SPECIAL PROJECT PROGRESS REPORT

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

Reporting year	2020
Project Title:	High-impact precipitation events prediction with convection-permitting models nested in the ECMWF ensemble: new tests with the MOLOCH and Meso-NH models
Computer Project Account:	SPITCAPE
Principal Investigator(s):	Valerio Capecchi (capecchi@lamma.rete.toscana.it)
Affiliation:	LaMMA Consortium - Environmental Modelling and Monitoring Laboratory for Sustainable Development
Name of ECMWF scientist(s) collaborating to the project (if applicable)	………………………………………………………….…… ………………………………………………………….……
Start date of the project:	2019
Expected end date:	2021

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

(if applicable)
Please answer for all project resources

		Previous year		**Current year**	
		Allocated	Used	Allocated	Used
High Performance Computing Facility	(units)	2 900 000	2 893 655	2 900 000	0
Data storage capacity	(Gbytes)	40 000	35 553	60 000	35 553

Summary of project objectives (10 lines max)

The SPITCAPE 2019-2021 Special Project (SP) was conceived to be the continuation of the activities carried out during the SPITCAPE 2016-2018 SP. The goal is to reforecast three heavy precipitation events (Cinque Terre October 2011, Genoa November 2011 and October 2014) by using two mesoscale models (the MOLOCH and Meso-NH models) for the convection-permitting ensemble simulations. The global ensembles produced using the IFS model cycle 41r2 at the spectral resolution TCo639, provide the initial and boundary conditions. The comparison between the results obtained with the WRF model during the first stage of the SPITCAPE SP and those obtained with the two additional models, contribute to the debate regarding the strengths and weaknesses of these three models with respect to: (i) the accuracy of the results for the three events considered, (ii) the integration with ECMWF products, (iii) the ease of implementation and (iv) the computational costs in view of a potential use for operational ensemble forecasting.

Summary of problems encountered (10 lines max)

none

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

In the rest of the current year, we will reforecast the third and last heavy precipitation event (the flooding of Genoa occurred on the 9th of October 2014). The limited-area models implemented will be the Meso-NH and MOLOCH. The numerical outputs will be analysed and compared with those obtained with the WRF model during the first stage of the Special Project (SPITCAPE 2016-2018).

List of publications/reports from the project with complete references

Capecchi, Valerio "Reforecasting Two Heavy-Precipitation Events with Three Convection-Permitting Ensembles", *Weather and Forecasting* 36, 3 (2021): 769-790, https://doi.org/10.1175/WAF-D-20-0130.1

Summary of results

If submitted **during the first project year**, please summarise the results achieved during the period from the project start to June of the current year. A few paragraphs might be sufficient. If submitted **during the second project year**, this summary should be more detailed and cover the period from the project start. The length, at most 8 pages, should reflect the complexity of the project. Alternatively, it could be replaced by a short summary plus an existing scientific report on the project attached to this document. If submitted **during the third project year**, please summarise the results achieved during the period from July of the previous year to June of the current year. A few paragraphs might be sufficient.

Following, we report a short introduction to the project's goals, a description of the methods and the main achievements (relevant results shown in Figures 8, 9, 13, 14 and 15). More detailed information are available in Capecchi, V. "Reforecasting Two Heavy-Precipitation Events with Three Convection-Permitting Ensembles", *Weather and Forecasting* 36, 3 (2021): 769-790, https://doi.org/10.1175/WAF-D-20-0130.1

Reforecasting two heavy-precipitation events with three convection-permitting ensembles

Valerio Capecchi*

LaMMA, Laboratorio di Meteorologia e Modellistica Ambientale per lo sviluppo sostenibile, Firenze, Italia

ABSTRACT

We investigate the potential added value of running three limited-area ensemble systems (with the WRF, Meso-NH and MOLOCH models and a grid spacing of approximately 2.5 km) for two heavy-precipitation events in Italy. Such high-resolution ensembles include an explicit treatment of convective processes and dynamically downscale the ECMWF global data, which have a grid spacing of approximately 18 km. The predictions are verified against rain-gauge data and their accuracy is evaluated over that of the driving coarser-resolution ensemble system. Furthermore, we compare the simulation speed (defined as the ratio of simulation length to wall-clock time) of the three limited-area models to estimate the computational effort for operational convection-permitting ensemble forecasting. We also study how the simulation wall-clock time scales with increasing numbers of computing elements (from 36 to 1152 cores). Objective verification methods generally show that convection-permitting forecasts outperform global forecasts for both events, although precipitation peaks remain largely underestimated for one of the two events. Comparing simulation speeds, the MOLOCH model is the fastest and the Meso-NH is the slowest one. The WRF model attains efficient scalability, whereas it is limited for the Meso-NH and MOLOCH models when using more than 288 cores. We finally demonstrate how the model simulation speed has the largest impact on a joint evaluation with the model performance because the accuracy of the three limited-area ensembles, amplifying the forecasting capability of the global predictions, does not differ substantially.

1. Introduction

[. . .]

This paper presents the results produced in the framework of two ECMWF Special Projects, the computational resources of which were granted during the years 2016-2018 and 2019-2021. The common goal of the two projects is to assess the added value of running a limited-area CP ensemble in terms of quantitative precipitation forecast (QPF). The accuracy of the cascade of state-of-the-art ensembles, from global to local, is evaluated by reforecasting past high-impact precipitation events and using three different mesoscale models. The dynamical downscaling method is chosen to start the regional ensembles, and the forecast lengths considered are longer than 24 hours.

[. . .]

The comparison of the results obtained with these three models contributes to the debate regarding their strengths and weaknesses with respect to (i) the accuracy of the results for the two events considered and (ii) the computational costs in view of the potential use for operational ensemble forecasting.

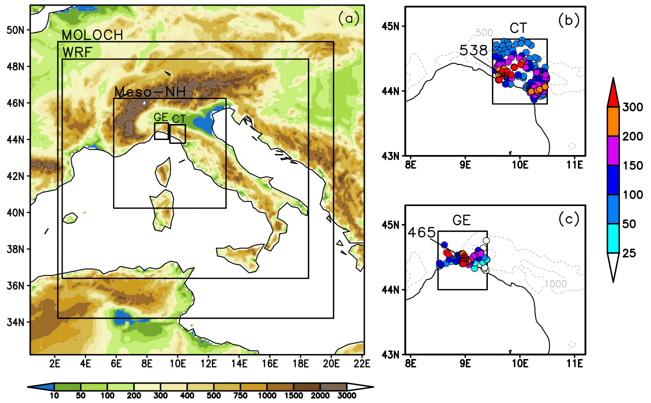


Fig. 1. Panel (a): topography and geographical extent of the three domains of integration for the limited-area models. Panel (b): area of interest of the CT case with observed rainfall data registered by rain-gauges on the 25th of October 2011. Panel (c): area of interest of the GE case with observed rainfall data registered by rain-gauges on the 4th of November 2011. The maximum observed value is reported for each event (units of mm/day). In panels (b) and (c), the gray dashed contours indicate elevations of 500 and 1000 meters above sea level.

