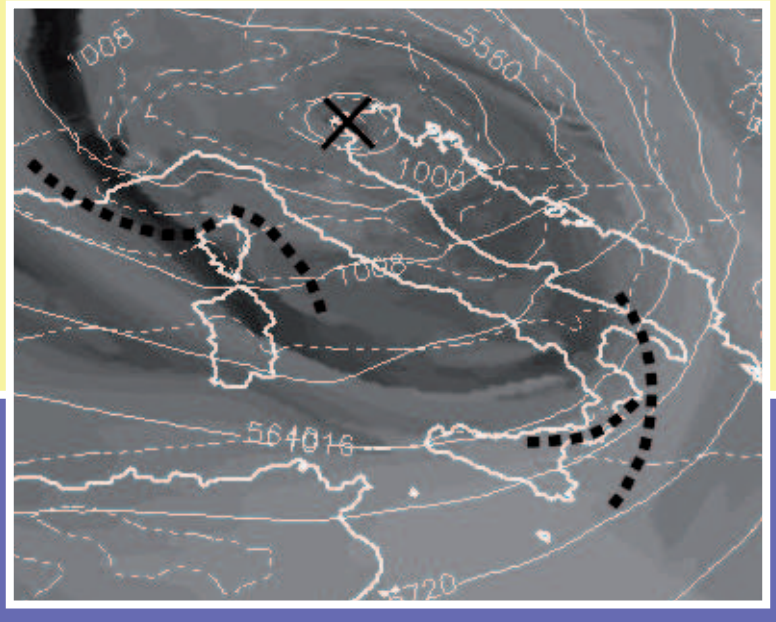
FORALPS

FORECAST VERIFICATION: A SUMMARY OF COMMON APPROACHES AND EXAMPLES OF APPLICATION

Stefano MARIANI, Marco CASAIOLI



Alpine Space
Interreg III B

FORALPS Project Manager:
Dino Zardi

Editorial Team:
Stefano Serafin, University of Trento, Italy
Marta Salvati, Regional Agency for Environmental Protection of Lombardia, Italy
Stefano Mariani, National Agency for Environmental Protection and Technical Services, Italy
Fulvio Stel, Regional Agency for Environmental Protection of Friuli Venezia Giulia, Italy

The project FORALPS has received European Regional Development Funding through the INTERREG IIIB Alpine Space Community Initiative Programme.

Cover design: Marco Aniello

Printed in Italy by: Grafiche Futura s.r.l.

Publisher:
Università degli Studi di Trento
Dipartimento di Ingegneria Civile e Ambientale

Trento, March 2008

ISBN 978–88–8443–234–6

Forecast verification: A summary of common approaches, and examples of application

Stefano MARIANI [(1)], Marco CASAIOLI [(1)]

Contributors:
Michela CALZA [(2)], Irene GALLAI [(3)], Alexandre LANCIANI [(1)], Paul RAINER [(4)], Marta SALVATI [(5)], Fulvio STEL [(3)], Michele TAROLLI [(6)], Christoph ZINGERLE [(7)]

[(1)] *National Agency for Environmental Protection and Technical Services, Rome, Italy*
[(2)] *Regional Agency for Environmental Protection of Veneto, Teolo, Italy*
[(3)] *Regional Agency for Environmental Protection of Friuli Venezia Giulia, Palmanova, Italy*
[(4)] *ZentralAnstalt für Meteorologie und Geodynamik, Klagenfurt, Austria*
[(5)] *Regional Agency for Environmental Protection of Lombardy, Milan, Italy*
[(6)] *Autonomus Province of Trento, Trento, Italy*
[(7)] *ZentralAnstalt für Meteorologie und Geodynamik, Innsbruck, Austria*

Contact: stefano.mariani@apat.it

Contents

1 Introduction

This manuscript, written in the context of the INTERREG IIIB project FORALPS, is intended to provide a brief overview of the forecast and verification activity performed by the project partners (PPs). This document is the final result of a survey about the verification techniques in use by the different PPs; it includes a brief review of verification methods and a description both of routinely performed operational verification activities, and of specific verification studies carried out within the FORALPS projects.

The survey highlights that PPs employ a heterogeneous set of verification methodologies, which includes both standard and scientific/diagnostic verification methods; such as, for instance, the "eyeball" verification, graphical summary techniques, parametric and non-parametric skill scores, spatial verification techniques, etc. Besides, it seems quite evident that some of PPs have a consolidated experience in forecast verification.

A general description of the issue of forecast verification can be found in two widely adopted textbooks: "Forecast Verification – A Practitioner's Guide in Atmospheric Science" by I.T. Jolliffe and D. B. Stephenson (Eds.; John Wiley & Sons, New York, 2003); and "Statistical methods in the atmospheric sciences: An introduction" by D.S. Wilks, (Academic Press, San Diego, 1995). Other up-to-date information is available at the website of the WWRP-WGNE Joint Working Group on Verification http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html. For a thorough review on verification methodologies, the reader can also refer to the bibliography reported at the end of the present report.

Besides verification approaches, our survey also includes information about the meteorological models employed by PPs (either operational at PPs' centres or available via official agreements), the meteorological forecast variables, usually considered in the forecasting activity, and the available observations (from weather stations, radar, radiosoundings, satellite, and analyses). Not all the meteorological forecast variables are subject to verification activities. The important issue of how to treat observational data for spatial analysis, modelling and verification purposes is not developed in the present report. It is instead dealt with in two other FORALPS reports, namely "Data quality control procedures in the regional meteorological services of the Alpine area" by Marta Salvati and Erica Brambilla (ARPA Lombardia) and "Spatial interpolation of surface weather observations in Alpine meteorological services" by Alexandre Lanciani (APAT) and Marta Salvati (ARPA Lombardia).

The structure of the document is as follows. Chapter 2 presents a brief review of the different purposes of forecast verification, according to the Brier and Allen (1951) classification, as well as a survey on the forecast verification process. In Chapter 3, a resume of the forecasting and verification activities performed by PPs is reported. In Chapter 4, results of the verification studies performed within the FORALPS project are discussed. Chapter 5 contains the list of the references cited in the text and of papers, reports and web sites dedicated to forecast verification.

2 Forecast verification

This chapter provides a brief review of the forecast verification process. As stated by Murphy and Winkler (1987) and by Doswell (1996), forecast verification is an essential, necessary component of a forecasting system, since it provides a "measure" of the quality and value[1] of a numerical forecast. Several types of verification methods, including the spatial verification plots, the summary continuous measures, the categorical skill scores and the diagnostic object-oriented methods, are reported here as a reference for the verification studies described in chapters 3 and 4.

2.1 *Verification classification*

According to the classification proposed by Brier and Allen (1951), at least three main purposes to perform forecast verification can be identified: **administrative**, **economic** and **scientific** (or **diagnostic**, following the definition by Murphy et al., 1989 and Murphy and Winkler, 1992).

The first reason is related to the monitoring of the performance of the operational NWP system that produces the forecast. The aim is to evaluate the improvement, through time, of the operational forecasting chain due, for instance, to the implementation of updated numerical schemes, or the porting of the forecasting system to another platform, etc. Thus, the aim is to judge, and financially justify, the changes and the improvement of the forecasting system. In this framework, it is possible to perform intercomparison studies in order to give a relative measure of the forecast performance with respect to competing forecasts.

The second reason focuses instead on the value of the forecast. It concerns the support that a correct forecast can give to a decision-making activity (e.g., civil defence activities, flooding management, agriculture, etc.) from an economic point of view. Therefore, it motivates a user-oriented approach, based on the consideration that a forecast could be "good" for a user and "bad" for another, depending on their different needs.

The last purpose of forecast verification focuses on observations and forecasts and their relationships, in order to underline the ability of a NWP system to correctly forecast meteorological events. Forecast verification provides, this way, skilful feedback to the operational weather forecasters, giving insight about how the atmospheric physical processes are modelled. Generally, such an approach requires the application of more sophisticated measures compared with the ones usually employed for administrative and economic verification tasks.

Other classifications of forecast verification can be done. An example is given by WWRP-WGNE (2008), which provides a useful classification of the different verification methods associated to the different types of forecast. The chosen verification approach should depend on the nature of the forecast (i.e., deterministic, worded, and probabilistic), its space-time domain (i.e., time series, spatial distribution, and pooled space and time), and its specificity (i.e., categorical, continuous, and object- and event-oriented). Hence, it is easy to understand that verification is a multi-faceted

[1] Quality and value are distinct concepts, although both refer to the "goodness" of a forecast (Murphy, 1993). In fact, the latter indicates the economic benefit given by a correct forecast, whereas the former refers to the ability of a model to correctly predict an event, that is, it refers to the degree of agreement between the forecasts and the corresponding observations.

process, since it depends not only on the type of forecast, but also on the reason for which the verification is done.

2.2 *Quantitative precipitation forecast verification*

When performing forecast verification, attention is mainly focussed on the quantitative precipitation forecast (QPF) verification. As pointed out by Mesinger (1996) and Ebert et al. (2003a, b), QPF is indeed considered by several operational weather services as a general indicator of the capability of a NWP model to produce an accurate forecast. Precipitation strongly depends on atmospheric motion, moisture content and physical processes.

When verifying precipitation, the first step is to define the accumulation time intervals to be treated and, if necessary, the spatial aggregation scale (due to the intermittency of the rainfall field). In any case, it is necessary to have a dense and homogeneously distributed network of rain gauges (with a resolution comparable to the model resolution) and to find the "optimum" area for QPF evaluation.

At this point, it should be recalled that traditional instruments, such as rain gauges, provide point measurements, whereas NWP models usually provide areal mean quantities (see, e.g., Haltiner and Williams, 1980). Hence, two approaches can be defined for QPF comparison. In the spatial verification approach, both observations and forecast data are reported on a common grid (sometimes, coarser than the model native grid). Aggregation/interpolation on a regular grid can be performed using several methods, including Kriging, Barnes objective analysis, optimal interpolation, etc. For a thorough discussion on data interpolation the reader may refer to the aforementioned report "Spatial interpolation of surface weather observations in Alpine meteorological services" by Lanciani and Salvati. In the punctual verification approach, it is instead desired to "transfer" the gridded forecast on the rain gauge location; so, model outputs are interpolated to a very high-resolution grid (e.g., 1 km) in order to match the rain gauge observation representation scale. The latter approach is actually origin of controversies not addressed in the present report. For major details about the spatial approach vs. the point approach, readers may refer, for instance, to Cherubini et al. (2002), Ebert et al. (2003a, b), Skelly and Henderson-Sellers (1996), WWRP-WGNE (2004) and references therein.

Finally, it is important to notice that temporal accuracy evaluations are as important as considerations on the spatial scale of the verification. In order to have a reliable idea about models' performance, different accumulation time intervals should be chosen for model validation, according to the purpose of the work. Only this way, a quantitative evaluation of the space and time components of the model error is possible.

2.3 *Eyeball comparison and continuous summary measures*

An eyeball comparison of forecast field against the corresponding observed field is always recommended before starting any statistical verification study, especially if only few forecast fields have to be compared, that is, if forecast verification is performed in a case-study context. This kind of comparison, based on subjective interpretations, is useful to recognize forecast errors, to detect possible observed errors, to qualitatively evaluate the spatial displacement of the forecast field with respect to the observed field, to identify missing data, etc.

After that, summary continuous[2] measures can be calculated in order to assess the differences between the values of the forecasts and the values of the observations (Wilks 1995). Different measures can be defined, such as the multiplicative bias, the mean error (or additive bias), the mean absolute error (MAE), the mean square error (MSE), the root mean square error (RMSE), the Pearson correlation (with confidence intervals assigned by means of the method proposed by Fisher, 1925), the ranked correlation, the S1 score which measures the correspondence between the forecast gradients and the observed gradients, and so on.

2.4 *Categorical skill scores*

A measure-oriented approach is also valid when verifying discrete forecasts, that is, when forecasts are expressed in terms of categories or classes (multi-categorical forecasts). These categories can be defined in terms of ranges of continuous forecast values by introducing several thresholds. When only one threshold is introduced, we are dealing with categorical dichotomous forecasts.

A categorical dichotomous statement is simply a "yes/no" statement; it means that forecast or observed data are below or above a pre-defined threshold. The combination of different possibilities between observations and the forecast defines a 2×2 contingency table (see, e.g., Wilks, 1995), where the forecast-observation pairs are classified into four categories: hits; false alarms; misses and correct non-events. The elements of the contingency table, which are usually indicated as a, b, c and d, respectively, indicates the frequency (not normalized) of each category. The sum of these frequencies represents the total number of the forecast-observation pairs ($n=a+b+c+d$).

Using this kind of categorization, a large number of different scores have been developed over time to evaluate different properties of the categorical forecast (see, e.g., Hanssen and Kuipers, 1965; McBride and Ebert, 2000; Schaefer, 1990; Stephenson, 2000; Wilks, 1995). For example:

- ACC, forecast accuracy. Fraction of the total forecast events when the categorical forecast correctly predicted event and non-event; ACC = 1 means a perfect forecast. Sometimes, this score is multiplied by 100% and it is referred to as the percent correct, or the percentage of forecast correct (PFC).
- POD, probability of detection. Fraction of the observed precipitation events that are also correctly forecast; POD = 1 means a perfect forecast.
- FAR, false-alarm rate. Fraction of forecast events that were not observed; FAR equal to zero means perfect forecast.
- BIA (or FBIAS), bias score (or frequency bias), ratio between the frequency of "yes" forecast (hits + false alarms) and the frequency of "yes" observations (hits + misses). BIA = 1 means that forecast is unbiased, that is, forecasts and observations have a value above a given threshold the same number of times. When the BIA is larger than unity, the model overestimates the frequency of precipitation over the threshold (overforecasting); in the

[2] The term "continuous" refers to the nature of the forecast. From a mathematical point of view, computers provide only a discrete representation of continuous real variables, however, the continuous assumption is reasonable when using computers with high enough machine precision (Jolliffe and Stephenson, 2003).

same sense, a BIA smaller than unity indicates that the model underestimates the frequency of events (underforecasting).

- TS, threat score. Fraction between the number of correct "yes" forecasts and the total number of times that event was observed and/or forecast. TS = 1 means a perfect forecast.
- ETS, equitable threat score. It is a modification of TS that takes into account also the weight of randomly correct forecast; ETS = 1 indicates a perfect forecast, whereas a small (or even negative) ETS indicates a poor (or random) forecast quality.
- HK, Hanssen-Kuiper score. It gives a measure of the accuracy both for events and non-events and it ranges from minus one to one. The main difference between this score and ETS is that HK emphasises in the same way forecast of events and non-events. A score value equal to or lower than 0 shows that model is unable to produce a significant forecast; HK = 1 means a perfect forecast.
- HSS, Heidke skill score. It measures the ability to predict both events and non-events, corrected in order to take into account both random hits and random correct non-rain forecasts. HSS = 1 means a perfect forecast.
- ORSS, the odds ratio skill score. It expresses the ratio between the odds of producing a good forecast to the odds of producing a bad forecast. ORSS is equal to zero when observations and forecasts are independent, whereas a score close to one means that forecasts and observations are associated variables

The indices are defined as follows:

$$ACC = \frac{a+d}{a+b+c+d} \tag{1}$$

$$POD = \frac{a}{a+c} \tag{2}$$

$$FAR = \frac{b}{b+d} \tag{3}$$

$$BIA = \frac{a+b}{a+c} \tag{4}$$

$$TS = \frac{a}{a+b+c} \tag{5}$$

$$ETS = \frac{a-a_r}{a+b+c-a_r},\ a_r = \frac{(a+b)(a+c)}{a+b+c+d} \tag{6}$$

$$HK = \frac{ad-bc}{(a+c)(b+d)} \tag{7}$$

$$HSS = \frac{a+d-T}{a+b+c+d-T},\ T = \frac{(a+b)(a+c)+(a+d)(c+d)}{a+b+c+d} \tag{8}$$

$$ORSS = \frac{ad-bc}{ad+bc} \tag{9}$$

When using these scores, some care should be taken in order to have unambiguous and statistically significant results. First of all, associating an "observed" value to a grid point forecast (see above) is a quite nontrivial task. In particular, precipitation is a strongly intermittent field, so that the mean areal value predicted by the model on the grid point and the point value observed by the rain gauge represent two truly different physical quantities (the so-called representation error).

Secondly, when considering score differences (e.g., between two competing models, or different versions of the same model, etc.), a measure of the statistical significance of the results should be provided. Since in general the probability distribution function of the score differences is unknown, ordinary hypothesis tests cannot be performed. Computer-based resampling techniques, such as the bootstrap method (Diaconis and Efron, 1983), can instead be used to perform non-parametric hypothesis tests, which provide confidence intervals to a binary comparison of model scores. However, some conditions should be respected in order to apply such techniques. This arises in constraints on the requested sample size, in terms of minimum time resolution (accumulation time), time series duration, extension of the sample area and spatial resolution of the grid. Such constraints depend on the space and time correlations of the sample, the data coverage and density and the representation error (Hamill 1999; Accadia et al. 2003a).

2.5 *Diagnostic spatial verification*

Since skill scores are meant to quantify a point-to-point matching between observation and forecast, they are strongly sensitive to localisation and timing errors up to the space and time resolution of the sample. Hence, the above-described non-parametric scores are prone to several shortcomings, which can be expressed with the so-called "double penalty effect": a spatially shifted, but otherwise perfect forecast, arises in two errors: a miss where rain is observed and a false alarm where it is predicted. Consequently, the definition of the verification grid may be a critical issue. Representations of the same precipitation field on different grids, either with different or with the same horizontal resolution, may have different statistical proprieties The effect of grid-to-grid transformations on the statistical properties of the gridded fields is a non-trivial aspect, which should be carefully considered when designing verification studies (Accadia et al., 2005, Lanciani et al., 2008, Mariani et al., 2008).

A deeper insight on the structure of the forecast error can then be provided by diagnostic verification methods as scatter-plots, box-plots, histograms and distribution functions. For instance, scatter-plots give information about the correspondence between forecasts and observations and offer the advantage of presenting in a synthetic way all the statistical information in the data set, differently from statistic summary measures, which compress all the information in few numbers. The histogram approach is instead used to analyse the correspondence between the experimental distribution of the observations and the modelled distribution of the forecasts. Furthermore, this approach gives information on similarities between location, spread, and skewness of forecast and observed distributions (box-plots also provide this kind of information).

Other sophisticated diagnostic methods have been developed for spatial forecast verification. These state-of-the-art methods focus on the realism of the forecast, by comparing features or "objects" that characterize both the observed and forecast fields. For instance, Ebert and McBride (2000) introduced an object-oriented methodology, the contiguous rain area analysis (CRA), that identifies and quantifies the forecast displacement error by "measuring" the spatial error between

observed and forecast precipitation objects. For QPF verification, these objects are identified by isolating precipitation patterns in both the forecast and observed fields, through the use of suitable rainfall thresholds. Once the horizontal displacement is determined (e.g., by minimizing the MSE between the forecast and observed objects, or by maximizing the correlation between the objects), the total forecast error can then be decomposed into its components: displacement error, rainfall volume error, and fine-scale pattern error.

