



On the forecast skill of a convection-permitting ensemble

Theresa Schellander-Gorgas, Yong Wang, Florian Meier, Florian Weidle, Christoph Wittmann, and Alexander Kann

Department of forecasting models, Central Institute for Meteorology and Geodynamics, 1190 Vienna, Austria

Correspondence to: Yong Wang (yong.wang@zamg.ac.at)

Received: 15 July 2016 – Published in Geosci. Model Dev. Discuss.: 17 August 2016

Revised: 3 November 2016 – Accepted: 22 November 2016 – Published: 3 January 2017

Abstract. The 2.5 km convection-permitting (CP) ensemble AROME-EPS (Applications of Research to Operations at Mesoscale – Ensemble Prediction System) is evaluated by comparison with the regional 11 km ensemble ALADIN-LAEF (Aire Limitée Adaption dynamique Développement InterNational – Limited Area Ensemble Forecasting) to show whether a benefit is provided by a CP EPS. The evaluation focuses on the abilities of the ensembles to quantitatively predict precipitation during a 3-month convective summer period over areas consisting of mountains and lowlands. The statistical verification uses surface observations and 1 km × 1 km precipitation analyses, and the verification scores involve state-of-the-art statistical measures for deterministic and probabilistic forecasts as well as novel spatial verification methods. The results show that the convection-permitting ensemble with higher-resolution AROME-EPS outperforms its mesoscale counterpart ALADIN-LAEF for precipitation forecasts. The positive impact is larger for the mountainous areas than for the lowlands. In particular, the diurnal precipitation cycle is improved in AROME-EPS, which leads to a significant improvement of scores at the concerned times of day (up to approximately one-third of the scored verification measure). Moreover, there are advantages for higher precipitation thresholds at small spatial scales, which are due to the improved simulation of the spatial structure of precipitation.

1 Introduction

The prediction of deep convection in mountainous terrain is known to be one of the greatest challenges in atmospheric modeling. The initiation and development of deep convection is dependent on small-scale orographic structures and related processes, which cannot be easily described by at-

mospheric models (Wulfmeyer et al., 2011; Barthlott et al., 2011; Weckwerth et al., 2014). Nevertheless, the estimation of the location, duration, and intensity of precipitation events is important, as Alpine areas are more exposed to natural hazards connected with heavy precipitation (landslides and flooding) than flat land (e.g., Rotach et al., 2009; Haiden et al., 2014).

Models with deep convection parameterization perform poorly in simulating heavy and highly localized precipitation, especially those with a grid spacing larger than 10 km (Weusthoff et al., 2010). One source of errors is that the applied convection schemes act independently in individual model grid columns. As a consequence, convectively generated cold pools that drive convective system propagation cannot be properly simulated, resulting in simulated system movement that is too slow. In weak synoptic forcing, for example, organized mesoscale convection systems (MCSs) are particularly challenging for convection-parameterizing models (Clark et al., 2007; Liu et al., 2006). Another drawback is that the inadequate descriptions of buoyancy and updrafts in a convection-parameterizing model often cause convection to initiate too early. This premature initiation of convection often results in timing and location errors as well as difficulty to simulate the diurnal cycle of rainfall (Clark et al., 2007). Detailed discussion on the convection initiation in a convection-parameterizing model can be found in Davis et al. (2003) and Bukovsky et al. (2006).

A solution for this kind of forecasting problem is offered by a new generation of numerical weather prediction (NWP) models, which have been developed during the last decade. Convection-permitting models with horizontal grid spacings of approximately 2–3 km offer new possibilities for estimating local impacts. The term “convection permitting” as used in this article (CP hereafter) means that a deep convection parameterization is not used in the model. It is assumed that

the horizontal resolution around 2–3 km is sufficient to depict the bulk properties of precipitating convective cells, but not to truly resolve the processes within precipitating convective cells such as turbulence and entrainment (Bryan et al., 2003). This is in accordance with Weisman et al. (1997), who suggested setting the upper limit for the range of convection-permitting resolutions at 4 km.

Despite the higher resolution and explicit simulation of deep convection, the exact prediction of location, intensity, and spatiotemporal extent of deep convection is still difficult. Recently, probabilistic approaches using convection-permitting ensembles have proven valuable, since they provide direct information on forecast uncertainty, which is often quite large for deep convection. An ensemble usually consists of a number of model runs, which differ in their initial and boundary conditions and/or model configurations. In order to produce a reliable probabilistic forecast, the individual ensemble member forecasts should be equally likely to occur and cover the range of future states. Following Clark et al. (2011), the ideal number of ensemble members is dependent on the point of diminishing returns, i.e., the ensemble size where no new information can be expected by additional members.

In recent years, several CP ensemble prediction systems (EPSs) have been developed and considerable experience has already been gained. To name but a few, there are the COSMO-DE-EPS (Consortium for Small-scale Modeling – EPS, Gebhardt et al., 2011; Peralta et al., 2012; Ben Bouallègue et al., 2013; Kühnlein et al., 2014) at the Deutscher Wetterdienst (DWD), the CP version of UK Met Office’s MORGES (Met Office Global and Regional Ensemble Prediction System, Bowler et al., 2008; Caron, 2013; Hanley et al., 2013; Tennant, 2015), a storm-scale ensemble forecast (SSEF) run by the Center of Analysis and Prediction of Storms (CAPS) at the University of Oklahoma (Xue et al., 2007, 2009; Clark et al., 2011; Schumacher et al., 2013; Schumacher and Clark, 2014), WRF-based CP ensemble at NCAR (e.g., Schwartz et al., 2015), and AROME-EPS (e.g., Vié et al., 2012; Bouttier et al., 2012) developed at Météo France. A common feature of all of these EPSs is that their horizontal mesh size is equal to or less than 4 km, but mostly between 2 and 3 km.