2. The CT and GE cases

Although detailed descriptions of the CT and GE cases can be found in the literature, we include a short summary of both to make this paper self-contained.

[. . .]

**Corresponding author*: Valerio Capecchi, capecchi@lamma.toscana.it

Table 1. Setup of the key characteristics of the limited-area and convection-permitting ensemble forecasts.

Model	Grid Spacing (km)	Rows×Columns	Vertical Levels	Grid Points	Time Step (sec)
	Δx	$R \times C$	L	$R \times C \times L$	Δt
WRF	3	400×440	55	≃ 9.7 million	18
Meso-NH	2.5	225×270	52	≃ 3.1 million	6
MOLOCH	2.5	514×614	50	≃ 15.4 million	30

3. Models and numerical setup

In the following, we provide a short overview of the models used in this study. We stress again the fact that they all are set with the explicit treatment of convective processes.

[. . .]

In Table 1, we summarize a few basic settings of the integration domain for the three models, namely, the grid spacing (expressed in km), the number of rows, columns and vertical levels, the resulting total number of grid points and the time step (expressed in seconds). The grid spacing is set to 3 km for the WRF model and 2.5 km for the Meso-NH and MOLOCH models. The extent of the horizontal grid is not the same among the three models; the extents are shown in Figure 1. The number of vertical levels spans from 50 (for the MOLOCH model) to 55 (for the WRF model). With these settings we obtain a number of grid points, for the three dimensional grid, ranging from approximately 3.1 million for the Meso-NH model to approximately 9.7 million for the WRF model and up to approximately 15.4 million for the MOLOCH model. To satisfy the Courant-Friedrichs-Lewy (CFL) stability condition, we set the time step to 18, 6 and 30 seconds for the WRF, Meso-NH and MOLOCH models, respectively.

We use different compilers and compilation options to build the executables on the ECMWF supercomputer. Some details are given in the Appendix.

4. Data and methods

[. . .]

Starting dates that have been considered for the CT (GE) case are from 00 UTC 23 October (2 November) 2011 to 12 UTC 24 October (3 November) 2011, every 12 hours. The ending dates are 00 UTC 26 October 2011 for CT and 00 UTC 5 November 2011 for GE so that the forecast length ranges from 72 hours to 36 hours for both cases; forecast lengths shorter than 36 hours were not considered. See Table 2 for a summary of the simulations and the codes adopted to name each forecast. In the following text, we use the acronyms WRF-ENS, MNH-ENS and MOL-ENS to refer to the CP ensembles produced using the WRF, Meso-NH and MOLOCH models, respectively, and using the ENS data as the initial and boundary conditions. The number of members of each CP ensemble system is the same as that in the ENS data (i.e., 50 members).

Table 2. Summary of the numerical simulations performed. The third column indicates the codes adopted to name the forecasts.

CT case	Starting date of the simulations	Forecast length to 00 UTC 26 Oct 2011	Forecast code
	00 UTC 23 Oct 2011	72 hours	CT+72h
	12 UTC 23 Oct 2011	60 hours	CT+60h
	00 UTC 24 Oct 2011	48 hours	CT+48h
	12 UTC 24 Oct 2011	36 hours	CT+36h
GE case	**Starting date of the simulations**	**Forecast length to 00 UTC 5 Nov 2011**	**Forecast code**
	00 UTC 2 Nov 2011	72 hours	GE+72h
	12 UTC 2 Nov 2011	60 hours	GE+60h
	00 UTC 3 Nov 2011	48 hours	GE+48h
	12 UTC 3 Nov 2011	36 hours	GE+36h

Table 3. Percentiles and maximum values of the daily rainfall data observed during the CT and GE cases.

	25th percentile	Median	75th percentile	Max
CT	59	110	168	538
GE	80	130	172	465

The QPF data are compared with observed precipitation amounts collected at the rain-gauges belonging to the inset boxes shown in Figure 1 (panels on the right). Such boxes are chosen subjectively by drawing a $1^\circ \times 1^\circ$ square around the areas for which the rain-gauges registered the highest precipitation amounts. The total numbers of rain-gauges are 149 and 55 for the CT and GE cases, respectively. The basic statistics of the daily accumulated precipitation observed during the two events are reported in Table 3. When presenting or discussing the results regarding the verification of the model predictions against observed data, we refer to the precipitation that occurred in the 24-hour period ending at 00 UTC 26 October (5 November) 2011 for the CT (GE) case.

To reduce the effects of the double-penalty error (Ebert 2009), when extracting the QPF values at rain-gauge locations, we picked the four nearest-neighbor grid values and averaged them to provide the forecast value at that location. The performance of the ensemble mean, chosen as the representative member of each ensemble system, is assessed by looking at the performance diagrams (Roebber 2009). Such diagrams plot four measures of the dichotomous forecast: probability of detection (POD), success ratio (SR), bias and critical success index (CSI). Using the 2 × 2 contingency table for the dichotomous (yes/no) forecast shown in Table 4, the four skill measures are defined as follows:

$$POD = \frac{A}{A+C},$$
$$SR = 1 - \frac{B}{A+B},$$
$$bias = \frac{A+B}{A+C},$$
$$CSI = \frac{A}{A+B+C}.$$

To estimate potential heavy rainfall, we evaluate the maps of the probability of precipitation (PoP) exceeding predefined thresholds. The PoP is a common ensemble-based product, which expresses the occurrence probability of an extreme event measured by the fraction of ensemble members that predict a value higher than a predefined threshold. The probabilistic skills of the CP ensembles are compared to those of ENS by constructing the receiver operating characteristic (ROC) curve and calculating the area under it (Mason 1982). The ROC curve contrasts the hit rate versus false alarm rate, using a set of increasing probability thresholds to make the yes/no decision. The area under the ROC curve is frequently used as an index of accuracy of an ensemble system in order to be able to discriminate between the occurrence and nonoccurrence of weather events; the higher the value is, the better it is, with 1 as the upper limit and values below 0.5 indicate no skill compared to a random forecast.

Following the notations of Coiffier (2011), the simulation speed of the generic model M is defined as the ratio between the forecast length H over the time T_M required to end the simulation. It can be expressed by the following relationship:

$$\frac{H}{T_M} = \frac{\Delta t_M \cdot S}{N_{v,M} \cdot N_{c,M}} \tag{1}$$

where Δt_M is the time step, which depends on the grid spacing Δx_M and has to satisfy the CFL condition. The numerator S is a measure of the computational speed (e.g., the number of processing elements or the floating operations per second). The term $N_{v,M}$ is the number of variables to be processed at each time step Δt_M and depends on the number of grid points (i.e., the number of rows, columns and vertical levels) of the integration domain. The term $N_{c,M}$ represents the number of calculations to be made at each time step Δt_M and is a function of the computational cost required by the numerical method used to solve the equations. In view of a possible use for operational ensemble forecasting, we compare how the simulation speed, defined by the left-hand side of equation 1, of model M scales as the computer speed S increases. The factor $\frac{H}{T_i}$ is taken as a measure of actual time-to-solution. In this study, the index M can assume the value of WRF, Meso-NH, or MOLOCH. We stress the fact that this evaluation is not biased towards either the number of grid points of the integration domain or the time step adopted. In fact, with the settings summarized in Table 1, the ratio $N_{v,M}/\Delta t_M$ is almost constant for all the CP models.