More recently, fuzzy-based techniques, which reduce the constraint about the spatial coherence between the observed and forecast fields, have been developed for forecast verification (see Ebert, 2008, and references therein). The forecast quality is evaluated by measuring the agreement between forecasts and observation within a spatial window, or neighbourhood, of the grid point analysed. Since suitable sizes for the neighbouring window are not known a priori, fuzzy techniques are applied at different spatial/temporal sizes, that is, forecast evaluation is done as a function of the window's size. For this reason, fuzzy methods are referred to be multi-scale verification techniques.

One of the underlying ideas of this approach is that discrepancies between the multi-scale statistical structure of the forecast and observed precipitation fields may result in an incorrect vision of the forecast performance if only traditional statistical measures are used. The power spectrum (Wilks 1995) can be an effective diagnostic tool to study the spatial structure of a gridded field and its scale dependency (Goody et al. 1998).

Given a real field $\Phi(x, y)$, the associated 2-D spectrum $E(k_x, k_y)$, where k_x and k_y are the wavenumber components, is formally defined as the Fourier transform of its autocorrelation function. However, according to the Wiener-Khinchin theorem, it can also be computed by multiplying the 2-D Fourier transform by its complex conjugate, that is:

$$E(k_x, k_y) = \frac{1}{2\pi} \left| \int e^{-i(xk_x + yk_y)} \phi(x, y) dx dy \right|^2 \tag{10}$$

This latter method is to be preferred, since it suppresses some computational noise. Moreover, a Hanning window can be previously used to filter the data and to reduce aliasing (Press et al., 1992).

The relationship between the model domain grid size Δx and the wavenumber grid size Δw is given by $\Delta w = (N\Delta x)^{-1}$, where N is the number of the model grid points. The largest wavenumber from which it is possible to extract meaningful information is given by the Nyquist frequency, $1/2\Delta x$. When comparing different models, those defined on grids with different grid steps will have a different wavenumber range.

The 2-D spectrum can be presented as an isotropic power spectrum $E(k)$, if it is averaged angularly and k is defined as the root-square of the sum of the squared wavenumber components, that is $k = (k_x^2 + k_y^2)^{1/2}$. The width of the bands where the average is made is chosen in order to smooth the isotropic spectrum without losing any significant information.

Scaling of the power spectrum occurs when it can be written as $E(k) \sim k^{-\beta}$. In other words, the spectrum shows scale invariance if it is linear in k on a log-log plot. The absolute value of the spectral slope, β, is an indicator of the field's smoothness. The higher β, the smoother and more organized is the structure.

3 A synthetic resume of the PPs' activities

3.1 *Forecasting activity*

Most FORALPS PPs perform subjective forecasting, providing also daily bulletins, and many of them perform some kind of routine verification activity. Forecasting activity based on numerical weather prediction (NWP) models is also performed as reported in Table 1.

Table 1. Resume of FORALPS PPs' forecast and verification activities.

PP	Forecast		Model configuration[3]			Daily meteo bulletin	Verification			
	Subj.	NWP	Operat.	Resear.	Other		Real time	Case studies	Long series	Inter-comp.
APAT	no	yes	yes	no	no	no	no	yes	yes	yes
ARPA Lombardia	yes	no	no	no	yes	yes	no	no	no	no
ARPAV	yes	yes	no	no	yes	yes	yes	yes	yes	yes
OSMER	yes	yes	no	yes	no	yes	yes	yes	yes	no
PAT	no	yes	no	no	yes	yes	no	yes	yes	yes
ZAMG	yes	yes	yes	no	no	yes	no	no	yes	no

Operational limited area models (LAMs) run at ZAMG (the Austrian version of the Aire Limitée Adaptation dynamique Développement InterNational – ALADIN), APAT (Quadrics BOlogna Limited Area Model – QBOLAM) and EARS (the Slovenian version of ALADIN). Forecasts from the ECMWF global model have been verified at OSMER, PAT and EARS; ALADIN outputs have been also verified at PAT. Moreover, PAT has also verified forecast data modelled by the Italian version of the Lokal Modell (LM), called LAMI, and the BOlogna Limited Area Model (BOLAM).

[3] **Operat.**: PP operationally runs a NWP model at its own Institution. **Resear.**: PP runs in a research configuration one or more NWP models. **Other**: Forecast fields from NWP models, operational in a different Institution, but used by PPs in their own forecast/verification activity.

Model intercomparison studies on case studies and on long time series have been also performed at APAT: QBOLAM forecasts have been compared with meteorological fields modelled by models operating in Italian and European meteorological and research centres (e.g., BOLAM, LAMBO, LAMI, MM5 and RAMS). Furthermore, PAT performed some intercomparison studies in the framework of the EU Project METEORISK (http://www.meteorisk.info/).

Moreover, in the framework of FORALPS, OSMER has implemented the Weather Research & Forecasting Model (WRF) in a research configuration.

3.2 *Standard verification methods: Eyeball verification, graphic plots and tables, time series*

The good old "eyeball" comparison is usually employed in the operational activity.

In the subjective forecast verification, all ZAMG compartments use eyeball verification and graphic tables to evaluate minimum and maximum temperature, precipitation (rain, snow and hail; at different time interval) and gusts. Graphic plots and time series verification is used for global radiation. Observations from both ground stations and radar are employed.

Always with a subjective verification approach, ARPAV employed graphic plots in the verification of the minimum and maximum temperature and the Scharlau (1950) index (an empirical bio-meteorological index, based on the temperature and relative humidity values, which evaluates the effect of temperature on health).

Eyeball verification is used by OSMER, together with graphical plots, tables and histograms to compare ECMWF forecast with radiosounding data at different levels for the following meteorological fields: temperature, geopotential height, relative humidity and horizontal wind.

APAT has mainly used the eyeball verification for analysing single case studies; in particular, for the comparison of precipitation, total column water vapour and temperature on 75 mg kg^{-1} specific humidity surface. It also used histogram as graphical display device in comparing 10-m wind fields modelled by QBOLAM with observation data from buoy network.

At EARS, subjective verification is usually done once per day during forecasters' meeting, including daily forecasts for the next two days, and an up-to-5-day outlook (intercomparison among available NWP model forecasts and verification against point, satellite and radar observations).

Finally, it is common, among PPs, to compare forecast and observation data plotting times series (e.g., APAT, OSMER and ARPAV, but only for verification of ECMWF surface temperature fields) or scatterplots (e.g., APAT; with confidence ellipse). Graphical display by means of plots of difference between observation and forecast data is also used; for instance, by ZAMG-I and OSMER.

3.3 *Continuous and Multi-category Statistics*

Summary measures as BIAS, MAE, MSE, RMSE and correlation are used for quantifying the forecast quality in term of bias, accuracy and association, too.

For example, PAT uses MAE and RMSE for the subjective forecast verification of minimum and maximum temperature; BIAS, MAE and RMSE are instead used by ZAMG compartments for verifying ALADIN forecasts for the following meteorological variables: temperature, geopotential height, relative humidity, wind, precipitation and MSL-pressure. APAT employs MSE and

correlation (with confidence intervals assigned using the Fisher method) in case-study verification of forecast precipitation. OSMER calculates RMSE and anomaly correlation for the extended field verification of the following ECMWF forecast fields: temperature, geopotential height and relative humidity. EARS employs MAE for minimum and maximum temperature forecast by forecaster over 5 locations (for next 24 and 48 hours) whilst, in a synoptic approach, employs BIAS, MAE, RMSE, and SD for verification of ALADIN and ECMWF forecast (ground variables against surface stations, any 3 hours; upper-level variables against radiosondes, any 6 hours).

3.4 *Dichotomous Forecast Verification*

This approach is widely used in the meteorological community to evaluate the categorical forecast quality, especially for verification of precipitation forecasts. The categorical verification data are usually displayed on 2×2 contingency tables, which are used to calculate non-parametric skill scores. Among PPs, only APAT, OSMER, and ZAMG compartments indicated this method in the survey tables.

For example, in verifying QBOLAM forecasts of precipitation (both on case studies and long time series) APAT uses different scores, such as BIA, ETS, HK, ORSS, FAR, etc. A more straightforward approach to skill score interpretation has been obtained by plotting the spatial distribution of the contingency table elements. Besides, in intercomparison studies, the bootstrap method has been also employed in order to assess the statistical significance of score differences between two "competing" models.

OSMER uses the contingency table approach in precipitation forecast verification, too; different skill scores are calculated, such as TS, BIA, HSS, HK, etc.

Verification by means of contingency tables (with specification on used skill scores only by ZAMG-S) is also made by ZAMG compartments for comparing precipitation and 10-m wind fields modelled by ALADIN model with observations.

Contingency tables are employed at EARS in order to compute POD and FAR for precipitation (forecaster) on 5 verification locations, and BIA, POD, FAR, ACC (in percentage) and HSS for precipitation and cloud cover (ALADIN and ECMWF forecasts) on a selected station set. Computation and visualisation of scores is available via web interface. Using the web interface, comparison with ALADIN results from other centres (Meteorological Services) is also possible.

3.5 *Second-order statistical analysis*

At the moment, APAT is employing, on a case-study basis, the power spectrum analysis in order to study the spatial structure of the compared forecast fields (i.e., fields predicted by different NWP models). In this way, it is possible to assess whether the differences, for instance in skill scores, are due to an actual difference of the performance of the competing forecasts or are only related to a different spatial structure. This analysis can also be used to assess the impact of post-processing methods on forecast verification.

3.6 *Probabilistic distribution approach*

A probabilistic distribution approach has been used by APAT in order to verify 10-m wind modelled by QBOLAM against the one observed by buoy stations.

ARPAV and OSMER employ the percentage of correct forecast for the 1-day ground forecast of minimum and maximum temperature. OSMER uses this measure also in the 1-day precipitation amount verification. Besides, a probabilistic verification by means of the diagram reliability is followed by OSMER for temperature, precipitation and sun duration forecasts.

3.7 *Spatial techniques*

State-of-the-art object-oriented techniques (CRA and Hoffman object-oriented method) have been employed by APAT to verify the QBOLAM precipitation and total column water vapour fields.

The S1 score is applied by OSMER to measure the accuracy of the forecast of temperature, geopotential height and relative humidity at different pressure levels,. The metric analysis is also employed in temperature and geopotential height forecast verification.

4 FORALPS verification studies

4.1 *APAT: Model intercomparison on FORALPS case studies*

A common verification activity on selected case studies, employing PPs' models and observations and several verification techniques, has been carried on in FORALPS.

Two case studies of intense precipitation events over the eastern Alpine range have been selected. Both events are connected with the passage of a cyclone over the Mediterranean region, although with a markedly different phenomenology.

Model simulations of the two events have been produced by three Limited Area Models (LAMs), which run, in operational or research mode, at the PP's institutions. Such models are the Slovenian version of the 11-km non-hydrostatic ALADIN model (http://www.cnrm.meteo.fr/aladin/) operational at EARS, the 0.1° hydrostatic QBOLAM model (Speranza et al. 2004, 2007) operational at APAT, and the 10-km WRF (http://wrf-model.org/) model running in its hydrostatic configuration at OSMER.

At APAT, model forecasts have been collected and verified against observations employing a wide range of subjective and objective techniques: eyeball comparison, non-parametric skill scores, CRA analysis and second-order statistics.

4.1.1 *Meteorological description of the two case studies*

The first event (Figs. 1a-c) was due to the passage of two subsequent weather systems – a synoptic trough and a cutoff low – over the eastern Alps. The complex surface pattern on day 17, reported in Fig. 1a, is associated with a large-scale trough aloft, and southerly warm advection over the Alps, with orographic precipitation over the eastern Alpine range. Afterwards, (Fig. 1b) the trough is stretched longitudinally by effect of a stationary high over the eastern Mediterranean Sea. A cutoff low is generated on day 18 (see Fig. 1c) over the western Mediterranean Sea. Its passage provides precipitation on the eastern Alps during days 18 and 19 November.

The second event is linked to the slow passage of a small cyclone over the eastern Mediterranean on days 7 to 10 September 2005. The instability lines in the analysis (see Fig. 1d) are seemingly connected with the observed precipitation. Lightning observations (not shown) provide evidence of a strong convection activity, reaching its maximum over the verification area on day 9 September. For major details on the events evolution, readers may consult Mariani et al. (2008).

4.1.2 *Meteorological models and rain gauge dataset*

The three models have a different domain size (see Fig. 2), covering from the Alpine area (ALADIN and WRF) to the Mediterranean Basin (QBOLAM).

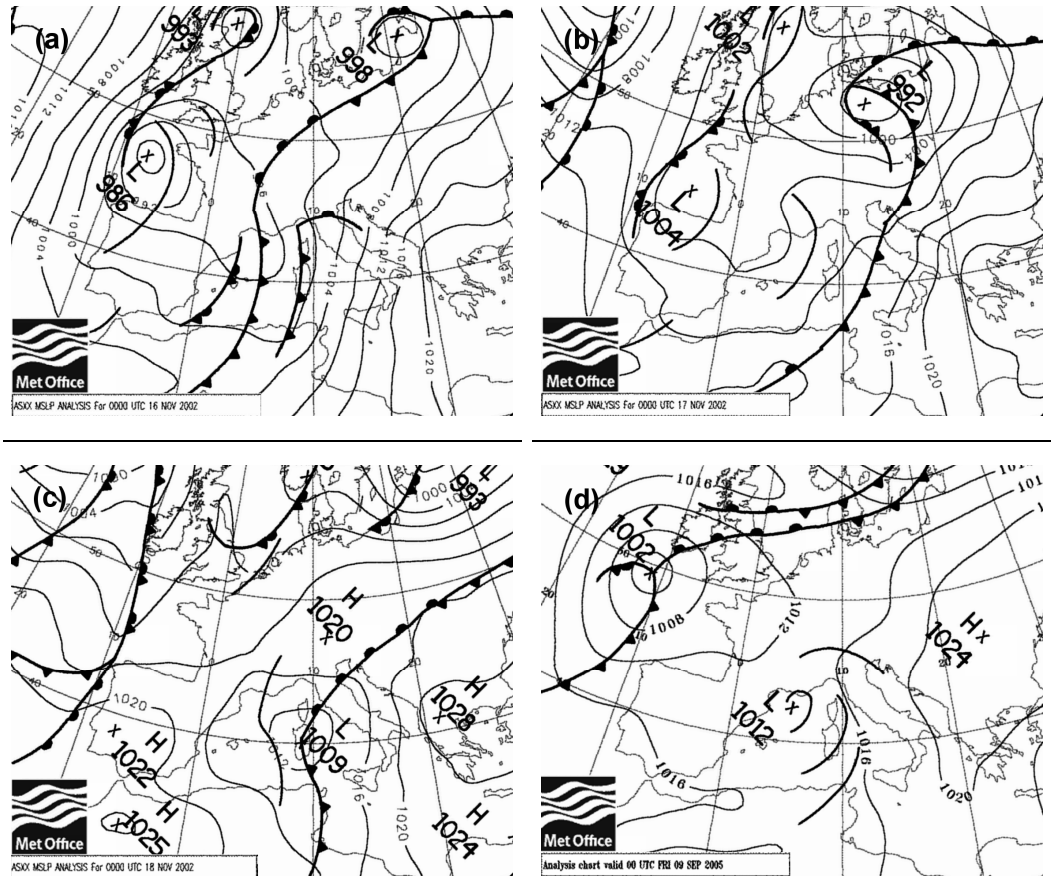


Figure 1. The 0000 UTC mean-sea-level pressure analysis over Europe, for the November 2002 event: (a) 16 Nov. (b) 17 Nov. (c) 18 Nov.; and the September 2005 event: (d) 9 Sep. 2002. Courtesy of the UK Met Office.

Models also differ for the parameterisation and discretisation schemes employed and about initial and boundary conditions. For instance, ALADIN is a spectral model; QBOLAM and WRF are instead finite-difference models.

The 1200 UTC analyses and forecast from the European Centre for Medium-range Weather Forecast (ECMWF) model are used for the initialisation of QBOLAM and WRF, whereas the 0000 UTC analyses and forecast from the Météo-France global model, called Action de Recherche Petite Echelle Grande Echelle (ARPEGE; http://www-pcmdi.llnl.gov/), are used for ALADIN.

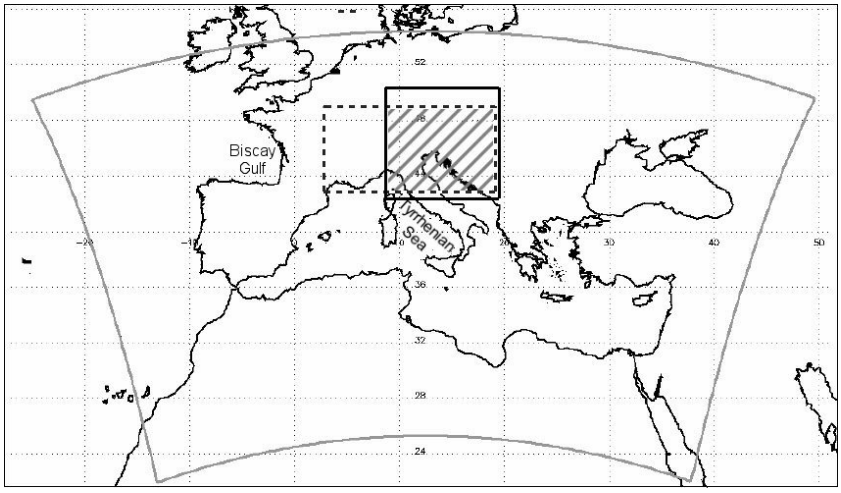


Figure 2. Models' domains: ALADIN (solid black line), QBOLAM (solid grey line), and WRF (dashed black line). The grey shaded indicates the verification area.

For the precipitation comparison, forecast data have been post-processed on two common verification grids (with grid size of 0.1° and 0.5°, respectively) by means of a simple nearest-neighbour average method, also known as remapping (Accadia et al. 2003). Remapping is to be preferred to bilinear interpolation as a post-processing procedure, since the latter tends to smooth the structure of the original field (Lanciani et al. 2008). Moreover, the bilinear interpolation does not conserve the total amount of precipitation forecast over the native grid (Accadia et al. 2003a). Forecasts have also been accumulated on a daily basis, starting from 0000 UTC.

For the November 2002 event, precipitation data have been obtained by the working rain gauges belonging to the networks of APAT (former Italian National Hydrographic and Marigraphic Service network – SIMN; 407 gauges over the northern Italy), the Regional Agency for Environmental Protection (ARPA) of Emilia-Romagna (DEXTER system; 147 gauges over the Emilia-Romagna region), ARPA of Liguria (24 gauges over the Liguria region), OSMER (25 gauges over the Friuli Venezia Giulia region), EARS (18 gauges over Slovenia), and ZAMG (145 gauges over Austria).

For the September 2005 event, precipitation data have been collected from the working rain gauges of APAT (147 gauges only over the northeastern Italy), ARPA of Emilia-Romagna (240

gauges), ARPA of Liguria (119 gauges), ARPA Lombardia (67 gauges over the Lombardia region), OSMER (25 gauges), EARS (21 gauges), and ZAMG (162 gauges).

In order to produce an adequate 24-h gridded analysis of observed rainfall (starting at 0000 UTC) over the two common verification grids, a two-pass Barnes objective analysis scheme has been used (Barnes, 1964, 1973), using the implementation proposed by Koch et al. (1983).

