COSMO/ICON-LAM evaluation over Common Areas: 2021-2022

COSMO WG5: Verification and case studies

Flora Gofa (fgofa@hnms.gr)

1 Overview

Verification results of statistical indices for main weather parameters are derived using the operational COSMO and ICON-LAM model implementations in each service. The domain (common), the resolution, the statistical scores/methods and the graphical representation approaches, are decided on an annual basis from WG5. A common verification software is used for both point wise and neighborhood approach verification which allows for a homogeneous, standardized and objective way to apply, calculate and present the verification scores. The outcome of this activity provides a basis to monitor the performance of the operational models implementation and track the systematic errors. Since the introduction of ICON-LAM in the operational forecast procedure of some services, special focus is given to the relative performance of the two models. In this report, statistical results of JJA-2021 up to MAM 2022 model performance are presented.

COSMO consortium has developed a standardized procedure for assessing the performance of its partners models, which involves evaluating verification scores across common areas, using the same observations and methods, and since this year, the same verification software. The verification results for key weather parameters, generated using the operational COSMO and ICON-LAM model implementations, are compared in each service, while the decision about the specific domain, resolution, statistical scores and methods, as well as the graphical representation of the scores is made on an annual bases by WG5. The results of this analysis, along with long-term trends, are presented every September during the GM plenary session, providing a means to track model performance. The use of common verification software ensures a standardized and objective approach to applying, calculating, and presenting verification scores, with observation data preparation and seasonal statistics calculated according to guidelines (`https://www.cosmo-model.org/content/tasks/verification.priv/common/reports/CP-2021-2022.pdf`)Guidelines 2021/22 developed annually by WG5. ICON-LAM models statistical results are included from any of the various services that use the model operationally.
The verification approach is primarily based on point-wise comparison that is performed with the use of Feedback Files generated by MEC software and analyzed from Rfdbk R-based libraries. For 2021-2022, conditional verification is also applied on temperature forecasts, meaning the interdependency of temperature and cloudiness with respect to model performance. Finally, neighborhood (spatial) methods are applied to precipitation fields and this year also to total cloudiness as it is shown on following paragraph. Selective verification results of COSMO and ICON-LAM models over Common Area 1 (ComA1) and Area 2 (ComA2) are presented below while the complete selection of statistical results can be found in `https://www.cosmo-model.org/content/tasks/verification.priv/common/reports/CP-2021-2022.pdf`Guidelines 2021/22.

2 Areas of Verification

The areas and specifications for model performance evaluation are presented below. In ComA-1, models with coarser resolution are included, while the higher resolution COSMO and ICON-LAM models are compared over ComA-2 (Table below).

doi:10.5676/dwd_pub/nwv/cosmo-nl_22_03

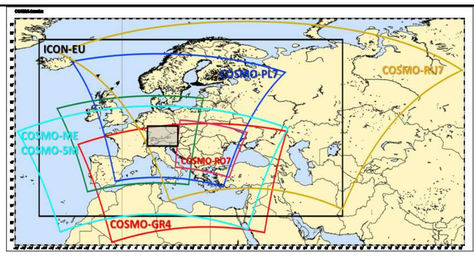
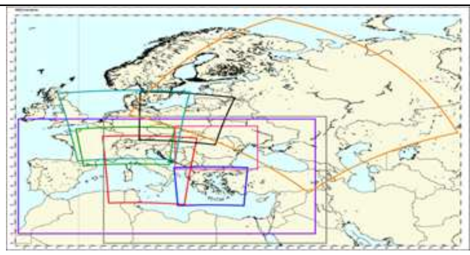
ComA-1 Area/Specs	
	00UTC Forecast runs Forecast Horizon: 72h Seasonal: JJA20, SON20, DJF21, MAM21
Models	**Global LAMS** ICON global DWD: ICON-EU COMET: COSMO-ME HNMS: COSMO-GR4 ARPA-E: COSMO-5M IMGW-PIB: COSMO-PL7
ComA-2 Area/Specs	
	W10.963, S46.597, E17.437, N49.550 00UTC Forecast run Forecast Horizon: 48h Seasonal: JJA20, SON20, DJF21, MAM21
Models	**LAMS** DWD: ICON-D2 COMET: COSMO-IT, ICON-IT HNMS: ICON-GR2.5 ARPA-E: COSMO-2I IMGW-PIB: COSMO-PL2.8, ICON-PL2.5 MCH: COSMO-1E, COSMO-2E IMS: ICON-IMS

