speed - FF (m/s), wind direction - DD (deg) and wind gust - Gust (1 hour, m/s). Scores for these parameters include the mean error (ME), root mean squared error (RMSE) and mean absolute error (MAE).

Categorical scores are also computed, for the following parameters: 6 hour accumulated precipitation - RR_6h (thresholds: 0.2, 1, 5, 10, 20 mm/6h), total cloud cover (thresholds >=1, >=4 and >=7) and 10 meter wind gust (thresholds: >=12.5, >=15, >=20 m/s). Dichotomic scores computed for these variables include the probability of detection (POD), false alarm rate (FAR), equitable threat score (ETS), frequency bias (FBI). Other scores based on the number of hits, misses, false alarms and correct negatives respectively are available but not shown.

Both for the standard verification and the dichotomic scores, the number of observations used in computations (LEN) can be shown. General information of scores and a detailed description of the common plot verification procedures can be found in [7]. The verification results presented below (figures 3-12) are a sample of the derived statistics that were obtained. A complete set of statistical scores obtained with the MEC-Rfdbk verification system for the three models considered here is available on the COSMO web-site [6].

3 Considerations Regarding the MEC-Rfdbk System

As mentioned before, currently, CP verification activities are carried out using the VERSUS verification software environment. Because of the technical limitations of VERSUS and lack of further development, it has been decided within the consortium that it should be replaced with the MEC-Rfdbk software. The MEC-Rfdbk system uses small files for fast calculation of verification scores, while the results can be browsed interactively online. The verification is based on the use of feedback files, that hold information on observations (including meta-data) and their usage in data assimilation, as well as the corresponding model analysis, first-guess and past forecasts, in NetCDF format. These files are produced for each valid time and observation type and can be used for various verification tasks.

MEC [1] is a Fortran 2003/2008 and C - based binary that produces feedback files using observations in netcdf format and model data in grib format. These files can be produced for any model (COSMO, ICON, IFS, etc.). Some mandatory parameters from the model of interest must be available on all model levels: PS, T, U, V, P, Q, while others are optional, depending on the available observations and user needs (T2M, TD2M, PS, N, FF, DD, Gust, RR, etc). MEC applies the observation operators from the data assimilation scheme to model fields and stores the results in feedback-files. The software can use as input observations, model runs (deterministic or ensemble), analysis runs or even another MEC run, depending on the user needs and can be applied to interpolate between two or more time periods [9].

The advantages of using the MEC software as part of the verification system for CP activities are related to data pre-processing (all data in one place) and ensuring observation and forecasts are correctly assigned to each other, with quality control done by data assimilation.

The Rfdbk package [3] is an R interface that aims to exploit the information contained in the feedback files and can be used to perform feedback file based verification. Rfdbk is the basis to a set of verification scripts that use as input the feedback files obtained from MEC (one file for each validity date and observation type) and outputs score files (again, for each validity date and observation type). Based on Rfdbk, verification scripts are available for various types of observations (SYNOP, radiosondes, radio occultation, aircraft, wind profiler and so on). Continuous scores are computed for various types of observations, while categorical scores are also available for SYNOP observations. Verification can be performed for deterministic runs (forecast or hindcast) or ensemble, for any model (COSMO, ICON, IFS, etc.), while cross model verification (e.g. COSMO vs. ICON vs. IFS) is also possible. Rfdbk -based scripts can be used for:

- domain average verification (function of forecast lead-time for a user defined verification period), including domain stratification,
- time series (function of valid time in the verification period or as a function of the forecast lead-time)
- station based verification (function of observation station).
- aggregation on sub-domains, height bins, levels or periods
- significance test
- conditional verification

The advantage of using Rfdbk based verification scripts is their flexibility, which means they can be modified and adjusted according to the needs of each user and (UNIX) system. Finally, the centralized, online, interactive visualization of the results on the COSMO web-site using the R Shiny server [10] allows for an easier evaluation of the results.

4 First Results

Results for continuous parameters - Common Area 1 (figures 3 and 4)

For 2 meter temperature, similar behaviour can be observed for all three models, that is overestimation of forecasted values during night hours and underestimation during the day. The amplitude of errors is of about 2 K, with slightly higher error amplitude for COSMO-NMA during the day (+15, +18 hours).

With regards to 2 meter dew point temperature, there is a similar behaviour for COSMO-D2 and COSMO-RO, mainly overall overestimation of forecasted values, with higher errors during the afternoon. Smaller errors are registered for COSMO-PL compared to the other 2 models, with overestimation of temperatures for the first forecast interval (up to +9hours) and for +18 hours anticipation, while for the remaining intervals the tendency of the model is that of underestimating the values for this parameter. The amplitude of errors slightly larger than that for 2 meter temperature, again with slightly higher error amplitude for COSMO-NMA during the day (+15, +18 hours).