The EPSs mentioned above differ regarding their number of ensemble members and their perturbation strategies and post-processing. Some of them apply an ensemble data assimilation (EDA) approach for perturbing the initial conditions (ICs) (Vié et al., 2012; Caron, 2013; Schumacher and Clark, 2014; Schwartz et al., 2015). The applied model perturbation methods range from a multiparameter approach (Gebhardt et al., 2011) to a stochastic physics scheme (Bouttier et al., 2012; Romine et al., 2014) and to using different dynamical cores (Schumacher et al., 2013). In order to increase ensemble size and to improve the representation of the ensemble distribution, some systems also apply the neighborhood method and/or lagged ensemble concepts (Ben Boual-

lègue et al., 2013). While the neighborhood method is based on ensemble probabilities derived from grid points of a defined environment (Theis et al., 2005; Schwartz et al., 2010), the lagged ensemble approach uses forecasts of successive ensemble runs (Ben Bouallègue et al., 2013).

A number of evaluative studies concerned with these CP EPSs have been conducted. They mainly focus on the investigation of the impact of CP ensemble configurations, for example, the generation of IC perturbation, representation of the model error, uncertainties from the lateral boundary conditions (LBCs), ensemble size, and spatial scale (Kong et al., 2006; Clark et al., 2009, 2011; Vié et al., 2012; Bouttier et al., 2012; Ben Bouallègue et al., 2013; Kühnlein et al., 2014; Schwartz et al., 2015; Schumacher and Clark, 2014; Romine et al., 2014; Tennant, 2015). There are few comprehensive studies on the evaluation of CP EPS, in particular, in comparison with the mesoscale regional EPS. Clark et al. (2009) compared a 5-member 4 km grid spacing convection-permitting ensemble with a 15-member 20 km grid spacing regional ensemble. Their case studies revealed that the convection-permitting ensemble generally provided more accurate precipitation forecasts than the coarser-resolution regional EPS. Le Duc et al. (2013) examined the ability to predict precipitation of two 11-member ensembles with 10 and 2 km horizontal resolution, with the fine model using direct downscaling of the coarser one. They could show that the 10 km ensemble was more reliable in predicting light rain, whereas the 2 km ensemble outperformed the coarser one in cases of heavier rain. Schwartz et al. (2009) combined subjective and objective verification approaches and found that a higher-resolution ensemble with 4 km produced better forecasts than a 12 km regional model. However, additional comparisons of control runs with 2 and 4 km resolution did not reveal further prognostic value for the lower-resolution model.

In this paper, we will evaluate the performance of a 16-member 2.5 km grid spacing convection-permitting EPS by comparing it with its driving 16-member and 11 km grid spacing mesoscale regional ensemble. Focus will be on the capabilities of the CP ensemble to quantitatively predict precipitation during a convective summer period over an area consisting of mountains and lowlands. Of interest here is the Alpine region, since the impacts of the mountainous terrain, such as windward/lee effects, the differential heating of valley, and mountain slopes can cause large inaccuracies in forecasting convective precipitation and pose a challenge for numerical models and their physical parameterizations (Richard et al., 2007; Wulfmeyer et al., 2008, 2011; Bauer et al., 2011). Therefore, an evaluation study is designed and conducted for a typical convective season (3 months, May–August 2011), i.e., a period, which is long enough to make at least basic statements about the significance of results. Naturally, this period length is not sufficient to enable statistically reliable statements on real hazardous events, such as landslides and flash floods. However, the investigations can be

regarded as a first step towards this aim. The CP ensemble, which is evaluated in this paper, is a version of AROME-EPS, developed at the Central Institute for Meteorology and Geodynamics in Austria (ZAMG). It is compared with its coarser driving regional EPS ALADIN-LAEF (Wang et al., 2011). The following questions are raised:

- Can a convection-permitting EPS provide an advantage over its coarser, driving regional EPS in complex terrain?
- Is there any difference in the performance for the compared EPSs between lowlands and mountainous areas?
- How well can CP EPS and lower-resolution regional EPS simulate the diurnal cycle of precipitation? Is the onset and development of convective precipitation realistic?
- Does a significant difference in performance for different weather regimes (i.e., days with weak and strong synoptic forcing) exist?

A verification study is designed and conducted to answer these questions and to establish whether AROME-EPS can outperform ALADIN-LAEF, a regional mesoscale ensemble with deep convection parameterization on a coarser grid. Wang et al. (2012) demonstrated the added value of ALADIN-LAEF as a regional mesoscale EPS to the global ECMWF-EPS (European Centre for Medium-Range Weather Forecasts). Hence, the present study extends this research by addressing the step between regional mesoscale and CP ensembles.

For the present paper, AROME-EPS is coupled to the 16 perturbed ALADIN-LAEF members. This is done to take advantage of the simulation of uncertainties used in ALADIN-LAEF. This uncertainty information is subsequently transferred to finer scales via the dynamical downscaling of the ALADIN-LAEF forecasts by AROME. This means that both IC perturbations and LBC perturbations are provided from the driving model and are thus consistent. No further IC perturbations and model perturbations are applied. Generally, the setup is kept as simple as possible to point out the pure effects of the downscaling: AROME-EPS is directly coupled to a daily ALADIN-LAEF run initiated at 00:00 UTC. There is no time lag between the ALADIN-LAEF and the AROME-EPS simulations, and the forecasts are evaluated for the first 30 h of the model runs, hence for a whole day and the subsequent night each.

The benefits of AROME-EPS compared to ALADIN-LAEF are revealed in the framework of a comparative verification study. Although the focus of the verification study is on the onset and development of precipitation, the performance of other surface weather parameters is considered. The verification methods are selected in such a way that the overall performance, in a deterministic and probabilistic

manner, and the abilities of the ensembles to reproduce spatial structures, can be investigated. Hence, ensemble-related scores are combined with spatial verification methods. Unintentionally, the strategy of this paper shows parallels to the verification study conducted by Le Duc et al. (2013), especially concerning the two ensembles (10 and 2 km resolution) coupled by direct downscaling. Further similarities are the complex terrain in which the study is conducted (Japan) and the use of traditional and advanced verification metrics. As a consequence, parallels in the results are mentioned in the results section.