Table 1: Specifications of ComA-1, ComA-2 verification areas

3 Selective Verification Results

3.1 Point-wise verification for Com-A1

The models are evaluated in terms of Mean Error and Root Mean Square Error indices for the continuous parameters, using SYNOP observations over the Common Areas 1 and 2 domains. Summary plots of main weather parameters for coarser models are shown in Figure 1, selectively for DJF2022 season. For ICON models (ICON, ICON-EU, ICON-RU), the bias diurnal cycle is weaker and RMSE values are reduced for T2m. RMSE is wind speed is slightly higher in warm hours of the day and comparable for all models in all seasons. The nighttime WS overestimation however which is a typical systematic error with COSMO models, is not apparent on ICON-LAM models, which exhibit a much weaker ME daily cycle, and an almost underestimation. RMSE for PS is reduced for ICON models. However, the tendency of PS RMSE increase with forecast time for all seasons appears also with ICON models while the bias in DJF, exhibits negative values for ICON models, in contrast to positive for COSMO. TCC RMSE values and bias diurnal cycles are similar in behavior with reduced RMSE values for ICON models with RMSE maximum values at night. With respect to ME however, ICON models produce higher underestimation in DJF during midday and slightly higher overestimation at night, for all seasons.

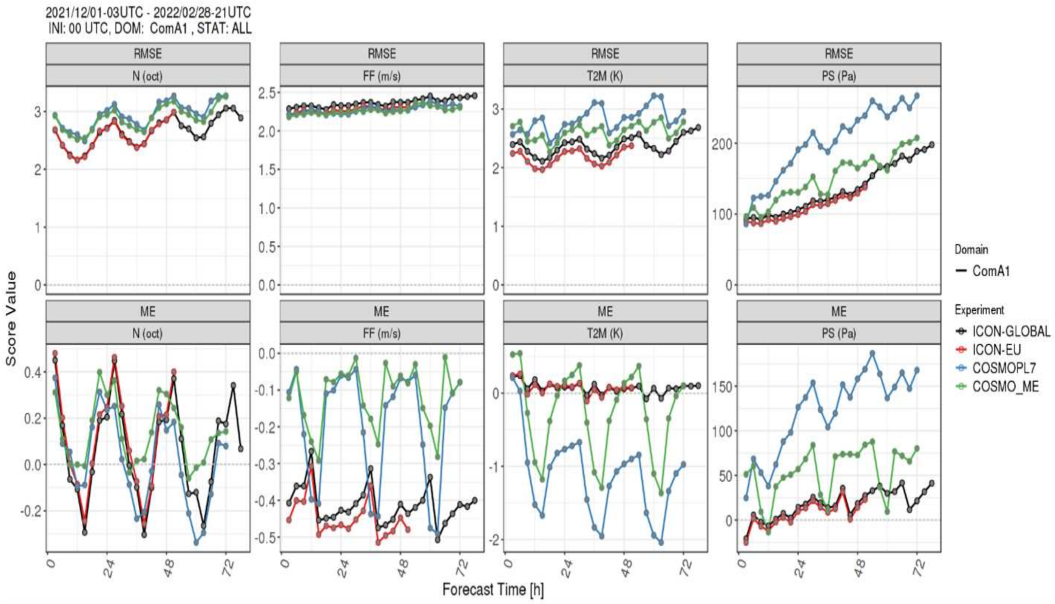


Figure 1: RMSE (first row) and ME (second row) indices for (from left to right) Total Could Cover, Wind Speed, 2m Temperaure and Pressure calculated over ComA-1 (DJF2022).

3.2 Point-wise verification for ComA-2

For a better comparison among COSMO and ICON-LAM higher resolution models over ComA-2, they are grouped together in order to detect general tendencies or differences that can be attributed to the various implementations. Below statistics for selected parameters are presented for DJF 2022.

