Technical Memo

ECMWF
European Centre for Medium-Range Weather Forecasts

901

How do ecPoint precipitation forecasts compare with postprocessed multi-model ensemble predictions over Switzerland?

Stephan Hemri[1], Timothy Hewson[2], Estibaliz Gascón[2], Jan Rajczak[3], Jonas Bhend[3], Christoph Spirig[3], Lionel Moret[3], Mark A. Liniger[3]

[1]University of Zurich, Department of Mathematics, Switzerland

[2]ECMWF, Forecast Department, United Kingdom

[3]Federal Office of Meteorology and Climatology MeteoSwiss, Switzerland

August 2022

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:

http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, Berkshire RG2 9AX, England

Abstract

Statistical postprocessing aims to reduce systematic biases and dispersion errors of forecast ensembles provided by state-of-the-art numerical weather prediction (NWP) models. Often, this also entails an implicit mapping from NWP forecasts valid for grid cell means to point forecasts valid at specific points of interest within a given grid cell. In general, statistical postprocessing increases forecast skill considerably. However, due to the strongly non-Gaussian distribution and low predictability of precipitation at points, statistical postprocessing of precipitation forecasts is particularly difficult. ecPoint, a postprocessing approach developed at the European Centre for Medium-Range Weather Forecasts (ECMWF) is tailored to point forecasting of precipitation and proved to increase forecast skill considerably. Moreover, ecPoint is data efficient in that it needs only one year of training data from the global ECMWF-IFS NWP model. The combination of data efficiency and forecast skill of ecPoint calls for a comparison with (postprocessed) probabilistic forecasts of precipitation provided by high-resolution limited area NWP models like COSMO-E, which is a MeteoSwiss configuration of the Consortium for Small-scale Modeling (COSMO) model.

In this study, we compare ecPoint forecasts for 12 h accumulated precipitation with ensemble model output statistics (EMOS) as a reference postprocessing method. We assess the performance of raw ecPoint and raw COSMO-E alongside EMOS applied to pooled ensembles constructed using either COSMO-E and ECMWF-IFS or COSMO-E and ecPoint with varying weights. Verifying the different forecasts on a set of about 850 gauge stations in Switzerland and neighboring areas confirms the good performance of ecPoint. For long lead times and heavy precipitation, ecPoint tends to be more skillful than EMOS. However, further research is needed to assess the impact of the lengths of the respective training periods on the relative skill of ecPoint compared with EMOS. Moreover, it will be beneficial to identify in which regions and in which meteorological regimes ecPoint ordinarily outperforms forecasts based on a high-resolution limited area model, and vice versa.

Plain Language Summary

Global weather models provide weather forecasts for different variables like temperature and rainfall on a grid with a rather large mesh size. The collection of runs of the same weather model with slightly different configurations is called an ensemble forecast. Ensemble forecasts allow us to quantify forecast uncertainty. However, ensemble forecasts are subject to biases, like forecasting too little or too much rainfall on average, and to being too certain or too uncertain about the future weather. In addition, global ensemble forecasts denote only “average weather” over pre-defined regions, that currently measure about 20km by 20km. In contrast, real rainfall can vary a lot over such a region, in particular for thunderstorms. These various issues can be tackled by applying correction algorithms to the ensemble forecasts. This process is called statistical postprocessing. The “ecPoint” postprocessing method, which has been developed at the European Centre for Medium-Range Weather Forecasts (ECMWF), is specially tailored to forecast for points rather than for regions. For rainfall, ecPoint forecasts have been shown to increase forecast quality considerably compared to ECMWF’s global weather model forecasts (ECMWF-IFS), in a timely manner and at low cost.

In this study, we compare ecPoint forecasts with forecasts provided by a limited area weather model ensemble (COSMO-E). Compared to ECMWF-IFS, COSMO-E has a considerably smaller grid mesh size (about 2km by 2km) that should allow it to better predict local weather variations. COSMO-E is run by MeteoSwiss, for a

domain covering Switzerland and neighbouring areas. We assess the relative quality of rainfall forecasts over Switzerland from ecPoint, from COSMO-E and from different combinations of COSMO-E, ECMWF-IFS and ecPoint. The combinations have been postprocessed using an alternative postprocessing method, which is often used as a standard benchmark method in the field of ensemble postprocessing. Our results confirm the good performance of ecPoint compared both to COSMO-E, and to model combinations postprocessed with the alternative method. ecPoint performs particularly well for forecasts several days into the future, and for heavy rainfall. Further research is needed to assess why forecast quality varies between the different approaches. It will also be helpful to identify in which regions and in which weather situations ecPoint ordinarily outperforms forecasts based on a high-resolution regional weather model, and vice versa.

1 Introduction

State-of-the-art probabilistic weather forecasts are based on ensemble runs of numerical weather prediction models (NWP). Despite their great success over the last decades, NWP ensembles still suffer from systematic biases and dispersion errors (e.g. Vannitsem et al., 2021). Hence, statistical postprocessing is increasingly applied to NWP ensemble predictions to correct both bias and dispersion.

Currently, NWPs' horizontal grid resolution ranges from about 20 by 20 km (Hewson & Pillosu, 2021) in state-of-the-art global models to about 1 by 1 km in very high resolution regional models. For instance, COSMO-E (Baldauf et al., 2011; Klasa et al., 2018), which we consider in this study, has a resolution of about 2 by 2 km. MeteoSwiss used to operationally run COSMO-E daily at 0 and 12 UTC until it was replaced by COSMO-1E and COSMO-2E (Kaufmann & Rüdisühli, 2019) in 2020. COSMO-1E is an 11-member ensemble extension of the formerly deterministic COSMO-1 configuration with a very high resolution of about 1 by 1 km. COSMO-2E is an updated version of COSMO-E, operationally running every 6 hours. The underlying COSMO model is being developed in the Consortium for Small-scale Modeling. Clearly, weather variables with a lot of sub-grid variability, i.e. variability at spatial scales smaller than the grid size, like convective precipitation cannot be represented that well by (global) NWP models. Rather, the predicted precipitation represents the average value for the whole grid cell. To derive forecasts valid for any point within a grid cell from the global integrated forecasting system of the European Centre for Medium-Range Weather Forecasts (ECMWF-IFS; Owens & Hewson, 2018) ensemble grid cell mean forecasts, Hewson & Pillosu (2021) have introduced ecPoint. ecPoint is a novel postprocessing approach, developed originally for precipitation, but could in principle also be applied to other weather variables like 2m temperature. In a nutshell, ecPoint is a member-by-member postprocessing method that utilizes meteorological expert knowledge to construct mapping functions that convert grid cell forecasts to point forecasts conditional on the "gridbox weather type". Unlike most postprocessing approaches, ecPoint needs only 1 year of training data, as it utilizes global observations in a non-local calibration procedure. Therefore, it is capable of weather type dependent postprocessing without needing, say, several years of reforecasts from an expensive high-resolution limited area NWP model.

Traditional postprocessing methods like ensemble model output statistics (EMOS; Gneiting et al., 2005) implicitly also map probabilistic grid cell mean forecasts to forecasts for a particular point in space, i.e. the location of the verifying gauge station. This raises the question of how ecPoint compares to a postprocessed high-resolution limited area model. To gain insights into this question for precipitation over Switzerland, we

analyze forecast skill of ecPoint, direct model output (DMO) of COSMO-E, and EMOS postprocessed forecasts based on COSMO-E. As multi-model combination is known to increase forecast skill (Weigel et al., 2008), we consider also EMOS based on different naïve two-model combinations comprising either COSMO-E and ecPoint, or COSMO-E and DMO of ECMWF-IFS. Considering also the latter combination helps to examine the extent to which the gain in skill of ecPoint over raw ECMWF-IFS is attributed to non-linear corrections to the forecast distribution, which cannot be reproduced by a linear postprocessing approach like EMOS.

In Section 2, we provide a brief description of the data, the study design and the methods used for postprocessing and verification. The results in Section 3 are followed by a discussion in Section 4 and the conclusions in Section 5.

2 Data and methods

2.1 Data and study design

Joint availability of COSMO-E, ECMWF-IFS, ecPoint, and gauge observations covered the period 14 March 2019 to 31 January 2020 at the time this study was performed. Observational data are available for about 850 automatic gauge stations, mainly in Switzerland, but also in neighboring areas in Austria, France, and Germany. The example forecasts in Figure 1 show the locations of the gauge stations. Due to the short study period of about 10 months, EMOS coefficients are estimated based on a 45-day moving window training period. Accordingly, pairs of observation and gauge measurement data from May 2019 to January 2020 are used for verification. We consider COSMO-E and ECMWF-IFS runs initialized daily at 00 UTC. The ensemble size is 21 and 51 for COSMO-E and ECMWF-IFS, respectively. No major changes have been applied to COSMO-E during the study period, ECMWF-IFS was updated from Cycle45r1 to Cycle46r1 in June 2019. However, we do not expect that this change has any relevant effect on our results. The horizontal resolutions are about 2 by 2 km and about 18 by 18 km for COSMO-E and ECMWF-IFS, respectively. The forecast horizon of COSMO-E is 5 days, while ECMWF-IFS mainly provides forecasts up to 15 days into the future. For this study, we consider 12 h accumulated precipitation for lead times 12, 24, …, 120 h, where the lead time denotes the last hour of the accumulation period.