Detailed characteristics of the compared models are described in Sect. 2 along with the verification data. The methods chosen for the evaluation of the two ensembles are described in Sect. 3. Section 4 comprises the verification results and Sect. 5 the summary and concluding remarks.

2 Ensemble systems and data

2.1 The regional ensemble ALADIN-LAEF

ALADIN-LAEF is the operational regional ensemble system of ZAMG and runs at ECMWF (Wang et al., 2010, 2011). It is based on the hydrostatic spectral limited area model ALADIN (Wang et al., 2009). ALADIN-LAEF has 16 members and is coupled to ECMWF-EPS (Weidle et al., 2013) with a horizontal grid spacing of 11 km. In operational mode, it runs two times per day at 00:00 and 12:00 UTC and provides probabilistic forecasts on a forecast range up to 3 days ahead, i.e., 72 h. In this study, however, evaluation is confined to the run at 00:00 UTC and a forecast range of 30 h ahead only. This is done in order to investigate the onset and development of convection in its diurnal cycle.

The 16 members of ALADIN-LAEF are not sufficient to represent the atmospheric state probability density function (PDF). However, Schwartz et al. (2014) have shown that similar verification scores can be obtained from a 50-member ensemble and subsets of 20–30 members. Hence, we can expect, at least, reasonable results from verification based on a 16-member ensemble.

The ALADIN-LAEF domain (Fig. 1) covers the whole European continent, Iceland, the whole Mediterranean Sea, Black Sea, Caspian Sea, and adjacent countries. The eastern margins reach the Ural Mountains and parts of Siberia. To deal with the atmospheric initial condition perturbation, ALADIN-LAEF applies a breeding–blending method for generating the IC perturbations for the upper levels. It uses large-scale perturbations from the driving global-ECMWF-EPS combined with small-scale perturbations from the ALADIN-breeding vectors (Toth and Kalnay, 1993). The blending method (Wang et al., 2014) ensures that inconsistencies between small- and large-scale perturbations are avoided. Therefore, a digital filter is applied on the low spectral truncations of both the breeding vectors and the fields

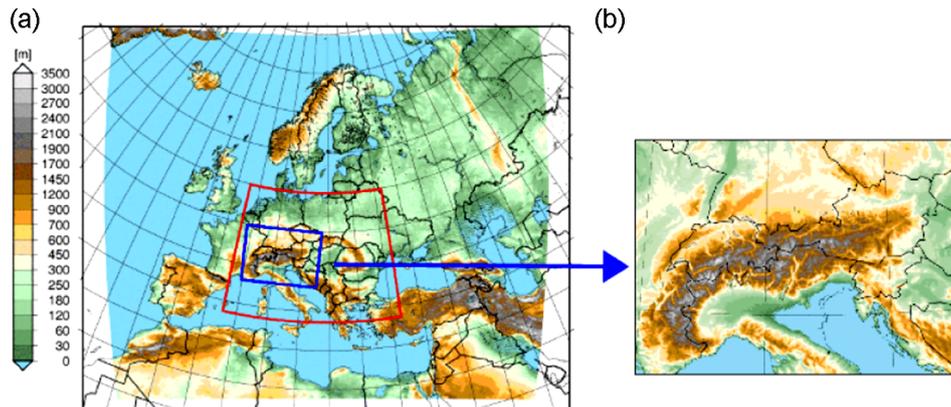


Figure 1. Geographic domains and topographies of (a) ALADIN-LAEF, where the red frame is the output domain used for the present study, and (b) AROME-EPS, which is shown by the blue frame in (a).

from the global model. Afterwards, the filtered breeding vectors on the full spectral resolution are subtracted from the original ones and added by the filtered global fields resulting in initial perturbations that are consistent with the regional EPS itself as well as with the driving global EPS.

To consider uncertainties arising from the initial surface conditions in ALADIN-LAEF, a surface data assimilation scheme based on optimum interpolation (CANARI – Code for the Analysis Necessary for Arpège for its Rejects and its Initialization, Taillefer, 2002) is implemented using randomly perturbed observations. To account for uncertainties in the model itself, a multi-physics approach is implemented in ALADIN-LAEF. The perturbed members use different model configurations with several combinations and tunings of schemes and parameterizations available in the ALADIN physics package. The main emphasis is put on the variation and tunings of the following schemes and parameterizations: the diagnostic convection scheme as described in Bougeault (1985); the prognostic deep convection scheme 3MT (modular multiscale microphysics and transport scheme; Gerard et al., 2009), and the connected microphysics scheme described in Geleyn et al. (2008) and Gerard et al. (2009); the radiation scheme based on Ritter and Geleyn (1992) or alternatively the scheme described in Mlawer (1997) and Morcrette (1991); the pseudo-prognostic TKE (turbulent kinetic energy) scheme described in Vana et al. (2008). Further details can be found in Wang et al. (2010). Authors are aware that the forecasts of the individual members produced by the multi-physics approach cannot be regarded as equally likely. However, a previous evaluation (apart from this study) of the multi-physics in ALADIN-LAEF revealed that some of the members showed larger biases and errors than the other members. The configurations of these worse members were changed accordingly. Hence, we can assume that the members now produce forecasts of comparable quality.