Table 4. The 2 × 2 contingency table.

		Event Observed	
		yes	no
Event Forecast	yes	A	B
	no	C	D

To jointly evaluate a numerical weather model M in terms of its simulation speed S_M and performance P_M, we heuristically define the linear integrated speed-performance ($LISP$) index as a linear combination of S_M and P_M, namely:

$$LISP_M(\alpha) = (1-\alpha)S_M + \alpha P_M, \tag{2}$$

where the scalar $\alpha \in [0,1] \subseteq \mathbb{R}$ is a weight such that if $\alpha = 1$, then the $LISP_M$ index is conditioned on having the more accurate forecast data, regardless of the time needed to accomplish the simulation, whereas if $\alpha = 0$ then the $LISP_M$ index weights the faster forecast (provided that the accuracy satisfies some minimum requirement, i.e. ROC area > 0.5). We note that, if we choose the ROC area as a measure of the performance P_M, we have that $LISP_M \geq 0, \forall \alpha \in [0,1]$ and the higher the value is, the better it is. If we have to evaluate models M and N on the basis of both the performance and speed, we can evaluate if:

$$LISP_M \geq LISP_N,$$

that is if:

$$P_M \geq P_N + \kappa(S_N - S_M).$$

where, for sake of simplicity, we set $\kappa = \left(\frac{1-\alpha}{\alpha}\right)$.

5. Results

a. Model performance: precipitation verification

In Figures 2 and 3, we show the QPF ensemble mean maps for CT+36h and GE+36h; longer-range forecasts do not provide substantial differences. The visual comparison of such maps with observed precipitation patterns shown in panels (b) and (c) of Figure 1 suggests that the predicted ensemble means strongly underestimate the rainfall for both cases. To quantitatively analyze such underestimation, in Table 5,

TABLE 5. Root mean square error (RMSE) and mean error (ME) between the observed and predicted 24-hour accumulated precipitation values for the CT and GE cases.

		WRF-ENS	**MNH-ENS**	**MOL-ENS**	**EN**
CT	RMSE (mm)	103	89	93	10
	ME (mm)	-64	-44	-56	-6
GE	RMSE (mm)	138	129	126	12
	ME (mm)	-94	-71	-59	-6

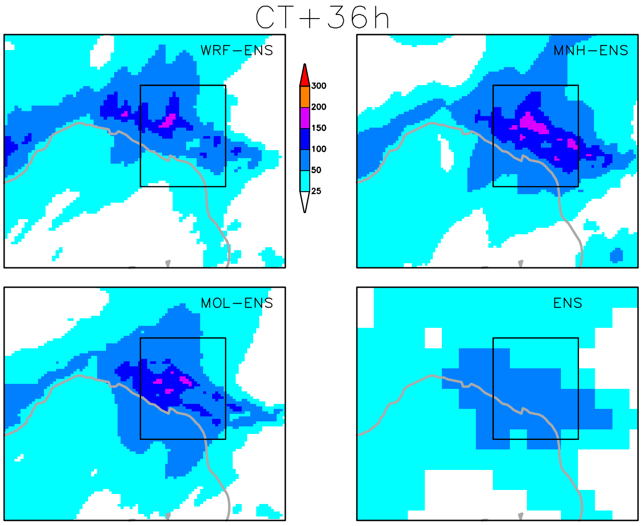


FIG. 2. CT case: 24-hour accumulated precipitations for the WRF-ENS (top-left panel), MNH-ENS (top-right panel), MOL-ENS (bottom-left panel) and ENS (bottom-right panel) ensemble mean forecast. The forecast length is 36 hours.

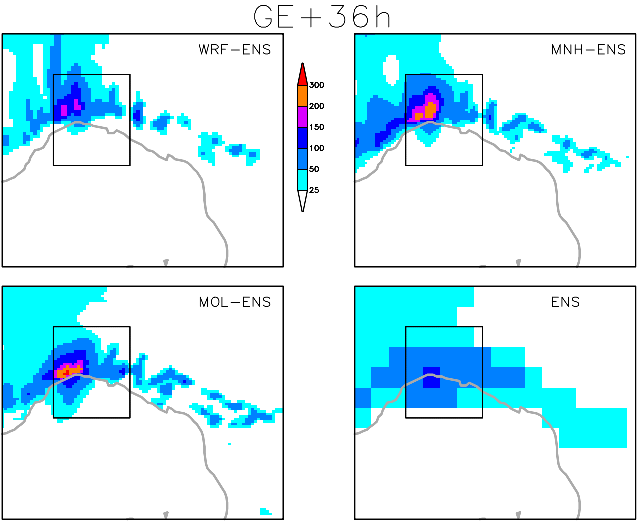


FIG. 3. As in Figure 2 but for the GE case.

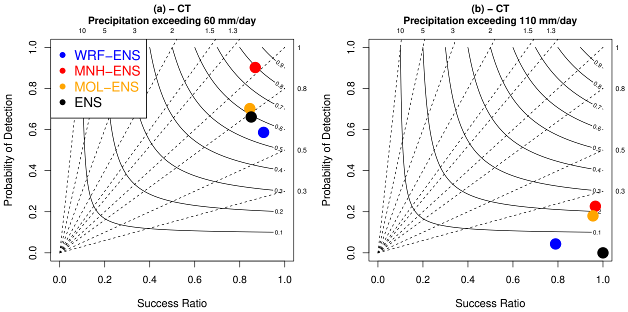


FIG. 4. CT case: performance diagram of the ensemble mean forecast of WRF-ENS (blue), MNH-ENS (red), MOL-ENS (orange) and ENS (black). The X-axis shows the success ratio (SR), the Y-axis shows the probability of detection (POD), the curved lines represent the critical success index (CSI) values, and the dashed diagonal lines represent the bias. Panel (a) shows the scores for the precipitation threshold corresponding to the 25th percentile of the observed accumulated rainfall. Panel (b) as in (a) but for the 50th percentile.