Furthermore, prior to computing the ensemble statistics used by EMOS, the raw ensembles to be considered (either COSMO-E and ECMWF-IFS or COSMO-E and ecPoint) are pooled together into a single ensemble. We have selected this pooling approach for two reasons: First, we can actively control and evaluate the effects of different raw ensemble weighting in EMOS. Second, this approach reduces the number of dynamic predictors to be estimated. Details on the construction of the different pooled ensembles are presented in Section 2.4.

2.2 ecPoint

Hewson & Pillosu (2021) and Owens and Hewson (2018) give a comprehensive introduction to ecPoint and ECMWF's associated point forecast product, respectively. We provide just a short summary of ecPoint here, which follows Hewson & Pillosu (2021) closely. As mentioned in Section 1 ecPoint is a postprocessed

probabilistic (precipitation) forecast product tailored to deliver point forecasts based on global ECMWF-IFS ensemble forecast data. For each grid cell for each time interval, ecPoint maps predicted precipitation of each ensemble member to a probabilistic point forecast PDF. The mapping function used for a specific ensemble member depends on the weather type assigned to that ensemble member (for the said gridcell, for the said time interval). Technically speaking, applying ecPoint postprocessing comprises the following steps:

1. Set up a decision tree based on meteorological expert knowledge. The decision rules governing the construction of the decision tree are universal in that they do not differ among grid points. However, adding grid point specific static predictors like regional topography to the decision tree would enable one to take account of regional effects induced by, e.g., coastlines or mountain ranges[1]. Subsequently, this decision tree assigns a weather type to each member at each grid cell for each time interval, conditional on the grid cell's NWP output for precipitation-related variables like wind speed at 700 hPa or the fraction of convective precipitation.

2. The above decision tree differentiates between K different weather types. For each weather type, a probability distribution function that maps grid cell precipitation forecasts onto point precipitation forecasts needs to be obtained. This is done based on the forecast error ratio (FER), which describes the difference of an observed precipitation value relative to grid cell (i.e. mean) forecast precipitation. Based on observed precipitation from gauges over a large domain, preferably the world, a FER probability density function (PDF) mapping function $M_k,\ k = 1, \dots, K$ can be constructed for each weather type k. Figure 2 shows an example of an empirical FER distribution.

3. For grid cell s and ensemble member i, $i = 1, \dots, m$, where m denotes the size of the raw ensemble, the probabilistic point forecast for precipitation r (for a given time interval) based on its FER PDF is given by

$$F_{s,i}(r) = \left(1 + M_{k|s.i}(FER)\right) G_{s,i}, \tag{1}$$

where $G_{s,i}$ denotes the predicted precipitation of raw ensemble member i and $M_{k|s,i}(FER)$ denotes its associated mapping function.

[1] Whilst this type of static predictor has been used in other ecPoint decision trees the real-time ecPoint output used here did not incorporate such factors.

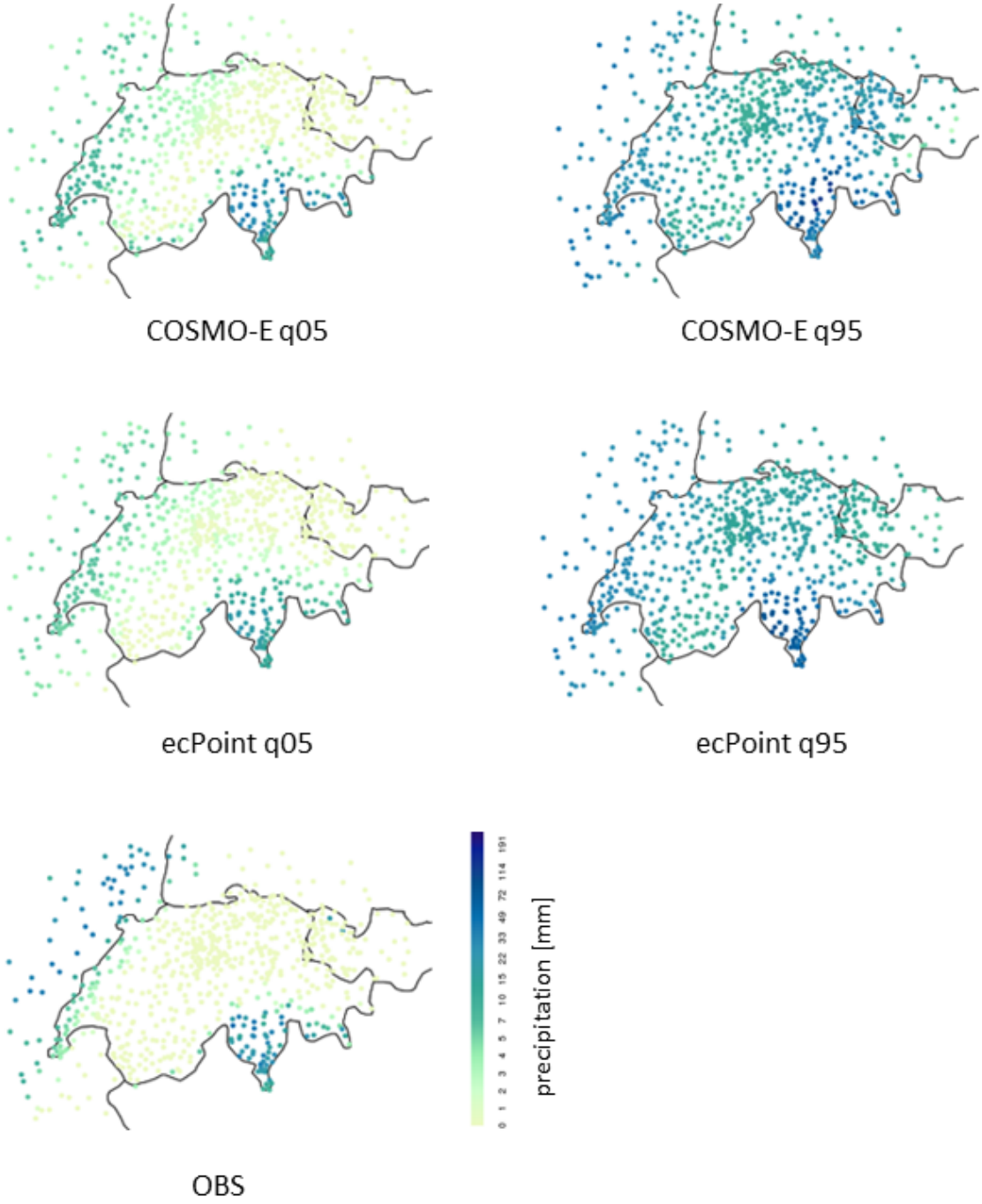


Figure 1: Example forecasts initialized on 16 October 2019 00 UTC valid for 12 h accumulation period 19 October 2019 12 UTC to 20 October 2019 00 UTC. The subfigures show the 5th and 95th percentiles of ecPoint and COSMO-E and the corresponding observations. For COSMO-E the driest and the wettest members at the corresponding station serve as proxies for the 5th and 95th percentiles, respectively.

4. The probabilistic forecast F_s for points in grid cell s can then be obtained by computing

$$F_s = \frac{1}{m}\left(\sum_{i=1}^{m} F_{s,i}(r)\right). \quad (2)$$

These steps are illustrated in Figure 3, which is copied from Figure 6 in Hewson & Pillosu (2021).

Applied to the 51 member ECMWF-IFS ensemble, ecPoint generates 5100 (51 members times 100 values sampled from the FER PDF per member) equi-probable precipitation forecasts for points in a specific grid cell. Ordered percentiles, i.e. the $1, 2, \ldots, 99$ percentiles, from the grid cells' distributions of 5100 forecast values are then stored as the final ecPoint forecasts (Hewson and Pillosu, 2021). In principle, the ecPoint approach works for different accumulation periods. Here, we follow Hewson & Pillosu and consider 12 h accumulation periods. Unlike DMO ensembles, the standard ecPoint output does not provide a set of spatio-temporal forecast scenarios[2], but rather a single calibrated, probabilistic point forecast distribution for any point in space for each 12 h accumulation period for lead times 12, 18, 24, …, 120 h.

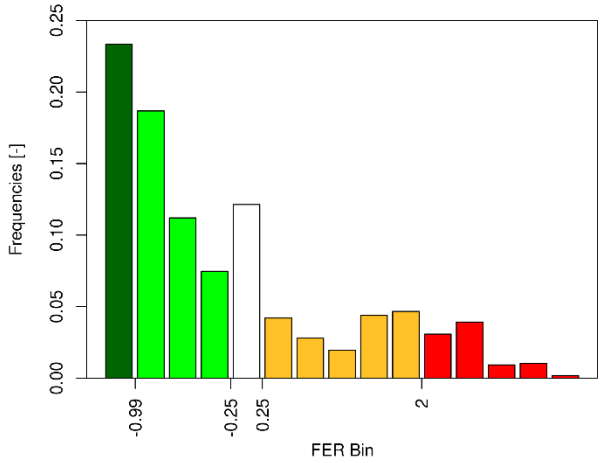


Figure 2: Simplified version of Figure 5 in Hewson & Pillosu (2021): An example of a FER based mapping function. The colours dark green, green, white, ochre and red, denote FER ranges mostly dry, 'over-prediction', 'good forecast', 'under-prediction' and 'strong under-prediction', respectively. Note the non-linearity in the x-axis.