2.2 The convection-permitting ensemble AROME-EPS

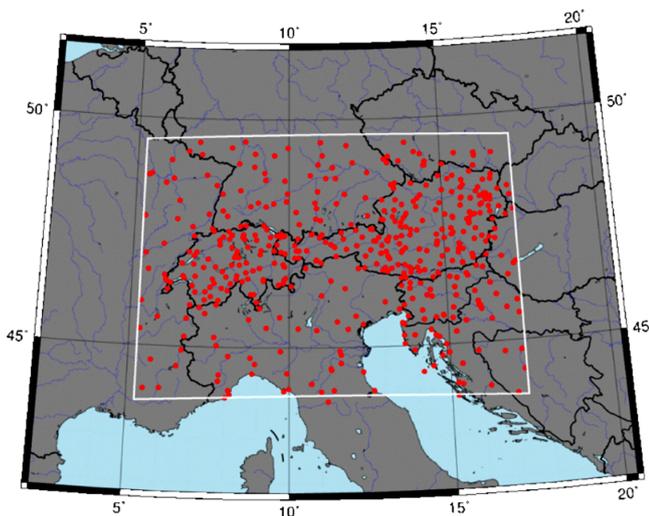
The model core of AROME-EPS is the non-hydrostatic, spectral limited area model AROME (Seity et al., 2011), which is especially designed to run at very high resolutions with a grid spacing of 2.5 km or lower. Deep convection is treated explicitly, while shallow convection is parameterized with a mass flux approach (Pergaud et al., 2009). The single-moment bulk microphysics scheme ICE3 for mixed-phase cloud parameterization (Pinty and Jabouille, 1998) can handle mixing ratios of five prognostic hydrometeor classes: cloud water, cloud ice, rain, snow, and graupel and also simulates complex interactions between them. AROME, by default, uses a three-layer soil model SURFEX (Surface Externalisé) with the effects of sea and urban areas parameterized using a tile approach (Masson, 2000).

At ZAMG, a deterministic version of AROME with 2.5 km grid spacing has been operational since January 2014 running every 3 h up to a lead time of 48 h. The domain for the model integration encompasses the Alpine region (Fig. 1). Table 1 summarizes the most important model characteristics of ALADIN-LAEF and AROME-EPS.

To run AROME-EPS, the same version of AROME with the same resolution is initialized by a dynamical downscaling of ALADIN-LAEF and coupled to the 16 members of ALADIN-LAEF. The ensemble runs with a forecast range of 30 h are initiated at 00:00 UTC each day, i.e., at the same time as ALADIN-LAEF. There is no time lag considered, as the pure impact of enhanced resolution and the convection-permitting configuration shall be investigated. Apart from the perturbations of initial conditions and lateral boundary conditions, no further perturbations (e.g., multi-physics parameterizations as in ALADIN-LAEF) are induced in the model integration. This comparatively simple configuration is used for several reasons: first, AROME-EPS has been set up quite recently at ZAMG and is still at an early stage of development. Secondly, the development of physics perturbations

Table 1. Main characteristics of the ALADIN-LAEF and AROME-EPS.

	ALADIN-LAEF	AROME-EPS
Ensemble size	16+1 members	16 members
Horizontal resolution	11 km	2.5 km
Vertical resolution	45 layers	60 layers
Model time step	450 s	60 s
Coupling model	ECMWF-EPS (first 16 members)	ALADIN-LAEF
Coupling update	6 h	3 h
No. of grid points	206 × 164	432 × 320
Forecast range	72 h	30 h
Runs per day	2 (00:00, 12:00 UTC)	1 (00:00 UTC)

**Figure 2.** Locations of meteorological surface observation stations within the evaluation domain.

in AROME-EPS will rather go towards a stochastic physics scheme or a combined stochastic–multi-physics scheme than towards pure multi-physics as currently used in ALADIN-LAEF. Thirdly, the aim of this study is to test the possible advantage of a CP EPS compared to the operational system of ALADIN-LAEF.

2.3 Verification data

Station observations are used for the evaluation of ALADIN-LAEF and AROME-EPS surface weather variables. Figure 2 shows the 517 surface stations in the AROME domain, providing observations at 6-hourly intervals for 2 m temperature, 2 m humidity, 10 m wind speed, and mean sea level pressure. The upper-level verification is achieved using ECMWF analyses reference data at four pressure levels: 925, 850, 700, and 500 hPa, which are adapted to the model resolutions of both AROME-EPS and ALADIN-LAEF.

The evaluation of precipitation forecasts is performed using the very high-resolution precipitation analyses of the ZAMG nowcasting system INCA (Integrated Nowcasting

through Comprehensive Analyses; Haiden et al., 2011). This is necessary as the average station distance of precipitation observations is too large to resolve the fine spatial structures of precipitation events. The advantage of the INCA analyses is that they use additional observations and are provided on a regular grid. Based on these gridded data, it is possible to apply enhanced verification methods on precipitation fields, which cannot be computed on a point-to-point basis.

The INCA system, developed at ZAMG, operates on a horizontal resolution of 1 km × 1 km. INCA blends data from automatic weather stations, remote sensing data (radar, satellite), forecast fields of numerical weather prediction (NWP) models, and high-resolution topographic data (Haiden et al., 2011). It provides hourly 3-D fields of temperature, humidity, wind, and 2-D fields of cloudiness, precipitation rate, and precipitation type with an update frequency of 15 min to 1 h. The precipitation analyses are provided for different accumulation periods. In the present study, the 1 h accumulated INCA precipitation analyses are used as a reference for the spatial verification of EPS forecasts. For these analyses, precipitation measurements from surface stations and radar data are accumulated to 1 h sums and algorithmically merged. Prior to the analysis procedure, the data are quality controlled and climatologically scaled (Haiden et al., 2011). In this way, the higher quantitative accuracy of the station data and the better spatial coverage of the radar data are utilized. The resulting analysis reproduces the observed values at the station locations while preserving the spatial structure provided by the radar data. The analysis error, which is computed from classical cross-validation, varies from case to case, and depends on precipitation type, e.g., large scale or convective, and on the accumulation period. The magnitude of analysis errors of grid point values can be quite large, but areal mean values are significantly more reliable (Haiden et al., 2011).

Amending the rain gauge–radar combination, the scheme includes elevation effects on precipitation using an intensity-dependent parameterization (Haiden and Pistotnik, 2009). A NWP model first guess is not required in the precipitation analysis; thus, such analyses are ideally suited as an independent reference to validate NWP models.

