IL NUOVO CIMENTO Vol. 26 C, N. 1 Gennaio-Febbraio 2003

Application of a statistical methodology for limited area model intercomparison using a bootstrap technique(*)

C. Accadia(1), M. Casaioli(1), S. Mariani(1), A. Lavagnini(1), A. Speranza(2)
A. De Venere(3), R. Inghilesi(3), R. Ferretti(4), T. Paolucci(5)
D. Cesari(6), P. Patruno(6) G. Boni(7), S. Bovo(8)
and R. Cremonini(8)

(1) *CNR, Istituto di Scienze dell'Atmosfera e del Clima, Area di Ricerca di Roma Tor Vergata Italy*
(2) *Dipartimento di Matematica e Fisica, Università di Camerino - Italy*
(3) *Dipartimento per i Servizi Tecnici Nazionali della Presidenza del Consiglio dei Ministri Roma, Italy*
(4) *Dipartimento di Fisica, Università dell'Aquila - Italy*
(5) *Parco Scientifico e Tecnologico d'Abruzzo - L'Aquila, Italy*
(6) *Servizio Meteorologico ARPA - Emilia Romagna, Bologna, Italy*
(7) *Centro di ricerca Interuniversitario in Monitoraggio Ambientale - Savona, Italy*
(8) *Regione Piemonte, Settore Meteoidrografico e Reti di Monitoraggio - Torino, Italy*

(ricevuto l'1 Febbraio 2002; revisionato il 2 Luglio 2002; approvato il 18 Luglio 2002)

Summary. — The forecast verification problem of precipitations is a complex task. Within the European project "INTERREG II C" a method designed to discriminate statistically significant differences between skill scores has been applied. This methodology uses a resampling technique (bootstrap) as hypothesis test. Three different operational Limited Area Models (LAMs) are evaluated over the Piedmont and Liguria Regions as a test of the statistical method.

PACS 92.60.Jq – Water in the atmosphere (humidity, clouds, evaporation, precipitation).
PACS 93.85.+q – Instrumentation and techniques for geophysical research.

1. – Introduction

Flood forecasting has been under focus in the recent past as a fundamental task for meteorological services, and particularly in the Mediterranean area. Numerical meteorological Limited Area Models (LAMs) are operating with this purpose in many forecasting centers in the area. In this context, verification of precipitation forecast is a key issue

(*) The authors of this paper have agreed to not receive the proofs for correction.

for the forecasting community since affects both research and development issues and operational forecast verification activities. Research on techniques to evaluate numerical model skill in predicting floods is one of the tasks of the EU program "INTERREG II C"—*Gestione del territorio e prevenzione dalle inondazioni* (land management and floods prevention).

A flood alarm procedure can require use of a selected alert threshold generally defined on a local basis. This can be done using categorical dichotomous forecasts [1]. A categorical dichotomous forecast (also called non-probabilistic dichotomous forecasts) is simply a *yes/no* statement, *i.e.* whether the precipitation forecast is below or above a defined threshold. The same kind of statement is also true for the observations. The combination of the occurrence possibilities of observation and forecast gives origin to a contingency table.

The elements of the contingency table are used to compute discrete measures for model evaluation. Use of non-probabilistic scores is widely acknowledged by the forecasting community. Used scores include the bias score, or BIA [1], the equitable threat score (ETS) [2], the Hanssen-Kuipers score (HK) [3], also known as the true skill statistics (TSS) [4] or Peirce skill score (PSS) [5,6], and the odds ratio skill score (ORSS) [6], also called Yule's Q [7]. The BIA is a measure of the relative "dryness" or "wetness" of the model forecast with respect to observed precipitation; the other indexes are skill scores, *i.e.* measures of the forecast accuracy, as seen better later.

For purpose of model comparison a measure of the uncertainty on the score should be requested. Performing a hypothesis test provides a confidence interval for the score difference between the two competing models. This is rarely done due to several problems. First of all, if the BIA of the two models differ sensibly the comparison of the relative skill scores can be ambiguous. In fact a "wet" BIA may result in a comparatively larger skill score than for the corresponding case with a "dry" BIA [8].

Moreover, commonly used hypothesis tests require conditions which are rarely satisfied for this kind of problems [9]. It is difficult to make assumptions about the probability distribution of the score differences, so there is no warranty that it is a known parametric distribution as required by commonly used hypothesis tests [10].

Caution is also required in applying hypothesis tests if the data sets are spatially or temporally correlated. Spatial correlations of forecast errors among single grid points may be significantly high. Hence single grid points cannot be treated individually.

Time correlation of forecasting errors can be non-negligible when performing a hypothesis test. All kinds of tests, including the bootstrap technique used in this study, require the assumption of temporal statistical independence of the sample. This assumption may be relaxed if the hypothesis test can take into account the time correlation of the sample. Ordinary hypothesis tests (*e.g.*, Wilcoxon signed-rank test, paired t test, etc.) are commonly applied to time series of scores, each one calculated from a daily contingency table. A drawback associated of the use of daily contingency tables is their sensitivity to small changes in the population of the elements of the contingency tables.

Since we are mostly interested in verification of extreme events forecasts, a special concern should be paid to the problem of instability of the histogram estimator (*i.e.* any non-probabilistic skill score) with respect to the tails of the distribution.

The hypothesis test design proposed by Hamill [9], as better seen later, overcomes the problem of space correlation of forecast errors, and reduces the sensitivity of the contingency tables to the threshold selection. More in detail, the method is a particular application of a resampling technique called bootstrap [1,11]. This is a computer-based non-parametric test. The basic idea is to build an artificial data set from a given sample

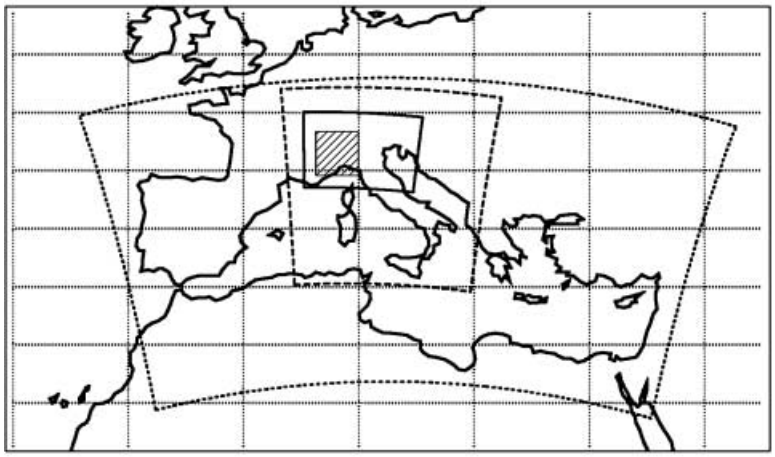

Fig. 1. – Verification area (shaded box) and domains of the selected models in cylindrical projection. Solid line: MM5 domain; dashed line: LAMBO domain; dotted line: QBOLAM domain.

of real data, by resampling the available collected data, in a way consistent with the null hypothesis.