4.1.3 *Subjective QPF verification*

Eyeball comparison for the two rainiest days of the 2002 event –16 and 18 November – gives quite different results. On 16 November, a very good agreement is found between observed and forecast 24-h accumulated rainfall patterns, for all three models (not shown).

A remarkable, and model-dependent forecast error is found instead on day 18 November (Fig. 3).

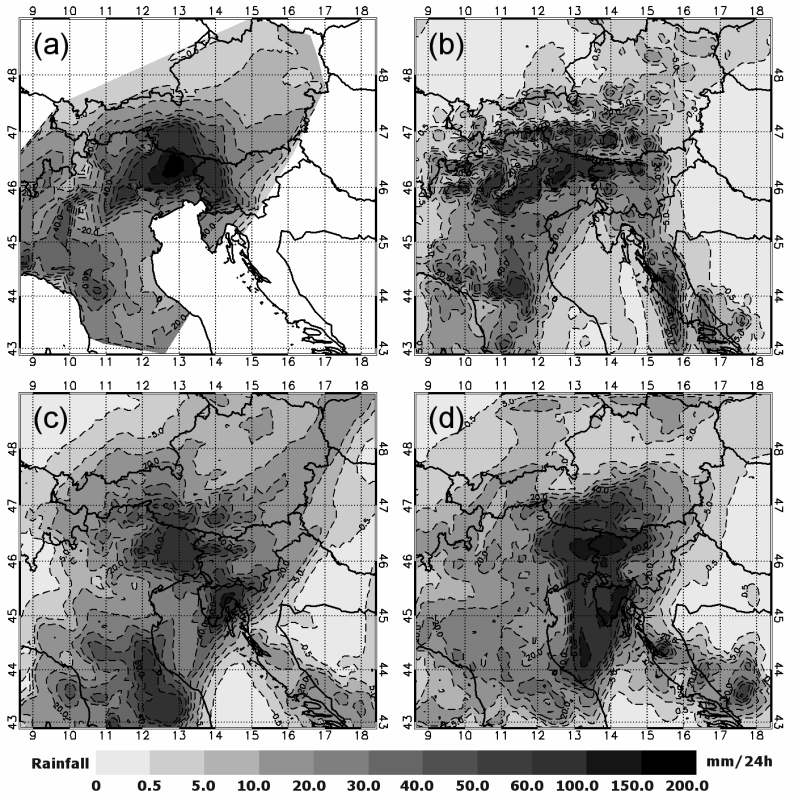


Figure 3. Contour, in mm $(24h)^{-1}$, of precipitation observed (a) and forecast by ALADIN (b), QBOLAM (c) and WRF (d) on 18 Nov. 2002. Forecasts are remapped on the 0.1° verification grid. For observations, the contours of the 0.1° Barnes analysis is masked over the sea.

Such a decrease in the overall model skill can be related to the nature and the scale of the phenomenon. The weather system evolution on days 15-16 November is, in fact, dominated by large-scale forcing whereas the trajectory and the structure of the subsequent cutoff low are sensitive to small-scale processes, which can be either unresolved in the initial analysis, or insufficiently reproduced by LAMs. However, the ALADIN forecast seems to be more effective than others in matching the overall precipitation patterns, even if the absolute maximum – seemingly underestimated by all the models – is better caught by the WRF forecast. Finally, pattern-shifting error can explain, at least partly, the QBOLAM and WRF mismatch with the observed field.

The same comparison for the 2005 event – when rainfall on the target area is concentrated on day 9 September – leads to quite different results (Fig. 4). The event is a miss for ALADIN, whereas QBOLAM seems to provide the best match. However, also in this case WRF displays the best ability in matching the maximum rainfall peak. A similar comparison for the less rainy days 8 and 10 November (not shown) is interesting too, since false alarms are present in WRF and QBOLAM forecast.

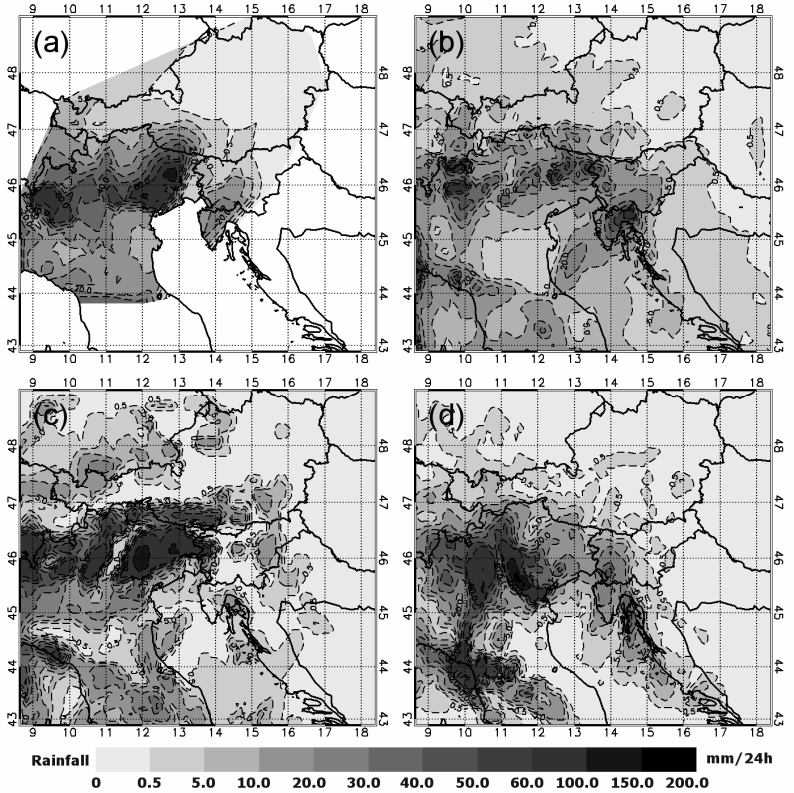


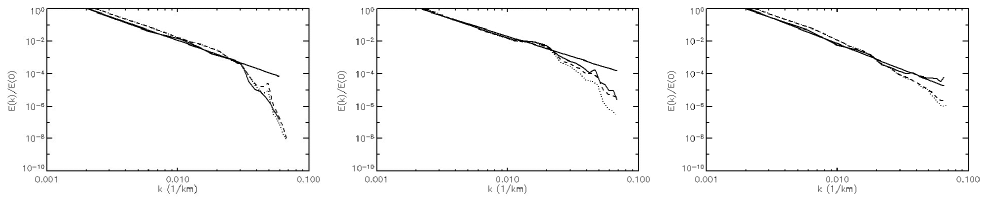
Figure 4. As in Fig. 3, but on 9 Sep. 2005.

4.1.4 *Second-order statistics: power spectral analysis*

In order to assess the comparability of the different models' forecast fields, the power spectral analysis, calculated using a Fast Fourier Transform (FFT) algorithm and the Hanning filter (Press et al. 1992), has been applied to the forecast fields on the shaded area in Fig. 2.

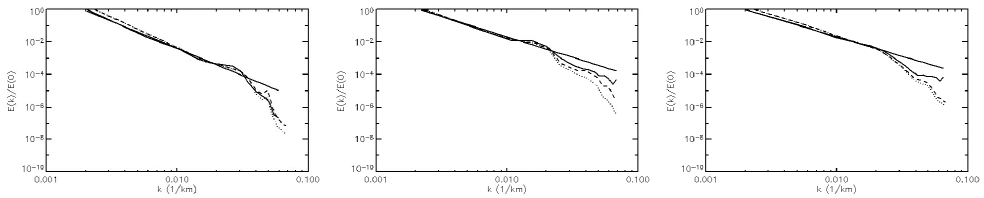
As an example, the spectra obtained using both the original forecast fields (i.e. the forecasts modelled on the native grids) and the post-processed forecast fields (mapped by means of bilinear interpolation and remapping on 0.1° and 0.5° grids) for days 16 November 2002 and 9 September 2005 are reported in Figs. 5 and 6, respectively.

The spectrum of the fields on their native grids displays scale invariance down to about 30 km, after which there is a fall off (see Figs. 5 and 6). Hence, 30 km can be taken as the minimal resolution of the grids insofar as the precipitation is concerned, although numerical considerations (e.g., the implementation of dissipative schemes in NWP models) may suggest that the actual resolution is even coarser (see, e.g., Chèruy et al. 2004).



(a) ALADIN (b) QBOLAM (c) WRF

Figure 5. Power spectrum of the 16 Nov. 2002 24-h accumulated precipitation forecast: original grid (solid line), bilinear interpolation (dotted line) and remapping (dashed line). A linear fit on the first part of the original spectrum is also shown.



(a) ALADIN (b) QBOLAM (c) WRF

Figure 6. Same as Fig. 5, but for the 9 Sep. 2005 forecast.

Both interpolation methods result in smoother fields, although bilinear interpolation slightly more so. This is more striking in the case of QBOLAM (see Figs. 5b and 6b). The scaling parameter β (see sect. 2.6) can be estimated from a linear fit on the first (scaling) part of the spectra. At least for the 2002 event, ALADIN's forecast on the original domain always has more structure than WRF's, which in turn has more structure than QBOLAM's.

Moreover, if we look at the spectra of the fields remapped on the same 0.1° and 0.5° common grids (Fig. 7), it is found that the differences among the models' spectra are greater on the former grid than on the latter. This suggest that, for QPF verification purposes, comparison with traditional skill scores should be performed on the coarser grid only; otherwise particular attention must be put in the discussion of the results, namely with regard to which model performs better.

To stress this point, a preliminarily evaluation of the ETS score was performed on both grids for the 2002 event (see Lanciani et al. 2008). It results that ALADIN is penalised in the higher-resolution intercomparison with respect to the lower-resolution one. This is not surprising since that model, which has more structure at smaller scales, suffer for a stronger double-penalty effect when verification is performed on a high-resolution grid. In general, comparatively smoother fields are less prone to double-penalty effect.

4.1.5 *Objective verification of dichotomous precipitation forecast*

A set of four contingency-table based, non-parametric skill scores (BIA, ETS, HK and ORSS) has been employed in order to provide a model skill intercomparison in terms of categorical precipitation forecast. The use of at least three indices provides a complete description of contingency tables' statistical properties (Stephenson 2000).

According to the second-order statistical analysis results (see above, sect. 4.1.4), skill scores' intercomparison has been performed over the 0.5° verification grid (including the ECMWF forecasts), where the models' structure is comparable.

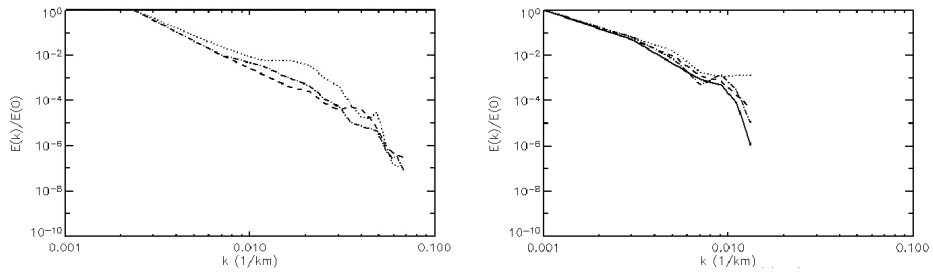


Figure 7. Spectra of the 18 November 2002 24h accumulated precipitation field interpolated on a 0.1° common grid (a) for ALADIN (dotted line), QBOLAM (dashed line) and WRF (dot-dashed line) and on a 0.5° common grid (b) for ALADIN (dotted line), QBOLAM (dashed line), WRF (dot-dash line) and ECMWF (solid line)

For the 2002 event, in fact, up to the 30 mm $(24h)^{-1}$ threshold, ALADIN displays the highest skill scores and the lowest bias among the models, overperforming ECMWF at all thresholds. This picture changes at the higher thresholds: here, the best performance is provided by the global model and, among the LAMs, by WRF (Fig. 8c), which is, nevertheless, affected by an increasingly wet BIA over 40 mm $(24h)^{-1}$ (Fig. 8a). Results are in agreement with the results of the visual verification, especially concerning day 18 November (see sect. 4.1.3).

Also in the 2005 case, the skill score intercomparison (Fig. 9) provides further insight and detail. In general, results confirm the subjective findings (see sect. 4.1.3). A close inspection shows the difficulty of the global model in resolving the structure of the event, evidenced by ETS and HK drop at increasing thresholds (Figs. 9b, c). The corresponding BIA drop (Fig. 9a) shows that ECMWF model underforecasts the rainfall peaks. In the ALADIN forecast this tendency is even more marked (but the model is initialised with ARPEGE); while the other two models catch the event, but are affected by wet BIA at all thresholds (Fig. 9a).

QBOLAM is the only model showing high skill scores on the highest thresholds, despite it has a comparatively worse performance below 30 mm $(24h)^{-1}$. WRF seems to be also penalised, on the highest threshold, by a larger number of false alarms.

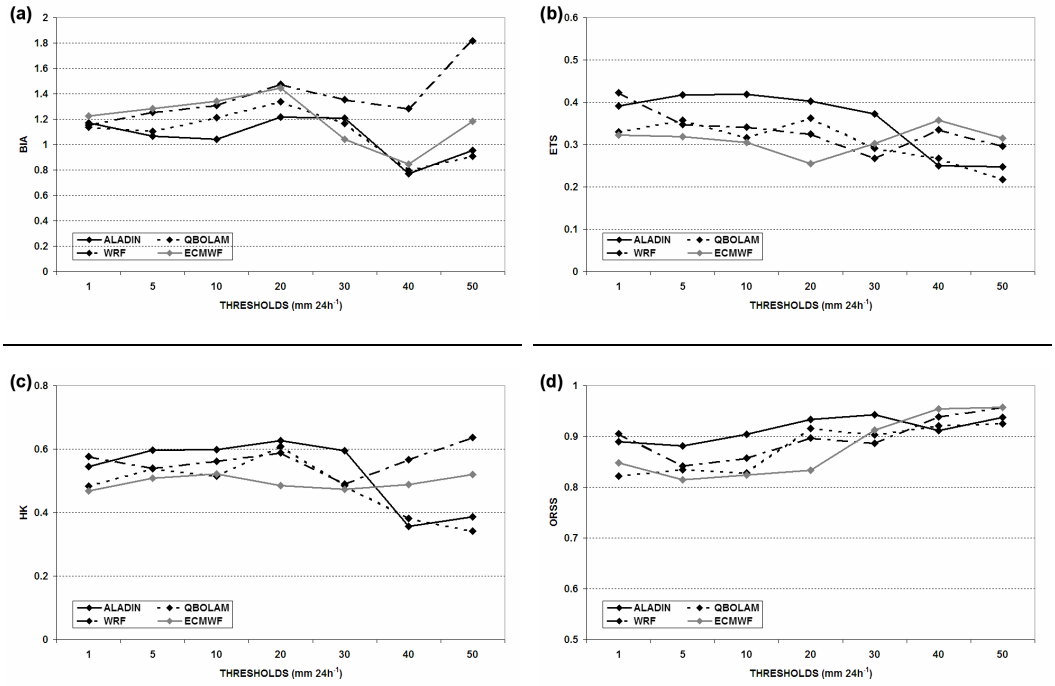


Figure 8. Skill scores for the 2002 event reported as a function of pre-defined thresholds: (a) BIA; (b) ETS; (c) HK and (d) ORSS.

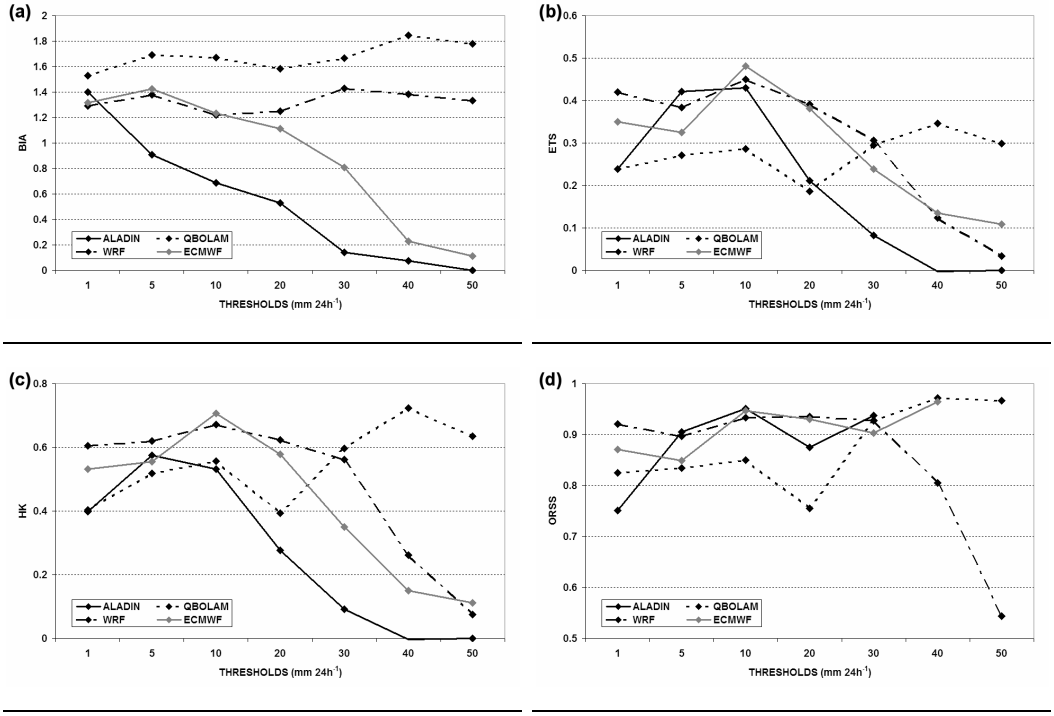


Figure 9. As in Fig. 8, but for the 2005 event.

4.1.6 *Object-oriented precipitation verification: CRA analysis*

In order to provide a quantitative assessment of the spatial properties of forecast error, as found in the subjective analysis (see sect. 4.1.3), the contiguous rain areas (CRA) analysis has been applied**.** Table 2 resumes the outcomes of the CRA analysis applied to the 0.1° remapped forecast fields.

For the 18 November 2002 event, results show that models forecast need to be shifted mainly in the west direction in order to obtain the best match with observations. More in detail, ALADIN shows a dramatic improvement of the correlation after correcting the location error. This is also confirmed by the high displacement error of ALADIN, about 42% of the total error in terms of mean square error (MSE), although pattern error represents the main source of error. For QBOLAM and WRF, the pattern error reaches instead a magnitude of about 80%.

For the 9 September 2005 event, it is worth to note that QBOLAM shows no location error and the pattern error is the quasi-totality of MSE. Pattern error plays a major role also for ALADIN and WRF. However, WRF forecast need to be moved eastward (two grid points) to achieve a best match with the gridded analysis, whereas ALADIN forecast needs instead to be shifted southeastward.

Table 2. CRA verification for 18 November 2002 and 9 September 2005. For each model, the best shift, the number of comparing grid points, the Spearman correlation coefficient (corr.) before and after CRA, and the mean square error (MSE) decomposition, in terms of displacement (displ.), volume (vol.) and pattern (patt.) errors, are indicated.