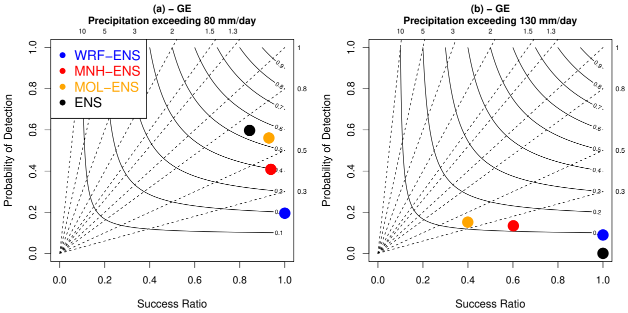


FIG. 5. As in Figure 4 but for the GE case.

we show the root mean square errors (RMSEs) and mean errors (MEs) between the predicted and observed precipitation values (Wilks 2011); the values are averaged among all the forecast lengths. For the CT case, the RMSEs range from 89 mm (for the MNH-ENS ensemble mean) to 103 mm (for the WRF-ENS ensemble mean). The MNH-ENS ensemble mean also provides the best ME (i.e., closest to 0). For the GE case, we obtain an average RMSE of approximately 129 mm, with the ENS ensemble mean providing the lower value (123 mm) and WRF-ENS providing the higher value (138 mm). The analysis of the ME produces similar conclusions.

To further assess the accuracy of the ensemble means, in Figures 4 and 5, we show the performance diagrams for the CT and GE cases, respectively. To make the yes/no decision, we chose two precipitation thresholds corresponding to the 25th percentile and the median of the observed precipitation data (see Table 3); the scores are averaged among all the forecast ranges. For the precipitation threshold equal to the 25th percentile (see panel (a) in the Figures), MNH-ENS provides more skillful predictions regarding CT; in fact, the score lies close to the top-right corner. For the GE case, ENS outperforms the other systems, having a bias of approximately 0.7 and relatively higher POD and CSI scores (approximately 0.6 for both). Considering the precipitation threshold equal to the median of the observed data (see panel (b) in the Figures), all the ensemble systems provide poor results (i.e. all the scores lie close to the bottom-right corner).

Since the QPF ensemble means for both cases are skillful only for the precipitation threshold equal to the 25th percentile of the observed values, we also evaluated some ensemble-based verification metrics. In Figures 6 and 7, we show the PoP maps for the four ensemble systems exceeding the medians of the observed values, which correspond to approximately 110 and 130 mm for the CT and GE cases, respectively. The forecast length is 36 hours, and longer-range forecasts do not provide results that differ substantially. Crosses represent the locations of rain-gauges where rainfall amounts greater than the threshold were actually registered. As regards the CT case (see Figure 6), the visual agreement between observations (Figure 1 panel (b)) and PoP patterns appears good for all the ensembles. On average, the PoP values extracted at the rain-gauge locations are approximately 24%, 51%, 35% and 10% for WRF-ENS, MNH-ENS, MOL-ENS and ENS, respectively. Only ENS fails to produce valuable information (i.e., PoP values <5%) for 27 out of 75 rain-gauges located in the northern part of the CT box (see the bottom-right panel in Figure 6). As regards the GE case, the PoP values are concentrated in a small portion of the domain and follow the pattern of the observations (see Figure 1 panel (c)). MNH-ENS and MOL-ENS produce darker shaded areas than WRF-ENS, causing higher false alarm ratios. In fact, we found that 1, 5 and 4 out of 26 locations that did not record 130 mm of rainfall were located within the PoP>50% contour for WRF-ENS, MNH-ENS and MOL-ENS, respectively. The ENS PoP map (bottom-right panel in Figure 7) has only one grid-point with a PoP value greater than 20%, but the misplaced position of this point causes both the yes and the no events to be incorrectly predicted.

To quantitatively evaluate the probabilistic skills of the ensembles, we show the area underneath the ROC curve for the CT and GE cases in Figures 8 and 9, respectively, by varying the precipitation thresholds on the X-axis and considering different forecast ranges. The upper limit on the X-axis is set to the 75th percentile of the observed rainfall amount (which corresponds to approximately 170 mm for both cases). In general, the CP forecasts outperform the ENS forecasts (i.e., the CP curves lie above the ENS ones) for all thresholds and for all lead times with a few exceptions (e.g., CT+48h and GE+48h). As regards the CT case (Figure 8), the ENS skill drops below the critical value of 0.5 approximately at the 130-mm precipitation threshold, whereas the CP ensembles provide valuable information (i.e., ROC area > 0.5) up to 170 mm and beyond (plots not shown). Concerning the GE case (Figure 9), the profiles of the CP ensemble systems are very similar to each other; however, WRF-ENS data provide better results for GE+72h and GE+48h (i.e., a higher ROC area of approximately 0.78 on average for both forecasts). The CP ensemble (ENS) curves approach the 0.5 horizontal line when evaluating precipitation thresholds in the interval 130-140 mm (90-100 mm). A summary of the ROC area analysis is reported in Table 6 in which the values are averaged among all the forecast ranges.

To assess the capability of the ensembles to predict rainfall peaks close to the actually observed peaks, we extracted the maximum QPF value predicted by each member of each

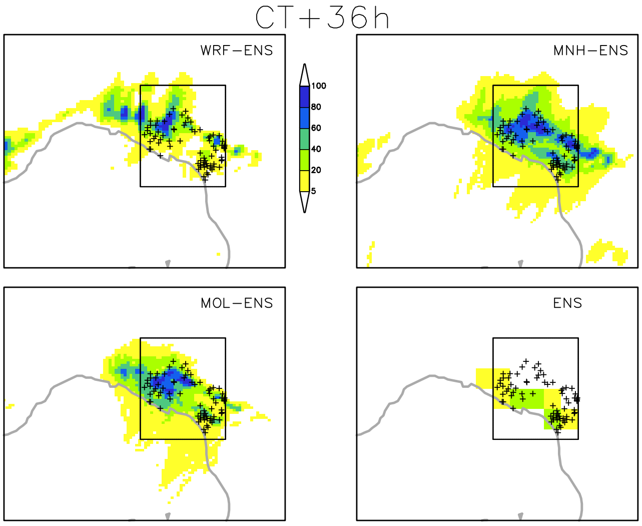


FIG. 6. CT case: probability of precipitation (PoP) in excess of 110 mm (corresponding to approximately the median of the observed rainfall) for the 24-hour period ending on the 26th of October 2011 at 00 UTC. The forecast length is 36 hours.

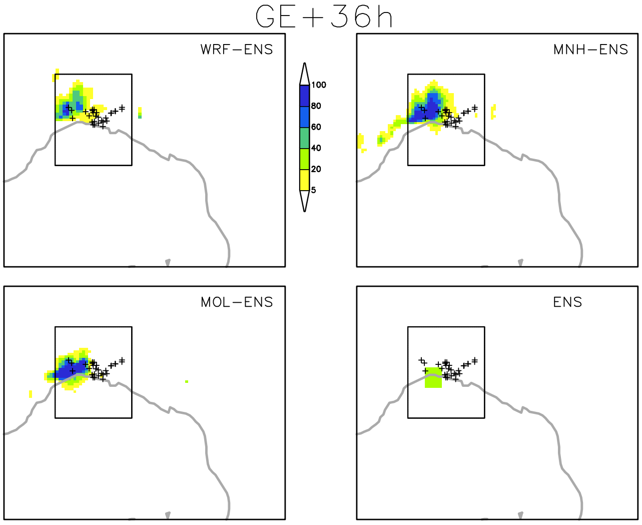


FIG. 7. As in Figure 6 but for the GE case. The precipitation thresholds corresponding to the median of the observed rainfall is approximately 130 mm.