In this work we investigate the application of this numerically based statistical methodology for intercomparison of LAMs operational in Italy, with respect to their ability to predict extreme precipitation over the Regions of Piedmont and Liguria, which were subject to dramatic flood events in the recent years. This study has been performed in the "INTERREG II C" framework, involving a factive collaboration among some Italian regional meteorological services, research and technical institutions.

These are: the Centro di ricerca Interuniversitario in Monitoraggio Ambientale (CIMA) of Savona (Liguria Region), the Settore Meteoidrografico e Reti di Monitoraggio (SMRM) of Turin (Piedmont Region), the Servizio Meteorologico ARPA of Emilia Romagna Region (SMR-ARPA) of Bologna, the Parco Scientifico e Tecnologico d'Abruzzo (PSTA) of L'Aquila, the Physics Department of L'Aquila University, the Atmospheric Physics Institute of Italian National Research Council (IFA-CNR) of Rome and the Dipartimento per i Servizi Tecnici Nazionali della Presidenza del Consiglio dei Ministri (DSTN-PCM) of Rome.

The aim of the work is an exploration of the methodological issues in the perspective of the development of a common procedure to be used by operational meteorological services in Italy.

The paper is organized as follows. In sect. **2** we describe the LAMs participating in the intercomparison and the observational and forecast data sets used in the study. Section **3** presents the verification measures suitable to evaluate the forecast quality. Section **4** examines in detail the bootstrap methodology used. In sect. **5**, the results of the application of the method to the data set are discussed. Section **6** summarizes the findings of the work, including recommendations for future works.

2. – Models and data sets

2˙1. *The selected models.* – Numerical weather prediction is performed by several Italian forecasting centers on regional or national basis, mostly by means of daily operating LAMs. Among these, three models were selected to provide forecast data for our work [12]. These are:

– the NCAR-Penn-State Fifth Generation Mesoscale Model (MM5) [13] operating in the Abruzzo Region by the PSTA;

– the ETA MODEL [14] operating at the SMR-ARPA (LAMBO) [15];

– the BOLAM [16] operating at DSTN-PCM (QBOLAM) [17] as a component of the POSEIDON wave-height and tide forecasting system.

All selected models are finite-difference LAMs, with a rotated horizontal grid and a *sigma* vertical coordinate. These models use operationally analysis and boundary conditions provided by the European Centre for Medium-range Weather Forecast (ECMWF). The domain extension of the different models in this study is depicted in fig. 1.

The MM5 model is running in a non-hydrostatic configuration. Model horizontal domain is organized with an Arakawa "B" staggered grid [18]. The MRF parameterization scheme [19] is used to represent turbulent fluxes, a "cloud-radiation scheme" [20, 21] calculates short- and long-wave radiation interactions with clouds, and the Kain-Fritsch scheme [22] is used for cumulus convection. The model is initialized daily at 12 UTC with the ECMWF analysis. After a 12 h spin-up time, a 24 h forecast is performed for the following day.

The model is run operationally over three nested domains, with grid step of 27 km (over the Central Mediterranean), 9 km (over the Central Italy) and 3 km (over the Abruzzo Region), respectively. Since these domains do not include the verification area (Piedmont and Liguria Regions, see below), we considered a different configuration, previously used by some of the authors in studies within the Mesoscale Alpine Program (MAP). It includes two nested (27 km and 9 km) domains; the higher-resolution one (shown in fig. 1) is centered on Northern Italy.

LAMBO is a hydrostatic, primitive-equations model derived from the University of Belgrade, National Meteorological Center-Washington model (UB/NMC, or ETA MODEL). The variables are staggered according to the Arakawa "E" grid [18]. Radiation parameterization is provided by the Geleyn scheme [23]; boundary layer fluxes are computed with a second-order closure scheme; convective processes are parameterized by Betts and Miller's relaxation scheme [24, 25].

The operating configuration includes two nested domains, with a horizontal grid step of 20 km and 10 km, respectively. These grid steps are calculated taking into account the Arakawa "E" staggering. The higher-resolution domain (shown in fig. 1) has an extension of 1400×1510 km, with 32 vertical levels, centered over Italy. Forecast timing is the same as MM5.

The QBOLAM model is a version of the hydrostatic, primitive-equation model BOLAM running on a massively parallel computer (QUADRICS). It is part of the POSEIDON system, used to calculate surface winds over the Mediterranean Sea as an input for the wave-height model WAM. For this reason also the highest-resolution domain has a quite large extension (fig. 1). The grid staggering follows the Arakawa "C" scheme. For computational reasons, due to the domain extension and the parallel structure of the code, simplified parameterization schemes are adopted for radiation [26] and cu-

TABLE I. – *Inner integration domain characteristics for the selected limited area models.*

Model	# longitude points	# latitude points	Grid step (km)
MM5	85	64	$\sim$ 9
LAMBO	201	217	$\sim$ 10
QBOLAM	386	210	$\sim$ 11

mulus convection [27]. Turbulent fluxes parameterization is provided by a similarity scheme [28].

The operating configuration includes two nested domains, with a horizontal step of 0.3 and 0.1 degrees, respectively (over the rotated grid). At the lower resolution, a 60 h forecast is performed, starting at 12 UTC. The higher-resolution 48 h run starts at 00 UTC after 12 h spin-up time.

2˙2. *Forecast and observation data.* – The verification data set include rain gauge observed precipitations over the Italian regions of Piedmont and Liguria (represented as a shaded box in fig. 1) and model forecast precipitation fields, for a 8 month time interval (from 1/10/2000 to 31/5/2001). This interval is suitable for a statistical study of precipitation as it corresponds to the Mediterranean wet season [9].

The model data are gridded forecast fields of total precipitation over the integration domains shown in fig. 1. The characteristics of the three domains are summarized in table I. While the domain extension is quite variable, the horizontal resolution is homogenous.