[2] Note that it is however possible to deconstruct any ecPoint forecast to deliver point forecast distributions for each ensemble member realisation, by using unadjusted gridbox precipitation forecasts, the diagnosed weather types which are archived, and the mapping function multipliers which are available offline.

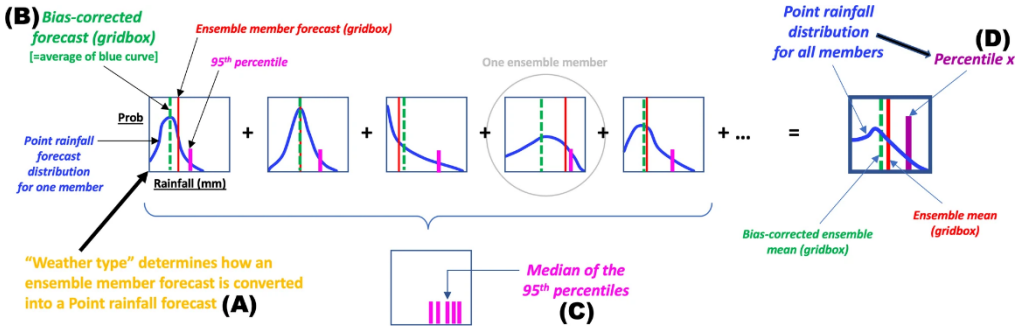


Figure 3: Copy of Figure 6 in Hewson & Pillosu (2021) illustrating the construction of an ecPoint precipitation forecast from gridbox ensemble member forecasts. The shape of the point precipitation forecast distributions (blue) in (B) depend on the 'weather type' (A) assigned to each ensemble member. (C) shows the 95th percentiles of the first 5 ensemble members' ecPoint distributions and the corresponding median. Averaging the ensemble members' ecPoint distributions leads to the aggregated ecPoint precipitation distribution based on all ensemble members (D).

2.3 COSMO-E and gEMOS

In this study, we use raw and global EMOS (gEMOS) postprocessed COSMO-E predictions as reference precipitation forecasts. Prior to any postprocessing COSMO-E forecasts have been interpolated from its 2x2 km grid to the locations of the gauge stations using nearest neighbor matching. The term global in gEMOS describes an EMOS procedure, which estimates only one single statistical model, i.e. one single set of coefficients for all stations in the study area, here Switzerland and some neighboring regions, simultaneously. To some extent, static predictors based on grid coordinates and topography allow one to model spatial differences in conditional errors. Following Friedli et al. (2021), we implement gEMOS as a heteroscedastic censored logistic regression model left censored at zero (Messner et al., 2014a, b). The predictors are listed in Table 1. Since we use a short training period, we need to restrict the number of predictors in gEMOS. Therefore, we selected only a minimal subset of the dynamic predictors, which proved to be beneficial for COSMO-E precipitation postprocessing at MeteoSwiss. The static predictors approximately discern different regions in Switzerland, for which we expect differences in precipitation biases: e.g. Swiss Plateau vs. Alps or northern slope of the Alps vs. southern slope of the Alps. Location and scale parameters of the censored logistic regression for 12 h accumulated precipitation are then modelled as

$$\mu = p_{mea} + p_{f0} + s_{lat} + s_{lon} + s_{alt} + s_{alt31} + s_{snd31} + s_{wed31} + s_{tpi31} + s_{tpi500}, \quad (3)$$

$$\sigma = p_{std} + s_{lat} + s_{lon} + s_{snd31} + s_{wed31} + s_{tpi31} + s_{tpi500}. \quad (4)$$

Due to the short time period covered by the dataset, only a moving window training approach is feasible. A window size of 45 days proved to produce reasonable results for gEMOS. We have used the R package `crch` to estimate the gEMOS model coefficients (Messner et al., 2016).

Table 1: List of predictors used in gEMOS. The static predictors latitude, longitude, and altitude refer to the gauge stations. The remaining static predictors are derived from a digital elevation model. TPI is a measure for the elevation of a grid cell relative to the neighbouring cells (Weiss, 2001). The dynamic predictors are derived from pooled ensembles of COSMO-E with ecPoint, or COSMO-E with ECMWF-IFS.

Short name	Description
p_mea	dynamic: precipitation ensemble mean
p_f0	dynamic: precipitation fraction zero
s_lat	static: latitude
s_lon	static: longitude
s_alt	static: altitude
s_alt31	static: altitude smoothed with 31 km kernel size
s_snd31	static: south north derivative (of altitude) smoothed with 31 km kernel size
s_wed31	static: west east derivative (of altitude) with 31 km kernel size
s_tpi500	static: Topographic position index (TPI) of DEM smoothed with 500 km kernel size
p_std	dynamic: precipitation standard deviation

2.4 Construction of pooled ensemble forecasts

As stated above, NWP-based gEMOS predictors are derived from the ensemble statistics of pooled ensembles of COSMO-E with ecPoint, or COSMO-E with ECMWF-IFS. This allows us to control the effect of ensemble model weighting. We replicate the raw ensembles using a different number of replicates for each model to obtain the desired weights. This leads to an implicit model weighting. The number of replicates to obtain the different combinations of COSMO-E with ecPoint or ECMWF-IFS are listed in Table 2 and Table 3, respectively.

Table 2: Weighting of ecPoint and COSMO-E for different pooled ensemble forecasts M1, …, M11. First row: number of ecPoint duplicates in pooled ensemble. Second row: number of COSMO-E duplicates in pooled ensemble. Third row: percentage of members stemming from COSMO-E in pooled ensemble.

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
# ecPoint duplicates	1	1	1	1	1	1	1	1	1	1	0
# COSMO-E duplicates	0	1	2	3	4	5	6	10	15	20	1
Percentage COSMO-E	0	18	30	39	46	51	56	68	76	81	100

Table 3: Weighting of ECMWF-IFS and COSMO-E for different pooled ensemble forecasts M1, …, M11. First row: number of ECMWF-IFS duplicates in pooled ensemble. Second row: number of COSMO-E duplicates in pooled ensemble. Third row: percentage of members stemming from COSMO-E in pooled ensemble.

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
# ECMWF-IFS duplicates	1	2	2	2	2	2	2	2	2	2	0
# COSMO-E duplicates	0	1	2	3	4	5	6	10	15	20	1
Percentage COSMO-E	0	17	29	38	45	51	55	67	76	80	100

2.5 Verification measures

We use the continuous ranked probability score (CRPS; Matheson and Winkler, 1976), which depends on both forecast spread and bias, in order to assess overall forecast skill. The CRPS is a negatively oriented scoring rule, the lower the score the better the forecast. Its discrete version that we use to verify ensemble forecasts favours larger ensembles. In order to be able to compare ensembles of different sizes (e.g. ecPoint: 99 members; COSMO-E: 21 members) we use the FairCRPS (Ferro et al., 2008; Ferro, 2014), which corrects for differences in ensemble size. Skill scores of the CRPS (or the FairCRPS), that is CRPSS (or FairCRPSS), are obtained by computing

$$CRPSS = 1 - \frac{CRPS_{forc}}{CRPS_{ref}}, \tag{5}$$

where $CRPS_{forc}$ and $CRPS_{ref}$ denote the CRPS of the forecast of interest and the reference prediction, respectively.

To evaluate forecast discrimination ability for different thresholds, we use receiver operating characteristic (ROC) curves. A ROC curve for a particular event, e.g. exceedance of a threshold, is generated by plotting the true positive rate (sensitivity) against the false positive rate (1 – specificity). The better a forecast in terms of sensitivity and specificity, the more the ROC curve lies in the upper-left part of the unit square. A perfect forecast would lead to a step function $1_{x>0}$, while a forecast with no skill leads to a ROC curve lying on the