Forecast verifications are performed at the observation locations for surface variables as 2 m temperature and humidity, 10 m wind speed, and mean sea level pressure, and on the INCA grid for precipitation. The model forecasts are interpolated bi-linearly to the station locations and INCA analysis grid points, respectively. Further, a height correction scheme is applied on 2 m temperature values based on atmospheric standard conditions. In doing so, the same number of forecast–observations pairs is available for the verification of each of the EPS models. This supports the comparability of the verification results.

3 Verification strategy

AROME-EPS and ALADIN-LAEF are evaluated over a 3-month summer period from 15 May–15 August 2011, which represents a typical convective summer season in central Europe.

Precipitation is one of the parameters for which the biggest improvement is expected from the convection-permitting models. Therefore, the evaluation of the ensembles focuses on the representation of the spatiotemporal structure of precipitation events in the forecasts. Nevertheless, the preconditions for the development and onset of precipitation are also considered. For this reason, other forecast parameters such as temperature, humidity, wind speed, air pressure, and geopotential height are also verified.

Precipitation forecasts are evaluated in both deterministic and probabilistic ways. The deterministic approach is directed towards predicting the correct precipitation amounts and the spatial distribution of the data. Probabilistic evaluation tests the capability of the ensembles to predict a predefined event with the probability which corresponds to its relative frequency, i.e., to produce a reliable probability density function (PDF) for the occurrence of the event. The events can be defined as, e.g., precipitation amounts exceeding a certain threshold. In this study, thresholds of 0.1 mm (threshold for the prediction of rain or no rain), 0.5, 1, 2, and 5 mm are chosen for 3-hourly accumulated precipitation amounts. These thresholds appear low, especially when taking into account convective precipitation events. However, the thresholds are selected according to the frequency of occurrence of the precipitation values in the individual grid cells of the 1 km × 1 km verification grid. They ensure that a sufficient number of observed events are available for evaluation over the 3-month test period. The two ways of deterministic and probabilistic evaluation reflect the main options for the efficient use of ensemble forecasts: first, as a conservative prediction of ensemble mean or median and, second, as a tool to estimate the uncertainty of the forecast and the probability of extreme values via the ensemble spread and PDF (e.g., Zhu et al., 2002).

A number of traditional point-to-point verification scores (see, e.g., Wilks, 2006) are computed for all evaluated pa-

rameters. In addition, significance tests for these scores are performed. Confidence intervals of the verification scores are estimated by a bootstrapping algorithm (Davison and Hinkley, 1997; Joliffe, 2007; Ferro, 2007) and confidence intervals of 90 %. The bootstrapping method uses 5000 random samples with a block length of 4 days (Hall et al., 1995).

In order to present the results concisely, three scores have been selected to describe the differences in forecast performance between AROME-EPS and ALADIN-LAEF: the ensemble mean bias (Eq. 1), the Brier score (BS), components derived from its decomposition, reliability, resolution, and uncertainty (Brier, 1950; Murphy, 1973; respectively, Eqs. 2–5), and the continuous ranked probability score (CRPS, Hersbach 2000; Gneiting and Raftery, 2007; Eq. 6).

The bias simply measures the mean deviation between the analyzed values (a) and the forecast values, in our case the ensemble means (f), averaged over n grid points with index i . Both positive as well as negative signs are possible. A perfect forecast has a bias of zero.

$$\text{BIAS} = \frac{1}{n} \sum_{i=1}^n (\bar{f}_i - a_i) \quad (1)$$

Like the bias, BS is also a measure for the accuracy of the forecasts, however, in probability space. It is the mean squared difference between the forecast probability ($p \in [0 : 1]$), e.g., derived from the distribution of ensemble members) for a predefined event (e.g., the exceeding of a threshold) and the analyzed truth x ($x \in [0, 1]$). The binary variable x is 1 if the event occurred, and 0 if the event did not occur. The minimal value of BS is 0. It is achieved for a perfect forecast, and the maximum value is 1 for the worst possible forecast.

$$\text{BS} = \frac{1}{n} \sum_{i=1}^n (p_i - x_i)^2 \quad (2)$$

According to Murphy (1973), the BS can be decomposed to three quantities which refer to the reliability, resolution, and uncertainty of the forecast (Eq. 3).

$$\text{BS} = \frac{1}{n} \sum_{k=1}^K N_k (p_k - \bar{x}_k)^2 - \frac{1}{n} \sum_{k=1}^K N_k (\bar{x}_k - \bar{x})^2 + \bar{x} (1 - \bar{x}) \quad (3)$$

The N_k values in Eq. (3) denote the sample sizes in K conditional subsamples pertaining to forecast probabilities p_k . The \bar{x}_k values (Eq. 4) are the conditional average observations and \bar{x} is overall average observation (Eq. 5).

$$\bar{x}_k = \frac{1}{N_k} \sum_{k \in N_k} x_i \quad (4)$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (5)$$

Reliability (first term of Eq. 3) measures how well a forecast system is calibrated, i.e., it is a measure of accuracy condi-

tional to a range of forecast values. Resolution (second term), on the other hand, describes the ability of the forecast to react differently to different weather situations or, in other words, to resolve them. While the value for a perfect forecast of the reliability term is zero, the resolution term is preferably large. The third term of Eq. (3), uncertainty, is not dependent on the forecast, but only on the variance of observations (here, the relative frequencies of the occurrence/non-occurrence of events). For a very comprehensible discussion of these quantities of forecast quality, see also Wilks (1995).