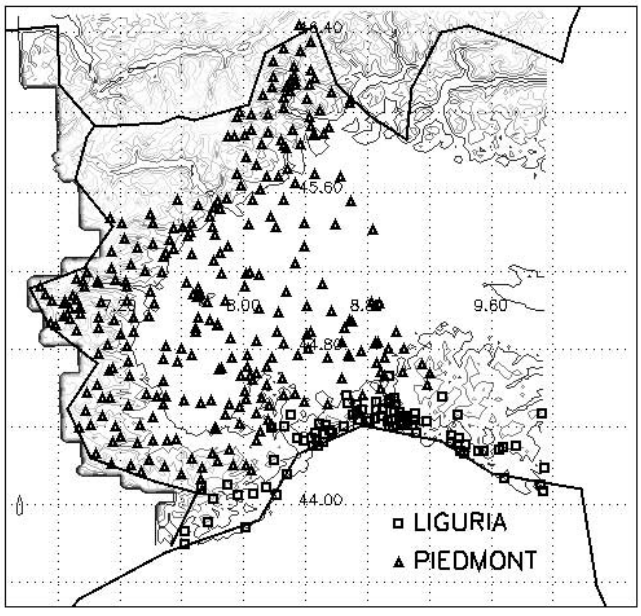

Fig. 2. – Distribution of the rain gauge stations over the Piedmont and Liguria Regions.

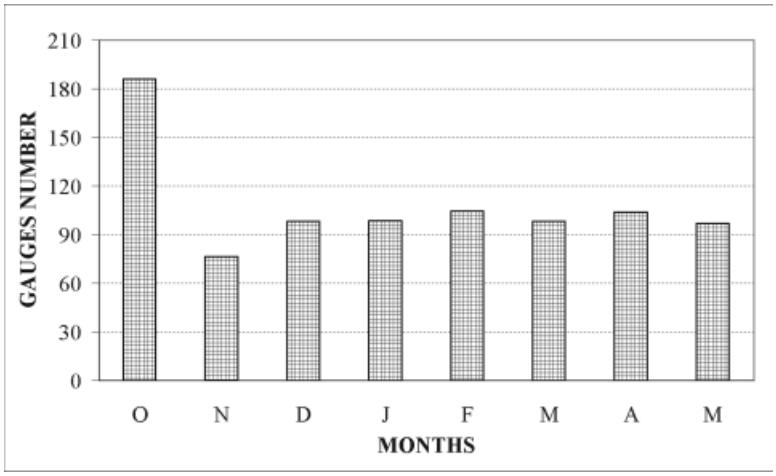

Fig. 3. – Average number of the accepted gauges for the time interval October 2000-May 2001.

LAMBO and QBOLAM operational simulations, archived at SMR-ARPA and DSTN-PCM, were used. As seen before, operational 9 km MM5 domain is not suitable for our study. Thus, daily simulations were performed newly over the MAP domain, with no other change with respect to the operational model configuration.

Daily precipitation forecast data are joined together to obtain for each model a continuous time series. Daily data are considered starting from 00 UTC to 24 UTC of the day after the initialization. These time series are cumulated up to 24 hours for the purposes of this work. The same is done for observation series.

The observed data are provided by the recently built automatic rain gauge networks of Liguria and Piedmont Regions. Data from 96 stations pertaining to the Rete Meteo-Idrologica Ligure (OMIRL) and from 294 stations of the Piedmont network were acquired. About 10% of the latter are snow sensors located at height above 1500 m; although these measurements might have large errors, a score sensitivity study has shown that these seem to not affect significantly the actual score values (not shown).

The distribution of the stations over the selected area is shown in fig. 2.

Not all the mentioned stations were active during the entire time interval considered. A quality check was performed, discarding as outliers the rainfall observations exceeding 25 mm in 5 minutes. The discarded values are considered missing values. After that, only stations with more than 10% of missing values (with respect to the cumulation interval) were discarded. This was done to preserve both an adequate number of included stations and a low time sampling error.

The number of accepted stations is variable and is depicted in fig. 3 as a function of time. Selected stations were associated with the respective nearest grid point in each

TABLE II. – *Precipitation thresholds for* 24 *h cumulation time.*

Cumulation time (h)	Thresholds (mm)				
24	2.4	24.0	40.0	60.0	90.0

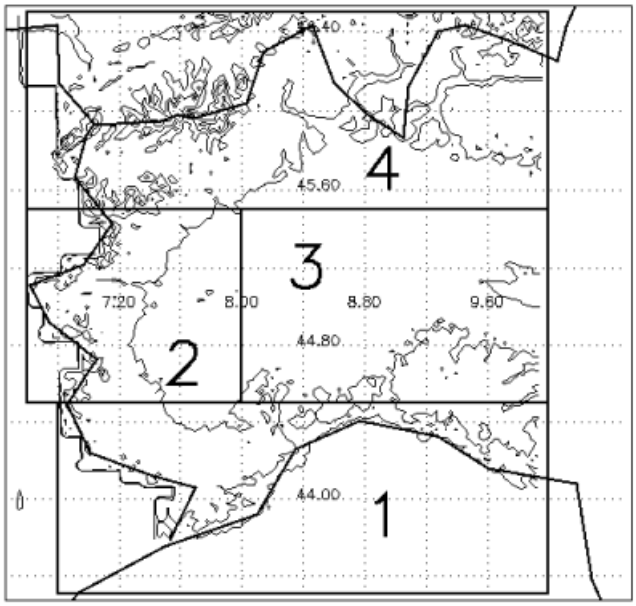

Fig. 4. – Subdivision of Piedmont and Liguria Regions into four subregions to study spatial correlations of forecast errors.

model's grid. Only grid points associated to one or more gauges were considered. For the points associated to more than one gauge, the mean precipitation value over such gauges was considered for the comparison.

The thresholds used in this study for the 24 h cumulation time are presented in table II.

Such kind of work may be considered incomplete without considering the possible time-space correlations of forecast errors.

First of all, the domain of study has been divided into four subregions as shown in fig. 4, in order to assess possible spatial correlation of forecast errors. The subdivision has been performed taking care of having roughly a comparable number of rain gauges in each subarea.

A Spearman rank correlation is performed between all possible couples of ETS score, each calculated over a subregion for the 24 h cumulation time. For each correlation coefficient, a two-side significance of its deviation from zero is computed. This value, called p value, ranges between 0 and 1. A rank correlation associated with a p value close to zero is significant. On the opposite, a p value close to one means that rank correlation is not meaningful.

Table III shows the results for 24 h cumulation time. A dash sign indicates that was impossible to determine a particular correlation for the above-mentioned reason. Non-negligible correlations (negative or positive) exist in the considered sample for all model forecasts. Hence, all grid points on a given day must be treated as a single sample.