Date	Model	[E, N] Shift	No. of comparing grid points	Corr.	Shifted Corr.	MSE displ. [%]	MSE vol. [%]	MSE patt. [%]
18 Nov. 2002	ALADIN	[–0.7, –0.1]	351	0.46	0.70	41.2	3.1	55.7
	QBOLAM	[–0.7, 0.1]	386	0.32	0.42	14.2	0.0	85.8
	WRF	[–0.4, –0.2]	392	0.50	0.57	12.6	2.2	85.2
9 Sep. 2005	ALADIN	[0.2, 0.0]	275	0.49	0.56	12.9	0.9	86.2
	QBOLAM	[0.0, 0.0]	399	0.48	0.48	0.0	0.5	99.5
	WRF	[–0.6, –0.2]	278	0.44	0.53	11.9	8.5	79.6

4.1.7 *Further qualitative verification: Forecast meso-synoptic dynamics*

Beyond verification of precipitation forecast, it is worth to investigate how model reproduce the evolution of the weather system responsible for precipitation; this includes, for instance, the origin of the shifting errors evidenced by quantitative and qualitative QPF verification. Some qualitative techniques, including comparison with METEOSAT-7 Water Vapour channel images, are suitable for this purpose; here, only results concerning QBOLAM and the 2002 event are presented (Fig. 10).

First, a comparison between the predicted mean-sea-level pressure field and the weather maps evidences a remarkable error in the surface minimum location on 18 November; moreover, for that day the two runs initialised on 16 November and on 17 November predict the minimum in different locations (not shown). This corresponds to the forecast precipitation shifting evidenced in Fig. 3 and in Table 2.

The role of initial conditions may be checked comparing, at a given time, ECMWF surface pressure and 500 hPa geopotential forecasts from subsequent daily runs. Results (not shown) suggest the existence of a "critical" time: if one or both forecasts start before that time while the comparison time is later, then local forecast differences are higher. This time is found around 1200 UTC 17 November, corresponding to the passage of the cut-off low over Algeria; thus suggesting that such passage is critical for the event predictability.

The forecast temperature field over the 75 mg kg^{-1} specific humidity isosurface (Casaioli et al. 2006) is suitable to be compared with METEOSAT-7 Water Vapour channel images, allowing to verify the QBOLAM forecast error growth (Fig. 10). This is evidenced by the dashed lines and crosses in the figure. Only 12 hours after initialisation, the predicted cyclone (Fig. 10c) is too little

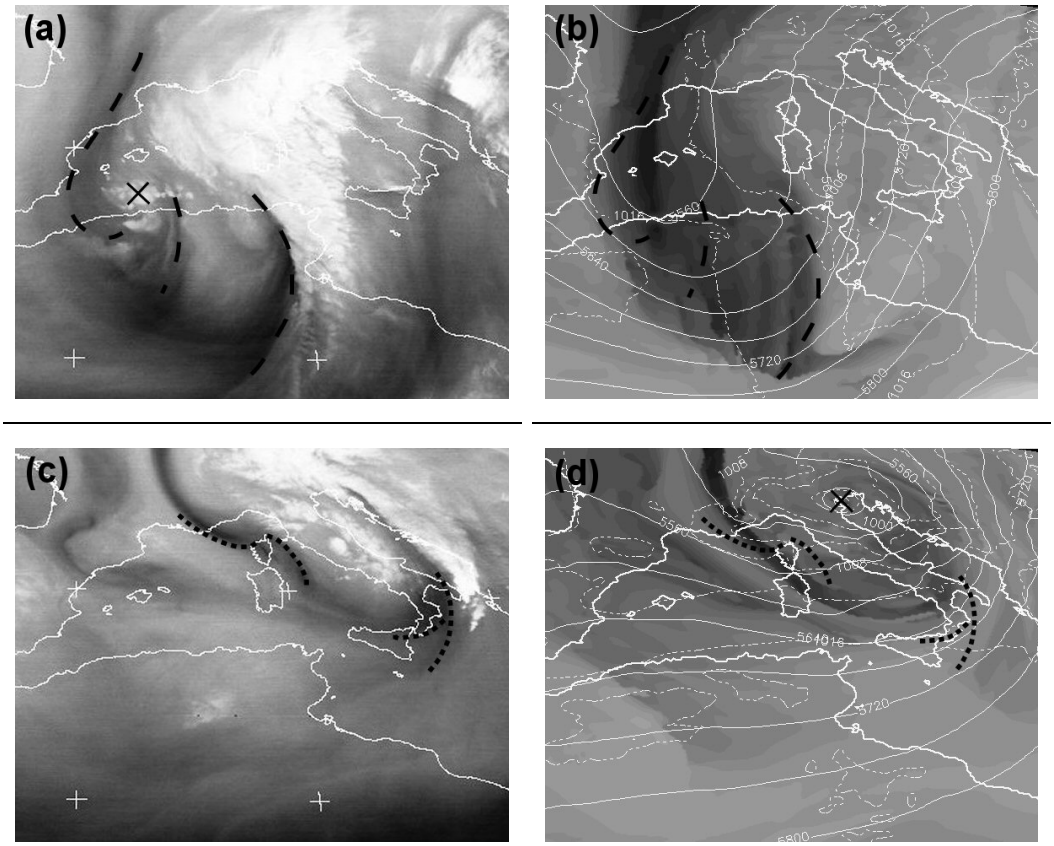


Figure 10. Comparison, at 0000 UTC 17 Nov (a) and (b), and at 0000 UTC 18 Nov, of METEOSAT-7 water vapour channel images (left) with QBOLAM forecast started at 1200 UTC 16 Nov. (right). White lines in QBOLAM plots: mean sea level pressure (dashed), and 500 hPa geopotential height (dotted).

developed on its eastern side; moreover, a secondary vorticity centre (marked by a cross in Fig. 10a) is not predicted. One day later (Figs. 10c, d) the model displays an error resulting from the evolution of the aforementioned two features. Firstly, the forecast cyclone has an insufficient elongation towards southern Italy; secondly, the predicted vorticity centre is badly located. This results in a deformation of the rain band pattern, which affects the rainfall distribution.

4.2 *ARPA Lombardia: Verification temperature forecasts*

From Monday to Friday the Regional Meteorological Service of the ARPA LOMBARDIA (RMS) issues daily temperature forecast, giving a numerical indication of expected lows and highs on the

low lying areas of the Po plain for the following days. To take into account predictability loss, the forecast for the current and following day are worded with the form: "Highs in the plain between x °C and y °C", while for d+2 the forecast are worded with the form: "Highs in the plain around x °C". After d+2 only the tendency is given.

Verifying worded forecast over an area is non trivial, as (1) there might be ambiguity on the wording and in the meaning, and (2) the observation set must, in some way, be defined.

Point (1) is partly take care of by past work done on the glossary of terms and by standardisation between forecasters, making it a very rare occurrence that that wording be different from what reported above.

Forecasters are, when expressing the forecasts, implicitly estimating limits of a distribution of temperatures which they believe significant at the regional scale (mesoscale-β). Consensus among forecasters is that the expressed limits represent the 10^{th} and 90^{th} percentile and the median.

The problem of choosing a significant observation to verify this forecast is common also to NWP verification and is usually taken care of by using an analysis field. This guarantees that local, non-representative scales are filtered out, and that results are less sensitive to the non-uniform station distribution (see for example Fig. 11).

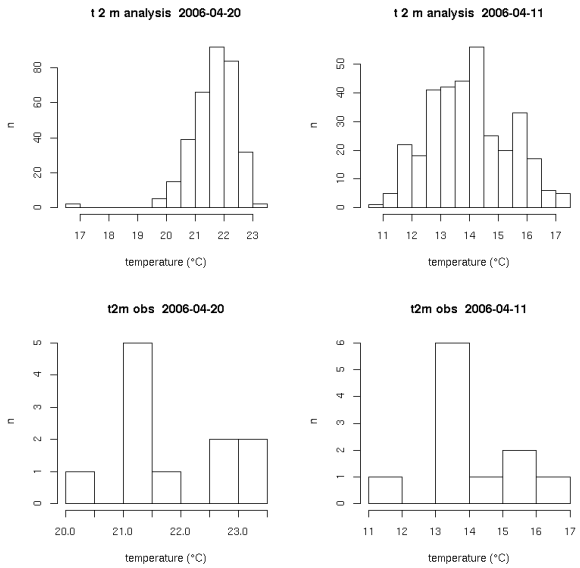


Figure 11. Two examples of temperature distribution from the analysis (above) and from the observation (below).

Fig. 12a shows an example of the hourly temperature analysis developed in the context of FORALPS project, and performed daily. From these, only grid values with height less than 150 m.a.s.l. (Fig. 12b) are kept, because the quantitative forecasts are issued only for the lowland. Of

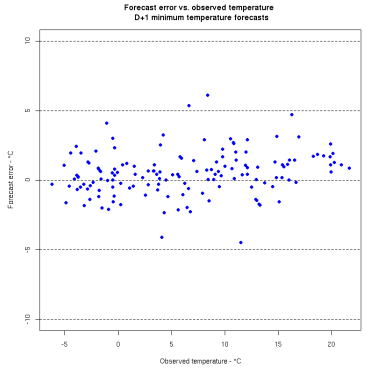
these values, only the grid points where the station density parameter is above 0.8 are considered sufficiently reliable for verification purposes (Lussana 2006). A distribution of ~5000 temperature values results from these selections, more than enough to have a good description of the temperature distribution in the chosen area. From this distribution, two values are extracted to be compared with each d+0 and d+1 forecasts, and one value is instead extracted to be compared with d+2 forecast. Non systematic tests have been performed using quartiles, 10th and 90th percentiles, median and middle point between the chosen percentiles; results have not shown great sensitivity to the choices made, though a more thorough study is needed. For synthesis, only the results of the verification of the median of the worded forecast against the median of the analysed distribution are presented here.

Figure 12. Temperature analysis map at 0100 CET 01 Jan. 2006 (a) and values used for verification, with $z < 150$ m and $xgs > 0.8$ (b).

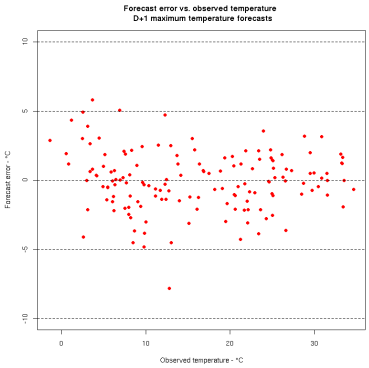
4.2.1 *Forecast and observation distributions*

Fig. 13 shows a notched box-plot of the distributions of observations and of different lead-time forecasts (Fig. 13a lows, Fig. 13b highs). The small (non-significant) difference in the overall distributions shows that the forecasts reproduce well the observed variability of temperature extremes.

Figure 13. Minimum temperature forecasts and observations (a) and maximum temperature forecasts and observations (b).



(a)



(b)

Figure 14. Forecast error ($f_i - o_i$) vs. observation: (a) lows; (b) highs.

Fig. 14 shows scatter plots of forecast error against observed lows (Fig. 14a) and highs (Fig. 14b) for day one. A slight warm bias on the winter highs is noticeable, balanced in the overall scores by a slight cold bias in the middle range of values. The same pattern is visible, though less marked, on d+0 and d+2. Large errors seem to be less frequent in the minimum temperature forecasts, which show, also for d+2, a slight warm bias in the forecast of summer lows.

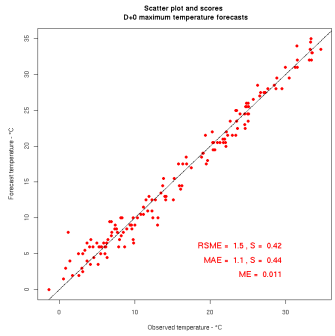
4.2.2 *Scores and skill*

As can be seen from the scatter plots of forecast against observation (Fig. 15), on the same day (panel a) large errors are infrequent, the forecast have almost no bias (excepting a very slight warm bias on winter highs), and the skill on persistence is very high (44%). On day 1, skill over persistence drops by 10% (though confidence intervals on the score still overlap, see Fig. 16); forecasts show a marked warm bias on summer lows and large errors become more frequent, particularly on high temperature forecasts. On day 2, large errors on highs are quite frequent (RMSE = 2.6 °C), the cold bias on highs becomes more visible, and the warm bias on summer lows is still present.

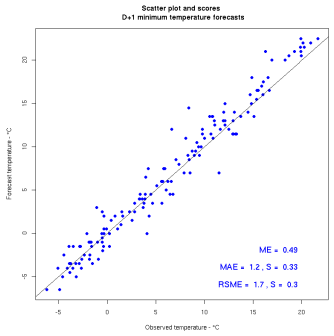
Note that both scores and skill over persistence for highs and lows separately decrease gradually with forecast lead-time (Fig. 16), but scores for lows are better than scores on highs, whereas skill for lows is lower that skill for highs (this could mean that persistence is a better forecasting technique for low temperatures than for high temperatures).

4.2.3 *Conclusions and future steps*

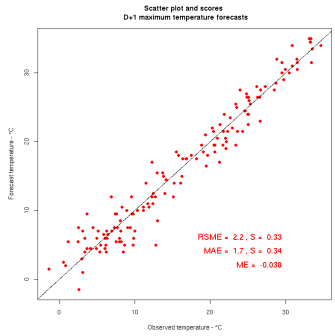
The present study is performed off-line as a preliminary test of automatic temperature verification. In the future an automated verification procedure (Fortran code for temperature analysis and R scripts for verification) will update seasonal and yearly tables and graphics to be published via web.



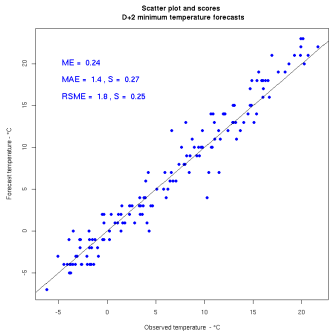
(a)



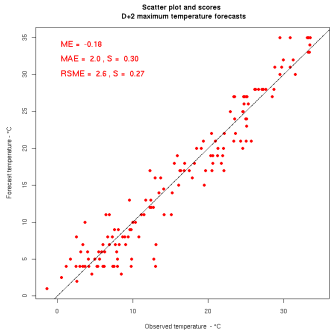
(b)



(c)



(d)



(e)

Figure 15. Scatter plots of forecast against observation. (a), (c), (e): highs; (b), (d): lows.

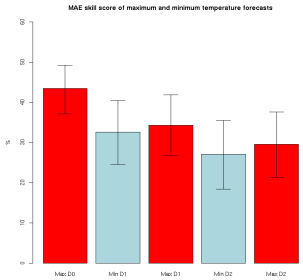


Figure 15. Skill over persistence and confidence interval of the statistic on the distribution (see Wilks 1995).

4.3 *ARPAV: A RADAR-based climatology of convective activity*

Convection is one of the most important meteorological phenomena during the warm season in northern Italy. Thermal convection is relevant close to mountains; moreover, potential instability along the Po Valley can generate several convective phenomena, sometimes associated with severe weather. The unique capability of weather RADARs to monitor precipitation with high spatial and temporal resolution is a well-known feature that allows a more detailed study of this kind of phenomena. A specific tool for weather RADAR data analyses was applied by ARPAV, focusing on the study of convection in the Veneto Region (north-eastern Italy).

The Storm Cell Identification and Tracking (SCIT) algorithm has been exploited to construct a detailed climatology of convection over the domain of Mt. Grande RADAR, managed by the Meteorological Center of Teolo (CMT). Cells identified by the algorithm were catalogued and

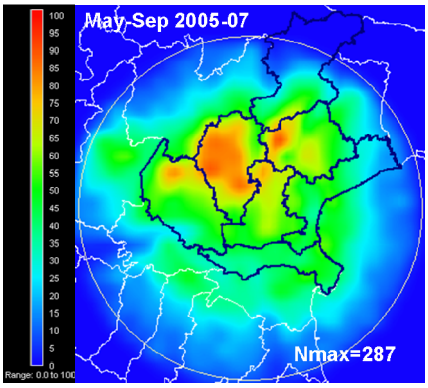


Figure 16. Cell density map for the warm seasons 2005, 2006 and 2007 performed on a lat/lon grid with a mesh size of 0.1°x 0.1° (8x11 km). The color scale is in percent of the maximum number of cells detected in one grid point (Nmax, reported in the upper right).

referenced in space and time; a cell density function was also derived. An off-line version of the SCIT algorithm has been implemented to collect and archive data in a systematic way. A flexible web-based analysis tool has been devised to inquire the SCIT database according to cell attributes. This tool allows the user to extract cells for selected periods of time and stratify them according to one or several of the about 40 parameters of SCIT.

RADAR volumes for the warm seasons 2005, 2006 and 2007 were analyzed to document the convective activity in terms of the cell density, i.e. the number of cells per unit area. On the overall, more than 56000 cell identifications were recorded. Preferred times of the day, geographical distribution, dependence from the month and tracks of convective storms were identified by mean of the SCIT. For example, the province of Vicenza, north west of the RADAR, was identified as the area with the highest frequency of convective activity. This area was hit 3 cells/km^2/3yr cells against a value of 1 cells/km^2/3yr relatively to the entire RADAR domain. The area with the least convective activity turned out to be the province of Rovigo in the south. A relative maximum in the overall cell density map has been found west of the Lake of Garda with 1.5 cells/km^2/3yr. This finding confirms that the Lake Garda FORALPS target area is a preferred region for convective activity; an overall maximum has been found for this area for hail-producing. It should also be noted that convective activity in the Garda region is likely to be underestimated, due to the beam-blocking exerted by the mountain barriers north of Verona and Brescia and to beam height increasing with distance from the RADAR.

RADAR data were used to carry out a preliminary verification of the convective component of the COSMO LAMI quantitative precipitation forecast (QPF) for the warm seasons (May-Sep) 2005, 2006 and 2007. As SCIT does not record quantitative precipitation estimates (QPE), the reflectivity information associated with the cells was converted to rain using a convective Z-R relation. This should give a rough estimate of the rainfall amounts produced by convection and be indicative for the rainfall distribution. This SCIT-derived QPE was compared with the convective component of the LAMI QPF and, for reference, with the rain gauge accumulations of the CMT observing network. An eye-ball verification revealed that the QPF maxima are located mostly in the Alpine and pre-Alpine areas of the Veneto region, whereas the minima are observed on the southern plains; this general distribution is in good agreement with the QPE derived from SCIT

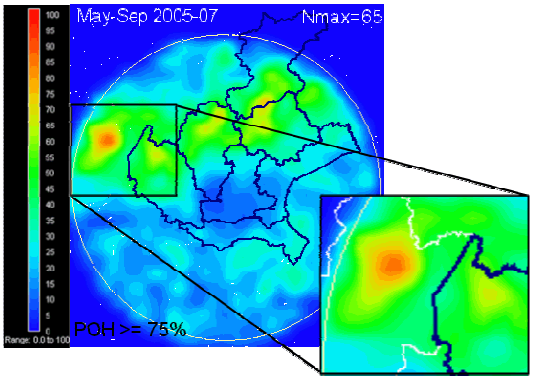


Figure 17. Cell density map with the condition probability of hail ≥ 75% for the warm seasons 2005, 2006 and 2007 performed on a lat/lon grid with a mesh size of 0.1°x 0.1° (8x11 km). The color scale is in percent of the maximum number of cells detected in one grid point (Nmax, reported in the upper right).

and the rain gauge network. SCIT-derived QPE maps agree with the rain gauge accumulations of the CMT observing network, at least regarding the geographical position of the maximum amounts of precipitation.