CRPS is related to BS insofar as it can be expressed as the integral of BS for all possible thresholds of the meteorological parameter ξ (Hersbach, 2000). The value for an ideal forecast of CRPS is zero as for BS.

$$\text{CRPS}_i = \int_{-\infty}^{\infty} [P_i(\xi) - P_i(\xi_a)]^2 d\xi \quad (6)$$

The continuous ranked probability score compares the cumulative distributions $P_i(\xi)$ (Eq. 7) and $P_i(\xi_a)$ (Eq. 8) of the forecast and the analyzed values at each grid point i .

$$P_i(\xi) = \int p_i(y) dy \quad (7)$$

$$P_i(\xi_a) = H(\xi - \xi_a) \quad (8)$$

$H(\xi)$ is the so-called Heaviside function (Eq. 9), which only takes the values 0 and 1.

$$H(\xi) = \begin{cases} 0 & \text{for } \xi < 0 \\ 1 & \text{for } \xi \geq 0 \end{cases} \quad (9)$$

In addition to the traditional statistical scores, precipitation forecasts are verified by spatial verification methods, which not only consider the exact match of forecast and verification values at individual points but also take into account the matching of forecasts and observations in terms of objects or spatial scales (Casati et al., 2008; Ahijevych et al., 2009; Gilleland et al., 2010). This is necessary as precipitation fields exhibit high spatial variability and discontinuity. Small deviations in space and time between forecast and verification data can lead to large errors in traditional point-to-point verification scores, which is also known as the double penalty problem (Nurmi, 2003).

3.1 Spatial verification methods

The selected spatial verification methods are the so-called SAL method (structure–amplitude–location method; Wernli et al., 2008) and the fractions skill score (FSS; Roberts and Lean, 2008).

SAL determines the forecast performance of precipitation in terms of structure (S), amplitude (A), and location (L). The method is object based. Precipitation objects in forecast and verification fields are contiguous areas of grid points ex-

ceeding a certain precipitation threshold.

$$A = \frac{\bar{R}_f - \bar{R}_a}{0.5[\bar{R}_f + \bar{R}_a]} \quad (10)$$

The amplitude score (Eq. 10) defines whether the domain-averaged amount \bar{R} of the precipitation field R is underestimated ($A < 0$) or overestimated ($A > 0$). Subscripts (f and a) denote forecast and analyzed fields, respectively.

The location score measures the agreement of the centers of mass in the analyzed and predicted precipitation fields together with the averaged distance between the center of mass and the individual objects. It is actually the sum of two components ($L = L1 + L2$) where both values are in the range [0, 1]. The first part $L1$,

$$L1 = \frac{|x(R_f) - x(R_a)|}{d_{\max}}, \quad (11)$$

is a measure of the distance between the mass centers x of the analyzed (R_a) and the predicted precipitation fields (R_f). d_{\max} is the longest possible distance in the domain.

As an identical mass center position does not necessarily mean that the forecast is perfect, the second component $L2$ (Eq. 12) is introduced:

$$L2 = 2 \frac{|r(R_f) - r(R_a)|}{d_{\max}}. \quad (12)$$

$L2$ takes into account the distance r (Eq. 13) between the mass center of each individual object R_n and the overall mass center and compares them between the observed and simulated precipitation field:

$$r = \frac{\sum_{n=1}^M R_n |x - x_n|}{\sum_{n=1}^M R_n}. \quad (13)$$

The L component has a range [0, 2], with $L = 0$ indicating a perfect forecast.

The structure score S ,

$$S = \frac{V(R_f) - V(R_a)}{0.5[V(R_f) + V(R_a)]}, \quad (14)$$

compares the weighted sums of the precipitation volumes $V(R)$,

$$V(R) = \frac{\sum_{n=1}^M R_n V_n}{\sum_{n=1}^M R_n}, \quad (15)$$

of the precipitation objects, where $V_n = R_n/R_{\max}$ describes precipitation sums scaled by their maxima. If $S < 0$, forecast

objects are too small and too peaked. In contrast, $S > 0$ indicates that the objects are too large and too flat.

The fractions skill score (FSS),

$$\text{FSS}(n) = 1 - \frac{\text{MSE}(n)}{\text{MSE}(n)_{\text{ref}}}, \quad (16)$$

evaluates the forecasts on different spatial scales. The scales are defined via neighborhoods, i.e., square boxes of length n grid spaces surrounding a selected grid point. The score compares the fractions of rain coverage of forecast and analysis in the neighborhoods. Depending on the precipitation event, small disparities of the coverage may lead to large forecast errors on fine scales, but to a better rating on a coarser scale. The aim of FSS is to identify scales for which the evaluated model can provide useful forecasts.

FSS is computed by assigning the grid points binary values 0 and 1 in each of the neighborhoods with subscripts (i, j) , according to a selected precipitation threshold. From these binary fields, the fraction of the points with value 1 are computed for analyses and forecasts as $A_{(n)i,j}$ and $F_{(n)i,j}$, respectively.

At each such defined scale n , the mean squared error (MSE),

$$\text{MSE}_{(n)} = \frac{1}{N_x N_y} \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} [A_{(n)i,j} - F_{(n)i,j}]^2, \quad (17)$$

is computed for the whole field of fractions and related to a reference (MSE_{ref}).

$$\text{MSE}_{(n)\text{ref}} = \frac{1}{N_x N_y} \left[\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} A_{(n)i,j}^2 + \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} F_{(n)i,j}^2 \right] \quad (18)$$

MSE_{ref} is the largest possible MSE which can be obtained from the underlying field. The skill score summarizes the performance in the whole field and ranges from 0 (complete mismatch) to 1 (perfect match).

3.2 Subdomains for precipitation verification

Verification is done for the whole domain of Austria. To account for the different topographic characteristics in the verification domain, two subdomains are chosen (Fig. 3). They comprise a mountainous area (hereafter region West) as well as a region with flat terrain (hereafter region Northeast). Due to the location of the Alps in Austria and the prevailing flow directions around the Alps, each of the subdomains has its own climatological properties which are also visible in the precipitation characteristics.