The lag 1 Spearman correlations of ETS and BIA score series are calculated for the selected models, in order to determine the time correlation of forecast errors. The Spearman correlation is calculated using data from 1st to 30th November, 2000, since climatologically November is the wettest month in the Mediterranean area. The results,

TABLE III. – *Spearman rank correlation of ETS score for* 24 *h cumulation time between two subregions of Piedmont and Liguria. A p value is assigned for each rank correlation.*

Threshold (mm)	Subregions association	MM5		LAMBO		QBOLAM	
		Rank correlation	p value	Rank correlation	p value	Rank correlation	p value
2.4	1 ⟶ 2	0.212	0.009	0.160	0.037	0.148	0.049
	1 ⟶ 3	0.147	0.091	0.157	0.076	0.174	0.029
	1 ⟶ 4	0.412	0.000	0.368	0.000	0.157	0.093
	2 ⟶ 3	0.247	0.002	0.118	0.179	0.259	0.000
	2 ⟶ 4	0.260	0.004	0.108	0.203	0.140	0.133
	3 ⟶ 4	0.165	0.083	0.264	0.005	0.119	0.220
24.0	1 ⟶ 2	0.296	0.093	−0.309	0.041	0.456	0.000
	1 ⟶ 3	0.286	0.086	0.180	0.273	0.415	0.000
	1 ⟶ 4	0.125	0.447	0.017	0.897	0.244	0.124
	2 ⟶ 3	0.181	0.329	0.342	0.043	0.300	0.014
	2 ⟶ 4	0.426	0.069	0.375	0.037	0.455	0.011
	3 ⟶ 4	0.434	0.038	−0.202	0.244	0.070	0.692
40.0	1 ⟶ 2	−0.380	0.313	0.061	0.804	−0.041	0.845
	1 ⟶ 3	0.200	0.512	−0.179	0.450	0.358	0.048
	1 ⟶ 4	0.069	0.766	0.286	0.107	0.068	0.782
	2 ⟶ 3	−0.712	0.047	0.220	0.352	0.061	0.719
	2 ⟶ 4	−0.512	0.130	−0.096	0.704	−0.048	0.865
	3 ⟶ 4	0.517	0.085	−0.100	0.702	0.192	0.431
60.0	1 ⟶ 2	−0.816	0.183	0.161	0.637	−0.703	0.016
	1 ⟶ 3	0.032	0.933	−0.225	0.481	0.280	0.312
	1 ⟶ 4	0.437	0.178	0.559	0.016	0.210	0.536
	2 ⟶ 3	−0.688	0.198	0.180	0.575	0.359	0.111
	2 ⟶ 4	−0.866	0.333	−0.525	0.147	−0.580	0.079
	3 ⟶ 4	−0.108	0.781	−0.460	0.154	0.000	1.000
90.0	1 ⟶ 2	–	–	−0.258	0.742	−0.516	0.236
	1 ⟶ 3	–	–	0.229	0.710	−0.500	0.391
	1 ⟶ 4	1.000	0.000	−0.076	0.871	–	–
	2 ⟶ 3	–	–	0.459	0.437	0.456	0.159
	2 ⟶ 4	–	–	–	–	–	–
	3 ⟶ 4	–	–	−0.889	0.111	–	–

presented in table IV, show little time correlations for all the considered models. HK and ORSS scores also have lag 1 Spearman rank correlation significantly close to zero (not shown).

3. – Verification measures

Verification of precipitation forecasts can be treated in many different ways. For determining the quality of non-probabilistic dichotomous forecasts, verification measures are computed using a contingency table (table V). A daily contingency table is generated

TABLE IV. – *Lag 1 Spearman rank correlation of ETS and BIA scores for* 24 *h cumulation time.*

Score	Threshold (mm)	MM5		LAMBO		QBOLAM	
		Rank correlation	p value	Rank correlation	p value	Rank correlation	p value
ETS	2.4	−0.10305	0.59478	0.04060	0.83438	0.24852	0.19361
	24.0	0.14762	0.44476	0.16429	0.39444	−0.06685	0.73043
	40.0	−0.11498	0.55257	−0.37328	0.04610	0.07857	0.68538
	60.0	−0.07398	0.70291	−0.07398	0.70291	−0.03973	0.83788
	90.0	−0.03572	0.85407	−0.07398	0.70291	−0.11498	0.55257
BIA	2.4	0.19726	0.87829	0.02974	0.30505	0.16091	0.40436
	24.0	0.12063	0.53306	0.14048	0.46733	0.21740	0.25728
	40.0	−0.11498	0.55257	0.19646	0.30704	0.12698	0.51156
	60.0	−0.07398	0.70291	−0.07398	0.70291	−0.15890	0.41032
	90.0	−0.03572	0.85407	−0.07398	0.70291	−0.11498	0.55257

from comparison between forecast and observed precipitation data sets for each selected threshold. The table elements define the absolute frequencies of the four possible combinations, that are *hits*, *false alarms*, *misses* and *correct non-rain forecasts* (respectively, a, b, c and d in table V).

The BIA [1] is the ratio between the frequency of *yes* forecast and the frequency of *yes* observed:

$$\text{BIA} = \frac{a+b}{a+c}. \tag{1}$$

A bias score equal to 1 means that forecast is unbiased, *i.e.* forecasts and observations have a value above a given threshold the same number of times. A BIA greater than one means that the model overestimates the frequency of the precipitations above the selected threshold ("wet" model). On the other hand, a BIA lower than one shows that the model underestimates the frequency of events ("dry" model).

The ETS skill score used in this study is a modification of the critical success index (CSI) [1] that takes into account the random forecast (see a_{r} in (2)) [2]. Despite this correction, the model BIA influence is not completely removed. The a_{r} tends to zero

TABLE V. – *Contingency table for categorical forecast verification.*

		Observed	
		Yes	No
Forecast	Yes	a	b
	No	c	d

while increasing the threshold value [29]. This score is defined as

$$\text{ETS} = \frac{a - a_{\,\mathrm{r}}}{a + b + c - a_{\,\mathrm{r}}}\,, \qquad \text{where } a_{\,\mathrm{r}} = \frac{(a+b)(a+c)}{a+b+c+d}\,. \tag{2}$$

A perfect forecast has an ETS equal to 1. A score value equal or lower than 0 shows that model is unable to produce a significant forecast.

Another score for determining the forecast accuracy is the Hanssen-Kuipers score [3]. This verification measure has a range between -1 and 1. Its expression is

$$\text{HK} = \frac{ad - bc}{(a+c)(b+d)}\,. \tag{3}$$

The HK can be also written as a sum, normalized to one, of the probability of detection (POD) and the non-events probability of detection (NPOD) [3]:

$$\text{HK} = \text{POD} + \text{NPOD} - 1 = \frac{a}{a+c} + \frac{d}{b+d} - 1\,. \tag{4}$$

Unlike the ETS, this kind of skill score emphasizes in the same way forecast events and non-events. Wilks [1] explains that the HK score is appropriate for verifying rare events too. A perfect forecast receives a score equal to 1, a random forecast has skill zero, while its value is -1 when $a = d = 0$.