An underestimation of surface pressure values can be observed during the day for all three models, especially COSM-PL The latter also exhibits a larger amplitude of errors in the afternoon.

On the other hand, total cloud cover values are overestimation by all three models, especially COSM-DE, while the amplitude of errors is similar for all models considered.

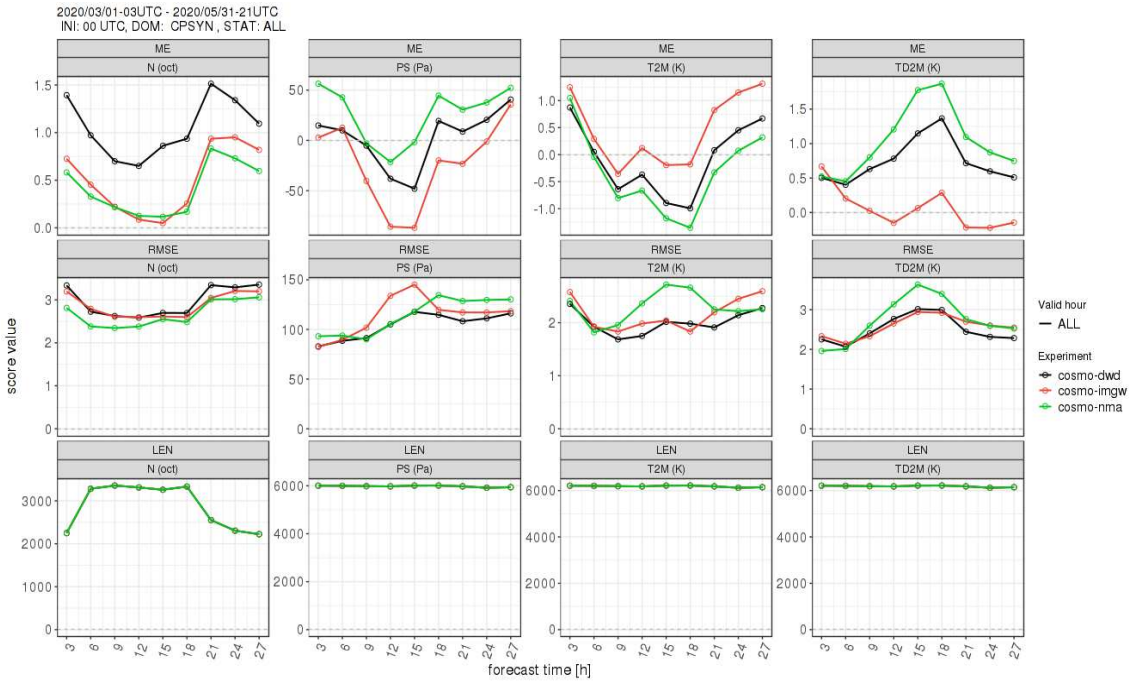

Figure 3: ME (top row), RMSE (middle row) and LEN (bottom row) values for CA1; left to right: N, PS, T2M, TD2M. COSMO-D2 (black), COSMO-PL (red) and COSMO-RO (green).

With regards to wind related parameters (FF, DD, Gust, figure 4), all three models exhibit a general tendency to overestimate forecasted values compared to observations, with slight underestimation of wind speed values during the day. In this case, the amplitude of errors is low and comparable for all three models. The varying number of observations (LEN) for DD is due to the limit of $FF > 3m/s$ which is imposed to eliminate unreliable wind direction observations in case of too little wind speed.

Results for continuous parameters - Common Area 2 (figures 5 and 6)

Again, for 2 meter temperature similar behaviour for all three models is observed also for CA2, mainly overestimation of forecasted values during night hours and underestimation during the day; with slightly higher error amplitude for COSMO-NMA during the day (+15, +18 hours).

Also for 2 meter dew point temperature there is a similar behaviour between all three models, with and overall overestimation. Generally, higher errors are registered in the afternoon. As for CA1, there are smaller errors from COSMO-PL compared to the other 2 models, with slight underestimation starting with +21 hours anticipation.