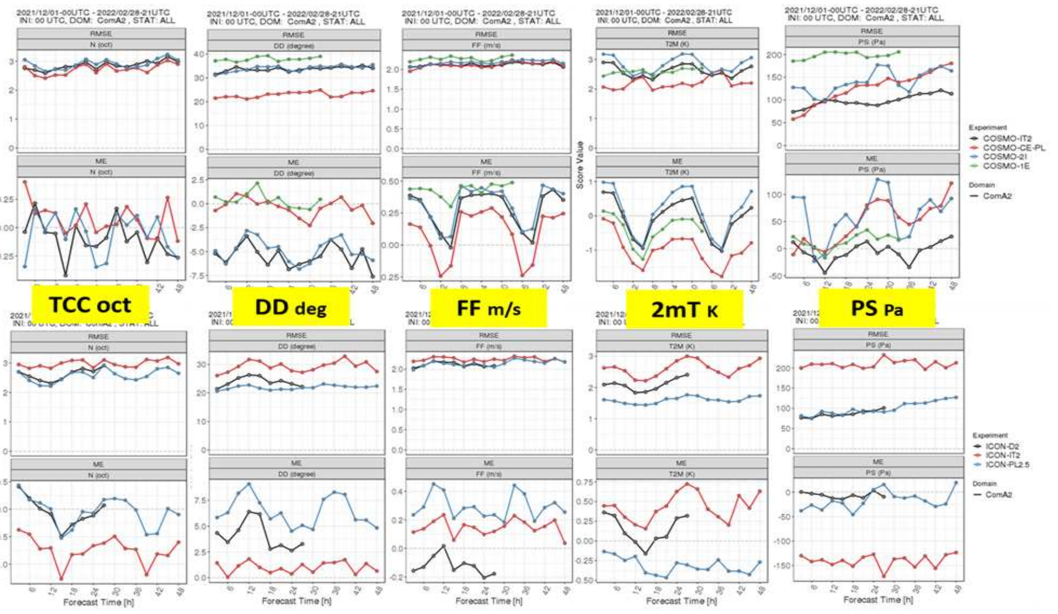


Figure 2: RMSE (first row) and ME (second row) indices for (from left to right) Total Could Cover, Wind dir, Wind Speed, T2M and Pressure calculated over ComA-2 (DJF2022)

For 2mT, one distinct difference, which is also consistent with coarser resolution models, is the ME reduced diurnal variability for ICON-LAM models, which is shown in DJF, with values closer to zero. The RMSE diurnal variation amplitude is comparable for the two sets of models with RMSE values for ICON models exhibiting an overall reduction.

For wind speed, The ME diurnal variation is weaker for ICON models and the tendency for overestimation is not as distinct with ICON-LAM implementations. Moreover, the bias diurnal variability is shifted among the two sets of models, with ICON overestimation in the early morning hours, while COSMO models bias is positive around evening hours.

The RMSE error cycle and range are similar for all models and no clear impact can be shown with either COSMO or ICON-LAM models. For wind direction, the reduction in error can be partially associated with Pressure error reduction. For Total Could Cover, the performance trend is not clear among ICON-LAMs as there is a significant spread while RMSE values are similar for both sets of models.

3.3 Wind performance

In this section, special focus is given to wind related properties in terms of their performance over ComA-1 and ComA-2 domains. Specifically, RMSE and ME indices are calculated for wind speed, wind direction and hourly wind gust.

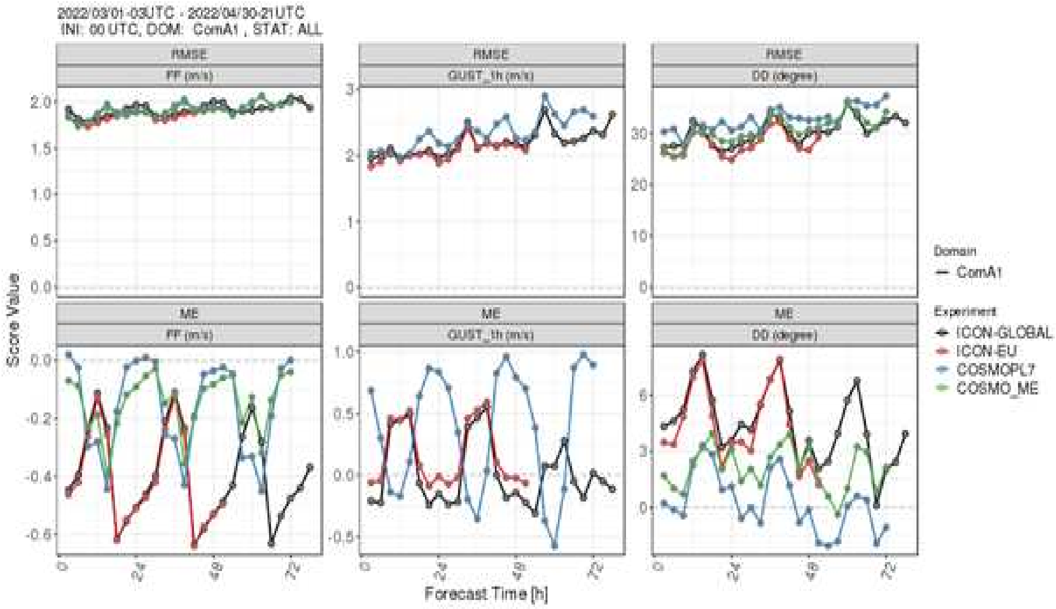


Figure 3: RMSE (first row) and ME (second row) indices for Wind speed, Wind Gust and Wind direction over ComA-1 (MAM2022).