In this study the odds ratio skill score (ORSS) has been considered for the forecast verification too. This measure is constructed from the odds ratio (ODDS) [6], that it is useful to measure the degree of association between observed and forecast precipitations, *i.e.* it summarizes the association in the joint conditional distribution of a categorical forecast. An event odds is calculated as the ratio between the probability that the event occurs and the probability that the event does not occur. In other words, if p is the probability that the event occurs, then the odds of the event is equal to $p/(1-p)$. The probability $p_{\,\mathrm{h}}$ $(5a)$ is the conditional probability associated to the observed *yes* given a *yes* forecast. The probability $p_{\,\mathrm{m}}$ $(5b)$ is the same quantity, but for *no* forecast. These probabilities are

$$p_{\,\mathrm{h}} = \frac{a}{a+b}\,, \tag{5a}$$

$$p_{\,\mathrm{m}} = \frac{c}{c+d}\,. \tag{5b}$$

The odds ratio is then defined as

$$\text{ODDS} = \frac{p_{\,\mathrm{h}}}{1 - p_{\,\mathrm{h}}} \cdot \left(\frac{p_{\,\mathrm{m}}}{1 - p_{\,\mathrm{m}}}\right)^{-1} = \frac{a}{b} \cdot \left(\frac{c}{d}\right)^{-1} = \frac{ad}{bc}\,, \tag{6}$$

the ODDS score is unity when observations and forecasts are independent, while a score value larger than one means that forecasts and observations are associated variables. The ORSS is obtained by the following transformation:

$$\text{ORSS} = \frac{\text{ODDS} - 1}{\text{ODDS} + 1} = \frac{ad - bc}{ad + bc}\,, \tag{7}$$

that normalize the ODDS score between -1 (*i.e.* ODDS $= 0$) and 1 (*i.e.* ODDS $\longrightarrow \infty$). This score has been introduced by Yule [7] as a measure of association. It depends on the joint conditional probabilities and it is not influenced by the marginal probabilities, hence "it strongly discriminates between the case with or without association" [6].

All the mentioned skill scores may be influenced by the BIA value. This effect cannot be neglected when forecast comparison is performed [8, 9]. The main reason for taking into account this effect is that it is important to discriminate the actual signals when two models are compared.

Hamill [9] proposed a BIA adjustment method to compare competing forecasts produced by models with different bias scores. Contingency tables are calculated introducing a new forecast threshold, that can be different from the observation threshold. An independent change of the forecast threshold can take into account the model tendency to overestimate or underestimate precipitations. This BIA adjustment is then applied for all selected thresholds independently.

This technique however does not remove the ambiguity that may still exist in model intercomparison. The application of a hypothesis test may help to discriminate whether the computed scores are different in a statistically significant way.

A preliminary simple model comparison was performed using a wide set of non-parametric scores (not shown), including the aforementioned ones. This was done in order to check the choice of the indexes and also to have a general description of the model forecast behaviour over the different indexes.

Results are summarized as follows. The BIA score evidences a "wet" tendency for QBOLAM and a "dry" one for MM5 as the threshold increases; while LAMBO exhibits an intermediate behaviour. The POD and FAR (False-alarm rate [1]) values are quite dependent on the BIA trend. The ETS differences over the three models are relatively small, except for the higher thresholds, where the score of the drier model is sensibly lower. The HK score seems to be more sensitive to the BIA differences among the models. The ORSS seemingly displays best values for the higher thresholds and very small differences among the models. The ODDS results reproduce the ORSS ones, as evident from the definition (7).

Three degrees of freedom are needed to completely summarize 2×2 categorical forecasts [6]. However, the score selection is not unique (dimensionality problem [1]). The bootstrap method will be then applied on the triplet BIA, ETS and HK; while the ORSS is included for a general evaluation.

4. – The bootstrap technique

The evaluation of the real differences in skill score of two competing forecasts can be supported by the use of a hypothesis test. As already mentioned, a bootstrap technique developed by Hamill [9] is applied in this study.

The existence of non-negligible space correlations among the elements of the sample set and forecast errors has been shown. Hence, the scores have been calculated without considering any geographical partition of the data. Data cumulated up to 24 hours are considered non-correlated in time, although significant time correlations are likely if data are cumulated at shorter time intervals. A particular application of the bootstrap technique (moving-block bootstrap), to deal with correlated series, will be the object of future studies.

The resampling technique needs the definition of a null hypothesis to be tested. In

this case the test is applied to check whether the following statement is true or not:

$$H_0 : \mathrm{S}_{\mathrm{M}_1} - \mathrm{S}_{\mathrm{M}_2} = 0\,, \tag{8}$$

where $\mathrm{S}_{\mathrm{M}_1}$ and $\mathrm{S}_{\mathrm{M}_2}$ are a general score S (*e.g.*, BIA, or ETS, etc.) calculated on a sum of contingency tables for two competing models M_1 and M_2. The alternative hypothesis H_{A} means that the equality statement (8) is not true.

For each model the elements of a daily contingency table can be expressed as a four-elements vector:

$$\mathbf{x}_{i,j} = (a\ b\ c\ d)_{i,j}\,, \qquad i = 1,\ 2 \ \text{ and } \ j = 1,\ \cdots,\ n\,, \tag{9}$$

where i is the model indicator, while j is the contingency table index and n is the number of case days. Thus, the statistic test is performed computing the difference

$$\left(\widehat{\mathrm{S}}_{\mathrm{M}_1} - \widehat{\mathrm{S}}_{\mathrm{M}_2}\right) \tag{10}$$

on the sum of the contingency tables

$$(a\ \widehat{b\ c}\ d)_{\mathrm{M}_1} = \sum_{j=1}^{n} \mathbf{x}_{1,j}\,, \tag{11a}$$

$$(a\ \widehat{b\ c}\ d)_{\mathrm{M}_2} = \sum_{j=1}^{n} \mathbf{x}_{2,j}\,. \tag{11b}$$

The resampling method to build a statistics consistent with the null hypothesis (8) is applied as follows. Let be I_j a random indicator that can be equal to 1 or 2, equivalent to randomly choose model M_1 or M_2, respectively, with $j = 1,\ \cdots,\ n$. Using the random index I_j, the resampled test statistic is obtained summing the shuffled contingency table vectors over the n available elements:

$$(a\ \widehat{b\ c}\ d)^{*}_{\mathrm{M}_1} = \sum_{j=1}^{n} \mathbf{x}_{I_j,j}\,, \tag{12a}$$

$$(a\ \widehat{b\ c}\ d)^{*}_{\mathrm{M}_2} = \sum_{j=1}^{n} \mathbf{x}_{(3-I_j),j}\,. \tag{12b}$$

Then it is possible to generate N_{B} (in our case $N_{\mathrm{B}} = 10000$) sample sums, defined by eqs. (12a) and (12b), by selecting each time a new set of random indexes I_j. The new N_{B} contingency table sums are used to compute the resampled statistics of the score difference

$$\left(\widehat{\mathrm{S}}^{*}_{\mathrm{M}_1} - \widehat{\mathrm{S}}^{*}_{\mathrm{M}_2}\right)\,, \tag{13}$$

that defines the sampling distribution consistent with the null hypothesis (8).