3.3 Temporal stratification

In order to investigate the influence of different weather regimes, the 92 days of the test period are classified into three

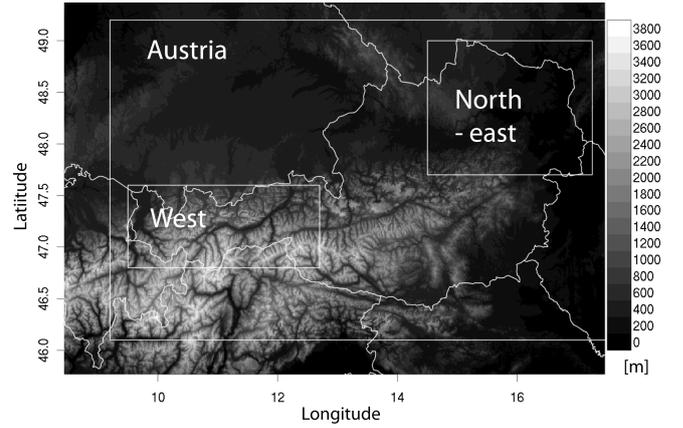


Figure 3. INCA domain and topography with the subdomains which are used for the evaluation.

bins according to the synoptic situation: strong synoptic forcing, weak synoptic forcing, and dry. Days are classified as dry (5 days) if the areal mean of the daily precipitation sum is below 0.05 mm. All other days, i.e., 87 days on which rain was reported, are assigned to the bins of weak (23 days) or strong synoptic forcing (64 days). For the classification, a method described by Done et al. (2006) and successfully applied by Kühnlein et al. (2014) is used, which is based on the temporal variability of CAPE (convective available potential energy) as a measure of atmospheric instability. According to Done et al. (2006), the approach helps to distinguish between days on which convection is predominantly at equilibrium or at non-equilibrium. This means that the destabilization of the atmosphere by large-scale synoptic forcing is balanced or unbalanced, respectively, by the stabilization through convection. The idea is that this balance or imbalance is related to the timescale in which CAPE is built up by large-scale processes and consumed by convection. On days with weak synoptic forcing, the consumption of CAPE is related to the diurnal cycle or to local triggering rather than to prevalent large-scale processes. In these cases, the convective timescale is long and CAPE is often not fully consumed by convection. In situations where CAPE is realized much faster by large-scale processes, i.e., in situations of strong synoptic forcing, convection is in equilibrium. In our study, the convective adjustment timescale t_c ,

$$t_c = \text{CAPE} \frac{d(\text{CAPE})}{dt}, \quad (19)$$

is calculated hourly from AROME-EPS CAPE forecasts using $\Delta t = 1$ h. Following the suggestion of Done et al. (2006), a specific day is assigned to weak synoptic forcing if the areal mean of t_c exceeds a threshold of 6 h at least once a day by at least three ensemble members. In order to test the method of Done et al. (2006), we compared the classification with alternative approaches, such as the temporal change of midtropospheric vorticity and convection related to pat-

terns in 500 hPa geopotential using archived ECMWF forecast and ERA-Interim reanalyses. The results were comparable to those of the equilibrium method.

4 Results

In the following, we present the evaluation of AROME-EPS and ALADIN-LAEF over a 3-month summer period. The focus is on the performance of near-surface parameters, in particular the precipitation forecast, which is of most interest to the users of convection-permitting and regional EPSs.

4.1 Evaluation of forecasts of temperature, wind, and humidity

The forecast performance of surface parameters (2 m temperature and humidity, 10 m wind speed, and mean sea level pressure, MSLP) and upper-level parameters (temperature, humidity, wind speed, and geopotential height) of AROME-EPS and ALADIN-LAEF are verified in this study, which form the background of the evaluation of precipitation.

A large number of verification metrics have been calculated for those near-surface and upper-air parameters. In general, there is no clear advantage either for ALADIN-LAEF or for AROME-EPS. Exceptions from this statement are solely constituted by biases in the forecasts, which are particularly found on the surface level. They form the most eminent differences in the performances of the EPSs: if the bias is low, the models provide good performance also for other scores.

For the surface level, we also found more results on a high level of significance (i.e., 90%). The verification results of the upper levels are less significant than for the surface and performance is more ambivalent. We used a large number of observations for both surface (station observations) and upper levels (ECMWF grid values). Hence, the lower significance of the results for the upper levels can be explained by the model setup rather than by the verification data. Near surface and on lower levels, AROME-EPS can add more information to the model simulation than on upper levels, compared to ALADIN-LAEF. This is due to the SURFEX soil scheme and the interaction between a refined representation of orography and the model physics schemes and dynamics. On the upper levels, however, there is less influence of the orography and the simulation resembles more the driving model. For this reason, surface results have been selected to highlight the main findings in the following.

Figure 4 compares the ensemble mean bias and the continuous ranked probability score (CRPS; see Wilks, 2006 for details) for 2 m relative humidity, 2 m temperature, and 10 m wind speed. CRPS compares the forecast PDF based on all ensemble members to the observed values of occurrence and non-occurrence, respectively. CRPS is sensitive to the difference between the forecast probabilities and observed values. The lower the difference, the better the forecast is rated.

Hence, the value of CRPS of a perfect forecast is zero. Due to the formulation of CRPS, variations of CRPS values are also reflected by many other scores, in particular those which are sensitive to deviations between the distributions of forecasts and observations. Thus, CRPS is useful for representing the results of this study exemplarily. It also shows the impact of biased forecasts.

Biases of 2 m relative humidity in Fig. 4a show noticeable diurnal variations. During the night and early morning, AROME-EPS is too dry, whereas ALADIN-LAEF is too moist during the day (12:00 and 18:00 UTC). The diurnal variations of the differences between AROME-EPS and ALADIN-LAEF are also reflected in CRPS in Fig. 4b. During the night, AROME-EPS and ALADIN-LAEF are at the same level, but for the daytime hours AROME-EPS shows better results. For 2 m relative humidity, most verification results are significant at a level of 90%. This is also true for the differences in forecast performance during the daytime hours. Results for 2 m temperature in Fig. 4c and d show an improvement for bias and CRPS at a significance level of 90% for AROME-EPS. This result is partially due to a large bias of ALADIN-LAEF temperatures. In contrast, there exist fewer deviations between the ensembles for wind speed (Fig. 4e and f) and MSLP (not shown). However, these results have only a low level of significance.