TABLE 6. Areas under the ROC curve for the three convection-permitting forecasts and the ENS global predictions. For each event, the values are averages among all the forecast lengths (from +72 to +36 hours) and among all the precipitation thresholds shown in Figures 8 and 9. For each event, the maximum value is highlighted in bold.

	WRF-ENS	MNH-ENS	MOL-ENS	ENS
CT	0.742	**0.815**	0.744	0.580
GE	**0.739**	0.683	0.653	0.558

ensemble. In Figures 10 and 11, we show the distributions (in the form of boxplots) of the QPF maxima for CT and GE, respectively. The boxplots demonstrated that ENS maxima are considerably lower than the CP maxima (approximately one-half for CT and one-third for GE). We also note that for GE, members of the CP ensembles provide QPF maxima close to or higher than the maximum observed values (indicated with the dashed horizontal line). For the CT case, none of the members provide QPF values close to the observed peak (≃ 538 mm).

To investigate the physical mechanisms underlying the CT and GE cases, in Figure 12, we show the 3-hour accumulated rainfall and the 10-meter wind speed and direction (averaged over the same time period) of a single member of MNH-ENS for CT+36h (panels on the left) and GE+36h (panels on the right). The black point indicates the location of the rain-gauge that registered the maximum rainfall rate. In both cases, a convergence line is visible over the Ligurian Sea and marks the initiation of convective rainfall (Buzzi et al. 2014). In the CT case, the precipitation band oscillates from the east (panel (a)) to the west (panels (b) and (c)), and the resulting rainfall pattern is widespread over the whole area of interest. In the GE case, the position of the convergence line is steady, and thus, the precipitation pattern is limited to a small portion of the Genoa area.

b. Model scaling

In light of the potential use for operational forecasting, in Figure 13 we show the scalability of the simulation speed, defined as the ratio of simulated time to elapsed wall-clock time, by varying (namely by doubling at each step) the number of cores used to realize a 36-hour long simulation (the CT+36h forecast). The values shown on the Y-axis are obtained by averaging the simulation speeds of five selected members, taken as representative of the speed of the whole ensemble system. The wall-clock time taken into account, considers only the period spent to compute the evolution of the state variables and not that spent for reading the initial conditions and postprocessing the model outputs. The MOLOCH model turns out to be the fastest, being on average approximately 2.3 times faster than the WRF model and approximately 5.3 times faster than the Meso-NH model. To visualize the improvement in speed performance, in Figure 14 we show the reduction, in percentage, of the wall-clock time when doubling the number of cores. The three models exhibit a fairly satis-

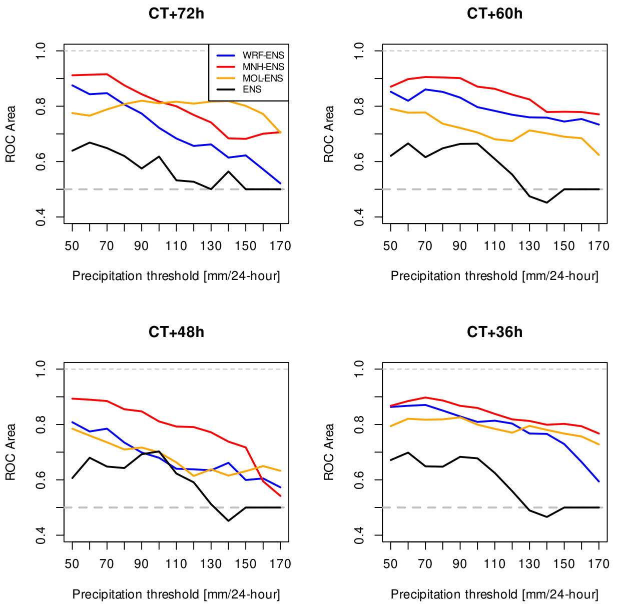


FIG. 8. CT case: area under the receiver operating characteristic (ROC) curve as a function of the precipitation threshold for WRF-ENS (blue), MNH-ENS (red), MOL-ENS (orange) and ENS (black) data. The forecast lengths are 72 hours (top-left panel), 60 hours (top-right panel), 48 hours (bottom-left panel) and 36 hours (bottom-right panel).

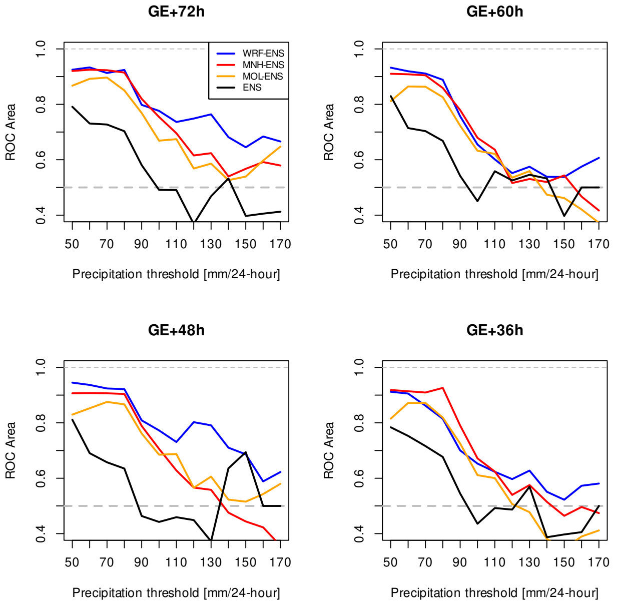


FIG. 9. As in Figure 8 but for the GE case.

factory reduction (i.e., $\simeq$ 50%) up to 288 cores, and then the gain degrades rapidly as the number of cores increases, dropping below 25% when using 1152 cores. Figure 14 also shows the number of horizontal grid points assigned to each core (say $N_{x,y}$) with the colored labels. In fact, domains are decomposed into horizontal patches, and each computing element (namely, core) is responsible for a single patch. The Meso-NH model is poorly sensitive to the patch size through $\simeq$ 200 $N_{x,y}$, whereas the MOLOCH model is strongly limited by the patch size when $N_{x,y}$ is less than $\simeq$1000. The WRF model has a good scaling up to $\simeq$ 600 $N_{x,y}$, and then the elapsed wall-clock time is further reduced by approximately 35% when $N_{x,y}$ is $\simeq$ 300.