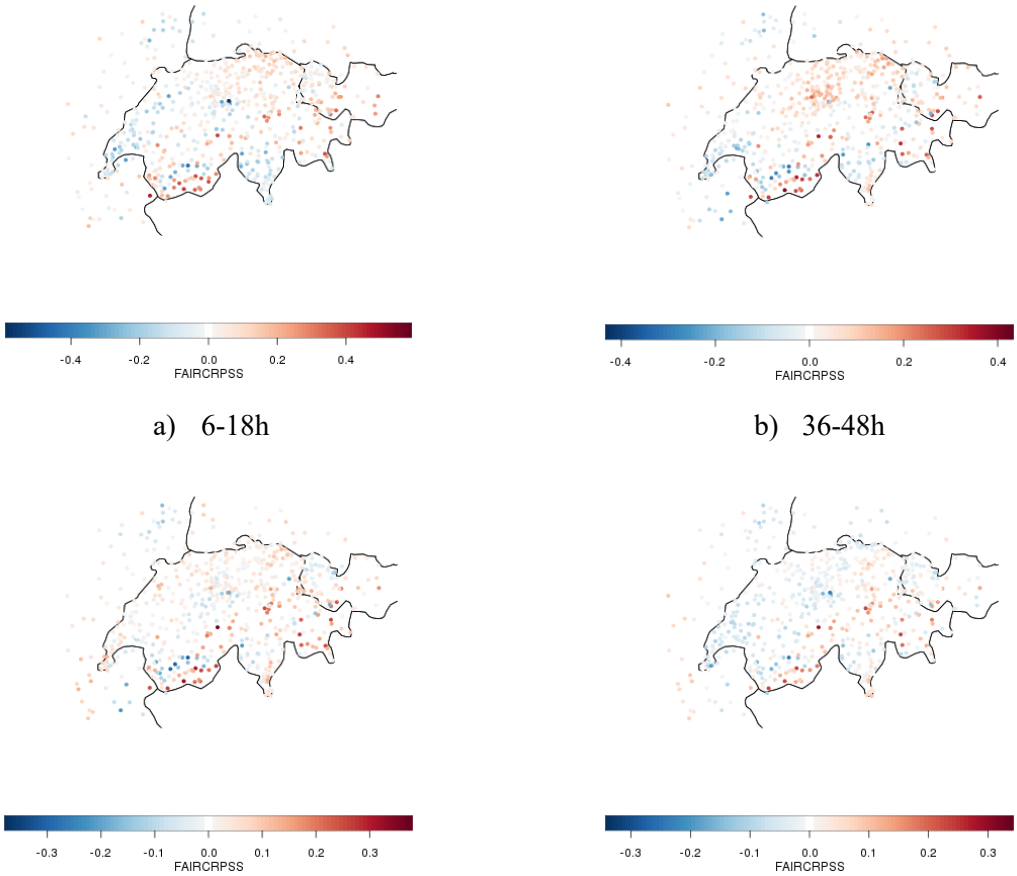
diagonal. We have also computed the trapezoidal area under the ROC (AUC), which we show as summary measures in the ROC figures, but do not discuss further, because trapezoidal AUC values may be misleading as discussed in Ben Bouallègue & Richardson (2021).

3 Results

First, we evaluate station-wise ecPoint forecast performance compared to COSMO-E in terms of FairCRPS. Figure 4 shows FairCRPS skill score (FairCRPSS) values of ecPoint relative to COSMO-E for accumulation periods 6-18h, 36-48h, 60-72h, and 108-120h. Apparently, for all accumulation periods considered, there are both regions where ecPoint outperforms COSMO-E (reds) and regions where COSMO-E outperforms ecPoint (blues). ecPoint exhibits particularly high values of FairCRPSS in Northeast Switzerland. Similarly, some alpine regions like Southern Valais show quite consistent outperformance of ecPoint compared to COSMO-E. However, it is difficult to explain any of the spatial patterns in FairCRPSS by topography.

Figure 5 and Figure 6 show the CRPS against lead time of gEMOS of pooled ensembles constructed from COSMO-E and ECMWF-IFS and from COSMO-E and ecPoint, respectively. Irrespective of the exact pooled ensemble variant, there is a daily cycle in gEMOS forecast performance in terms of CRPS. CRPS values are higher (worse prediction) in the afternoon/evening than in the morning. Most likely, this is due to convection-induced precipitation, which predominantly occurs during the summer half year in the afternoon/evening. For short lead times up to 30 h, gEMOS on pooled ensembles with a considerable proportion of COSMO-E (55 % when combined with ECMWF-IFS and 51 % or even 46 % when combined with ecPoint) performs best. At lead times 36 h and beyond gEMOS of pooled ensembles with COSMO-E proportions of about 30 to 45 % perform best. When combined with ecPoint even using a considerably lower COSMO-E proportion of about 18 % leads to well performing predictions. For a lead time of 24 h (12-24 h accumulation period), gEMOS based on pooled ensembles with high ECMWF proportions leads to a particularly poor performance, using ecPoint instead seems to mitigate this issue.

Now, let us have a look at gEMOS forecast skill in terms of CRPSS with ecPoint as reference forecast. As shown in Figure 7, gEMOS based only on COSMO-E does not outperform ecPoint. In fact, ecPoint exhibits considerably better CRPS for lead times beyond 30 h. gEMOS based on the pooled ensemble consisting of 38 % COSMO-E and 62 % ECMWF-IFS shows positive skill compared to ecPoint for lead times up to about 55 h. At longer lead times ecPoint performs better. Figure 8 shows the corresponding CRPSS values for gEMOS based on pooled ensembles consisting of COSMO-E and ecPoint. For most lead times gEMOS applied to a pooled ensemble consisting of 30 % COSMO-E and 70 % ecPoint is among the best performing gEMOS models. Nevertheless, at lead times beyond 70 h raw ecPoint forecasts tend to outperform all gEMOS variants. Overall gEMOS based on COSMO-E and ecPoint performs slightly better than gEMOS based on COSMO-E and ECMWF-IFS in terms of CRPSS. This is in line with the optimal proportion of ecPoint in COSMO-E/ecPoint pooled ensembles tending to be larger than the optimal proportion of ECMWF-IFS in COSMO-E/ECMWF-IFS pooled ensembles.



a) 6-18h b) 36-48h

c) 60-72h d) 108-120h

Figure 4: FairCRPSS for ecPoint minus FairCRPSS for COSMO-E, for 12h accumulation periods over lead times a) 6-18h, b) 36-48h, c) 60-72h, and d) 108-120h.

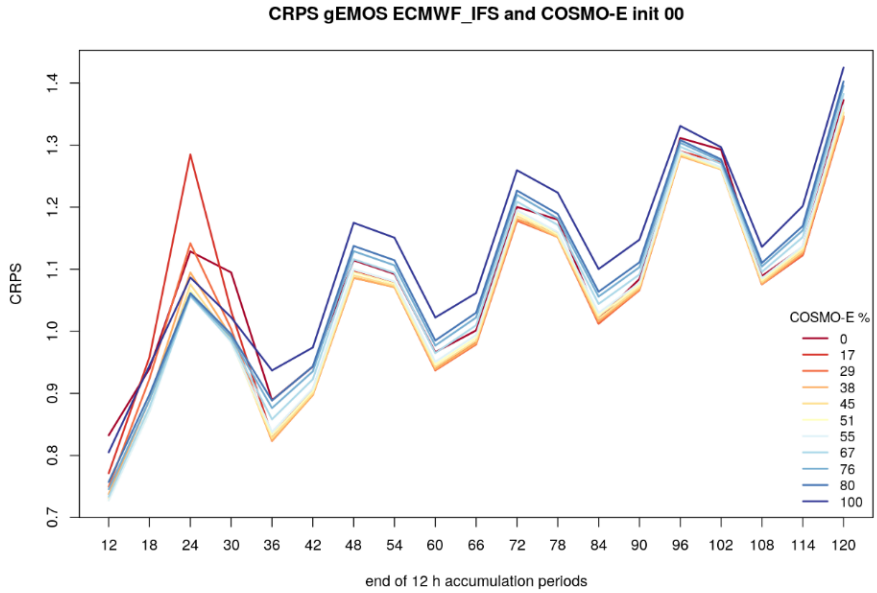


Figure 5: CRPS of gEMOS applied to pooled COSMO-E and ECMWF-IFS ensembles with different COSMO-E ratios against lead time (here denoted as end of 12 h accumulation period). Lower CRPS values are better.

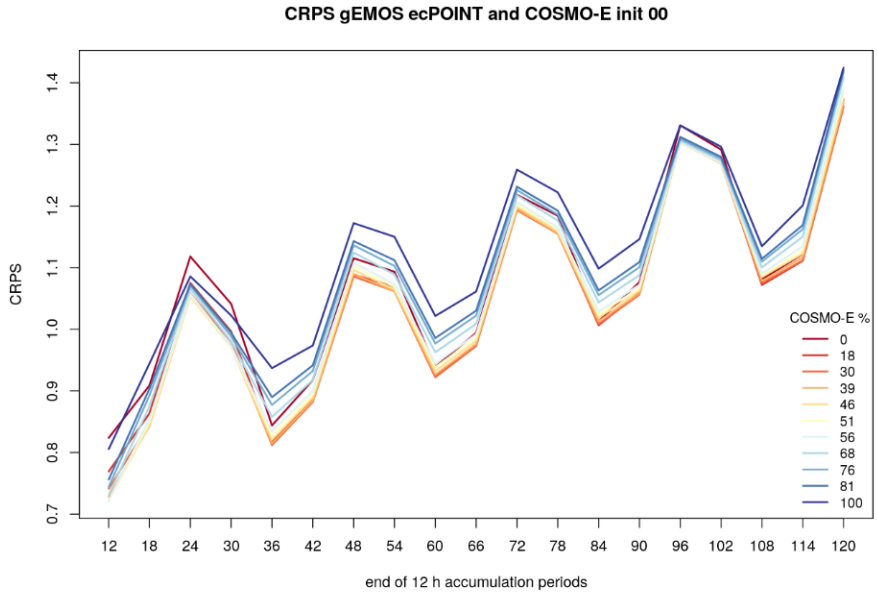


Figure 6: CRPS of gEMOS applied to pooled COSMO-E and ecPoint ensembles with different COSMO-E ratios against lead time (here denoted as end of 12 h accumulation period). Lower CRPS values are better.