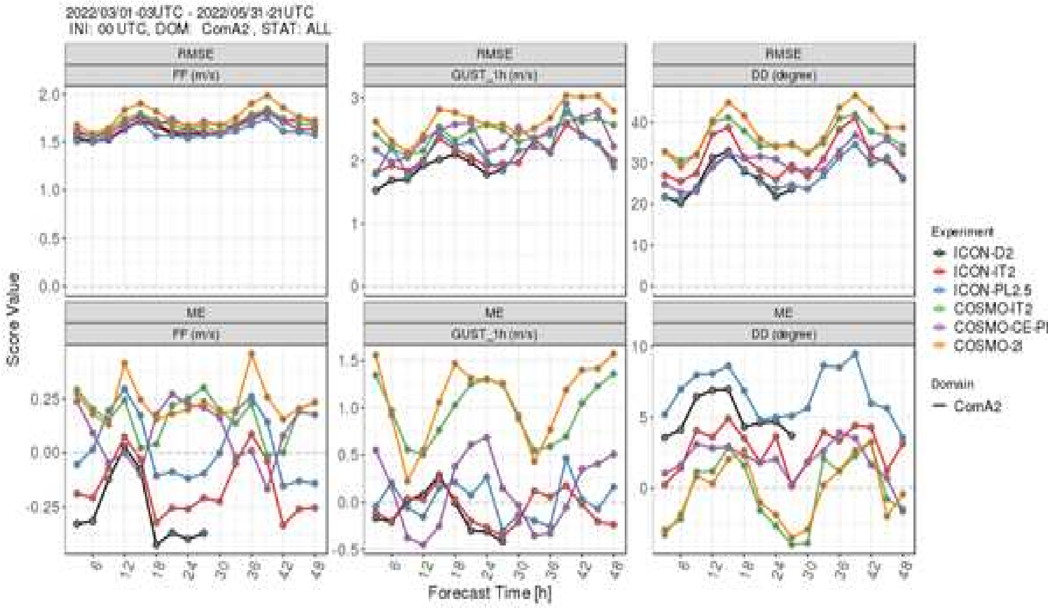


Figure 4: RMSE (first row) and ME (second row) indices for Wind speed, Wind Gust and Wind direction over ComA-2 (MAM2022).

As expected, the resolution-dependence is very clear for wind speed and direction. With respect to model effect on wind properties, all three parameters performance is grouped accordingly. For wind gust, a more significant reduction of the overprediction tendency with higher resolution ICON-LAM models while with wind speed the common trend of overprediction for coarser models is only changed in phase with ICON models (max over evening hours). In higher resolution models, wind speed is overpredicted with COSMO models and underpredicted with ICON-LAMs. Large diurnal cycle for ME for all wind properties is exhibited that can be related to boundary layer mixing in higher resolutions, with differences in phase for the maxima among COSMO/ICON implementations.