Further details on the SCIT algorithm and on the results obtained can be found in the FORALPS report "A RADAR-based climatology of convective activity in the Veneto region" by Michela Calza, Alberto Dalla Fontana , Francesco Domenichini, Marco Monai and Andrea M. Rossa.

4.4 *OSMER: ECMWF forecast verification at OSMER*

4.4.1 *Introduction*

One of the main activities of the WP7 of the FORALPS Project was the study of statistical and deterministic methods for objective model verification, with the aim to produce operational tools for verification of numerical models. Tools to produce a daily verification of the ECMWF model outputs have been developed at OSMER – ARPA Friuli Venezia Giulia.

4.4.2 *Methodology*

The administrative region has been partitioned in six distinct areas with similar geographical and climatic characteristics, and each of the OSMER synoptic stations has been assigned to one of these areas according to its location. In this way the calibration refinement approach (Giaiotti et al. 2007 and references there in) has been adopted. This choice, alternative to the likelihood-rate base approach (Giaiotti et al. 2007), was made to adopt the point of view of end users who are planning their activities on the basis of the weather forecast, and would like to know how well the forecast matches reality. The zones identified, shown in Fig. 19, are the Coastal area (blue triangles), the Plain area (in green), the Carnic (in red) and Julian (in pink) pre-alpine areas, and the Carnic (in black) and Julian (in yellow) Alpine area. Different areas might, in any case, be chosen following different criteria, thanks to the software flexibility.

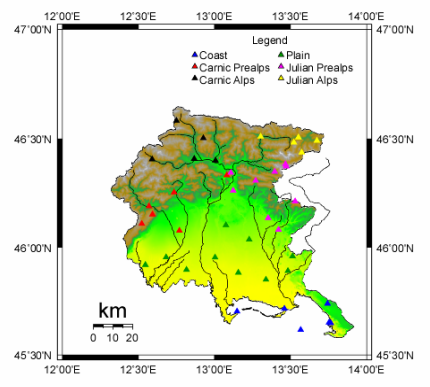


Figure 18. OSMER stations network; each colour corresponds to different verification areas.

About the **observed** value to be compared to the numerical forecast, each area has been assigned the mean value between the data of all the stations belonging to it.

Concerning the **forecast** value at each area, it has been calculated as the mean value of the model output extracted at the grid-points closest to it. The data extraction routines allow the user to select the run of the model, if more than one is available, and the forecast lead-time. It has been chosen to verify the same temporal window (see Fig. 20) for all the runs available, and not the same lead-time. So, in our case, having the 0000 UTC and 1200 UTC runs available, when talking of "one day forecast" (DD1) we take into account the outputs between +24h and +48h for the former, and those between +36h and +60h for the latter, being this one started 12 hours in advance.
Numerical model forecast verification is currently performed at OSMER both daily (as soon as ECMWF outputs and measurements are available), and yearly, at the end of the meteorological year, performing a verification of model outputs on annual and seasonal basis, and making some comparisons with the past years results. The variables verified at the moment are: i) rain occurrence (yes/no); ii) rain amounts; iii) 2-m minimum temperature; iv) 2-m maximum temperature. Other variables can be verified, if desired, according to the data available. Currently at OSMER the verification is made over daily data (accumulated rain and extreme temperature values over 24 hours), but the same procedure can be applied to data with different time range, only by changing some parameter.

The verification procedures include the calculation of statistical indexes from contingency tables, plotting temporal series of data and indexes and producing categorical diagrams (Wilks 2005; Jolliffe and Stephenson 2003). More precisely, in the case of **dichotomous forecasts** that, in our case, refers only to rain occurrence forecasts, contingency tables are produced, and from these the classical statistical indexes are calculated. The indexes selected are BIA, POD, FAR, HIT, TS and KSS. The criterion used for the definition of "rainy day" is the same used for the subjective forecasts issued by OSMER, using a threshold value of 1 mm of daily rain. This definition is not a standard, but is historically used at OSMER; furthermore, it permits to exclude dew effects which are difficult to discriminate from rain using automatic stations.

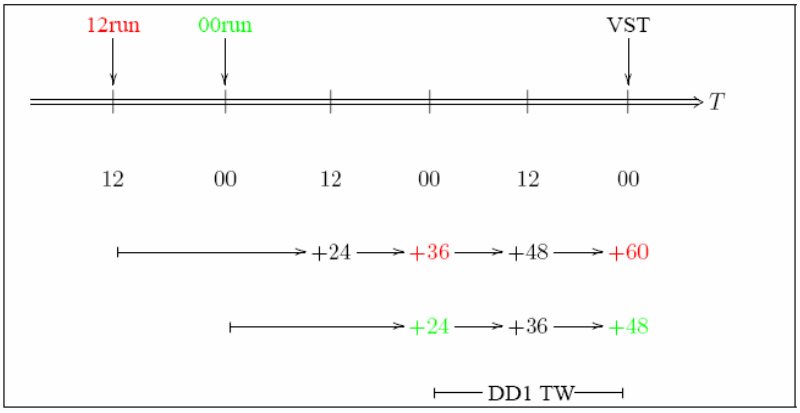


Figure 19. Time scheme for "one day forecast" (DD1) verification. VST stands for Verification Start Time and the range, indicated as DD1 TW, is the temporal window of a DD1 forecast.

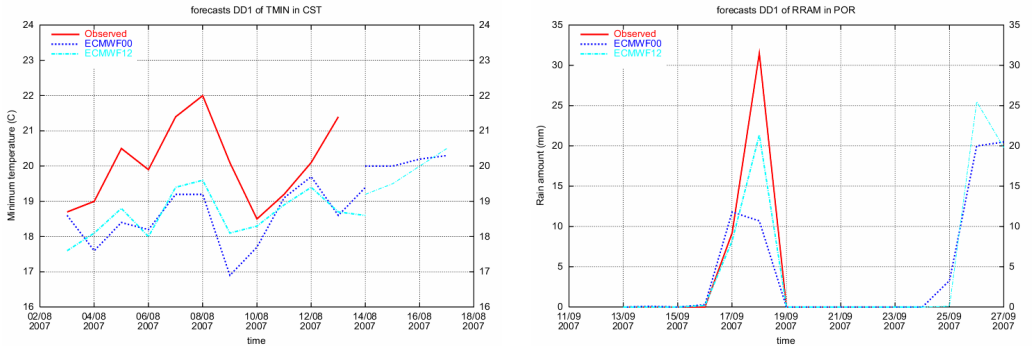
Concerning the forecasts of **continuous variables**, as for example rain amounts and minimum and maximum temperatures, they are evaluated using time series of the percentage of forecast falling in a reasonable range around the observations. In other words, first of all, the continuous variable is transformed in a categorical one, splitting all the possible variable values in a finite number of specific intervals. Once intervals are defined, both forecasts and observations are assigned to one of them, and the forecast is considered correct only if it lies in the same interval of the correspondent observation.

As an example, if the rain amount forecast belongs to the range 15-30 mm day^{-1}, it is considered correct only if the observed rain amount falls between those extremes. The ratio between the number of correct forecasts and the total number of valid forecasts is used to calculate the percentage of correct forecasts that is then displayed using temporal series. Note that the absolute value of this percentage is not very significant, being mostly characterised by the number of no rain forecasts. What is most significant is its trend, which shows any tendency to an improvement or a worsening in the model ability to get the right events intensity.

Furthermore the root mean square error (RMSE) and the half Brier score (HBS) are calculated using all the valid forecast/observation pairs.

Another tool, used mainly for the daily verification, compares forecasts and observations of the last 10 days, plotting their time series in the same diagram (see Fig. 21). Since this tool is used in the operational verification, daily performed at OSMER, in order to give to the forecast also some "prognostic" information, the two (DD2) and three (DD3) days forecast of the latest runs available are also appended at the end of the series.

Finally, with the aim to improve the verification quality itself, categorical diagrams are also produced for this type of variables (see Fig. 22). In these diagrams, for each category, the mean value of observations corresponding to all the forecast falling in a category is plotted versus the central value of the category itself. Considering the points distance from the "calibrated forecast" area, delimited by dotted lines that are obtained plotting the intervals extremes, it is possible to make some considerations over the forecast quality: a point falling over the area is a signal that



(a) **(b)**

Figure 20. Comparison between observations (solid line) and DD1 forecast data (dashed lines) for minimum temperature (a) and rain amount (b); in the same diagram the forecasts for the next days coming from the latest runs available are displayed as well.

generally the forecasts in that class underestimate the event, whereas points under the dotted lines indicate an overestimate. Where dotted lines exist with no points of observation, one or more events were missed. These plots are produced both for the daily and for the long-term verification, the only difference being the sample used for the verification. Annual analysis uses all the data covering a meteorological year, which goes from December to November; the seasonal analysis uses a three-month sample; whereas the daily verification spans the previous 30 days starting from the moment of verification. In any case, all the programs could be customised by the user, simply by changing some parameter value as, for example, verification starting and ending time, lead-time, areas, etc.

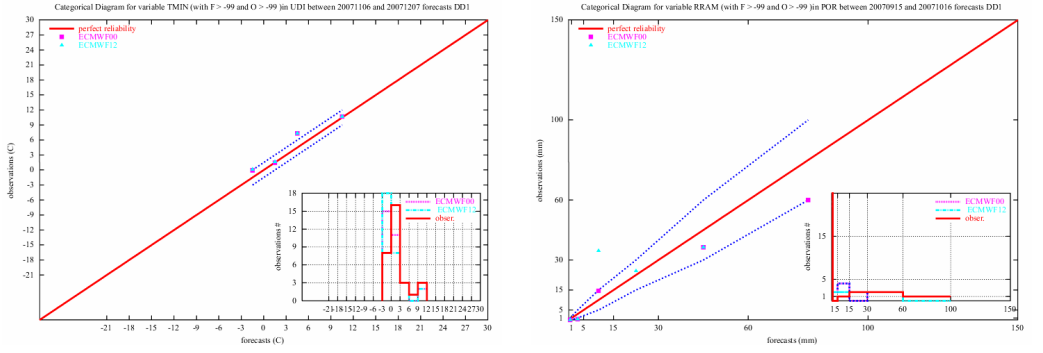
All programs used for forecast verification are written in a Linux environment using programs and tools freely available to the users community: bash and perl scripts, GrADS and Gnuplot. The daily products are made available to OSMER forecasters in a dedicated intranet web page, as soon as all the necessary data are available

4.4.3 *2005-2006 analysis*

Both the runs available in the OSMER database, 0000 UTC and 1200 UTC, have been verified, for the years 2005 and 2006 and for one- and two-day lead times. The variables for the years 2005 and 2006 are only two: i) rain presence (yes/no); ii) rain amount. Maximum and minimum temperatures have not been verified because the 2-m temperature field is available in the OSMER database only starting from the year 2007. The areas in exam are the "Udine" area, representing the plain; the "Trieste" area, representing the coast; and the "Carnic" and "Julian" pre-alpine and a alpine areas. If requested, there is the possibility to verify also different areas.

Presence-absence rain forecasts

First of all, we consider the verification of forecasts in the Udine area. First, from the indexes calculated over an annual sample (Fig. 23), it is possible to notice that there is not a large difference in the performances between the 1200 UTC and 0000 UTC runs. As inferred from BIA and FAR (Figs. 23b and 23d), both the model runs have a tendency to slightly overestimate the rainy days number (BIA >1) and to give a relatively high false alarm rate. The POD (Fig. 23a), instead, has rather elevated values (~80% for one day forecasts and ~70% for two days forecasts),



(a) **(b)**

Figure 21. Categorical diagrams for minimum temperature (a) and rain amount (b).

that is, the number of correct rain forecast is high. FAR and BIA have slightly increased going from 2005 to 2006: this is more evident for the two-day outputs.

Similar charts (not shown) allow to draw analogous conclusion for the other areas in the Friuli Venezia Giulia region. For the coastal area (Trieste) the tendency is similar to what is seen for the plain, maybe even more stressed, since all indexes are some decimal point higher. The POD is slightly reduced going from 2005 and 2006, but is still very high. Decreasing or constant are also FAR and BIA.

For the Carnic Alps, considerations similar to those made so far are still valid: POD is high but FAR and BIA are growing with consequential TS reduction. Moreover, it is possible to notice that there are no large differences between +24h and +48h forecasts. In the Julian Alps area, the same behaviour as for the Carnic area is observed but with a false alarm rate even higher, reaching in 2006 values of 0.5 for two days forecasts. The POD is still high, and constant over the two years considered. For the Carnic pre-Alps area the POD is always near to 80%, but the false alarm rate increased of one decimal point (from 0.3 to 0.4) from 2005 to 2006. Also the TS diminished from 0.6 to 0.5. In the Julian pre-Alps, the trend of the indices for the one-day forecasts is nearly constant, whereas two-days forecasts get slightly worse.

Rain amount forecast

Concerning the forecast of continuous variables, as said before, for the years 2005-2006 only the

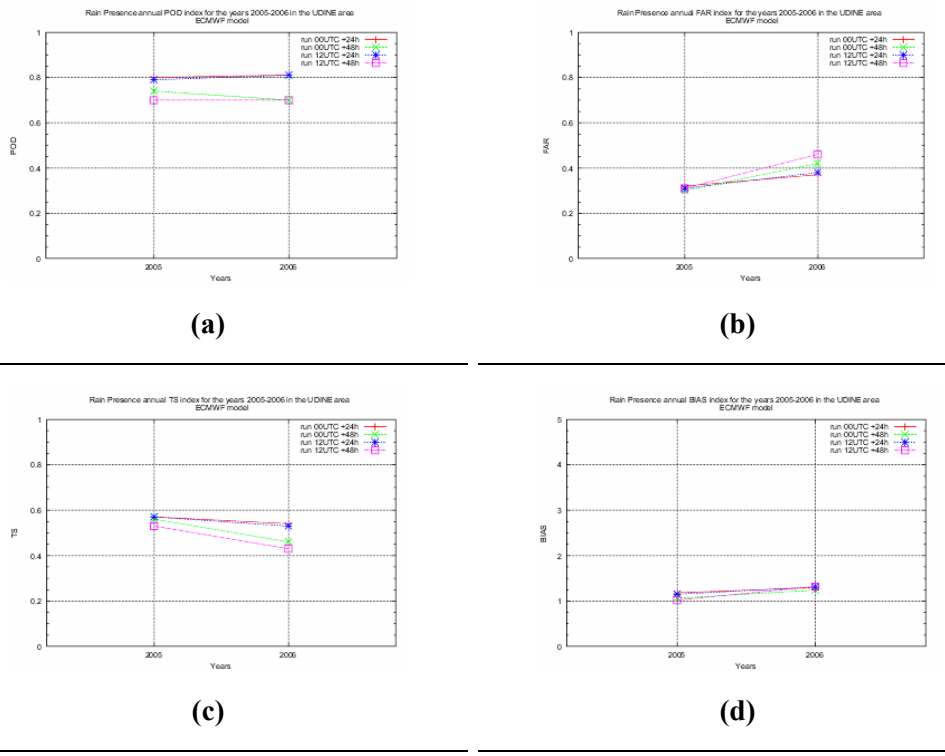


Figure 22. Annual indexes for the rain presence-absence over Udine: POD (a); FAR (b); TS (c) and BIA (d).

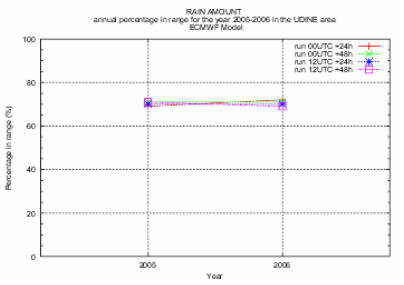
rain amounts have been considered. The temporal trend over the Udine area has been constant in the last two years, without relevant differences between different runs and different lead times (Fig. 24). From the point of view of categorical diagrams, it is observed that in average the models (runs at 0000 UTC and 1200 UTC) are well calibrated, both in the one-day forecasts and in the two-day ones, for both the considered years. Only for the 2005, there is some evidence of one or more "missed alarms" for events falling in the 60-100 mm range. This means that the model has never forecast events in this range, missing them completely or underestimating them.

Similar charts lead to a verification of rainfall amount forecasts in other areas. The coastal area (Trieste) shows for the percentage of correct forecasts a constant trend, with values of about 70%, similar to those of the plain area (Udine). There is a slight increase (of about 5%) for the 0000 UTC run one-day forecast. As for the plain, the differences between one- and two-day forecasts and for the two models are relatively low. Once more this could be an effect of the statistical weight of the non-rainy days. From the point of view of categorical diagrams it is observed that, in average, the rain amounts forecast are well calibrated, with the exception of the 30-60 m day^{-1} category, that is on average overestimated. The 15-30 mm day^{-1} category is also slightly overestimated. As for the plain, in 2005 there was one episode with rain amounts higher than 60 mm, which was not forecast or underestimated by the model.

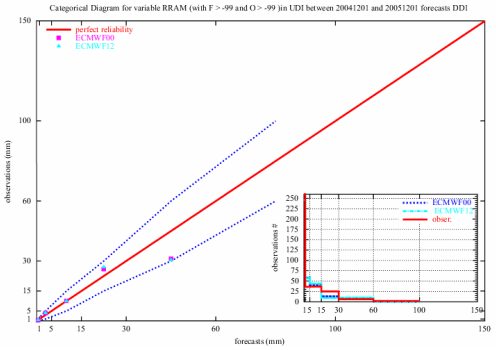
For the pre-Alpine and alpine Area, only the categorical diagrams have been taken into account, looking in detail to the seasonal trend. Concerning the Carnic Alps, on average, the best calibration occurred during the spring 2005. During the same year, in winter, there were some missed alarms for the 15-30 mm day^{-1} and 60-100 mm day^{-1} categories, for both model runs, and the 30-60 mm day^{-1} category for the 1200 UTC run. In any case, also for the 0000 UTC run, this last category is not well calibrated, showing a tendency to underestimate the events. However, it should be taken into account that the number of events of this category is very low (only 5 cases with rainfall over 30 mm day^{-1}). The rain amounts over 15 mm day^{-1} in summer are slightly underestimated, whereas in winter the two runs' behaviour is different: the 1200 UTC underestimates, whereas the 0000 UTC overestimates when forecasting the category 15-30 mm day^{-1}. For the 2006 there are still some "missed alarms", for the 30-60 mm day^{-1} category, for one or both runs during winter, spring and summer. Worth of more attention is the 2006 autumn when the models highly underestimated an event greater than 100 mm day^{-1}.

Also for the Julian Alps area the missed alarms are more frequent for the higher categories (i.e., the 30 and 60 mm day^{-1} thresholds). For the other categories, the calibration seems to be good enough. In any case, there are no events characterised by a rain amount higher than 60 mm day^{-1}.