c. Model performance vs scaling

In Figure 15, we show an analysis based on the $LISP$ index defined in equation 2. Panel (a) refers to the CT case and panel (b) to GE. As proxy data for the performance of the forecasts (P_M in equation 2), we selected the ROC area values shown in Figures 8 and 9. Panel (a) takes into account the average ROC values of the four panels in Figure 8; panel (b) as in panel (a) but averaging the data across the four panels in Figure 9. As a measure of the simulation speed (S_M in equation 2), we selected the simulation speed of the CP systems when running the CT+36h forecast with 288 cores (see Figure 13). Both S_M and P_M values were normalized to constrain them in the interval $[0,1] \subseteq \mathbb{R}$. In panel (a) of Figure 15, we evaluate, varying the precipitation thresholds on the X-axis, the more accurate forecasts (as regards the CT case) against the fastest ones, namely we show $P_{MNH-ENS}$ (red line) and $P_{MOL-ENS}+\kappa(S_{MOL-ENS}-S_{MNH-ENS})$ (orange line). Panel (b) as in panel (a) but looking at the GE case. We set $\kappa = 1/9$, that is we give more importance to the performance of the ensemble than to its speed (90% vs. 10%). Panel (a) shows that, for precipitation threshold higher (lower) than 120 mm, looking at the MOL-ENS (MNH-ENS) forecasts is more reliable considering both the performance and the speed of the ensemble. From panel (b) we can appreciate how looking at the slower WRF-ENS predictions represents a better trade-off between performance and speed than looking at the faster MOL-ENS predictions, for all the precipitation thresholds greater or equal to 120 mm.

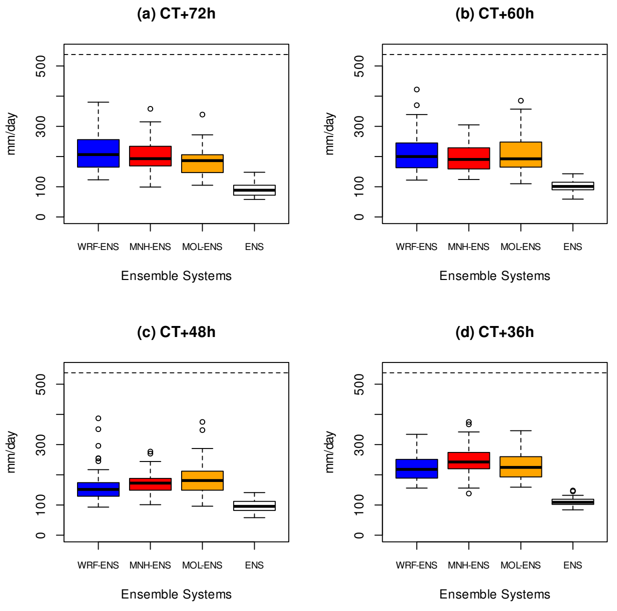


FIG. 10. CT case: boxplot of the QPF maxima provided by each EPS in the area of interest depicted in Figure 1. The lower and upper bounds of each box indicate the 25th and 75 percentiles, respectively, and the thick black line indicates the median. The upper (lower) whisker adds (subtracts) 1.5 times the interquartile difference to the 75th (25th) percentile. Points indicate outliers. The dashed horizontal line indicates the observed rainfall peak.

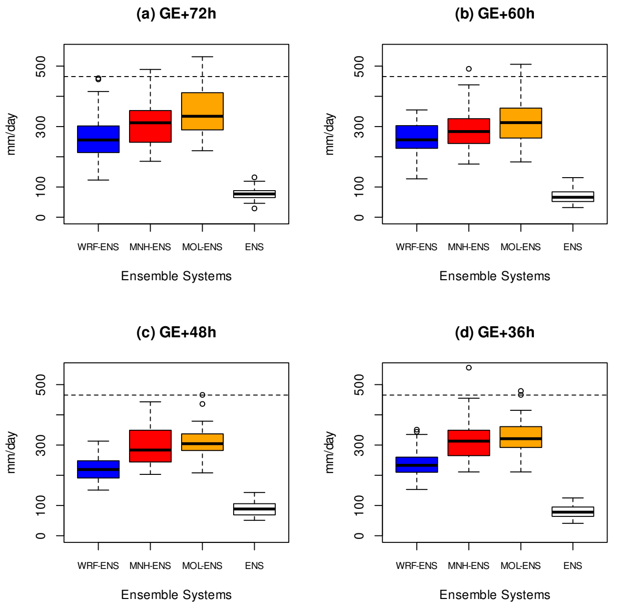


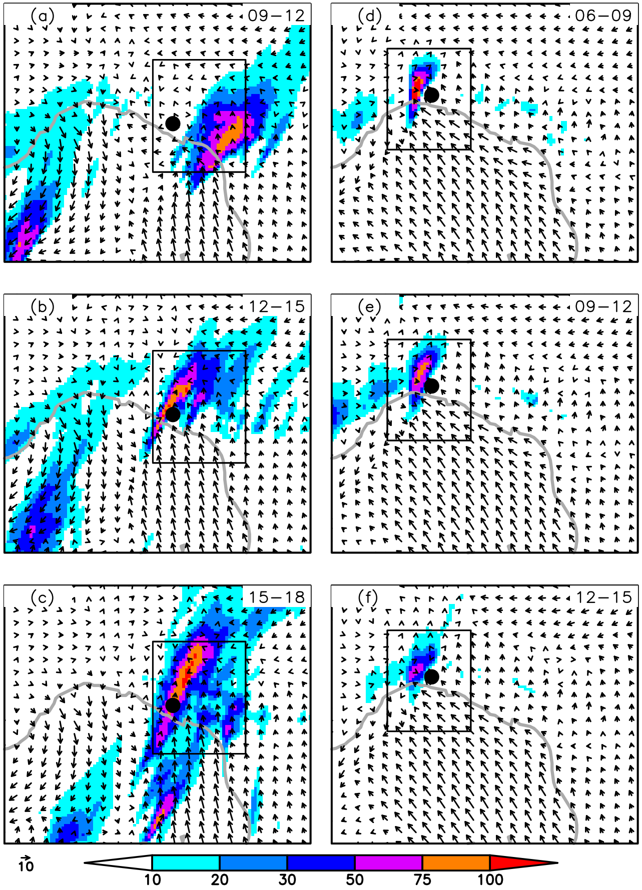
FIG. 11. As in Figure 10 but for the GE case.

6. Discussions and conclusions

Reforecasting past extreme weather events is an essential tool to understand the information content of current forecasting systems and monitor the progress achieved in weather modeling, both at the global and regional scale. By using recent versions of the ECMWF IFS model and three regional convection-permitting models, we showed the results obtained for the ensemble reforecasts of the CT and GE heavy-precipitation events that occurred in Italy in autumn 2011. Daily precipitation amounts registered at rain-gauges located in the two areas of interest provided the ground truth to assess the quality of such predictions.