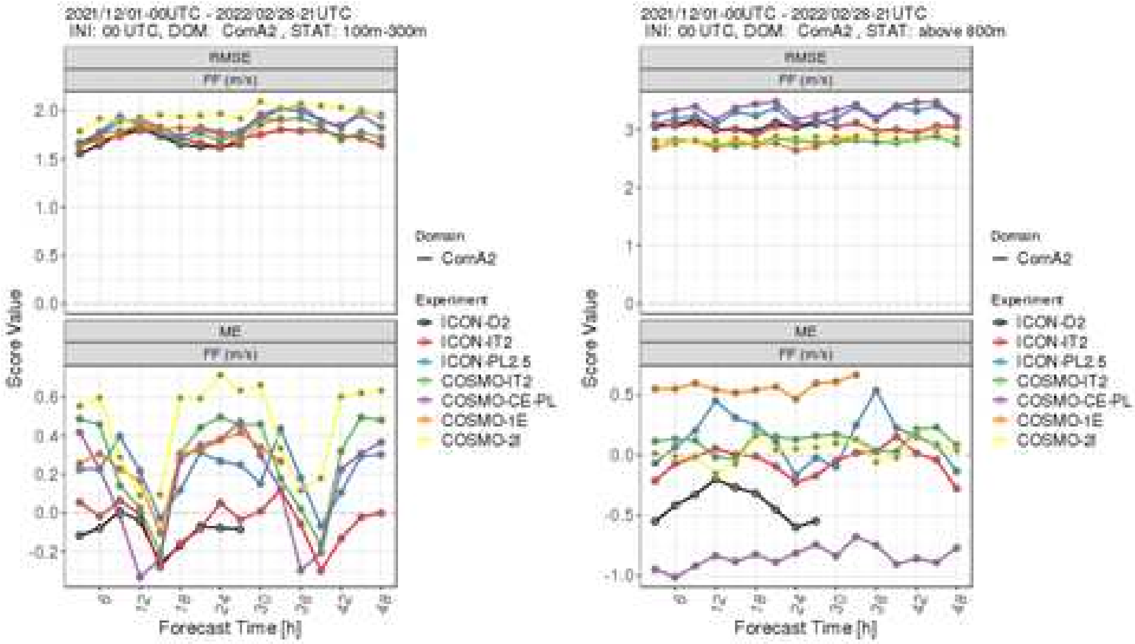


Figure 5: RMSE (first row) and ME (second row) indices for Wind speed, Wind Gust and Wind direction over ComA-2 (MAM2022).

Focusing in two different stratifications, one below 300m and one above 800m, there is a clear altitude dependence in performance for all seasons. The RMSE grows significantly in higher elevation points with sightly higher values for ICON-LAMs, with a general underestimation of wind speed while for lower elevations COSMO models tend as shwn also before to overestimate values some trend that is reduced with ICON-LAMs.

3.4 Temperature/Cloudiness performance dependence

The dependency of 2m temperature performance to total cloudiness is analysed in this section. Specifically, statistical indices are calculated for temperature when cloudiness is less than 25% (near clear sky conditions) and for cloudiness higher that 75% (near overcast conditions). Both conditions are imposed on observations. The outcome of this test selectively for one season (DJF2022) is presented in Figure 6. The RMSE values of 2mT are found in clear sky conditions, and lower errors when overcast conditions are examined compared to the total sample results. The diurnal variability of error is stronger for COSMO models when few/no clouds are present. When the relative performance of models is analysed, it is clear that there is a significant improvement of 2mT forecasts with ICON models in all seasons and cases. For the summer (not shown here), there is a distinct overestimation of 2mT mainly during cloudy days, which seems to be higher in some ICON-LAMs. Worst warming happens during midday while at night the effect is reverse in clear days with cooler models.

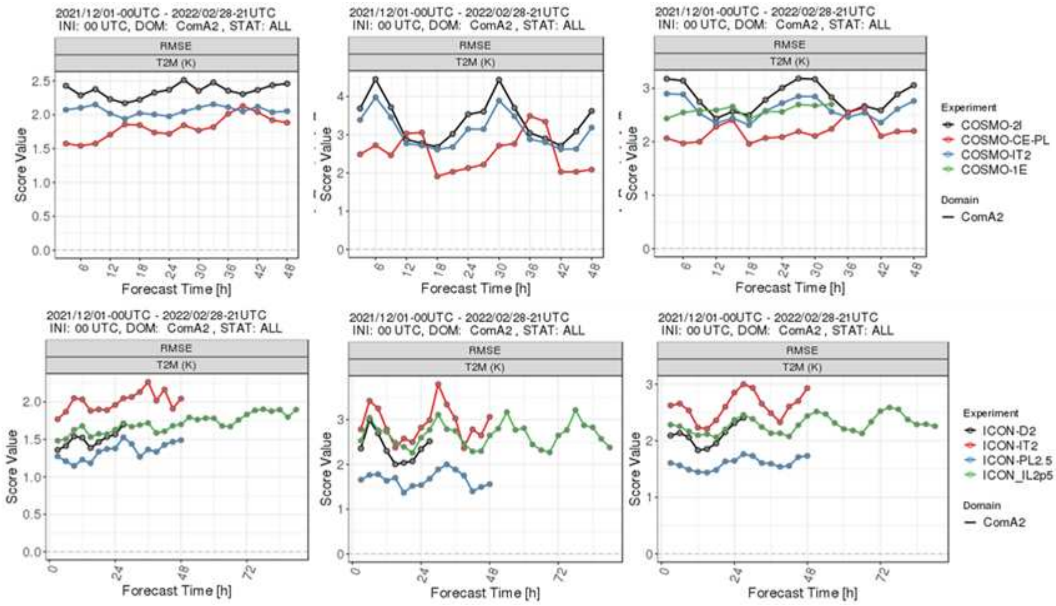


Figure 6: RMSE (top) and ME (bottom) for 2mT for overcast conditions (left), clearsky conditions (middle) and for all cases (DJF2022).

3.5 Cloudiness performance

Verifying cloud forecasts from NWP models has proved to be a difficult task because of the complex three dimensional structures of clouds, and lack of routine observations adequate for the purpose. Furthermore, it is not just of interest in its own but also has a major impact on other parameters, such as temperature and solar radiation. As it is highly variable in terms of time and location, it is difficult to forecasted but also verified. The main reason is the spatial representativeness mismatch between forecasts and SYNOP observations that are widely and almost exclusively used for this purpose. The area covered by visual observation typically varies between 10 and 100 km around a station, depending on visibility and topography. In this section, cloudiness is evaluated both point-wise and spatially with the use of satellite estimates.