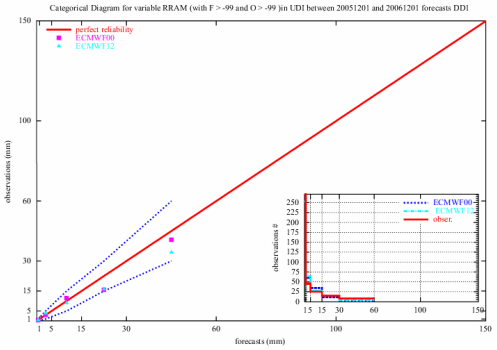
For the Carnic Prealps, we note a tendency by the models to overestimate during summer and to underestimate during autumn and winter. Also here an autumn 2005 event, with rain amounts grater than 100 mm fallen in one day, was strongly underestimated. Also for the Julian Prealps the models hardly forecast heavy rain amounts, with consequently large underestimates. During autumn 2006 underestimates are present also for forecasts falling into the 15-30 mm day^{-1} category, which are generally well-calibrated. Another missed event is in the class over 100 mm day^{-1}, in autumn. Events with great rain amounts were observed also in winter and spring for the same year, events that the model did not miss but underestimate.



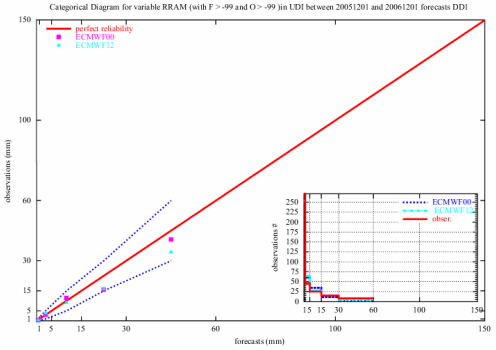
(a)



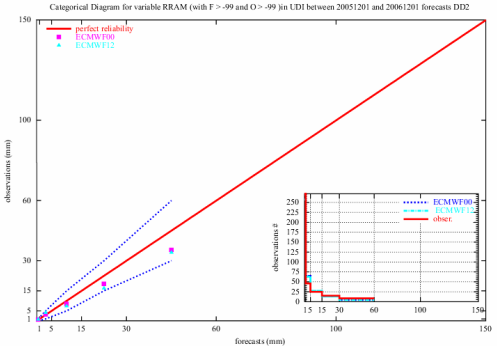
(b)



(c)



(d)



(e)

Figure 23. Annual percentage of forecasts in range (a) and categorical diagrams for 2004 (b) and (d) and for 2005 (c) and (e).

4.5 *PAT: Verification of meteorological forecasts*

The efforts of the verification study carried out by PAT in the framework of FORALPS have been focused primarily on the verification of meteorological forecasts, which may be either objective (by a model) or subjective (produced by a forecaster). The analysis has been performed over the territory of the Province of Trento, where 57 weather stations have been selected to guarantee a representative network of observations over the whole territory. These stations are, in fact, evenly distributed both in area and in altitude.

The time period of interest in this study is the whole two-year span between the 1st of January 2004 and the 31st of December 2005. An update related to the extension of the verification period from January 2004 to December 2006, has been recently made available online on the website: http://www.webalice.it/verifica_previsioni/. The website allows also a larger public to have a look at the general quality of the forecasts emitted by Meteo-Trentino.

4.5.1 *Objective forecasts*

Three forecast models have been taken into account: the global model of the ECMWF, the local model LAMI and the local model BOLAM. The global model ECMWF, run at Reading (UK), has a grid size of approximately 40 km (until the 1st of February 2006) and it produces forecasts every 6 hours. The local model LAMI is run by the Hydro-Meteorological Service of the ARPA of Emilia Romagna region and by the Meteorological Service of the Italian Airforce. It has a horizontal resolution of 7 km and provides forecasts up to the next 72 hours (at 3 hours steps). The local model BOLAM used in this study is run in Genoa at the Physics Department of the local University. Its horizontal resolution is comparable to that of LAMI, but its forecasts extend only up to the next 36 hours. The final aim of this study was to make a comparison between these three models. The variables taken into consideration were temperature and precipitation.

4.5.2 *Technique of analysis*

The main problem was to carry out a significant comparison between the values observed at the stations and the values (referred to the same time) foreseen by the model. The first difficulty encountered is that the station-point and the grid-point are located in different places. It is therefore necessary to transfer the data onto the same position in order to make a proper comparison between their respective values. Unfortunately this transfer (interpolation) introduces an error (the magnitude of which depends on the technique used to interpolate the values). Furthermore, the weather stations give values that are registered in certain points while the data of the models are box averaged. A further obstacle arises because the final aim of this study is to compare not only different forecast models, but also model outputs whose grids have different grid-box sizes (Reading has an horizontal resolution of ~40 km, LAMI of ~7km and BOLAM of ~7km).

To deal with these problems, two interpolation approaches have been in this study: 1) interpolation on the nearest station point; 2) remapping of the LAMI and BOLAM points on the Reading points, and averaging on the grid point of the observations registered (in each cell) by the stations. The first choice allows one to directly compare the values of the models to the values of the stations. This approximation is acceptable in the case of the values of temperature, while the situation becomes more complicated when considering the precipitation data. This parameter is, in fact, deeply influenced by the topography of the territory and presents a large spatial variability. Close

locations can easily be influenced by different events, so it becomes more difficult to compare values that refer to different areas (i.e. the comparison might be not meaningful). So for this variable, instead of interpolating on the nearest station point, the data from the models are firstly "remapped" onto a different grid (data from LAMI and BOLAM are remapped and converted into values averaged on the Reading grid).

4.5.3 *Forecast verification indices*

Several scalar measures are commonly used to assess the quality of a set of forecasts. Being scalar measures of a multidimensional problem, they only partially describe the various aspects of the forecasts. For this reason a combination of these measurements is often used. Within this study, it has been decided to employ two of the most commonly used indexes: the mean-squared error, for the values of temperature, and the equitable threat score, for the precipitation values. The first index represents the average squared difference between the forecast and observation paired values. The second index is equal to 1 in the case of perfect forecasts. The reference (random) forecasts have an ETS = 0. So a value of ETS > 0 indicates an improvement over the reference forecasts.

Table 3. Table 3. ECMWF and LAMI RMSE – interval time every 6 hours.

1200 UTC	Mean Average Error (°C)											
Model	6	12	18	24	30	36	42	48	54	60	66	72
ECMWF	-	1.93	-	1.99	1.88	1.90	2.12	2.00	1.91	1.96	2.17	2.08
LAMI	2.76	2.71	2.94	2.98	2.73	2.84	2.96	2.99	2.79	2.93	2.99	3.03
BOLAM	2.15	2.34	3.00	2.83	2.15	2.48	-	-	-	-	-	-

4.5.4 *Temperature results*

This whole analysis puts into evidence that the Reading model is more accurate than the two local models, LAMI and BOLAM (see Table 3). The differences between the observed and foreseen values range between about 2 °C for Reading and 3 °C for LAMI. The BOLAM model is in an intermediate position, with values of error being about 2.5 °C. Besides, performing a monthly analysis of the forecasts, one can see that the differences between the models are at a minimum between April and October, while they increase in the cold season. In addition, the analysis shows that the 1200 UTC's runs are generally more accurate than the 0000 UTC's runs. This might be helpful in improving the forecasts' quality.

4.5.5 *Precipitation results*

Several ETS were calculated for different values of the threshold of contingency tables. This allows us to analyse how well the models predict different types of events. The verification procedure shows that, for low levels of rain, Reading provides better scores than LAMI (see Fig. 25). When the threshold value increases, the discrepancy between the models decreases. Even if local models seem to provide higher score values for really important events, the uncertainty of the

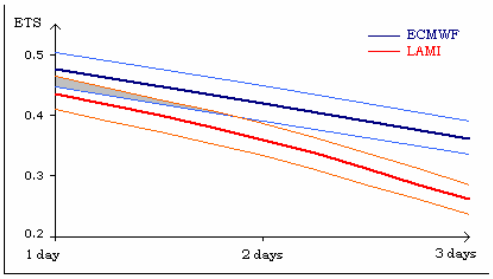


Figure 24. ETS for 0000 UTC's runs, with threshold equal to 5 mm.

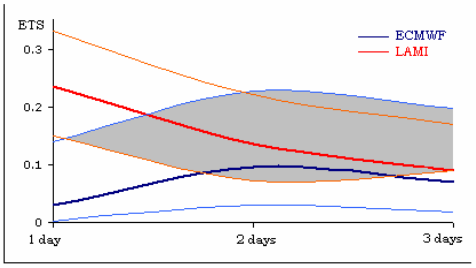


Figure 25. As in Fig. 25, but with a threshold equal to 40 mm.

skill assessment is high (Fig. 26). From a seasonal point of view, autumn is the season characterised by the best forecasts, while summer displays the worst ones.

4.5.6 *Subjective forecasts*

This analysis refers to the probabilistic bulletins, emitted daily by Meteo-Trentino and published on its website (http://www.meteotrentino.it). Again, verification has been conducted on the entire amount of data and on selected sub-periods, in order to evaluate the possible differences in the various seasons of the year. This kind of analysis has allowed to evaluate directly the forecast quality and suggested some modifications to the forecasters (particularly in the indexes scale associated to the different events).

4.5.7 *Conclusions*

The first year of this study has finally allowed the forecasters to acquire several pieces of information about the "behaviour" of the forecast models (how they usually perform and how they predict particular events). In addiction, it was possible to evaluate the quality of the forecasts they generated, and to suggest some modifications (particularly in the indexes scale associated to the different events). The local Civil Protection units also showed interested in this research, having to prevent accidents and danger outcomes due to possible meteorological disasters. It is clear to see how it was interesting for them to evaluate, for each variable, the success rate (up to the next three days).

4.6 *ZAMG Klagenfurt: Verification of precipitation forecasts for Carinthia*

4.6.1 *Introduction*

This study presents a comparison (and a ranking) of the numerical model outputs and the meteorologists' forecasts with "real" precipitation data as measured by rain gauges.

Verification of precipitation modelled by weather prediction models has been one of the main tasks of the WP7 FORALPS. A verification tool has been created to accomplish this task. The area of investigation is the Carinthia region, which is situated in the southern part of Austria. For verification, the Carinthia area was divided in 7 sub-areas, as reported in Fig. 27.

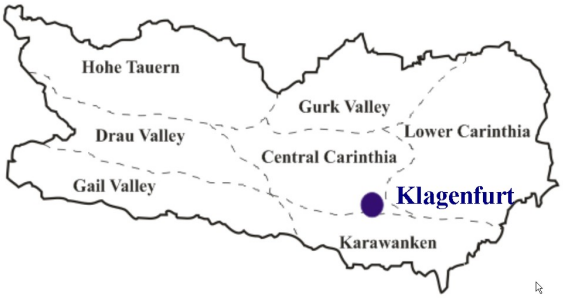


Figure 26. Map of sub areas in district of Carinthia.

The verification study has been performed taking into account the following models: the 40-km ECMWF global model, AUSTROMOS (model output statistic, based on ECMWF), the Deutsche Wetterdienst (DWD) 40-km global model, named GME, the Austrian version of the limited area model ALADIN (with a grid spacing of 7 km), the German limited area model LM (with a grid spacing of 7 km) and finally the meteorologists' forecasts.

From ECMWF, the interpolated values (QFA) at the geographical position of existing SYNOP stations have been taken into consideration. AUSTROMOS combines ECMWF model output data with observed data from SYNOP stations with statistic methods. Both ALADIN and LM provide the mean precipitation for all sub-areas.

The verification was carried out also in precipitation classes (0.1–0.9 mm; 1.0–4.9 mm; 5.0–14.9 mm; $\geq$ 15.0 mm). The verification of model outputs for weak-moderate-strong precipitation episodes for Carinthia should optimise quantitative precipitation forecasts both for electrical power industry and as input to hydrological discharge models. A subtopic was a comparison of stability indices for some special areas in Carinthia with a case study.

4.6.2 *Verification Tool*

For the areas reported in Fig. 27, ZAMG Klagenfurt provides a daily quantitative 48-h precipitation forecast for two regional energy providers. A set of eight 6h-intervals is one of the basic inputs for their power management.

The areas have different topographic characteristics:

- **Hohe Tauern**, mountains up to 3800m get stau wind during advective weather from north and south.
- **Gailtal** (Gailvalley) gets stau wind only from south. Being the primary "stau-region" it represents the area with the highest daily precipitation amounts during single episodes and the highest monthly mean values in Carinthia.
- **Drautal** (Drauvalley) is also influenced from south but at a weaker amount.
- **Karawanken** mountain region is a "stau-region" during flow from southwest and south, but partly shaded by the adjacent Julian Alps in the south.
- **Mittelkärnten** (Central Carinthia) **and Unterkärnten** (Lower Carinthia) is frequently influenced by lee effects.
- **Gurktal** (Gurkvalley) is an area with pronounced convection and thunderstorm activity during summertime.

A tool (Fig. 28, excel application) has been created for online verification for seven regions in Carinthia, and is used in operational work. The model data are saved on a local database in time steps of 6 hours. For the online verification a linear interpolation is calculated in time steps of 10 minutes. For each region, it is possible to obtain details about the verification comparison (see Fig. 29): the values observed by rain gauges (area mean) and forecast by models and forecasters, the relative error (in mm and in percentage) and a ranking, to get a quick overview about which model is the best and how was the performance of the forecaster.

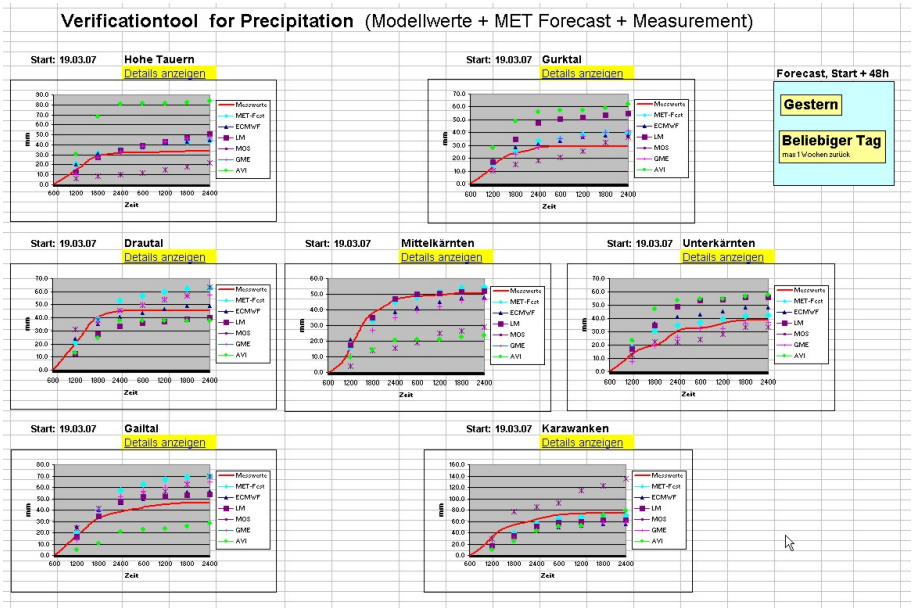
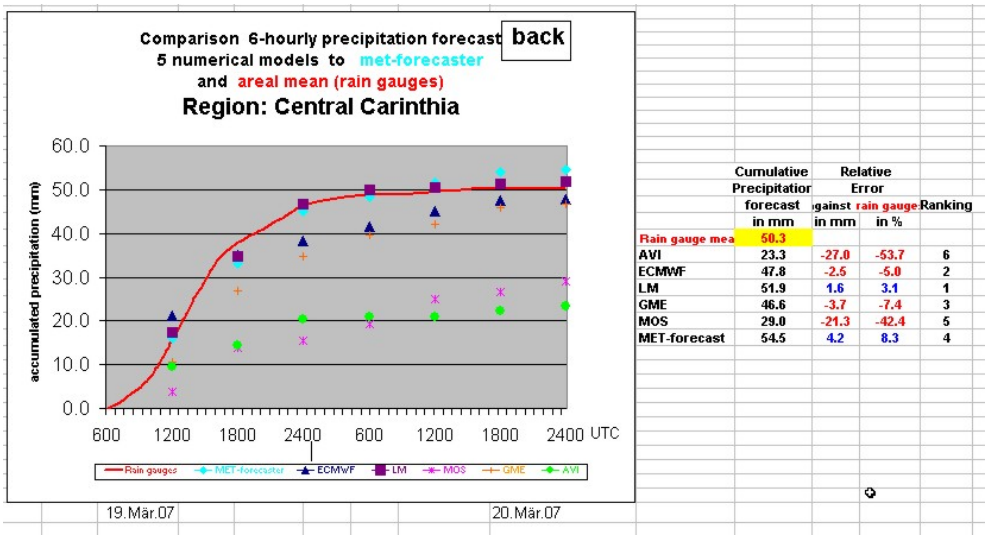


Figure 27. Verification Tool with graphics of 48-h numerical precipitation forecasts for the seven Carinthian sub areas.



	Cumulative Precipitation forecast in mm	Relative Error against rain gauge in mm	Relative Error against rain gauge in %	Ranking
Rain gauge mea	50.3			
AVI	23.3	-27.0	-53.7	6
ECMWF	47.8	-2.5	-5.0	2
LM	51.9	1.6	3.1	1
GME	46.6	-3.7	-7.4	3
MOS	29.0	-21.3	-42.4	5
MET-forecast	54.5	4.2	8.3	4

Figure 28. Verification Tool with graphic for a single sub area with precipitation values, statistic parameters and ranking.

Rain gauge measurements from automatic weather stations were used for the verification. Twenty of them are owned and managed by ZAMG, and twenty-one by the Hydrologic Service of Carinthia (Fig. 30). The METEORISK project brought us very useful data exchange between ZAMG and the Hydrologic Service of Carinthia. From each region, precipitation mean was calculated and compared with the results of the models and with the predicted values of the forecaster.

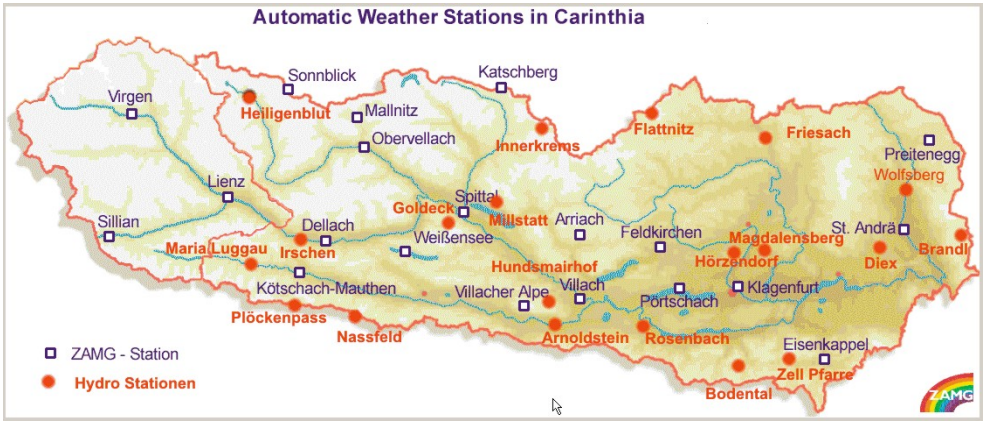


Figure 29. Map of 41 Automatic weather stations in Carinthia (20 ZAMG, 21 Hydro Service).