Both COSM-PL and COSMO-D2 mostly underestimate surface pressure values during the day, while COSMO-RO exhibits a behaviour of overestimation for this parameter for the entire forecast period, also with slightly

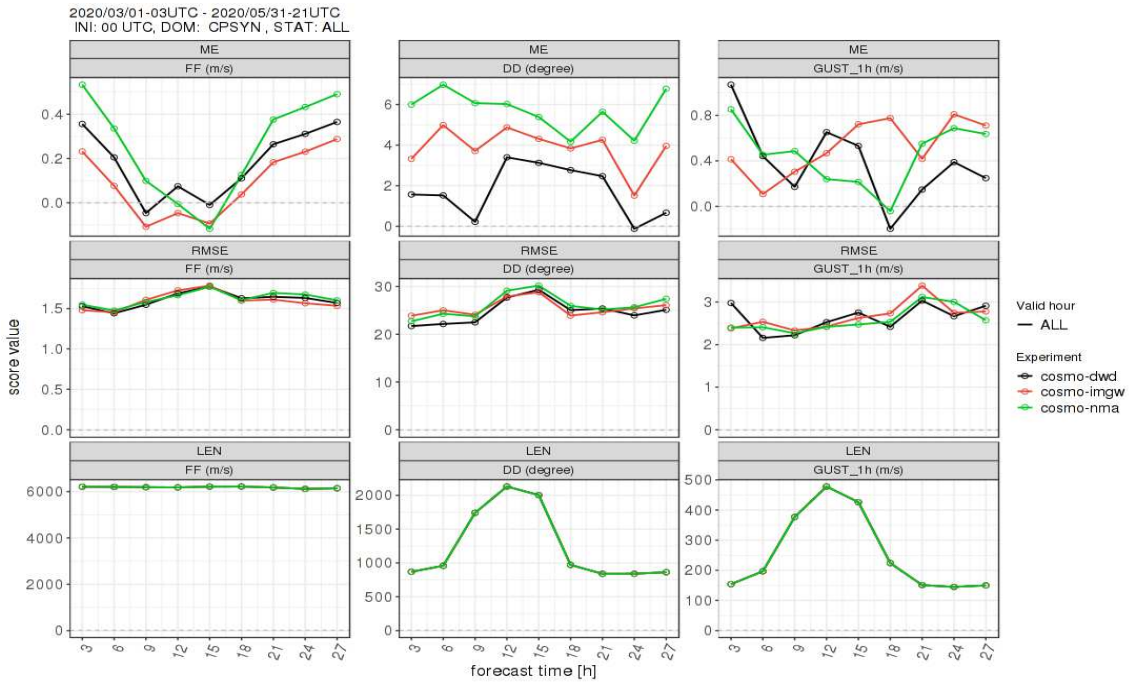

Figure 4: ME (top row), RMSE (middle row) and LEN (bottom row) values for CA1; left to right: FF, DD, Gust. COSMO-D2 (black), COSMO-PL (red) and COSMO-RO (green).

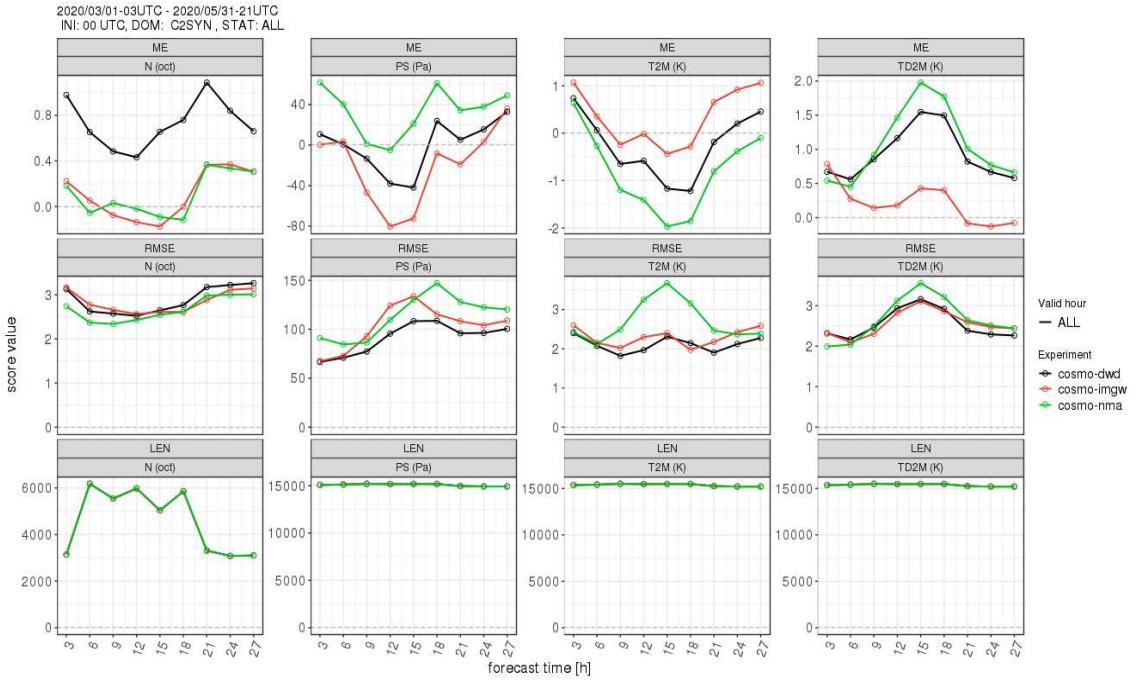

Figure 5: ME (top row), RMSE (middle row) and LEN (bottom row) values for CA2; left to right: N, PS, T2M, TD2M. COSMO-D2 (black), COSMO-PL (red) and COSMO-RO (green).

larger amplitude of errors.