3.5.1 Verification against SYNOP

As with other continuous parameters, COSMO and ICON-LAMS are evaluated against SYMOP observations over the common areas systematically. The seasonal statistical Mean Errors for winter and summer 2022 are presented in Figure 7, grouped separately for the two models. The TCC bias difference among the two sets of models is clear, with ICON models exhibiting a diurnal cycle with constant underestimation of observed values especially during the warm hours for both seasons. In the contrary, COSMO models exhibit mainly TCC overestimation while ICON models behavior is ambiguous. On the other hand, the RMSE diurnal cycle is similar for both sets, while higher values of errors are present during night hours.

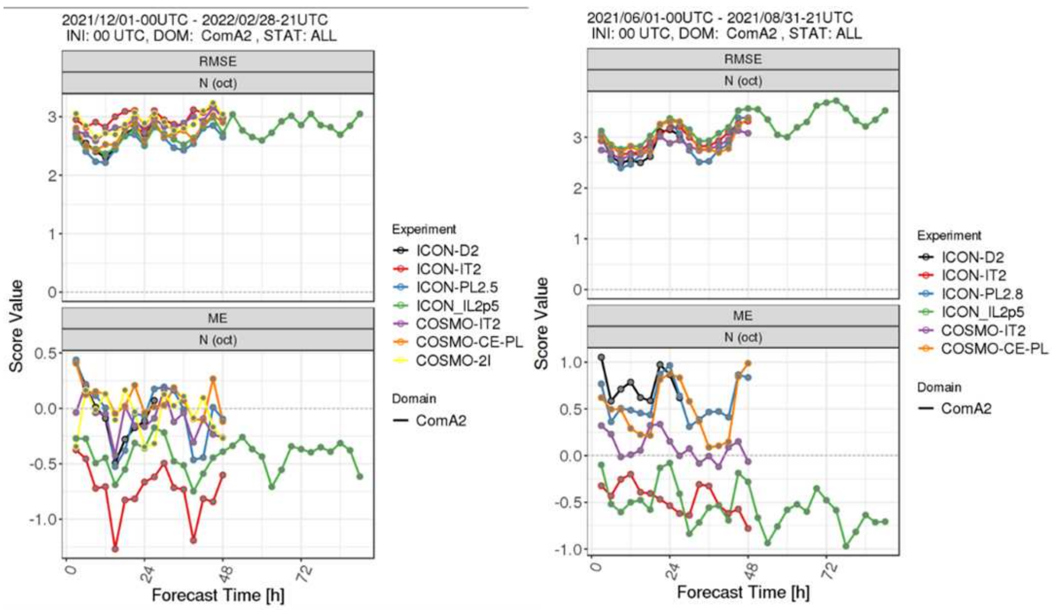


Figure 7: RMSE (top) and ME (bottom) for Total Cloud Cover for DJF 2022 (left column) and JJA2021(right column) over ComA-2.

3.5.2 Verification against NWC-SAF

Due to restrictions that were mentioned before regarding the representativeness of cloudiness SYNOP observations, a first approach evaluating this parameter against satellite estimates was initiated as part of WG5 common plot activity. In detail, NWC-SAF cloud mask fields (0.025 degrees) were retrieved and converted to TCC octants to be used in the gridded application of FSS score.

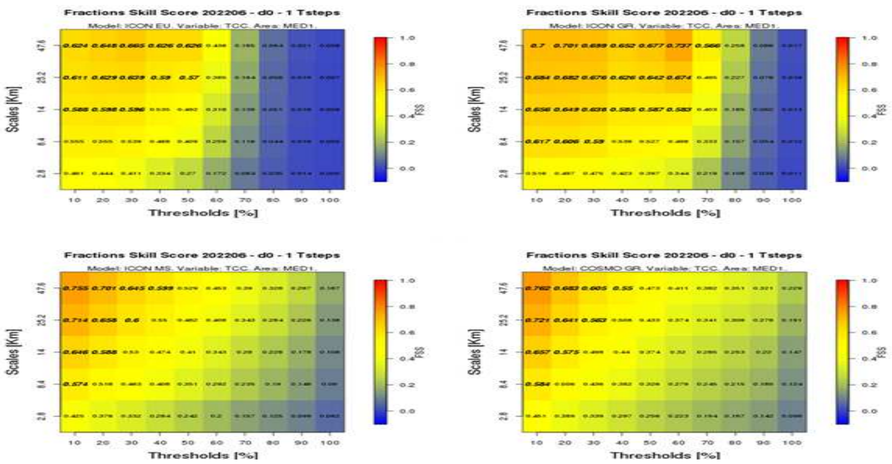


Figure 8: FSS for TCC for June 2021 over the east Mediterranean area.