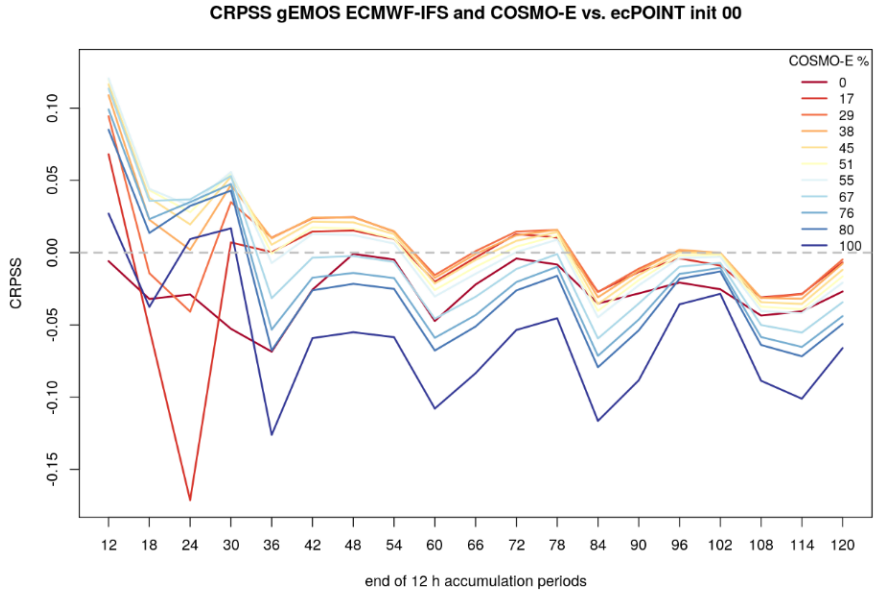


Figure 7: CRPSS of gEMOS applied to pooled COSMO-E and ECMWF-IFS ensembles with different COSMO-E ratios with ecPoint as reference forecasts against lead time (here denoted as end of 12 h accumulation period). Positive values mean the combination beats unadjusted ecPoint. Negative values mean unadjusted ecPoint is better.

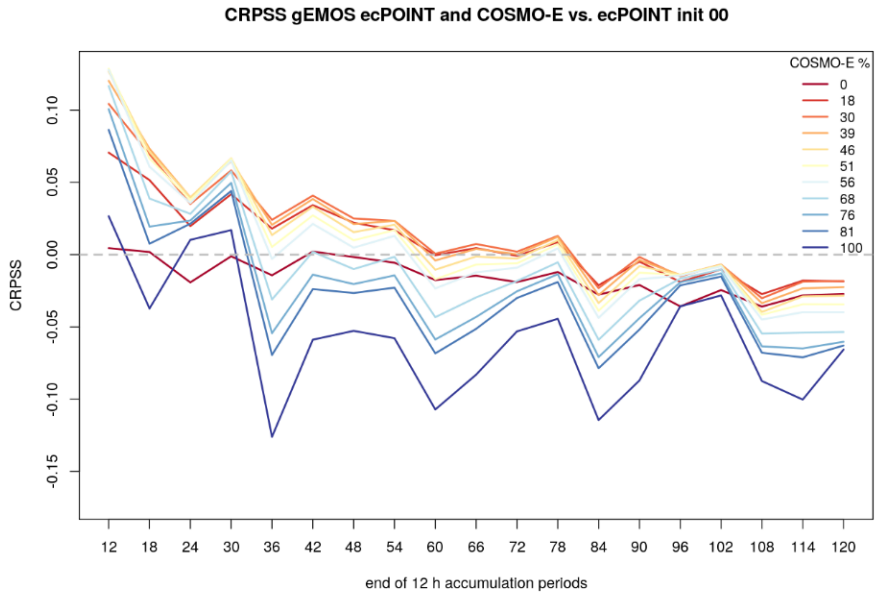
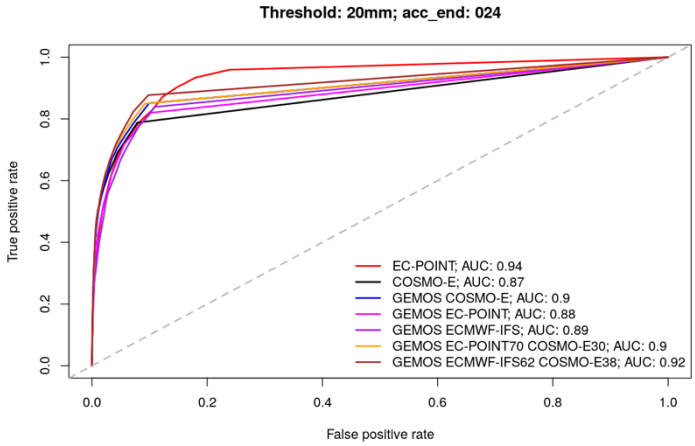


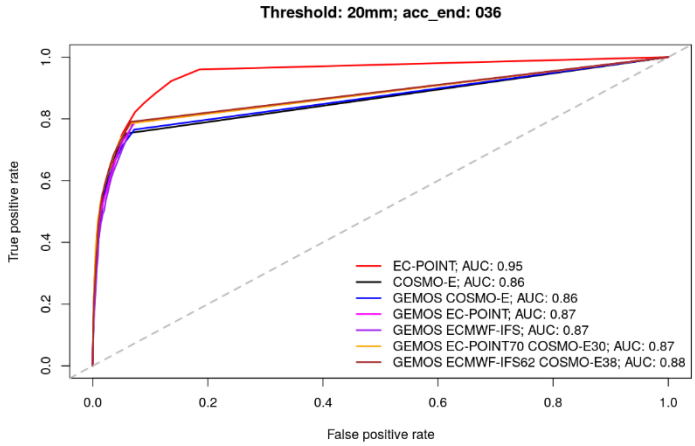
Figure 8: CRPSS of gEMOS applied to pooled COSMO-E and ecPoint ensembles with different COSMO-E ratios with ecPoint as reference forecasts against lead time (here denoted as end of 12 h accumulation period). Positive values mean the combination beats unadjusted ecPoint. Negative values mean unadjusted ecPoint is better.

Besides average skill, for precipitation in particular it is important to issue skillful predictions for thresholds like rain/no rain or some alert levels for heavy precipitation and flooding. Accordingly, we analyze discrimination ability in terms of ROC for thresholds of 1, 10, 20, and 50 mm / 12 h. Figure 9 toFigure 11 show example ROC curves for the threshold of 20 mm / 12h for different lead time accumulation periods. In each figure, subpanel a) refers to 12 to 24 UTC validation times, while b) refers to 0 to 12 UTC. Note that we do not consider the parts of the ROC curves lying to the right of the lowest probability threshold in order to avoid any misleading conclusions (Ben Bouallègue & Richardson, 2021).

For the short lead time accumulation period 12 to 24 h the gEMOS combinations perform best, i.e. they are to the left and above for the relevant probability thresholds. For longer lead time accumulation periods ecPoint and the gEMOS combinations perform quite similarly. Selected ROC plots for the other thresholds are shown in Appendix A1. In general, both the gEMOS combinations and ecPoint outperform the other forecast variants. The gEMOS combination tend to perform particularly well for low thresholds, while ecPoint catches up for higher thresholds, which is in line with the CRPS comparisons above. At higher thresholds a diurnal cycle in the ROC seems to exist, in that ROC curve are slightly better for 12 to 24 UTC validation times than for 0 to 12 UTC validation times. This cycle is more pronounced in raw COSMO-E and gEMOS than in ecPoint. Similar results were found in the MISTRAL (Meteo Italian SupercompuTing poRtAL) project for Italy (Gascón et al., 2022). The poorest ROC values occur in the second half of the night and in the morning. This is not in line with the diurnal cycle of CRPS. It looks like raw COSMO-E and gEMOS perform best in terms of CRPS at times of day without convection, while discrimination ability in terms of ROC curves is relatively poor at the same times of day.