The verification tool for 7 regions in Carinthia is also available in Internet, in the so-called weather box system of ZAMG-Klagenfurt. The address is: http://wetterbox.kso.ac.at/, where the user id is ***vop*** and the password is ***vop***. For each day, in the weather box the verification is presented with respect to the forecast from the previous day and the forecast for the present day. The last picture for each region is a statistic in 4 classes (0.1–0.9 mm, 1.0–4.9 mm, 5.0–14.9 mm, ≥ 15 mm) for the time range from 01 April 2004 to 17 April 2006.

4.6.3 *Verification of precipitation in classes:*

In co-operation with the FORALPS partners in Friuli Venezia Giulia, these classes for daily precipitation were defined:

0.1 to 0.9 mm	(314 Cases)	Very weak precipitation
1.0 to 4.9 mm	(134 Cases)	Weak precipitation
5.0 to 14.9 mm	(79 Cases)	Moderate precipitation
≥ 15.0 mm	(56 Cases)	Abundant precipitation

Mean absolute error and mean relative error were calculated for the aforementioned four classes in the seven regions in Carinthia, considering the time steps from 6 to 24 hours and from 24 to 48 hours. Examples of such analysis are reported in Figs. 31-32.

For the class 1.0 to 4.9 mm (time steps 6 to 24 hours), MOS presents the highest relative error in all regions, with a maximum in the Karawanken Range (7.2mm). This means that the MOS model overestimates very often precipitation (Fig. 31, left). Again from Fig. 31, also GME has a tendency to predict too much precipitation. ALADIN shows good results in the regions of Central Carinthia, Lower Carinthia, Gurk Valley and in the Gail Valley. In the Karawanken Range LM is the best model. In general, human forecasters perform well, but there is not any particular region in which they are the best.

Looking at the class with a precipitation mean greater equal than 15.0 mm (Fig. 31, right), the mean relative error shows a high underestimation in almost all the regions and for almost all the models, including the meteo forecasters, with the highest values in the Karawanken Range. ECMWF shows the greatest relative error (14 mm). Only in the Hohe Tauern range (ECMWF and LM) and in the Drau Valley (GME) overforecasting of precipitation sometimes occurs. In contrast to the class 1.0 to 4.9 mm, MOS shows a better performance especially in the regions of Hohe Tauern Range, Gail Valley, Central and Lower Carinthia. The meteo forecasters, in almost all the regions, obtain a noticeable underestimation of the precipitation value; only in the Hohe Tauern Range and in Lower Carinthia the error is low.

The relative error gives only information about systematic errors (bias), however the presence of compensating errors may result in an incorrect evaluation of the model. So, quantitatively, a better skill score is provided by the absolute error. In the example class 0.1 to 0.9 mm, forecast ranges 24 to 48 hours (Fig. 41, left), the absolute error shows highest values from MOS. The other models show in all regions a significant better performance with values in most cases lower than 2 mm.

Looking at the class with more than 15.0 mm (Fig. 32, right), the mean absolute error shows values between 7.7 mm (MET-forecaster, Lower Carinthia) and 16.6 mm (LM, Drau Valley). The best performance is provided by the MET-forecaster, with the best result in the Hohe Tauern Range, Drau Valley, Lower Carinthia, Gail Valley and Karawanken Range, and the second range

in Gurk Valley and in Central Carinthia. The highest mean absolute error in the Hohe Tauern Range comes from ALADIN; in the Gurk Valley, Central und Lower Carinthia the greatest mean absolute error was calculated by the GME model; in the Drau Valley the worst model output comes from the LM; in the Gail Valley the highest mean absolute error comes from the MOS and in the Karawanken Range ECMWF was at the end of the ranking, but there are only small differences from model to model.

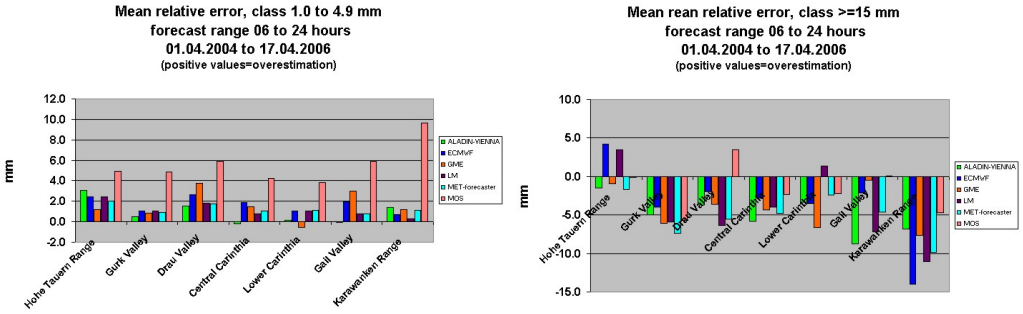


Figure 30. Left: Mean relative error of precipitation class 1.0 to 4.9 mm, forecast range 06 to 24 hours. Right: same indices for precipitation ≥ 15 mm..

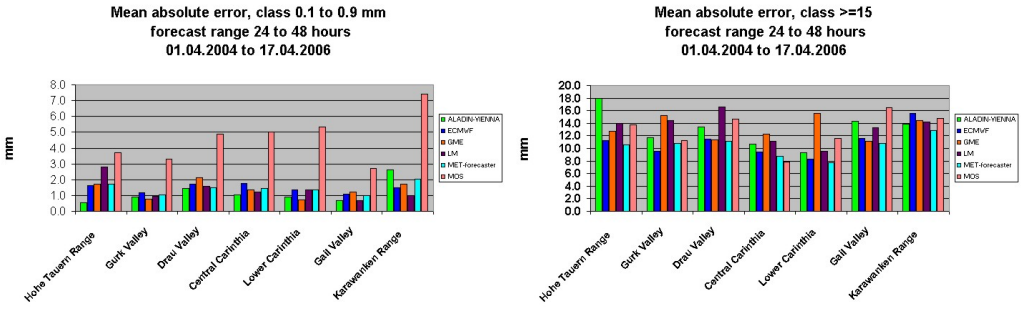


Figure 31. Left: Mean absolute error of precipitation class 0.1 to 1.9 mm, forecast range 24 to 48 hours. Right; same indices for the precipitation class ≥ 15 mm..

4.6.4 *Comparison of stability indices*

A climatology of lightnings was first developed in the region of Carinthia (Fig. 33) using data from 1995 to 2006. The investigation was done for 6 sub-areas with different forms in the landscape. In the Hohe Tauern region high mountains dominate, with glaciers and long snow coverage in spring. In the Carnian Alps, in the southwest of Carinthia, moist air from the Adriatic Sea causes more thunderstorms than in the central Alps and Hohe Tauern regions.

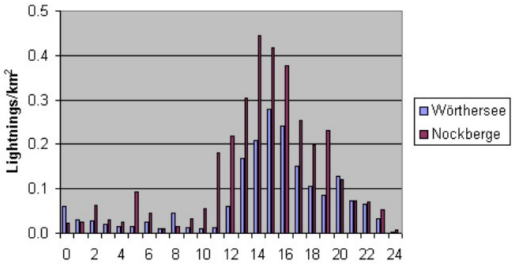


Figure 32. Mean frequency of lightnings per year and square kilometres in six Carinthian regions (1996-2006).

The highest probability for thunderstorms in Carinthia is found in the area of the Nockberge and in the region of the Lavent Valley, with more than three lightning per year per km^2. These mountains, with heights around 2000 meters and a lot of wooden areas, offer very good conditions to start a convective system. In the central area of Carinthia, around the lake Wörthersee, thunderstorms do not occur very often. The Karawanken Range is oriented west to east similarly to the Carnian Alps, but is affected by fewer lightning per year and km^2: the influence of the Adriatic Sea is not so high, since these mountains are situated in the lee side of the higher Julian Alps. The average occurrence of lightning in the course of a day in the Wörthersee region and in the mountain region of the Nockberge is given in Fig. 33.

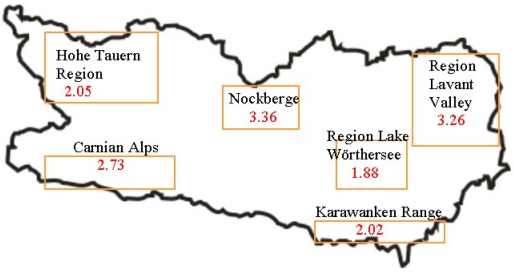


Figure 33. Mean lightning activity during the day in the mountain region Nockberge and around the lake Wörthersee (1996-2006).

In the mountain region lightning begins earlier with its maximum at 14 o'clock; in the basin of lake Wörthersee the maximum is one our later. Very often thunderstorm begins above the mountains, while the sun still shines above the lake. Later on, but not so frequently, the thunderstorms reach also the region of the lake.

To assess the potential of a thunderstorm, the Showalter Index (SWI) can be used. Thunderstorms are unlikely if SWI > 4, and they are possible only if strong triggering occurs if SWI is between 1 and 3. The chance for thunderstorms increases if SWI is between 1 and -2. The atmosphere becomes very unstable, and has a good potential for heavy thunderstorms, if SWI is between -2 and -5. Tornadoes are possible due to extreme instability if SWI is lower than -5. Fig. 35 shows the Showalter Index calculated from the ECMWF model forecast in comparison with the lightnings observed in the mountain area of Nockberge and in the region of Wörthersee in summer, for the period from 1 May 2006 and 30 September 2006.

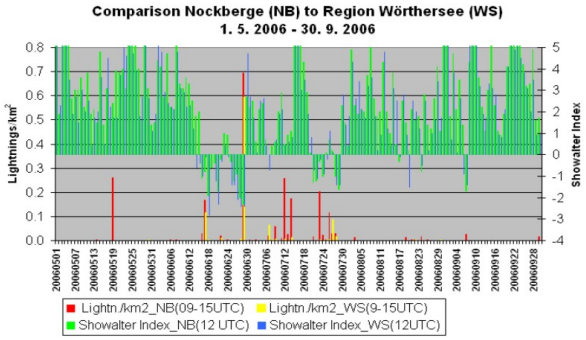


Figure 34. Showalter Index from ECMWF and lightnings, from 01 May 2006 to 30 September 2006, in the regions of Nockberge and Wörthersee.

Another index to assess the potential of a thunderstorm is the KO index. KO > 6 generally means no thunderstorms. KO 6 to 2 should be a signal for scattered thunderstorms. No thunderstorms at all should occur if KO < 2. Fig. 36 shows the KO index calculated from the ECMWF model in comparison with the lightnings in the mountain area of Nockberge in summer, for the period from 1 May 2006 to 30 September 2006.

Also the Total Totals (TT) index can be used to assess the potential of a thunderstorm. A TT greater than 44 is generally associated to thunderstorms. TT greater than 48 correspond to severe thunderstorms. If TT > 50, tornadoes are possible. Fig. 37 shows the Total Totals calculated from the ECMWF model in comparison with the lightnings in the mountain area of Nockberge in summer, for the time period from 1 May 2006 to 30 September 2006.

Finally, the Darkow Index (DI) can be used to assess the potential of a thunderstorm as well. A value of DI > 0 is typical of a condition with no thunderstorms. With DI > –1 light thunderstorms are possible; with DI > –2 thunderstorms are likely; with DI < –2 heavy thunderstorms should be expected. Fig. 38 shows the Darkow Index calculated from the ECMWF model in comparison with the lightnings in the mountain area of Nockberge in summer, for the time period from 1 May 2006 to 30 September 2006.

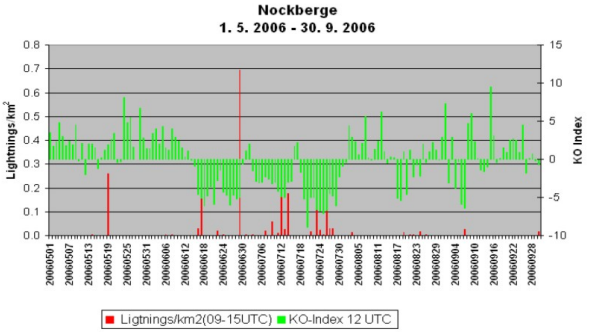


Figure 35. KO Index from ECMWF and lightnings, from 01 May 2006 to 30 September 2006, in the region Nockberge.

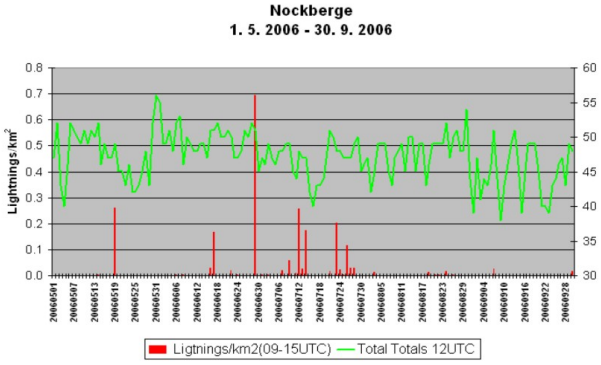


Figure 36. Total Totals from ECMWF and lightnings, from 01 May 2006 to 30 September 2006, in the region Nockberge.

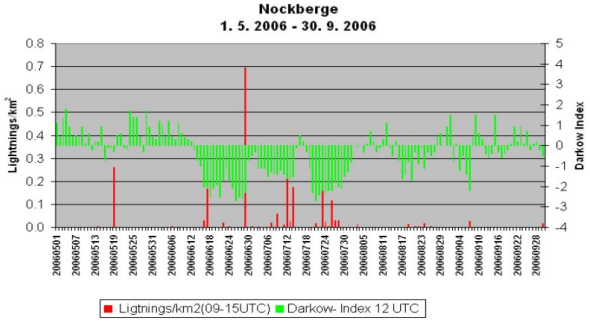


Figure 37. Darkow Index from ECMWF and lightnings, from 01 May 2006 to 30 September 2006, in the region Nockberge.

4.6.5 *Case Study*

On 29 June 2006, heavy thunderstorms occurred in Carinthia. Unusually, first heavy thunderstorms began in the morning (Fig. 39, left) in the northern part with a high density of lightnings and with heavy hail in the Gurktal Valley at 8 o'clock in the morning (Fig. 39, right). The reason was a weak trough in high levels; the instability comes from a cooling in the height from this trough. In the 500 hPa chart (Fig. 40, left) this trough at 0000 UTC was above France. During the day, intensive sunshine heated up the air, so that in the unstable air in the afternoon heavy thunderstorms developed again with storm and hail. The radar picture (Fig. 40, right) shows a high intensity. All stability indices (Showalter Index, KO Index, Total Totals and Darkow Index) showed on this day a high instability.

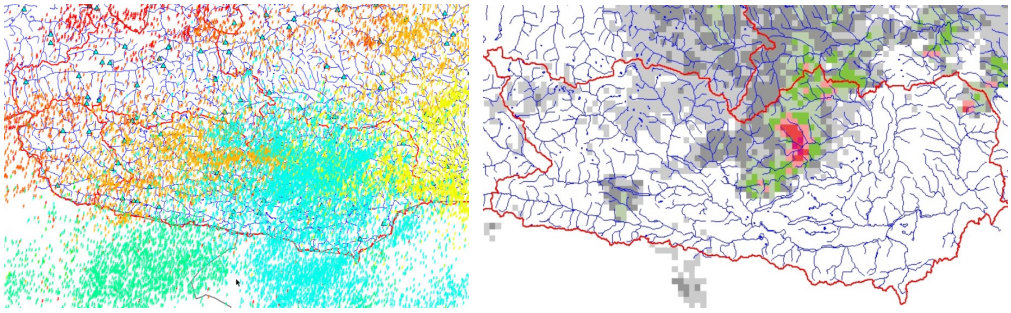


Figure 38. Left: heavy thunderstorms on 29 June 2006 caused a very high density of lightnings in Carinthia. Right: Radar image for the morning (0600 UTC) of 29 June 2006 with local heavy hail in the northern parts of Carinthia.

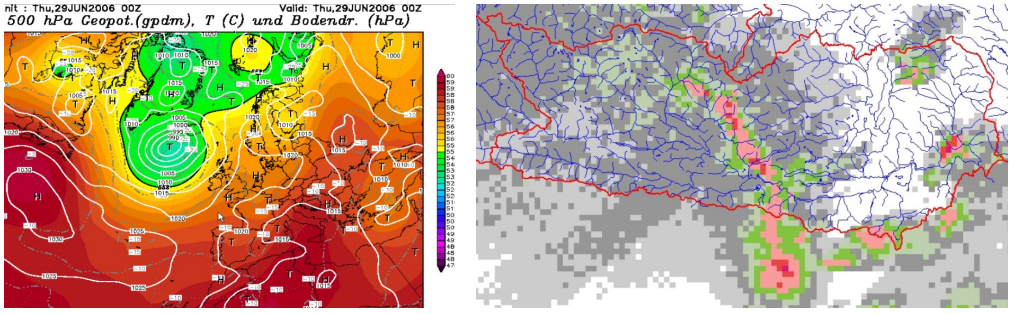


Figure 39. Left: 500 hPa and mean sea level pressure at 0000 UTC 30 June 2006. Right: Radar image for the afternoon (1400 UTC) of 29 June 2006 with local heavy hail and storm central und in the southeast of Carinthia.

4.7 *ZAMG Innsbruck: Verification of precipitation forecasts for Tirol*

The code used for forecast verification at ZAMG Innsbruck has been developed with the "R" geostatistical software, and is still in a development phase (therefore not operational). Currently, a database of forecast and observation fields has been filled with data since March 2005. Precipitation is the only parameter considered in the analysis. Five automatic weather stations within the Tyrol area, representative of five local climates have been considered (Figure 41), namely two stations situated in the northern Alpine slopes (Reutte, Kufstein) and three inner Alpine stations (Landeck, Innsbruck, Lienz).

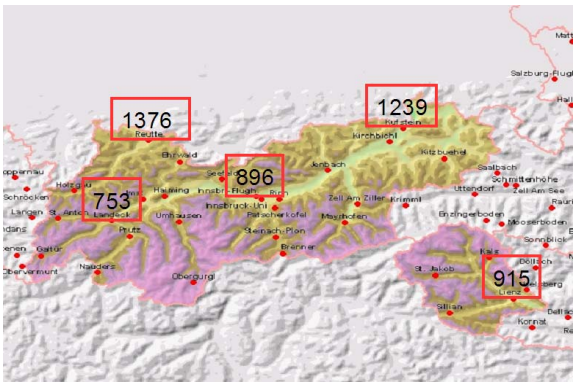
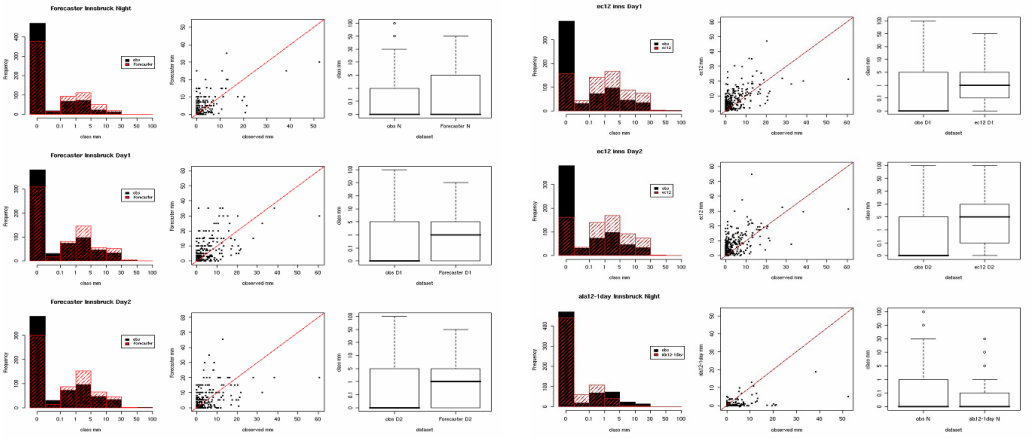


Figure 40. Location of the five rain gauge stations considered for forecast verification.