Overestimation of total cloud cover values from COSM-DE is shown for the entire interval; for the other two models, there is a slight underestimation of total cloud cover values during the day, while amplitude of errors is comparable between the three.

For 10 meter wind direction (for CA2), an overestimation from COSMO-PL and COSMO-RO for the entire forecast interval is exhibited, while for COSMO-D2 this can be seen only during the day. 10 meter wind

speed forecasts have a very low error. The general tendency is that of overestimation of the values forecasted for this parameter from COSMO-D2. COSMO-RO overestimated the values forecasted for this parameter only during the day, while from COSMO-PL we notice underestimations. Overestimation of wind gust values is seen during the night and early morning for all three models; during the day, the general tendency is that of overestimation of values from COSMO-PL and underestimation for COSMO-RO. For these last three parameters, the amplitudes of errors are comparable in all analyzed models.

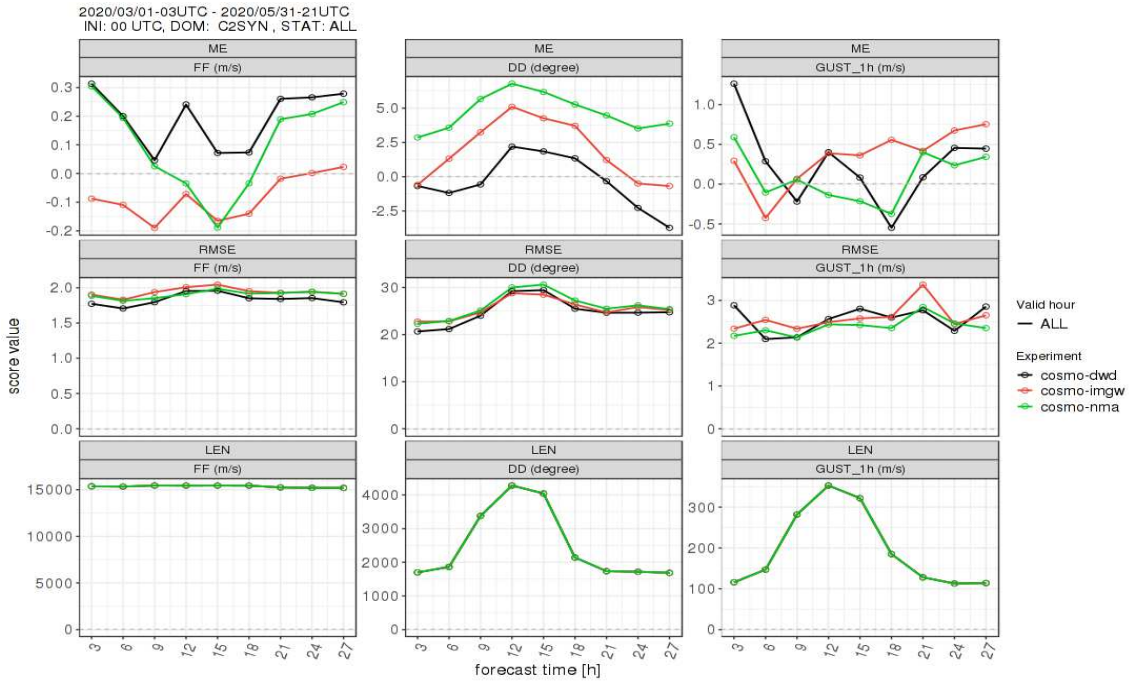

Figure 6: ME (top row), RMSE (middle row) and LEN (bottom row) values for CA2; left to right: FF, DD, Gust. COSMO-D2 (black), COSMO-PL (red) and COSMO-RO (green).

As mentioned before, the observations used to compute the scores are the same for all three models, with lower number of observations for N, DD and especially Gust.

Categorical scores for cloud cover (figures 7 and 8) suggest a a high probability of detection from all three models, for both common areas (all thresholds). A similar behaviour can be observed from all three models, with slightly better performance from COSMO-D2, especially for the highest threshold. POD values are consistent with results for FAR, that show a low false alarm rate, again for all three models and both common areas. Slightly higher false alarm rates were obtained when forecasting high cloud cover values. FBI values show a good performance from all three models for this parameter, with a slight tendency to overcasting especially for the higher cloud cover values.