A confidence level $\alpha = 0.05$ is assumed. A two-tailed test, the percentile method [1], is applied to determine the $(1 - \alpha)\%$ confidence interval, by finding the values of the

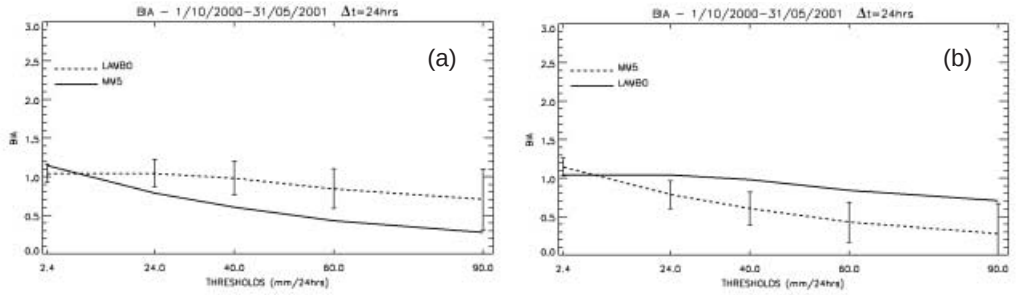

Fig. 5. – Transpose symmetry of the bootstrap technique applied to the 24 h BIA score intercomparison. (a) LAMBO as "reference" model (dashed line) and MM5 as "competitor" model (solid line). (b) MM5 as "reference" model (dashed line) and LAMBO as "competitor" model (solid line).

score differences defining the largest and smallest $N_{\mathrm{B}}\alpha/2$ of the N_{B} bootstrap samples. Thus, two values $\widehat{t}_{\mathrm{L}}$ and $\widehat{t}_{\mathrm{U}}$ are determined such that

$$\P = \left[\left(\widehat{\mathrm{S}}^{*}_{\mathrm{M}_1} - \widehat{\mathrm{S}}^{*}_{\mathrm{M}_2}\right) < \widehat{t}_{\mathrm{L}}\right] = \frac{\alpha}{2}\,, \tag{14a}$$

$$\P = \left[\left(\widehat{\mathrm{S}}^{*}_{\mathrm{M}_1} - \widehat{\mathrm{S}}^{*}_{\mathrm{M}_2}\right) < \widehat{t}_{\mathrm{U}}\right] = 1 - \frac{\alpha}{2}\,, \tag{14b}$$

then the null hypothesis H_0 is refused if the observed statistic (10) is outside the interval $(\widehat{t}_{\mathrm{L}},\, \widehat{t}_{\mathrm{U}})$.

5. – Results and discussion

As previously discussed in the preliminary study, the score differences among the models generally increase for higher thresholds. This observation should be assessed with some caution, for the comparatively smaller sample size at higher thresholds. Moreover, a skill scores comparison may be affected by the BIA difference among the models.

More in general, these results may be not suitable for a clear statement in model intercomparison, since no kind of statistical measure of uncertainty is provided. Hence, hypothesis test has been performed using the bootstrap methodology. This procedure can be applied only between two models, so it is necessary to define a "reference" model to evaluate the others, that are indicated as the "competitors".

LAMBO has been chosen as the "reference" model since it has an intermediate BIA trend among the three models. This choice does not compromise the analysis results. In fact, the bootstrap is a transpose symmetric technique and it is invariant swapping the "reference" and "competitor" role as shown in fig. 5.

The bootstrap methodology has been also used to evaluate the scores after the application, over the "competitor" model, of the BIA adjustment technique (see sect. **3**).

Figures 6 and 7 show the results of the comparison procedure for 24 h cumulation time. In the left columns, the scores are compared without the application of the BIA adjustment technique; while in the right columns, the same scores are presented after the BIA difference reduction obtained by the BIA adjustment technique. The dashed line indicates the "reference" model, while the solid line indicates the "competitor" model.