As the quite restricted area of ComA-2 was too small for the analysis of cloud field, an extended verification area over east Mediterranean was used for a stratification of the operational models that covered this domain, while the time period was June 2022. The spatial verification approach can reveal the scales that cloudiness forecasts can be useful, by relaxing the criterion of strict point to point comparison.

Overall, for this experiment the useful scales proved to be for windows that forecast information is averaged in higher than 14km windows while and cloud coverage less than 30% cloudiness the performance was most successful.

The most striking result that was consistent with the unclear performance of models against SYNOP data, was that COSMO models at high percentages of TCC, outperform ICON-LAMs.

3.6 6h Precipitation performance evaluation

The station-based 6h accumulated precipitation forecasts are evaluated in terms of categorical indices for different thresholds. JJA2021 and DJF2022 results for ETS, POD and FAR are presented in Figures 9a,b.

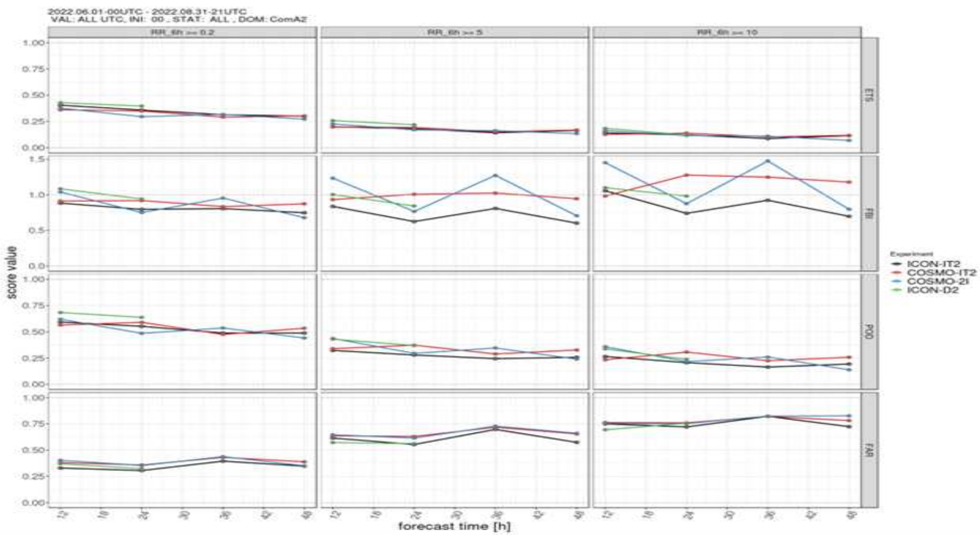


Figure 9: 6h accumulated precipitation indices for different thresholds. From top to bottom (ETS, FBI, POD, FAR). From left to right (0.2, 5, 10mm) for JJA2021.

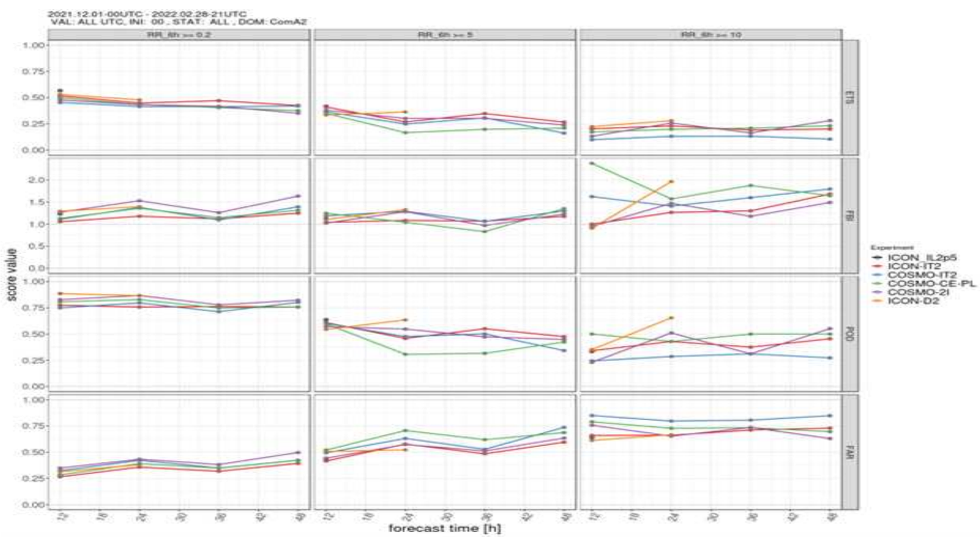


Figure 10: 6h accumulated precipitation indices for different thresholds. From top to bottom (ETS, FBI, POD, FAR). From left to right (0.2, 5, 10mm) for DJF2022.