Categorical scores for 10 meter wind gust (1 hour) computed for CA1 (figure 9) show a generally higher probability of detection for the 12.5 m/s threshold than for the 15m/s one. For the first threshold, highest POD values are obtained during the day, with lower values starting with +21 hours anticipation, especially for COSMO-D2. For CA2 (figure 9), a lower probability of detection is observed especially for the 15m/s threshold. A high false alarm rate from all three models can be seen for the 15m/s threshold, especially from COSMO-RO for CA1 and all three models for CA2 in the second part of the considered forecast interval. For this threshold, a slightly better behaviour is exhibited by COSMO-RO for the first hours of forecast, while for the 12.5m/s threshold, the behaviour of the three models is similar for both areas. Values for the FBI score suggest a general tendency of overcasting from the three models for both areas, especially for the 12.5m/s threshold, with some undercasting of wind gust frequencies from COSMO-RO for the 15m/s threshold for the first anticipations, especially for CA2. For both areas, the behaviour of the model (mainly for the 12.5 m/s threshold) is more similar between COSMO-D2 and COSMO-RO.

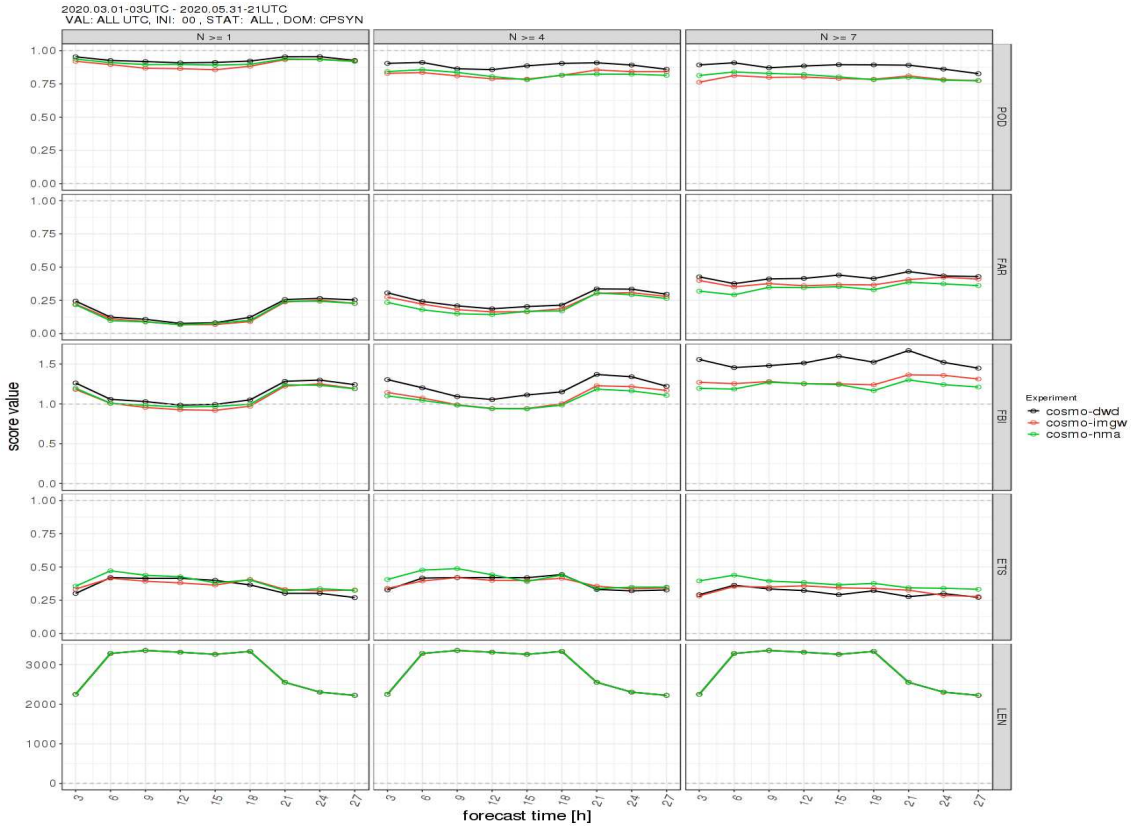

Figure 7: Categorical scores for N (CA1); top to bottom: POD, FAR, FBI, ETS and LEN for CA1; left to right: cloud cover >=1, >=4, >=7. COSMO-D2 (black), COSMO-PL (red) and COSMO-RO (green).