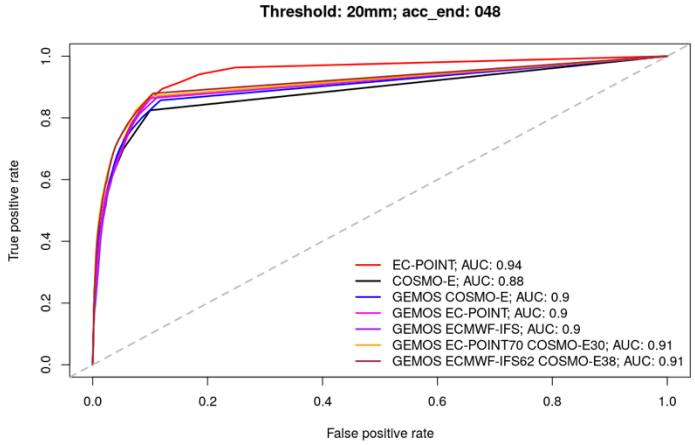


a) 12-24h

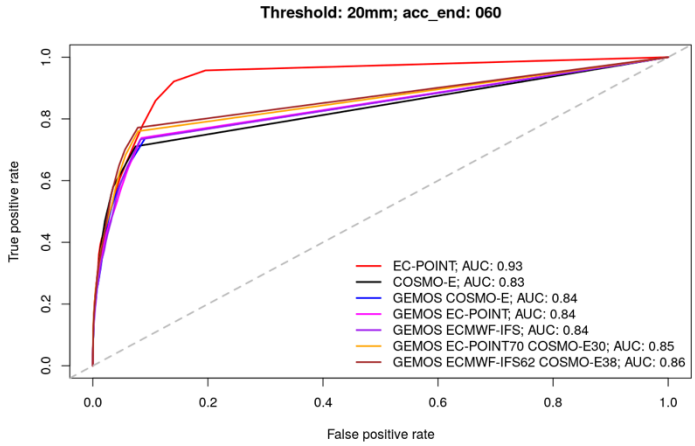


b) 24-36h

Figure 9: ROC curves and associated AUC values for ecPoint, COSMO-E, and different gEMOS postprocessed forecasts for a) accumulation period 12-24h and b) accumulation period 24-36h. gEMOS results are shown for the single model ensemble COSMO-E, ecPoint, ECMWF-IFS and for the pooled combinations of ecPoint or ECMWF-IFS with COSMO-E that performed best in terms of CRPS. The threshold is 20 mm/12h.

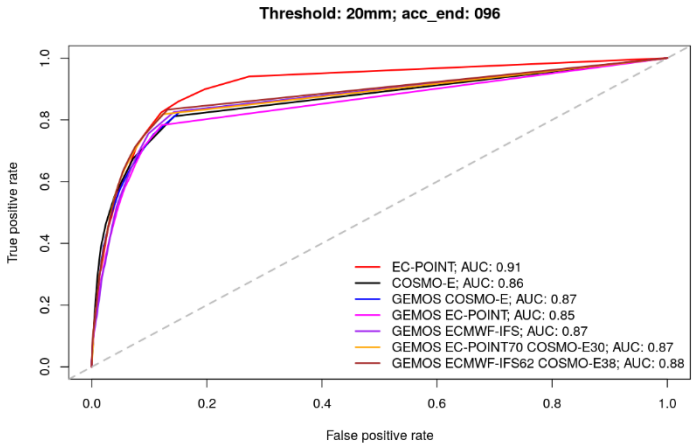


a) 36-48h

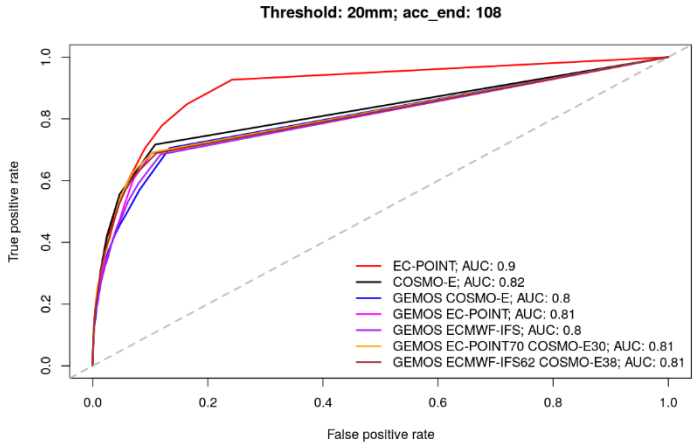


b) 48-60h

Figure 10: ROC curves and associated AUC values for ecPoint, COSMO-E, and different gEMOS postprocessed forecasts for a) accumulation period 36-48h and b) accumulation period 48-60h. gEMOS results are shown for the single model ensemble COSMO-E, ecPoint, ECMWF-IFS and for the pooled combinations of ecPoint or ECMWF-IFS with COSMO-E that performed best in terms of CRPS. The threshold is 20 mm/12h.



a) 84-96h



b) 96-108h

Figure 11: ROC curves and associated AUC values for ecPoint, COSMO-E, and different gEMOS postprocessed forecasts for a) accumulation period 84-96h and b) accumulation period 96-108h. gEMOS results are shown for the single model ensemble COSMO-E, ecPoint, ECMWF-IFS and for the pooled combinations of ecPoint or ECMWF-IFS with COSMO-E that performed best in terms of CRPS. The threshold is 20 mm/12h.

4 Discussion

Despite being based solely on the global ECMWF-IFS model, ecPoint performs well over Switzerland's complex topography. The higher the threshold considered and the longer the lead time, the better ecPoint performs compared to postprocessed COSMO-E predictions. Besides verifying ecPoint, this study confirms also that combining multiple models, here done by ensemble pooling prior to applying gEMOS, can improve forecast skill. When combining COSMO-E with ECMWF-IFS, the best gEMOS forecasts in terms of CRPS have been obtained with a pooled ensemble consisting of about two third ECMWF-IFS and one-third COSMO-E, despite the considerably coarser resolution of ECMWF-IFS. The same holds for the combination COSMO-E with ecPoint. The only difference is that the weights assigned to ecPoint for construction of the pooled ensemble resulting in the lowest gEMOS CRPS are slightly higher than the corresponding weights assigned to ECMWF-IFS in the former pooled ensemble.

As mentioned above, the diurnal cycles in CRPS and ROC are not in line. Poorest ROCs are observed in the second half of the night and in the morning while CRPS exhibits the poorest values for the accumulation period spanning from the afternoon to 00 UTC. On the one hand, one would expect poorer ROCs in the afternoon because of low predictability of convective precipitation. On the other hand, ECMWF-IFS's predictions of afternoon convective rainfall perform quite well, but evening and morning suffer a bit because of diurnal cycle errors. Another possible explanation of this behavior may be related to the short verification period and a diurnal cycle of climatological precipitation. That is, a specific threshold refers to climatological quantile that is probably different in the morning to what it is in the afternoon. This may affect ROC. Moreover, the verification period is rather short. Therefore, ROC for high thresholds may also be dominated by only a few events.

Despite its advantages (Table 1 in Hewson and Pillosu, 2021), ecPoint also entails drawbacks. First, though being computationally cheap compared to running a limited area NWP model, gEMOS is even cheaper. Second, the forecast percentile output of ecPoint is not related to specific ECMWF-IFS ensemble members. Hence, it is not straightforward to create a physics-based template of spatio-temporal dependence structure from ecPoint forecast output in its current form. The generation of realistic scenarios based on ecPoint using ensemble copula coupling (Schefzik et al., 2013) might need a template from another source, e.g. COSMO-E or ECMWF-IFS. Third, ecPoint forecasts the same predictive distribution for any point within a grid cell. It does not account for systematic intra grid-cell variability, e.g. induced by local topography. So, for events that are strongly driven by local topography, like e.g. orographic precipitation, we presume that a high resolution local model like COSMO-1E (Kaufmann & Rüdisühli, 2019) outperforms ecPoint. Fourth, very short accumulation periods below 6 h are not covered by ecPoint. Hewson & Pillosu (2021) considered 12 h accumulation periods. ECMWF produces also 6 h ecPoint output, running twice daily. For 6 h ecPoint, topography and local solar time have been included as predictors in the decision tree. Therefore, we expect 6 h ecPoint to perform well over complex topography. Hence, we suggest considering also 6 h accumulation periods in a follow-up study.

As the scope of this preliminary study was limited, we have not performed any in depth analyses of the spatial and temporal pattern of ecPoint skill relative to COSMO-E and gEMOS. However, identifying regions and weather regimes in which ecPoint's skill is particularly good or poor may be beneficial. Further, since this study is based only on 10 months of data, the results need to be viewed with some caution. As ecPoint is now available for a longer period, it would be worth to perform another analysis with a longer verification period.

This would also allow testing alternative gEMOS training periods, which may lead to a more skillful benchmark. In particular, the different sizes of the training data sets, which are 45 days of data (over a limited area) for gEMOS and a full year of global data for ecPoint, may have affected the results of this study. As COSMO-1E and COSMO-2E have replaced COSMO-E at MeteoSwiss in the meantime, we suggest considering these issues in a follow-up study based on COSMO-1E and COSMO-2E.