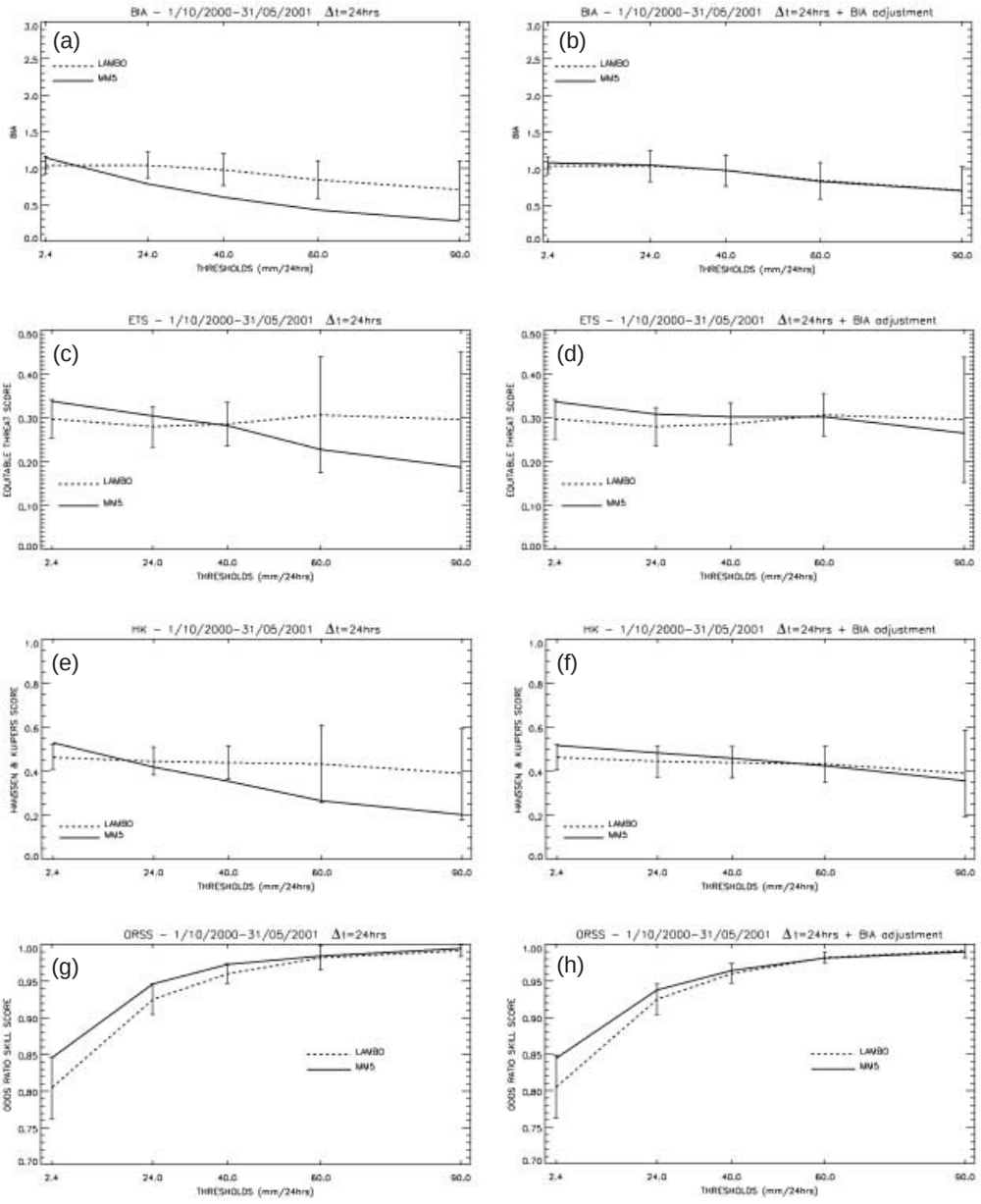

Fig. 6. – Bootstrap results between the "reference" LAMBO (dashed line) and the "competitor" MM5 (solid line) for 24 h cumulation time. In the left column, BIA (a), ETS (c), HK (e) and ORSS (g) are shown without the BIA adjustment. In the right column, BIA (b), ETS (d), HK (f) and ORSS (h) are shown with the BIA adjustment.

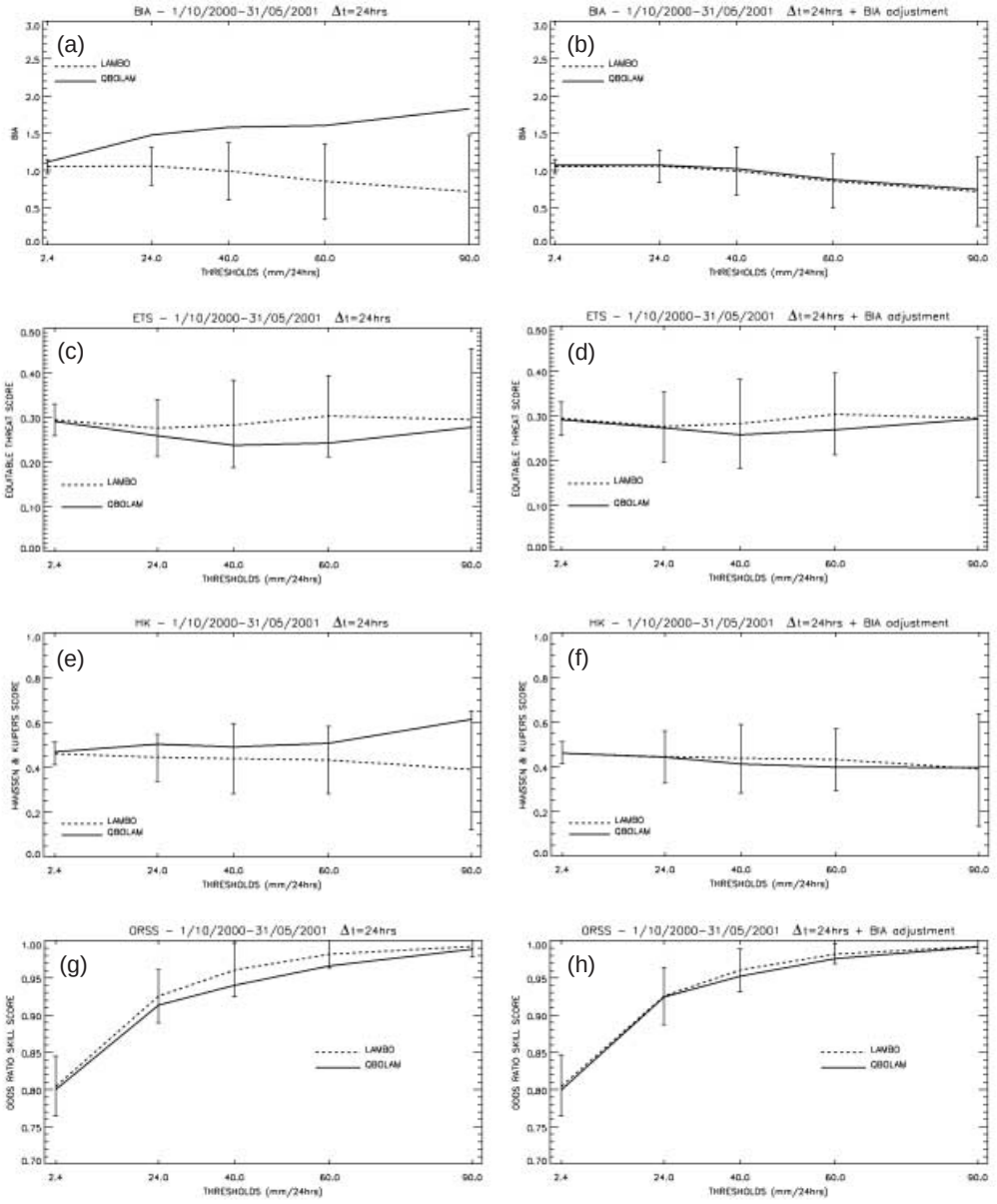

Fig. 7. – Bootstrap results between the "reference" LAMBO (dashed line) and the "competitor" QBOLAM (solid line) for 24 h cumulation time. In the left column, BIA (a), ETS (c), HK (e) and ORSS (g) are shown without the BIA adjustment. In the right column, BIA (b), ETS (d), HK (f) and ORSS (h) are shown with the BIA adjustment.

We first describe the results obtained without applying the BIA adjustment technique. Significant differences in the BIA score, between the competing models, are found in almost all cases. In other words, the hypothesis test shows that the aforementioned differences in the "wetness" of the models are reliable. This is the only remarkable difference among the models behaviour evidenced by applying the bootstrap technique.

For all the verification measures, with some exceptions, the confidence intervals tend to increase for higher thresholds (also true after adjusting the BIA).