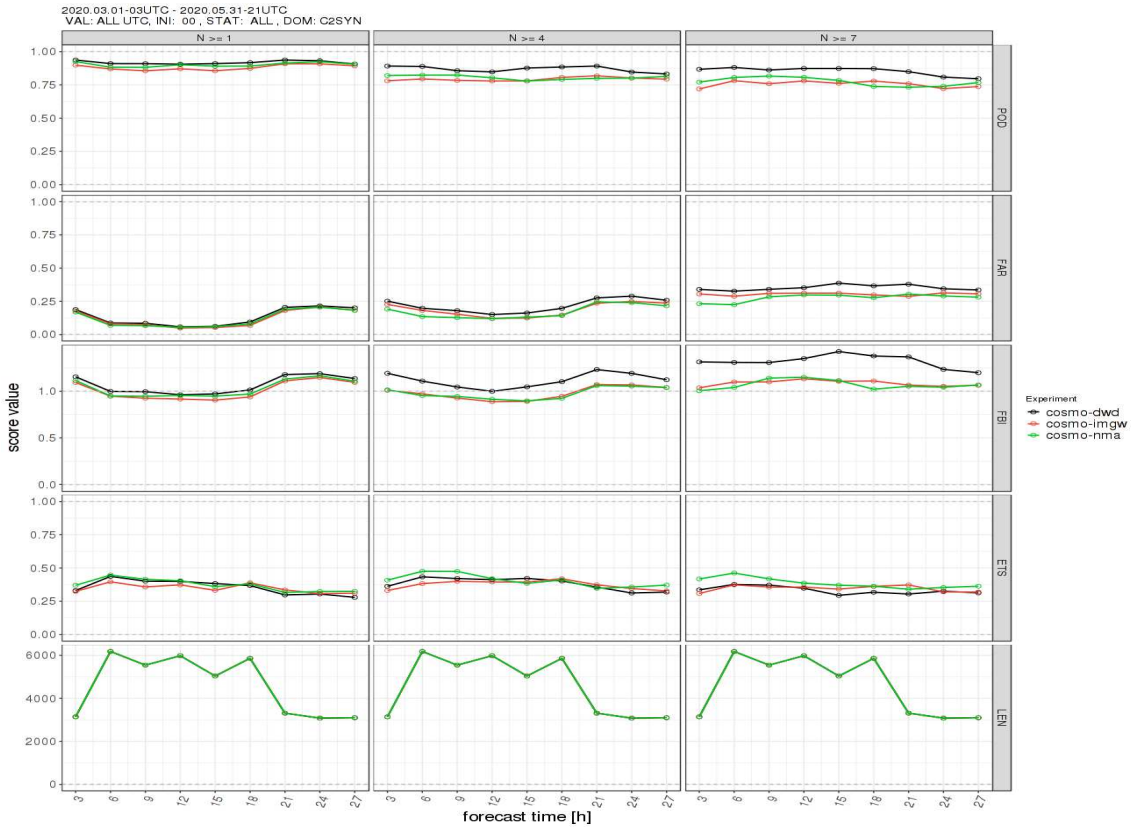

Figure 8: Categorical scores for N (CA2); top to bottom: POD, FAR, FBI, ETS and LEN for CA2; left to right: cloud cover >=1, >=4, >=7. COSMO-D2 (black), COSMO-PL (red) and COSMO-RO (green).

Figure 9: Categorical scores for Gust (CA1); top to bottom: POD, FAR, FBI, ETS and LEN for CA1; left to right: cloud cover >=1, >=4, >=7. COSMO-D2 (black), COSMO-PL (red) and COSMO-RO (green).

Figure 10: Categorical scores for Gust (CA2); top to bottom: POD, FAR, FBI, ETS and LEN for CA2; left to right: cloud cover >=1, >=4, >=7. COSMO-D2 (black), COSMO-PL (red) and COSMO-RO (green).

Categorical scores for 6-hour accumulated precipitation are similar for both common areas (figures 11 and 12). Probability of detection values are better for the lower thresholds (0.2 mm/6h, 1 mm/6h) and worsen with the increase of the thresholds. POD values are consistent with results for FAR, indicating a low false alarm rate for the lower thresholds, with slightly better scores for COSMO-PL. FAR scores also worsen with the increase of threshold (both areas), with slightly smaller false alarm rates from COSMO-RO. FBI values suggest a general tendency to overcasting from all three models, with some undercasting from COSMO-PL and COSMO-RO for the higher thresholds and a slightly better performance from COSMO-D2. ETS values indicate a low skill for the upper thresholds (10mm/6h, 20mm/6h) and forecast quality drops significantly.