Another issue in this study is related to how we have constructed the ROC curves. As stated by Ben Bouallègue & Richardson (2021) interpretation of ROC, and in particular the trapezoidal AUC values, may be misleading if not done appropriately. The gEMOS and the ecPoint forecasts used for generating the ROC curves in this study are based on the same discretization, i.e. same number of percentiles of the forecast distribution. As a simple increase in the discretization of gEMOS may improve the parts of the curve to the right of the lowest probability threshold considerably, the ROC curves shown in this study can only be interpreted up to the lowest available probability threshold. On top of that, the discretization innate to COSMO-E is much coarser, i.e. 21 members. However, as running a hypothetical 100 member COSMO-E ensemble would be computationally very expensive, one could argue that it is fair to use trapezoidal AUC, which tends to penalize small ensemble size. Note that the issue of small COSMO-E ensemble size when computing its ROC could be alleviated by improving the ROC using the ensemble mean based correction approach as described in Ben Bouallègue & Richardson (2021). Moreover, forecast bias affects also ROC and AUC. For the postprocessed forecasts, gEMOS and ecPoint, we assume that bias should not be a big issue. However, the effects on COSMO-E ROC are unclear. Accordingly, we hypothesize that our ROC based comparisons of discrimination ability of the different postprocessed forecasts is still valid as long as one focuses on the part of the ROC up the lowest probability threshold only. The comparison with raw COSMO-E ROC shows potential for postprocessing and increasing ensemble size. Hewson and Pillosu (2021) and Gascón et al. (2022) compared the AUC values also to AUC of a climatological reference. Due to the very restricted scope of this study and its focus on the comparison between ecPoint and COSMO-E, we did not include the climatological reference. We suggest including it and also considering the former AUC related issue in a follow-up study based on COSMO-1E and COSMO-2E.

As mentioned above, ecPoint performs well in comparison with gEMOS for high thresholds over Switzerland. Accordingly, we suppose that ecPoint may provide valuable information for warning applications. In particular, warnings associated with convective precipitation events and related flash floods, which occur at small spatial scales, may benefit from ecPoint. Moreover, ecPoint output could serve as a low-cost method to obtain calibrated probabilistic precipitation forecasts, which would not require any additional statistical postprocessing to be performed by MeteoSwiss.

5 Conclusions

This study shows that ecPoint forecasts for 12 h accumulated precipitation perform well in terms of overall skill. Additionally, ecPoint exhibits a decent discrimination ability particularly for higher thresholds and longer lead times. At those thresholds and lead times, it tends to perform well in comparison with gEMOS postprocessed pooled COSMO-E and ECMWF-IFS ensembles. Moreover, our analysis of ROC stresses also the benefits from the large “ensemble size” of ecPoint. To be able to draw conclusions on discrimination ability that are more reliable we suggest performing a follow-up study that takes the guidelines by Ben Bouallègue &

Richardson (2021) into account. Additionally, ecPoint is efficient in terms of the need for reforecasts in that its training is based only on one year of forecasts from a global model. gEMOS with a moving window training period needs even less reforecasts, but at least in our study it relies on the availability of a high-resolution limited area NWP ensemble model like COSMO-E. Furthermore, we presume that ecPoint performs equally well in other regions besides Switzerland.

Moreover, our study is limited in that we did not stratify the verification on different locations and weather regimes. Hence, we suggest comparing ecPoint with postprocessed forecasts from local high-resolution models in a stratified manner in order to detect when and where forecast quality might not benefit from using ecPoint. Additionally, we suggest testing longer gEMOS training periods and increasing the length of the verification period in order to obtain more reliable results, in particular for extreme events.

Acknowledgements

This study has been performed within the framework of a short-term secondment to ECMWF. The first author is grateful to F. Pillosu, A. Montani, Z. Ben Bouallègue and D. Richardson for detailed and valuable discussions during his stay at ECMWF. The authors thank D. Richardson also for very important comments in the ECMWF internal review process. Moreover, we are grateful to D. Nerini and M. Schaer for helpful comments.

References

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. *Monthly Weather Review*, *139*(12), 3887-3905.

Ben Bouallègue, Z., & Richardson, D. S. (2022). On the ROC area of ensemble forecasts for rare events. *Weather and Forecasting*, 37(5), 787-796.

Ferro, C. A. T. (2014). Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, *140*(683), 1917-1923.

Ferro, C. A. T, Richardson, D. S., & Weigel, A. P. (2008). On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, 15(1), 19-24.

Friedli, L., Ginsbourger, D., & Bhend, J. (2021). Area-covering postprocessing of ensemble precipitation forecasts using topographical and seasonal conditions. *Stochastic Environmental Research and Risk Assessment*, *35*(2), 215-230.

Gascón, E., Montani, A., & Hewson, T. (2022). Post-processing output from ensembles with and without parametrised convection, to create accurate, blended, high-fidelity rainfall forecasts. In preparation.

Gneiting, T., Raftery, A. E., Westveld III, A. H., & Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, *133*(5), 1098-1118.

Hewson, T. D., & Pillosu, F. M. (2021). A low-cost post-processing technique improves weather forecasts around the world. *Communications Earth & Environment*, *2*(1), 1-10.

Kaufmann, P. & Rüdisühli S. (2019). Dispersion model uncertainty simulation using a limited area ensemble model. *19th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, 3-6 June 2019, Bruges, Belgium.*

Klasa, C., Arpagaus, M., Walser, A., & Wernli, H. (2018). An evaluation of the convection-permitting ensemble COSMO-E for three contrasting precipitation events in Switzerland. *Quarterly Journal of the Royal Meteorological Society*, *144*(712), 744-764.

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management science*, *22*(10), 1087-1096.

Messner, J. W., Mayr, G. J., Wilks, D. S., & Zeileis, A. (2014a). Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Monthly Weather Review*, *142*(8), 3003-3014.

Messner, J. W., Mayr, G. J., & Zeileis, A. (2016). Heteroscedastic Censored and Truncated Regression with crch. *The R Journal*, 8(1), 173.

Messner, J. W., Zeileis, A., Broecker, J., & Mayr, G. J. (2014b). Probabilistic wind power forecasts with an inverse power curve transformation and censored regression. *Wind Energy*, *17*(11), 1753-1766.

Owens, R. G., & Hewson, T. D. (2018). ECMWF forecast user guide. *Reading: ECMWF, doi: 10.21957/m1cs7h*

Schefzik, R., Thorarinsdottir, T. L., & Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical science*, *28*(4), 616-640.

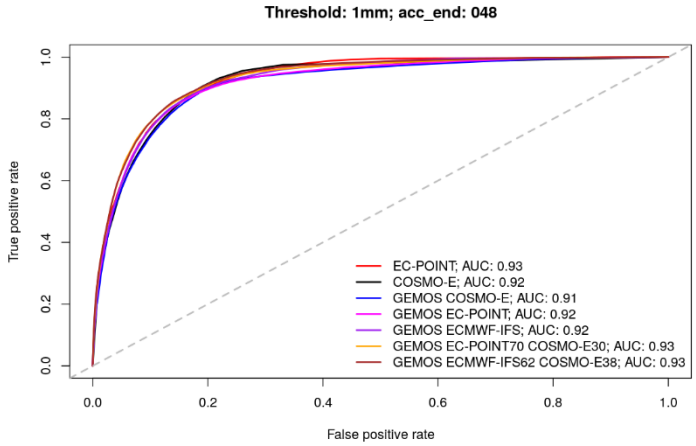
Vannitsem, S., et al. (2021). Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World. *Bulletin of the American Meteorological Society*, *102*(3), E681-E699.

Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2008). Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *134*(630), 241-260.

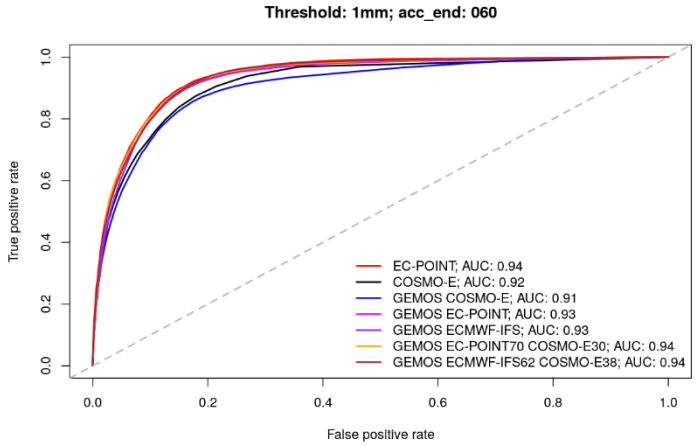
Weiss, A. (2001). Topographic position and landforms analysis. In *Poster presentation, ESRI user conference, San Diego, CA* (Vol. 200).