Collective results suggest the potential benefits of running the high-resolution CP ensembles. In fact, using objective verification methods, we demonstrated that CP forecasts outperform ENS predictions for the CT case. This is true in both deterministic (see Figure 4 and Table 5), and probabilistic terms (see Figure 8 and Table 6). As regards the GE case, the results are more controversial. The ENS ensemble mean is more skillful than CP forecasts for the 80-mm threshold (see panel (a) in Figure 5). We speculate that this happens because the precipitation maxima were observed in a very small portion of the area of interest (namely the Bisagno catchment having an area approximately 100 km^2; see Hally et al. 2015). It is known (Gallus Jr 2002) that in these contexts coarse-resolution models may provide more skillful QPFs than higher-resolution models. For the precipitation threshold equal to 130 mm (see panel (b) in Figure 5), any ensemble mean forecast fails to provide useful information. However, looking at the ROC-area profiles for GE (shown in Figure 9), CP ensembles have a better probabilistic precipitation forecast skill than ENS. In fact, the ENS ROC-area profile drops below the critical threshold of 0.5 for precipitation amounts greater than 90-100 mm, whereas CP ensemble ROC-area profiles are greater than 0.5 for precipitation amounts up to 130-140 mm (which approximately is the median of observed rainfall).

FIG. 12. Output of a single member of the MNH-ENS system: precipitation accumulated every 3 hours (unit of mm) and wind speed and direction averaged over the same period for the CT (panels on the left) and GE (panels on the right) case. The time period (in UTC hours) over which the data are accumulated and averaged is indicated in the top-right corner of each panel. The forecast length is 36 hours. The black point in each panel indicates the location of the rain-gauge that registered the maximum rainfall rate.



As regards the comparison of CP ensemble predictions, the precipitation pattern of the CT case is, in general, best simulated by the MNH-ENS (see Figure 8 and Table 6). This could be because the time step used for the simulation is lower than that used for the other two CP models. Indeed, the shorter the time step, the more accurate the prediction (Coiffier 2011). As regards the GE case, the results produced by the CP ensembles are very similar to each other (see Figure 9), and for some forecast lengths (namely, GE+72h and GE+48h), the WRF-ENS outperforms the other two systems due to small-scale position errors that lead to a reduced number of false alarms. In fact, as Figure 12 demonstrates, incorrect positioning of the QPF maxima by a few kilometers induces a double-penalty error and impacts the forecast quality assessed by traditional verification statistics. This suggests a few considerations regarding the predictability of the CT and GE cases. Although they share similar synoptic and mesoscale features (see Section 2**??**), as stressed by Davolio et al. (2015) the CT case is characterized by a greater instability with a level of free convection close to the surface, whereas the GE case exhibits higher levels of convective inhibition, which is overcome by orographic uplift. As a consequence, the rainfall pattern of CT is widespread, whereas that of GE is relatively concentrated along the Ligurian coast and Apenine Mountains. As Figure 7 shows, ENS data provide only one grid-point with a PoP value higher than 20%; this leads to an overconfident prediction for the GE case that none of the CP ensembles are able to mitigate. One may argue that the use of the simple dynamical downscaling method is not suitable to initialize CP ensembles. In fact, the uncertainties in the small-scale features, that are not captured by the large-scale models can lead to the rapid growth of the errors such that the predictability is strongly limited (Hohenegger and Schär 2007). However, the contamination of the small-scale uncertainty on the whole integration domain depends on the synoptic situation and it is overwhelmed by the influence of lateral boundary conditions when strong synoptic forcings are met (as in the GE case, Rebora et al. 2013). Looking ahead in the near future, when higher-resolution global data assimilation tecniques will produce more accurate analyses, the uncertainty, even at the meso-α and meso-β scales, can be better addressed by sampling the members of the global ensemble. Our approach is also justified by recently published papers. For instance, Schwartz (2019) investigated the value of a CP ensemble directly nested into the NCEP's

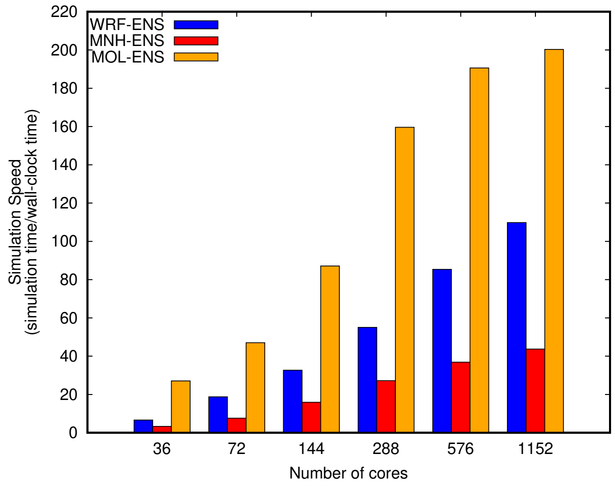


FIG. 13. Simulation speed, defined as the simulation length (36 hours in the present time) over the elapsed wall-clock time, as a function of the number of cores. For each model, the simulation speed is the average of five selected members of the CT+36h simulation.

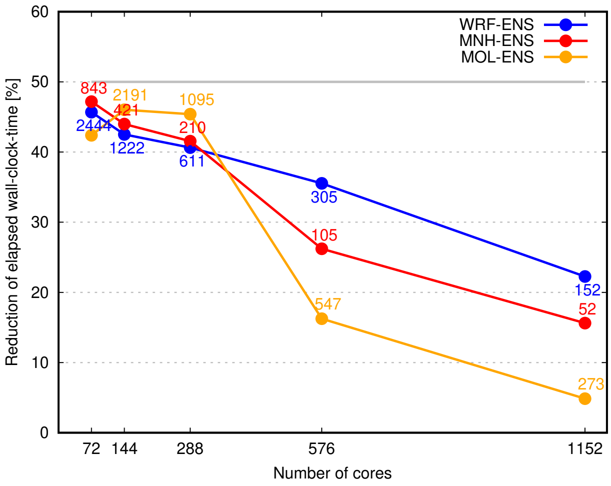


FIG. 14. Reduction (in percent) of the elapsed wall-clock time when doubling the number of computing cores. In a strong scaling regime, the colored lines should approximate the 50% horizontal line. The colored labels indicate the number of horizontal grid points assigned to each core.

operational global ensemble data and found that the 3-km ensemble outperforms coarser-resolution ensembles. The author tested his system in the “extended short-range”, that is, for lead times longer than 24 hours and shorter than 120 hours. Consequently, we deduce that our approach (dynamical downscaling of high-resolution global ensembles) is reasonable in this time range. On the other hand, for forecast lengths shorter than 36 or 24 hours, the design of any CP ensemble should adopt a strategy to perturb the initial conditions to account for small-scale uncertainties.

Figure 10 shows that all the members of the four EPS predict QPF maxima that largely underestimate the maximum rainfall amount observed during the CT case. This confirms the findings of similar studies (Buzzi et al. 2014; Davolio et al. 2015; Capecchi et al. 2015). However, we note that some members provide QPF maxima close to 400 mm (see, for instance, WRF-ENS and MOL-ENS in panel (b)), comparable to the outputs of deterministic forecasts run at a much higher horizontal resolution (≃ 1 km grid spacing, see Cassola et al. 2015). This confirms (Schwartz 2019) that for forecast lengths in the extended short-range (e.g., from +36 hours to +72 hours), it is worth running probabilistic predictions with a grid spacing of approximately 2.5-3 km to achieve results similar to higher-resolution forecasts in the short-range (forecast lengths less than 24 hours). Figure 11 demonstrates that some members of the CP ensembles (see for instance MNH-ENS and MOL-ENS in panel (d)) provide QPF maxima that approach or exceed the maximum observed rainfall value during the GE case. This is consistent with what is shown in Figure 12 (panels (d), (e) and (f)), which confirms (Buzzi et al. 2014) that correctly predicting the position of the convergence line over the Ligurian Sea is crucial for generating a rainfall band carrying a large amount of precipitation over the same area.