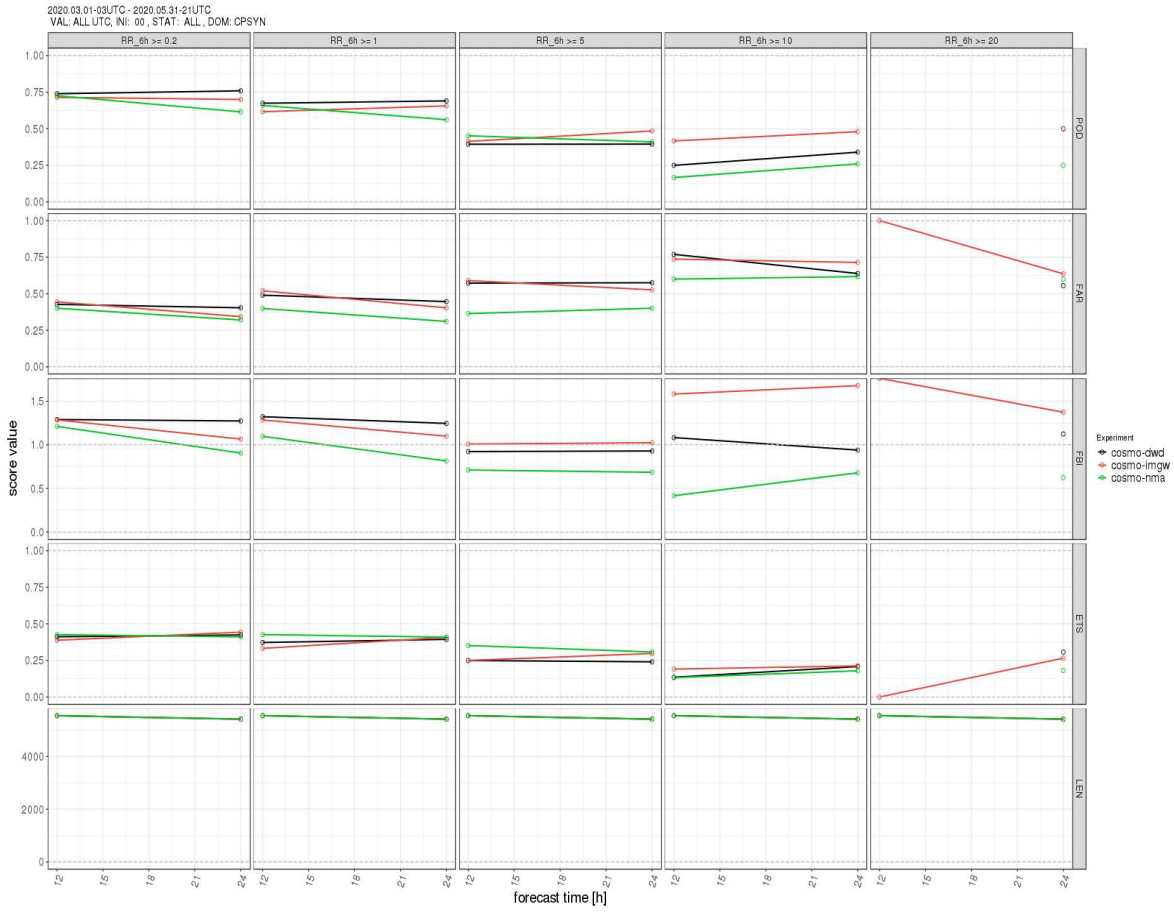

Figure 11: Categorical scores for 6-hour accumulated precipitation (CA1); top to bottom: POD, FAR, FBI, ETS and LEN for CA1; left to right: cloud cover >=1, >=4, >=7. COSMO-D2 (black), COSMO-PL (red) and COSMO-RO (green).