The score differences between the competing models are inside the confidence interval, or very close to its bounds, for the ETS, HK and ORSS score. Differences close to the bounds are found especially for the HK and ORSS score (as seen before, these measures are the most sensitive to the BIA trend). A significant difference in skill scores is found only between LAMBO and MM5 at the lowest threshold. We note that, for this threshold, MM5 has a "wet" BIA, in opposite to its general "dry" trend.

As an effect of the BIA adjustment, the score differences between the competing models are mostly reduced. This reduction is quite strong in most cases, making most of the differences smaller than the confidence interval, though reduction of this one may also occur. The correction is mostly effective when the BIA differences are larger and over the skill scores which display a higher sensitivity to the BIA value (see, for example, fig. 7c *vs.* fig. 7d). In particular, all the differences previously close to the confidence interval bound (as in fig. 6e or in fig. 6g) loose any significance, except for the aforementioned difference between LAMBO and MM5 at the lowest threshold (best evidenced in fig. 6).

6. – Conclusions

The application of the bootstrap method for score comparison over the Piedmont and Liguria Regions shows a statistically significant BIA difference among the models, while generally the selected skill scores appear to be statistically equivalent. This result is true also after the BIA adjustment. The confidence interval width generally increases at greater thresholds. Hence, the comparison is not particularly informative even when the score differences are large. In fact the selection of high thresholds reduces the numbers of successes, and the application of the bootstrap (with random swapping of daily tables of two models) tends to increase the score differences, widening the confidence intervals. The non-hydrostatic MM5 model shows a behavior that is statistically comparable to the other two hydrostatic models. This characteristic of the MM5 model does not play here a particularly important role, since MM5 results were used with a horizontal resolution of 9 km and with parameterized convection.

The BIA difference existing between QBOLAM and LAMBO seems not to affect strongly the simple ETS comparison (without considering confidence bars and without BIA adjustment), while HK, that is sensitive to correct non-rain forecasts, is higher for QBOLAM although not in a statistically different way.

The skill scores comparison between LAMBO and MM5 shows a more consistent BIA dependence of skill scores. The application of the hypothesis test performing the BIA adjustment confirms the statistical equivalence of the results. Moreover, the most evident differences that appear in a simple skill score comparison are due to the BIA differences.

The ORSS score increases proportionally to threshold values, becoming less informative at higher thresholds, because it is dominated by the correct non-rain forecasts.

Future works will study the possibility of application of the bootstrap method to shorter cumulation intervals (6 h, 12 h) time series. Some modifications of the methodology will be probably required, to deal with possibly larger forecast errors autocorrelation values of such time series.

* * *

We would like to thank Dr. T. HAMILL for his very useful suggestions and comments.

REFERENCES

[1] WILKS D. S., *Statistical methods in the atmospheric sciences*, in *International Geophysics Series*, edited by R. DMOWSKA and R. J. HOLTON, Vol. **59** (Academic Press, London) 1995.
[2] SHAEFER J. T., *Weather Forecasting*, **5** (1990) 570.
[3] HANSSEN A. W. and KUIPERS W. J. A., *Meded. Verh.*, **81** (1965) 2.
[4] FLUECK J. A., *A study of some measures of forecast verification*, preprint, in *X Conference on Probability and Statistics in Atmospheric Sciences*, American Meteorological Society (1987).
[5] PEIRCE C. S., *Science*, **4** (1884) 453.
[6] STEPHENSON D. B., *Weather Forecasting*, **15** (2000) 221.
[7] YULE G. U., *Philos. Trans. R. Soc. London, Ser. A*, **194** (1900) 257.
[8] MASON I., *Aust. Meteorol. Mag.*, **37** (1989) 75.
[9] HAMILL T. M., *Weather Forecasting*, **14** (1999) 155.
[10] WILKS D. S., *J. Climate*, **10** (1997) 66.
[11] DIACONIS P. and EFRON B., *Sci. Am.*, **248** (1983) 116.
[12] CASAIOLI M., MANTOVANI R., MARIANI S., NICASTRO S. and LAVAGNINI A., *Studio preliminare per il confronto dei modelli meteorologici ad area limitata operanti in configurazione di servizio in Italia*, C.N.R.-IFA Internal Rep. 1999-4, December 1999.
[13] GRELL G. A., DUDHIA J. and STAUFFER D. R., *A Description of the fifth-generation Penn State/NCAR Mesoscale Model (MM5)*, *NCAR Technical Note*, NCAR/TN-398+STR, June 1994.
[14] LAZIC L. and TELENTA B., *Documentation of the UB/NMC ETA MODEL*, *WMO/TMRP Techn. Rep.*, 1990.
[15] PACCAGNELLA T., TIBALDI S., BUIZZA R. and SCOCCIANTI S., *Meteorol. Atmos. Phys.*, **50** (1992) 143.
[16] BUZZI A., FANTINI M., MALGUZZI P. and NEROZZI F., *Meteorol. Atmos. Phys.*, **53** (1994) 53.
[17] NICASTRO S. and VALENTINOTTI F., *Lect. Notes Comput. Sci.*, **1401** (1998) 151.
[18] ARAKAWA A., *J. Comput. Phys.*, **1** (1966) 119.
[19] TROEN I. and MAHRT L., *Boundary Layer Mereorol.*, **37** (1986) 129.
[20] STEPHENS G. L., *J. Atmos. Sci.*, **35** (1978) 2123.
[21] STEPHENS G. L., *Mon. Weather Rev.*, **112** (1984) 826.
[22] KAIN J. S. and FRITSCH J. M., *Convective parameterization for mesoscale models: The Kain-Fritsch scheme*, in *The Representation of Cumulus Convection in Numerical Models*, edited by K. A. EMANUEL and J. D. RAYMOND, *Meteorol. Monogr.*, **46** (American Meteorological Society, Boston) 1993, pp. 165-170.
[23] RITTER B. and GELEYN .J F., *Mon. Weather Rev.*, **112** (1992) 303.
[24] BETTS A. K., *Q. J. R. Meteorol. Soc.*, **112** (1986) 677.
[25] BETTS A. K. and MILLER M. J., *Q. J. R. Meteorol. Soc.*, **112** (1986) 693.
[26] RUTI P. M., CASSARDO C., CACCIAMANI C., PACCAGNELLA T., LONGHETTO A. and BARGAGLI A., *Beitr. Phys. Atmos.*, **70** (1997) 201.
[27] KUO H. L., *J. Atmos. Sci.*, **31** (1974) 1232.
[28] LOUIS J. F., TIEDTKE M. and GELEYN J. F., *A short history of the PBL parameterization at ECMWF*, in *Proceedings of ECMWF workshop on PBL parameterization, Reading 25-27 November 1981*, edited by ECMWF (ECMWF, Reading) 1982, pp. 59-80.
[29] MESINGER F., *Bull. Am. Meteorol. Soc.*, **77** (1996) 2637.