Figure 13 shows that the MOLOCH model is the fastest while Meso-NH is the slowest, but Figure 14 demonstrates that the number of grid points per core $N_{x,y}$ influences the scalability of the three CP models. This is not a new assessment; our result agrees well with previously published papers regarding the WRF model. Kruse et al. (2013) found that the WRF model scales approximately linearly through ≃ 650 grid points assigned to each core. Furthermore, the authors concluded that when $N_{x,y}$ is further reduced, the time required to perform the calculations on the perimeter of each patch overwhelms the computational time. For the Meso-NH and MOLOCH models, we found that the gain in elapsed wall-clock time is limited when using 576 cores or more; this occurs when the number of horizontal grid points per core $N_{x,y}$ is less than 105 (547) for the Meso-NH (MOLOCH) model. To the author’s knowledge, this an unprecedented assessment regarding these two models. However, we acknowledge that there is no abrupt shift from the strong scaling regime to the weaker regime. The above thresholds can be better defined by smoothly increasing the number of computing elements (instead of doubling it).

When investigating different model simulations for the same weather event, it is not straightforward to assess which model provides more reliable information on the basis of both the performance (shown in Figures 8 and 9) and the computational speed (shown in Figures 13 and 14). In general, the optimal trade-off between these two terms depends on the end-user requirements. If a large number of cores is available, then the model outputs that, on average, provide the more accurate data should be further analyzed.

On the other hand, if the rapid availability of forecasts is crucial for taking the most appropriate action to save lives, properties or for feeding models downstream, then the outputs of the faster model should be examined (provided that its performance is sufficiently fair based on some minimum requirements, i.e., ROC area > 0.5). In Figure 15 we showed the results on the $LISP$ index obtained by setting $\kappa = 1/9$, which means that the weights of the performance and speed in equation 2 are 0.9 and 0.1, respectively. As regards the CT case (panel (a)), data demonstrate that looking at the MOL-ENS predictions should be preferred for all the precipitation thresholds more than 120 mm. On the other hand, for the precipitation thresholds in the interval [50-120] mm, MNH-ENS provide the more reliable information. As regards the GE case (panel (b)), data show that for precipitation thresholds greater than 120 mm the WRF-ENS ensemble are the more accurate, owing to the better localization of QPF maxima, and they are also the more reliable taking into account the time to realize the predictions. However, we stress the fact that the analysis based on the novel $LISP$ index is strongly influenced by the simulation speed of the MOL-ENS system, which is much higher than that of the other two systems. In fact, the simulation speed of MOL-ENS is, on average, approximately 2.3 and 5.3 times faster than the WRF-ENS and MNH-ENS speeds, respectively. On the other hand, the performances of the three CP ensembles are close to each other, since they amplify the forecasting capability of the global predictions. If we set $\kappa = 1/4$, that is the weights of the performance and speed are 0.8 and 0.2 respectively, the MOL-ENS ensemble turns out the more reliable in both cases and for all the precipitation thresholds (maps not shown).

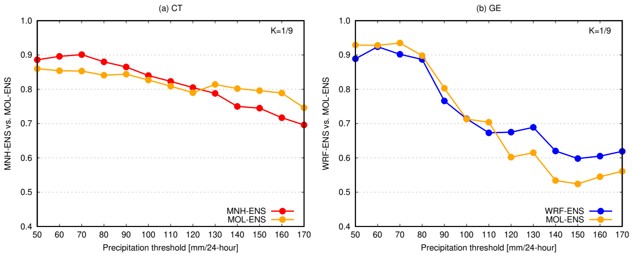


FIG. 15. Joint evaluation of the performance and simulation speed of the CP ensembles based on the $LISP$ index (the higher the better) defined in equation 2. Panel (a): CT case, MNH-ENS (red line) evaluated against MOL-ENS (orange line). Panel (b): GE case, WRF-ENS (blue line) evaluated against MOL-ENS (orange line).

We note that the experimental setup is not the same across the three CP models (see the settings reported in Table 1), which may impact both the model performance and the simulation speed. These experimental setups represent the trade-off between the limited computational resources available and the settings on the horizontal resolution and the extent of the integration domain. To draw meaningful assessments, the horizontal resolution has to be comparable with that of the state-of-the-art regional CP ensembles (see the references cited in Section 1) and high enough to partially resolve convective processes. The integration domain has to cover all of Italy to estimate the computational effort needed to deploy a CP ensemble system at the national level. Furthermore, we have to take into account the constraint on the time step, which has to satisfy the numerical stability criterion, and this constraint is not the same across the three CP models. These considerations led to the choices summarized in Table 1. We claim that the difference between the WRF grid spacing (3 km) and the Meso-NH and MOLOCH grid spacing (2.5 km) does not remarkably impact the model performance. In fact, as outlined in Buzzi et al. (2014), relevant improvements in the MOLOCH model accuracy are achieved only when the grid spacing is increased to 1.5 km. On the other hand, we underline how the use of a larger domain would be beneficial for the predictions based on the Meso-NH model (Davolio et al. 2020). Because of the small time step needed to guarantee numerical stability, the Meso-NH domain is the smallest one and we speculate that the results presented here most likely underestimate the potential accuracy of the MNH-ENS forecasts. Future developments will evaluate the performance of the three CP models in simulating the heavy precipitating event that affected Genoa's town center on the 9th of October 2014. For this event, Fiori et al. (2017) concluded that meso-γ processes played a crucial role in triggering the convection over the sea in front of the city. Small scale uncertainties grew into upscale uncertainties and contaminated the whole domain (Hohenegger and Schär 2007). For this case, the simple dynamical downscaling appears inappropriate and different strategies should be adopted to start the CP ensembles.

Data availability statement. The ENS data used in this study are accessible through the Meteorological Archival and Retrieval System (MARS) of ECMWF with experiment IDs: gmm2 (CT case), b09s and b09y (GE case).

WRF-ENS, MNH-ENS and MOL-ENS data are available on Zenodo: http://doi.org/10.5281/zenodo.3895551 (CT case), http://doi.org/10.5281/zenodo.3895561 (GE case).

Acknowledgments. Acknowledgment is made for the use of ECMWF's computing and archive facilities within the framework of the SPITCAPE Special Projects during the years 2016-2018 and 2019-2021.

The author would like to express great appreciation to Simon Lang (ECMWF) for his assistance with the IFS model runs and to Roberto Buizza (Scuola Superiore Sant'Anna) for his valuable suggestions during the planning of this study.