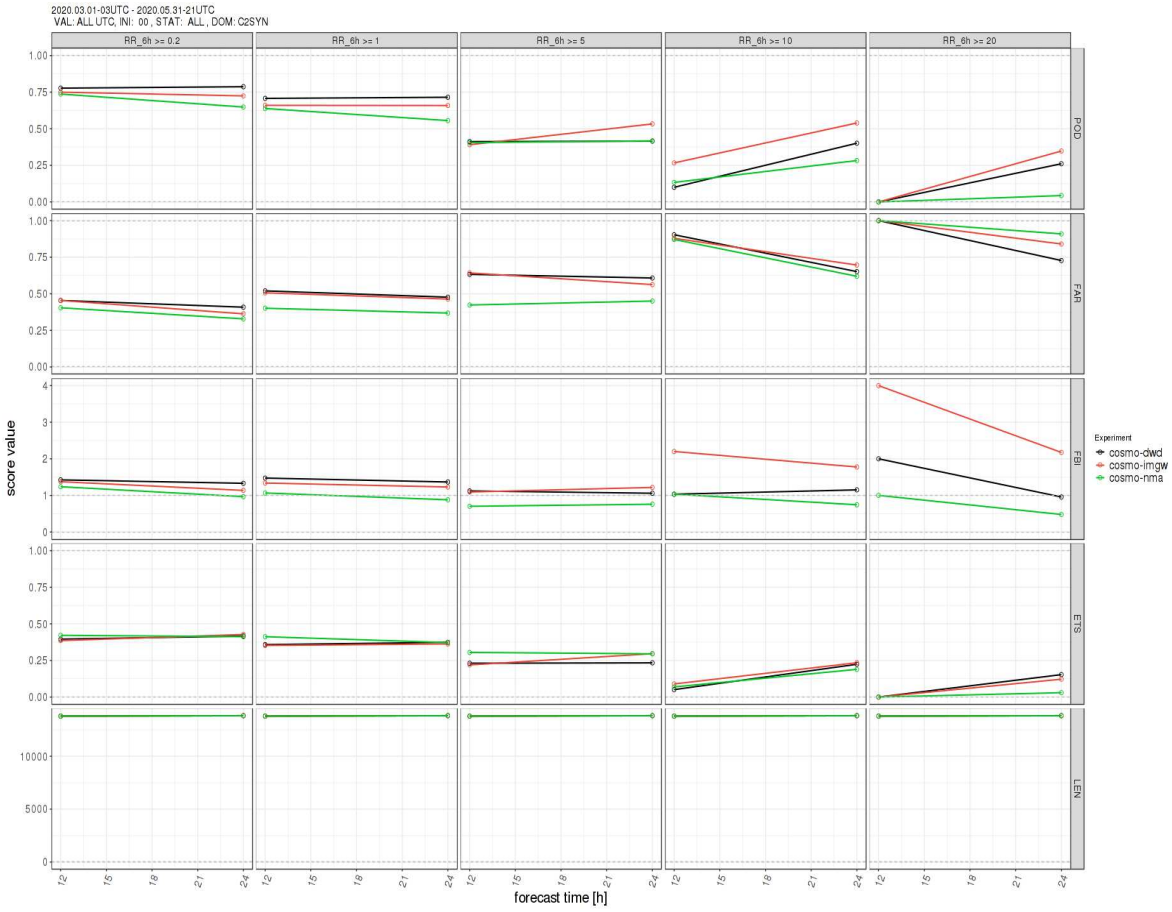

Figure 12: Categorical scores for 6-hour accumulated precipitation (CA2); top to bottom: POD, FAR, FBI, ETS and LEN for CA2; left to right: cloud cover >=1, >=4, >=7. COSMO-D2 (black), COSMO-PL (red) and COSMO-RO (green).

5 Conclusions and Outlook

Evaluation of operational COSMO (or ICON-LAM) model implementations in each service is part of the Common Plots Verification Activity, currently performed with the VERSUS software. These activities offer the opportunity to assess the performance of various operational COSMO (and ICON) model implementations over the same geographical areas, with the same set of observations. The existing VERSUS verification software environment is being replaced with the MEC-Rfdbk system as a Common Verification Software, with the latter being currently under implementation in all the member countries of the consortium.

For this purpose, cross model verification for the Common Plot activities for the 2020 spring season (MAM 2020) were performed with the MEC-Rfdbk software, with comparative results from three models already available. The verification activities with MEC-Rfdbk are carried out following the criteria and evaluation scores from the Common Plot tasks, in order to test the implementation of the new verification system.

The first results presented in this study show the comparative evaluation of three (COSMO) operational suites, while more operational COSMO (and ICON) model implementations will be further included in order to obtain a performance overview similar to that currently available from the Common Plot activities.

References

[1] Rhodin, A., 2015: MEC Manual, *DWD*

[2] Feedback-file definition, 2012: Supplementary Documentation, *DWD*

[3] Fundel, F., 2020: Feedback File Verification Suite at DWD, *DWD*

[4] Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., Reinhardt, T., 2011: Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities. *Mon. Wea. Rev.*, 139, **3887–3905**, https://doi.org/10.1175/MWR-D-10-05013

[5] Zängl, G., Reinert, D., Ripodas, P., Baldauf, M., 2015: The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core. *Quart. J. of the Royal Met. Soc.*, 141(687), **563–579**, https://doi.org/10.1002/qj.2378

[6] Consortium for Small-scale Modeling web-site: www.cosmo-model.org

[7] Consortium for Small-scale Modeling Verification tasks page: http://cosmo-model.org/content/tasks/verification.priv/default.htm

[8] Patruno, P., Cesari, D., 2011: Wreport-bufr2netcdf: a free library and tools for decoding BUFR reports and creating input files for COSMO-Model assimilation, 13th COSMO General Meeting, 5-9 September 2011, Rome (Italy), Parallel session: WG1 and PP KENDA

[9] Potthast, R., 2018: Introduction to Using MEC, datool, BACY and DACE, *DWD*

[10] Shiny Server page https://rstudio.com/products/shiny/shiny-server/