Forecast data have been provided by:

– Forecasters: night (18–06), day 1 (06–06), day 2 (06–06);
– ECMWF 0000 UTC run: night, day 1;
– ECMWF 1200 UTC run: day 1, day 2;
– ECMWF 1200 UTC run of the previous day (w.r.t. the one analysed): night, day 1;
– ALADIN (Austrian version) 0000 UTC: night;
– ALADIN (Austrian version) 1200 UTC of the previous day (w.r.t. the one analysed): night;

So far verification has been carried out on grid point basis. An "R" tool was coded that enables verification of continuous and categorical precipitation. First of all, a general view on the forecast and observation data is provided via histograms, scatter plots and box plots, which all show the potential of a forecasting system to reproduce the span of possible atmospheric states. For continuous variable, ME, MAE, MSE and BIAS are calculated (see Figures 42-43). To compute categorical skill scores, several thresholds are used: 0, 0.1, 1, 5, 10, 30, 50, 100 mm for night (18–06), day 1 and day 2 (06–06). The computed skill scores are the percent correct (or percentage of forecast correct), HSS, PSS and GS. Within each category there is also the possibility to plot POD, POFD, frequency bias, and threat score (i.e., the critical success index – CSI). An example is reported in Figure 43. The tool can be quite easily extended to other stations, different categories and more scores.



(a) **(b)**

Figure 41. Example of diagnostic spatial verification by means of histograms, scatter plots and box plots for the Innsbruck location: (a) forecaster; (b) ECMWF 1200 UTC forecast.

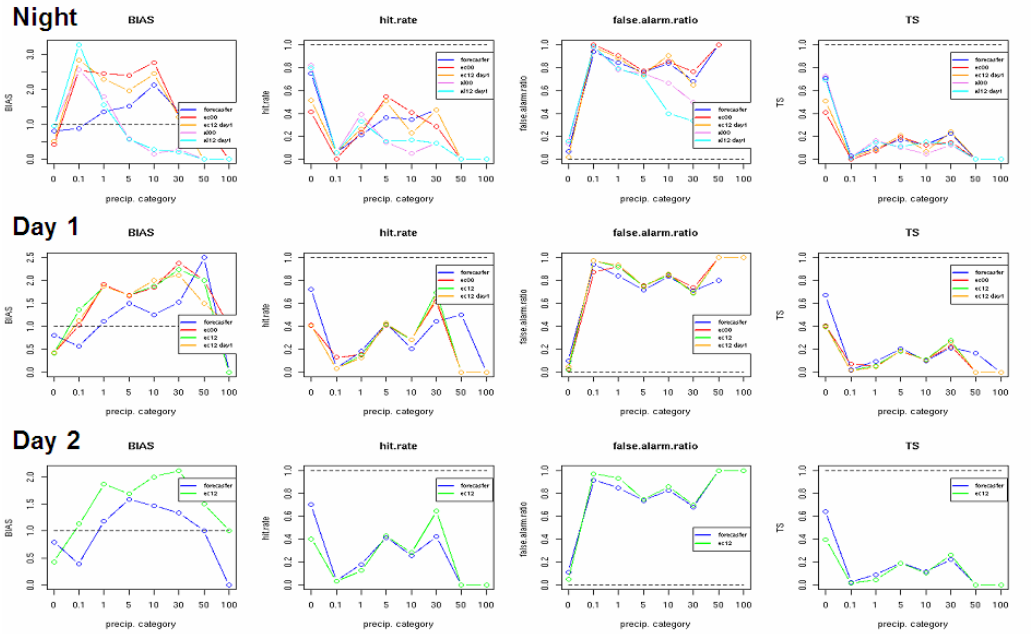


Figure 42. Example of continuous summary measures (a) and categorical scores (b) applied to forecaster, ECMWF and ALADIN forecasts.

5 Reference list and relevant bibliography

5.1 *Articles, books and technical reports*

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2005: Verification of precipitation forecasts from two limited area models over Italy and comparison with ECMWF forecasts using a resampling technique. *Wea. Forecasting*, **20**, 276–300.

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003a: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932.

Accadia, C., M. Casaioli, S. Mariani, A. Lavagnini, A. Speranza, A. De Venere, R. Inghilesi, R. Ferretti, T. Paolucci, D. Cesari, P. Patruno, G. Boni, S. Bovo, and R. Cremonini, 2003b: Application of a statistical methodology for limited area model intercomparison using a Bootstrap Technique. *Nuovo Cimento C*, **26**, 61–77.

Barnes, S. L., 1973: Mesoscale objective analysis using weighted time-series observations. NOAA Tech. Memo. ERL NSSL-62, National Severe Storms Laboratory, Norman, OK 73069, 60 pp. [NTIS COM-73-10781.]

Barnes, S. L., 1964: A technique for maximizing details in numerical weather map analysis. *J. Appl. Meteor.*, **3**, 396–409.

Brier, G. W., and R. A. Allen, 1951: *Verification of weather forecasts.* In Compendium of Meteorology, Malone, T. F., Ed., American Meteorological Society, Boston, pp. 841–848.

Casaioli, M., S. Mariani, C. Accadia, N. Tartaglione, A. Speranza, A. Lavagnini, and M. Bolliger, 2006: Unsatisfying forecast of a Mediterranean cyclone: a verification study employing state-of-the-art techniques. *Adv. Geosci.*, **7**, 379–386.

Casaioli, M., C. Accadia, S. Mariani, M. Bolliger, A. Lavagnini, and A. Speranza, 2004: Evolution of an autumnal mesoscale Mediterranean cyclone: A diagnostic study of a LAM precipitation forecast error, *Proc. of 1st Voltaire Workshop*, Barcelona, Spain, Voltaire Project (FP5), 75–83.

Cherubini, T., A. Ghelli, and F. Lalaurette, 2002: Verification of precipitation forecasts over the Alpine region using a high-density observing network. *Wea. Forecasting*, **17**, 238–249.

Chèruy, F., A. Speranza, A. Sutera, and N. Tartaglione, 2004: Surface winds in the Euro-Mediterranean area: The real resolution of numerical grids. *Annales Geophysicae*, **22**, 4043–4048.

Colle, B. A., and C. F. Mass, 2000: The 5–9 February 1996 flooding event over the Pacific Northwest: sensitivity studies and evaluation of the MM5 precipitation forecasts. *Mon. Wea, Rev.*, **128**, 593–618.

Colle, B. A., C. F. Mass, and K. J. Westrick, 2000: MM5 precipitation verification over the Pacific Northwest during the 1997–1999 cool seasons. *Wea. Forecasting*, **15**, 730–744.

Cressie, N. A. C., 1993: Statistics for spatial data. Revised Edition. Wiley series in probability and mathematical statistics, John Wiley & Sons, New York, 900 pp.

Deidda, R., 2000: Rainfall downscaling in a space-time multifractal framework. *Water Resour. Res.*, **36**, 1779–1794.

Diaconis, P., and B. Efron, 1983: Computer-intensive methods in statistics. *Sci. Amer.*, **248**, 116–130.

Doswell III, C. A., 1996: Verification of forecasts of convection: Uses, abuses, and requirements. *Proc. of the 5th Australian Sever Thunderstorm Conference*, Avoca Beach, New South Wales, Australia.

Ebert, E. E., 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteorol. Appls.*, in press.

Ebert, E. E., and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.*, **239**, 179–202.

Ebert, E. E., U. Damrath, W. Werner, and M. E. Baldwin, 2003b: Supplement to the WGNE assessment of short-term quantitative precipitation forecasts. *Bull. Amer. Meteor. Soc.*, 84, ES10–ES11.

Ebert, E. E., U. Damrath, W. Werner, and M. E. Baldwin, 2003a: The WGNE assessment of short-term quantitative precipitation forecasts. *Bull. Amer. Meteor. Soc.*, **84**, 481–492.

Fisher, R. A., 1925: Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, 239 pp.

Gallus, W. A., 2002: Impact of verification grid-box size on warm-season QPF skill measures. *Wea. Forecasting*, **17**, 1296–1302.

Giaiotti B. D., Steinacker R., and Stel F., 2007: Atmospheric convection; research and operational forecasting aspects. Springer, New York, 222 pp.

Göber, M., E. Zsótér, and D. S. Richardson, 2008: Could a perfect model ever satisfy the forecaster? On grid box mean versus point verification. *Meteorol. Appls.*, in press.

Goody, R., J. Anderson, and G. North, 1998: Testing climate models: An approach. *Bull. Amer. Meteor. Soc.*, **9**, 2541–2549.

Grams, J. S., W. A. Gallus, S. E. Koch, L. S. Wharton, A. Loughe, and E. E. Ebert, 2006: The use of a modified Ebert-McBride technique to evaluate mesoscale model QPF as a function of convective system morphology during IHOP 2002. *Wea. Forecasting*, **21**, 288–306.

Haltiner, G. J. and R. T. Williams, 1980: Numerical Prediction and Dynamic Meteorology. John Wiley & Sons, New York, 477 pp.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167.

Hanssen, A. W., and W. J. A. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *Meded. Verh.*, **81**, 2–15.

Harris, D., E. Foufoul-Georgiou, K. K. Droegemeier, and J. J. Levit, 2001: Multiscale statistical proprieties of a high-resolution precipitation forecast. *J. Hydrometeor.*, **2**, 406–418.

Jolliffe, I. T., and D. B. Stephenson (Eds.), 2003: Forecast verification: A practitioner's guide in atmospheric sciences. John Wiley & Sons, New York, 254 pp.

Kluepfel, C., 2004: Equitable Skill Score Use in the U.S National Weather Service. International Verification Workshop, Montreal, Quebec, Canada, Sep. 15–17, 2004.

Koch, S. E., M. desJardins, and P. J. Kocin, 1983: An interactive Barnes objective map analysis scheme for use with satellite and conventional data. *J. Climate Appl. Meteor.*, **22**, 1487–1503.

Lanciani, A., S. Mariani, M. Casaioli, C. Accadia, and N. Tartaglione, 2008: A multiscale approach for precipitation verification applied to the FORALPS case studies. *Adv. Geosci.*, in press.

Lussana, C., 2006: Oral communication, *6th EMS Conference*, Ljubljana, Slovenia.

Mariani, S., M. Casaioli, C. Accadia, A. Lanciani, and N. Tartaglione, 2008: Verification and intercomparison of precipitation fields modelled by LAMs in the alpine area: Two FORALPS case studies. *Meteorol. Atmos. Phys.*, under revision.

Mariani, S., M. Casaioli, C. Accadia, M. C. Llasat, F. Pasi, S. Davolio, M. Elementi, G. Ficca, and R. Romero, 2005: A limited area model intercomparison on the "Montserrat-2000" flash-flood event using statistical and deterministic methods. *Nat. Hazards Earth Syst. Sci.*, **5**, 565–581.

Mason, I., 1989: Dependence of the Critical Success Index on sample climate and threshold probability. *Aust. Meteor. Mag.*, **37**, 75–81.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430.

McBride, J. L., and E. E. Ebert, 2000: Verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia. *Wea. Forecasting*, **15**, 103–121.

Mesinger, F., 1996: Improvements in quantitative precipitation forecasting with the Eta regional Model at the National Centers for Environmental Prediction: The 48-km upgrade. *Bull. Amer. Meteor. Soc.*, **77**, 2637–2649.

Meteorological Office, 1993: *Forecasters' Reference Book*. UK Met Office, Bracknell, second edition.

Morrissey, M. L., 1991: Using sparse raingages to test satellite-based rainfall algorithms. *J. Geophys. Res.*, **96**, 18,561–18,571.

Murphy, A. H, 1991: Forecast Verification: Its Complexity and Dimensionality. *Mon. Wea. Rev.*, **119**, 1590–1601.

Murphy, A. H, 1988: Skill scores based on mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424.

Murphy, A. H, 1971: A Note on the Ranked Probability Score. *J. Appl. Met.*, **10**, 155–156.

Murphy, A. H, and R. Winkler, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting*, **7**, 435–455.

Murphy, A. H, and R. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330–1338.

Murphy, A. H, B. Brown, and Y.-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Wea. Forecasting*, **4**, 485–501.

Panofsky, H. A., and G. W. Brier, 1958: Some applications of statistics to meteorology. University Park, Pennsylvania State University, 224 pp.

Pettersen, S., 1956: Weather analyses and Forecasting. Mc Graw Hill, New York, 46 pp.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, 1992: Numerical Recipes in FORTRAN: The Art of Scientific Computing. Cambridge University Press, Cambridge, 1002 pp.

Romero, R., and C. A. Doswell III, 2000: Mesoscale numerical study of two cases of long-lived quasi-stationary convective systems over Eastern Spain. *Mon. Wea. Rev.*, **128**, 3731–3751.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575.

Scharlau, K., 1950: Einführung eines Schwülemasstabes und Abgrenzung von Schwülezonen durch Isohygrothermen, *Erdkunde*, **4**, 188–201.

Schönwiese, C.-D., 1985: Praktische Statistik für Meteorologen und Geowissenschafter. Stuttgart, Berlin, 231 pp.

Skelly, W. C. and A. Henderson-Sellers, 1996: Grid box or grid point: What type of data do GCMs deliver to climate impacts researchers? *Int. J. Climatol.*, **16**, 1079–1086.

Smith, R. B., 1979: The Influence of Mountains on the Atmosphere. *Advances in Geophysics*, **21**, 169–193.

Speranza, A., C. Accadia, S. Mariani, M. Casaioli, N. Tartaglione, G. Monacelli, P. M. Ruti, and A. Lavagnini, 2007: SIMM: An integrated forecasting system for the Mediterranean area. *Meteorol. Appl.*, **14**, 337–350.

Speranza, A., C. Accadia, M. Casaioli, S. Mariani, G. Monacelli, R. Inglesi, N. Tartaglione, P. M. Ruti, A. Carillo, A. Bargagli, G. Pisacane, F. Valentinotti, and A. Lavagnini, 2004: POSEIDON: An integrated system for analysis and forecast of hydrological, meteorological and surface marine fields in the Mediterranean area. *Nuovo Cimento C*, **27**, 329–345.

Steinacker, R., 1983: Diagnose und Prognose der Schneefallgrenze. *Wetter und Leben*, **35**, 81–91.

Stephenson, D. B., 2000: Use of the ‘‘odds ratio’’ for diagnosing forecast skill. *Wea. Forecasting*, **15**, 221–232.

Tartaglione, N., S. Mariani, C. Accadia, A. Speranza, and M. Casaioli, 2005: Comparison of raingauge observations with modeled precipitation over Cyprus using contiguous rain area analysis. *Atmos. Chem. Phys.*, **5**, 2147–2154.

Weldon, R., and S. J. Holmes, 1991: Water vapor imagery: Interpretation and application to weather analysis and forecasting. NOAA Tech. Rep., 213 pp.

Wexler, R. R., J. Reed, and J. Honig, 1954: Atmospheric cooling by melting snow. *Bull. Amer. Meteor. Soc.*, **35**, 48–51.

Wilks, D. S., 1995: Statistical Methods in the Atmospheric Science. Academic Press, San Diego, 467 pp.

Xie, P., and P. A. Arkin, 1995: An intercomparison of gauge observations and satellite estimates of monthly precipitation. *J. Appl. Meteor.*, **34**, 1143–1160.

Zepeda-Arce, J., E. Foufoula-Georgiou, and K. K. Droegemeier, 2000: Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *J. Geophys. Res.*, **105**, 10,129–10,146.

5.2 *Web sites*

Baldwin, M. E.: Quantitative Precipitation Forecast Verification Documentation.

http://wwwt.emc.ncep.noaa.gov/mmb/ylin/pcpverif/scores/docs/mbdoc/pptmethod.html.

Göber, M., and C. Wilson: Why, when a model resolution is improved, do the forecasts often verify worse?

http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/Goeber_Wilson/Double_penalty_combined.html

MetWatch: Documentation.

http://www.metwatch.de/doc/index.htm.

Weygandt, S. S., A. F. Loughe, S. G. Benjamin, and J. L. Mahoney, 2004: Scale sensitivities in model precipitation skill scores during IHOP.

http://ruc.fsl.noaa.gov/pdf/AMS-Avx-SLS-Oct2004/Oct04-Weygandt-precip-scale.pdf.

Wilson, C.: Review of current methods and tools for verification of numerical forecasts of precipitation. COST717 Working Group Report on Approaches to verification.

http://www.smhi.se/cost717/doc/WDF_02_200109_1.pdf.

WWRP/WGNE Joint Working Group on Verification, cited 2008: Forecast Verification – Issue, Methods and FAQ.

http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html.

WWRP/WGNE Joint Working Group on Verification, 2004: Recommendations for the verification and intercomparison of QPFs from operational NWP models.

http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/WGNE/QPF_verif_recomm.pdf.

NOTES

NOTES

NOTES

Foralps Partnership

UniTN (Lead partner)	University of Trento Department of Civil and Environmental Engineering	www.ing.unitn.it
APAT	Italian Agency for Environmental Protection and Technical Services	www.apat.gov.it
ARPALombardia	Regional Agency for Environmental Protection Lombardia Meteorological Service-Hydrographic Office	www.arpalombardia.it
ARPAV	Regional Agency for Environmental Protection Veneto	www.arpa.veneto.it
EARS	Environmental Agency of the Republic of Slovenia	www.arso.gov.si
OSMER	Regional Agency for Environmental Protection Friuli-Venezia Giulia Regional Meteorological Observatory	www.osmer.fvg.it
PAB	Autonomous Province of Bolzano Hydrographic Office	www.provincia.bz.it
PAT	Autonomous Province of Trento Office for Forecasts and Organization	www.meteotrentino.it
RAVA	Valle d'Aosta Autonomous Region Meteorological Office	www.regione.vda.it
ZAMG-I	Central Institute for Meteorology and Geodynamics: Regional Office for Tirol and Vorarlberg	
ZAMG-K	Central Institute for Meteorology and Geodynamics: Regional Office for Carinthia	
ZAMG-S	Central Institute for Meteorology and Geodynamics: Regional Office for Salzburg and Oberösterreich	
ZAMG-W	Central Institute for Meteorology and Geodynamics: Regional Office for Wien, Niederösterreich and Burgenland	www.zamg.ac.at

The project FORALPS pursued improvements in the knowledge of weather and climate processes in the Alps, required for a more sustainable management of their water resources. The FORALPS Technical Reports series presents the original achievements of the project, and provides an accessible introduction to selected topics in hydro-meteorological monitoring and analysis.

www.foralps.net

UNIVERSITÀ DEGLI STUDI DI TRENTO
Dipartimento di Ingegneria Civile e Ambientale

ISBN 978–88–8443–234–6