Appendix

A1 Additional ROC figures

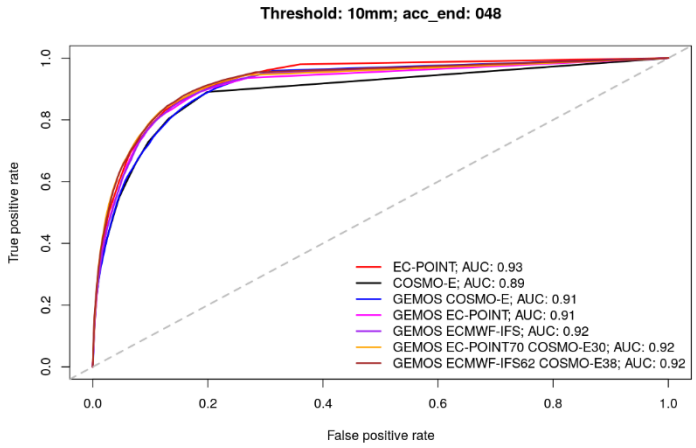


a) 36-48h

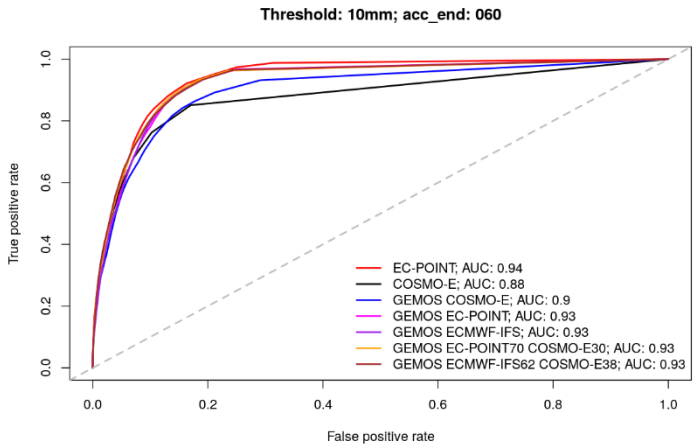


b) 48-60h

Figure A1: ROC curves and associated AUC values for ecPoint, COSMO-E, and different gEMOS postprocessed forecasts for a) accumulation period 36-48h and b) accumulation period 48-60h. gEMOS results are shown for the single model ensemble COSMO-E, ecPoint, ECMWF-IFS and for the pooled combinations of ecPoint or ECMWF-IFS with COSMO-E that performed best in terms of CRPS. The threshold is 1 mm/12h.

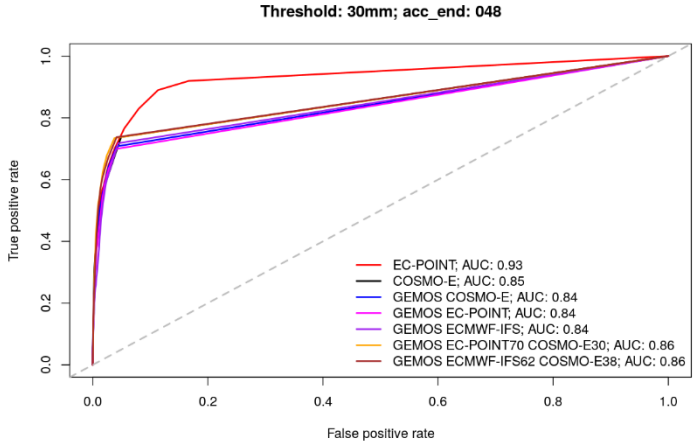


a) 36-48h

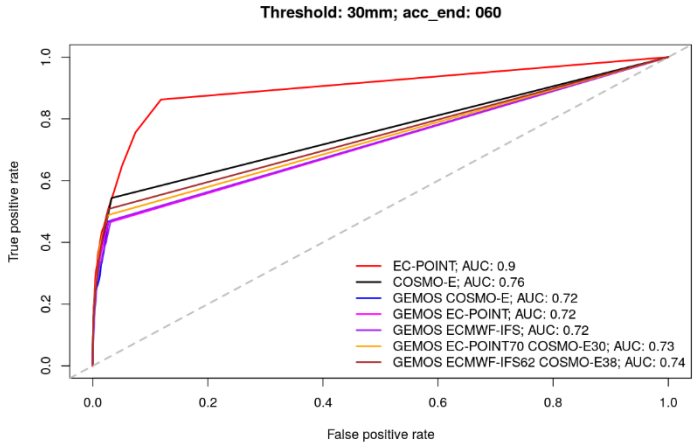


b) 48-60h

Figure A2: ROC curves and associated AUC values for ecPoint, COSMO-E, and different gEMOS postprocessed forecasts for a) accumulation period 36-48h and b) accumulation period 48-60h. gEMOS results are shown for the single model ensemble COSMO-E, ecPoint, ECMWF-IFS and for the pooled combinations of ecPoint or ECMWF-IFS with COSMO-E that performed best in terms of CRPS. The threshold is 10 mm/12h.

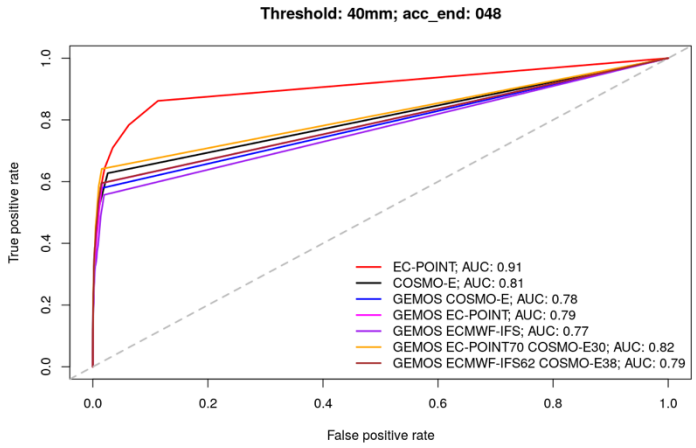


a) 36-48h

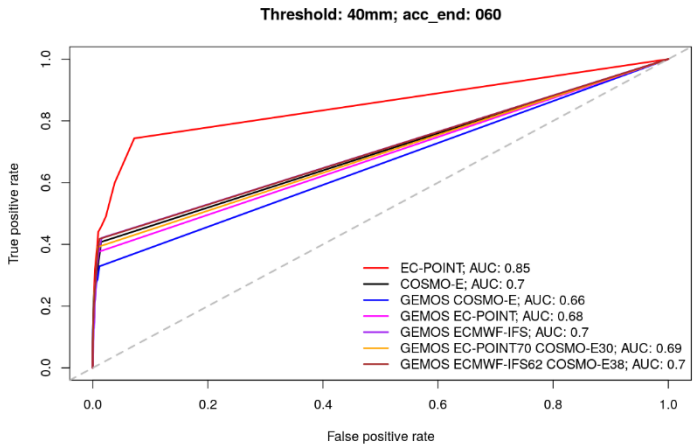


b) 48-60h

Figure A3: ROC curves and associated AUC values for ecPoint, COSMO-E, and different gEMOS postprocessed forecasts for a) accumulation period 36-48h and b) accumulation period 48-60h. gEMOS results are shown for the single model ensemble COSMO-E, ecPoint, ECMWF-IFS and for the pooled combinations of ecPoint or ECMWF-IFS with COSMO-E that performed best in terms of CRPS. The threshold is 30 mm/12h.

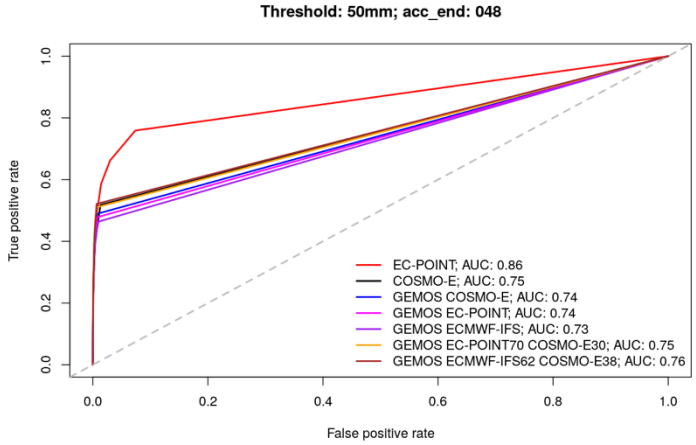


a) 36-48h

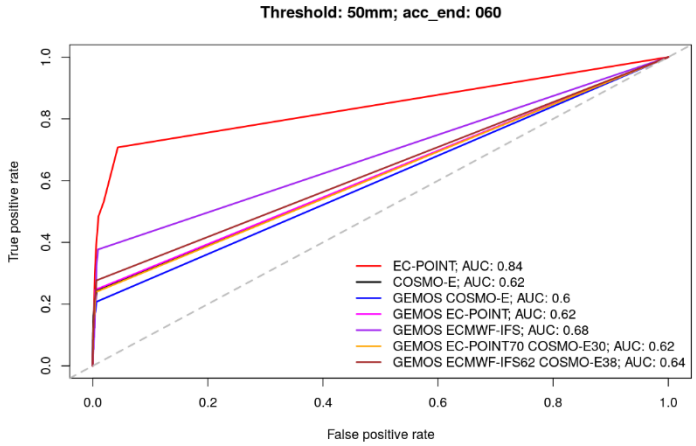


b) 48-60h

Figure A4: ROC curves and associated AUC values for ecPoint, COSMO-E, and different gEMOS postprocessed forecasts for a) accumulation period 36-48h and b) accumulation period 48-60h. gEMOS results are shown for the single model ensemble COSMO-E, ecPoint, ECMWF-IFS and for the pooled combinations of ecPoint or ECMWF-IFS with COSMO-E that performed best in terms of CRPS. The threshold is 40 mm/12h.



a) 36-48h



b) 48-60h

Figure A5: ROC curves and associated AUC values for ecPoint, COSMO-E, and different gEMOS postprocessed forecasts for a) accumulation period 36-48h and b) accumulation period 48-60h. gEMOS results are shown for the single model ensemble COSMO-E, ecPoint, ECMWF-IFS and for the pooled combinations of ecPoint or ECMWF-IFS with COSMO-E that performed best in terms of CRPS. The threshold is 50 mm/12h.