As in previous years, scores trend worsen slightly with forecast time and significantly with increasing threshold. Differences among ICON and COSMO models are not so clear but separating the two different model groups there is a distinct small improvement in all scores with ICON-LAMS. More significant is the improvement during the summer period that was analysed.

3.7 Fuzzy 6h-Precipitation verification

This section presents fuzzy verification scores for ComA-2 compared to the OPERA network radar composites. The VAST COSMO software, which is based on Beth Ebert's fuzzy verification IDL code, was used for this task. The VAST main code utilizes txt gridded files for each weather parameter, and a preprocessing of input files is available with the help of LIBSIM software. The main indices used to summarize the spatial verification results are FAR, Fraction Skill Score (FSS), and POD, as shown in Figure 10.
These scores compare the forecast and observation (radar) 3-hour gridded precipitation fields on continuously increasing spatial windows and for varying precipitation thresholds. The results for three different thresholds (0.1, 5 and 10mm/3h) are presented for the first forecast day and for the three seasons (JJA21, SON21, DJF22).
The spatial verification approach shows a relatively improved skill of ICON-LAM models compared to COSMO ones in precipitation forecasts, especially with respect to FAR and FSS scores. However, for the POD score, as also extracted from the point-wise verification mainly for the smaller thresholds, the scores are slightly worse in some cases.

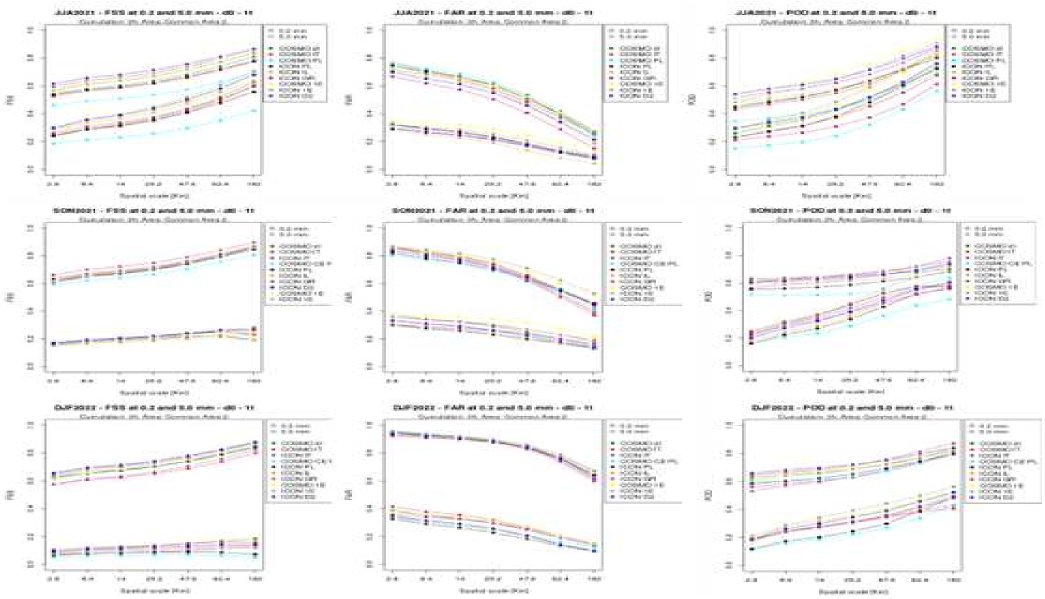


Figure 11: FAR (left), FSS (middle) and POD (right) 3h Precipitation scores for forecast day 1 for JJA21 (top row), SON21 (middle row) and DJF22 (lower row) calculated over ComA-2.

4 General Concerns

The WG5 common plot activity provides a good basis for fulfilling the minimum requirements of monitoring operational model performance of all participating services.
A detailed presentation of the verification findings were presented during General Meeting in September 2022 while the complete range of plots are available on `http://www.cosmo-model.org/content/tasks/verification.priv/default.htm`Verification tasks. As a general conclusion, improvement in performance in most cases/parameters analyzed with ICON-LAMs compared to COSMO implementations was derived from this year activity. There are however components that further model development is needed, as long-term biases are still present. The deviation among model performance is greater in ICON-LAMs than in COSMO models, revealing the need for further model tuning especially in high resolution scales.