4.2 Evaluation of precipitation forecasts

Precipitation is evaluated by 3-hourly INCA analyses on a regular 1 km × 1 km grid. A first insight into the strengths and weaknesses of the ensembles in forecasting precipitation is offered by a comparison of the daily variability of precipitation intensities. Figure 5 compares the 3-hourly precipitation sums of INCA and both EPS models for different regional domains and for days with strong (left panels) and weak (right panels) synoptic forcing.

Errors occur in terms of over- and underestimation of the maximum intensity and in terms of time shifts. The daily maximum of 3 h precipitation is overestimated by AROME-EPS for regions West and Austria and both types of synoptic forcing by 20–50%. In ALADIN-LAEF, the maximum of the ensemble mean in these regions is approximately at the same level as analyzed by INCA. Hence, the conditions of ALADIN-LAEF that are too moist near the surface in Fig. 4a are not directly reflected in the precipitation sums. For region Northeast, the ensemble mean of AROME-EPS simulates the maximum amount of precipitation quite well for strong synoptic forcing and only slightly overestimates it for weak synoptic forcing, whereas ALADIN-LAEF is too low for both types of forcing.

Considering the days with strong synoptic forcing in Fig. 5 (left panels), the highest precipitation sums are detected around 18:00 UTC. AROME-EPS describes the temporal maximum quite well, whereas the maximum in ALADIN-LAEF occurs too early (−3 h time shift). In the case of weak

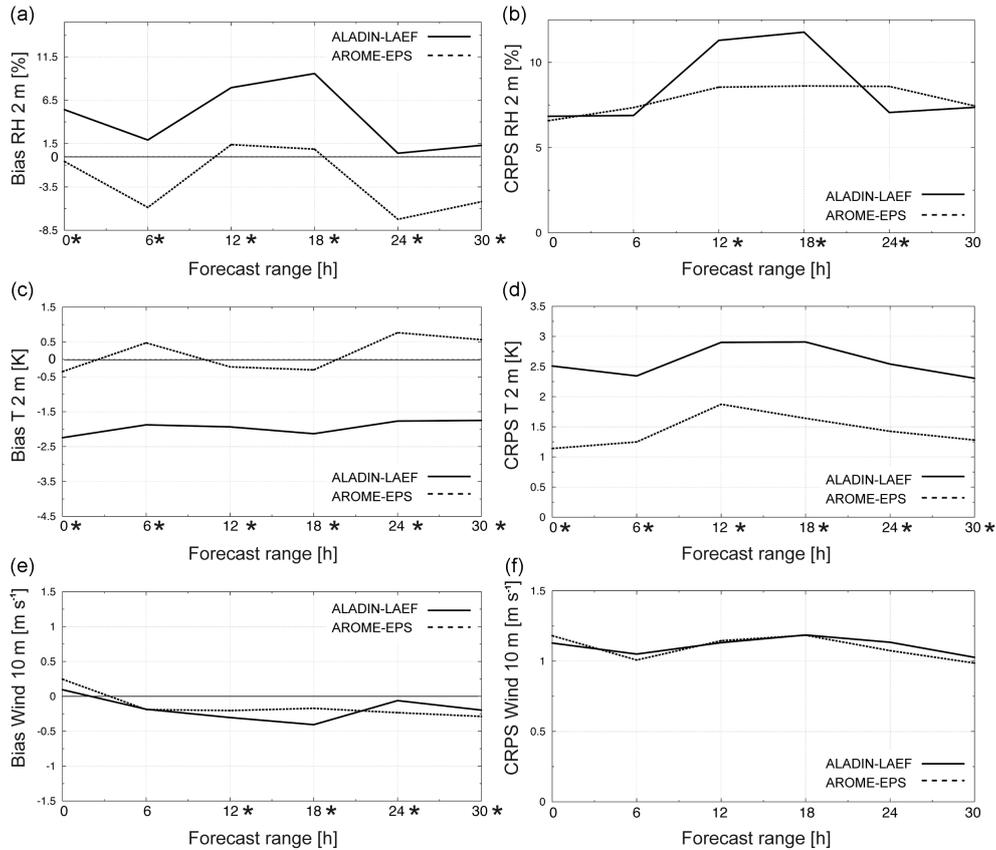


Figure 4. Bias of the ensemble means (left panel) and CRPS (right panel) for 2 m relative humidity (top), 2 m temperature (middle), and 10 m wind speed (bottom) for the period of 15 May–15 August 2011 of AROME-EPS (dotted line) and ALADIN-LAEF (solid line), both verified over the AROME domain. Lead times, which are marked with asterisks (*) indicate results with significant differences between the ensembles.

synoptic forcing shown in Fig. 5 (right panels), the precipitation maxima are observed later than for the other cases in region West (e.g., 21:00 UTC instead of 18:00 UTC). This is not reflected by the EPS models, which reach the maximum intensity of precipitation at 15:00 UTC (ALADIN-LAEF) and 18:00 UTC (AROME-EPS). Only for region Northeast and weak synoptic forcing does the maximum of precipitation occur too late in AROME-EPS. The characteristic that ALADIN-LAEF and AROME-EPS tend to trigger moist and deep convection over complex orography too early is well known (Wittmann et al., 2010). However, according to Fig. 5, running a model or an EPS on CP scales is beneficial for predicting the daily maximum of the convective diurnal cycle, at least over mountainous terrain. With respect to the timing of the maxima, AROME-EPS shows a time shift of -3 h, with ALADIN-LAEF showing a time shift of -6 h for weak synoptic forcing in regions Austria and West (panels b and d), respectively. Because of the limited framework of this study we can only speculate that this behavior might be due to differences caused by the deep convection scheme in ALADIN-LAEF, which is one of the reasons that causes an early onset of precipitation (Bechtold et al., 2013) and, respectively, the

explicit simulation of deep convection in AROME. Another reason, which we cannot exclude, could be that ALADIN-LAEF and AROME apply different physical parameterizations. The different dynamical cores, hydrostatic and non-hydrostatic, might also contribute to the differences to some extent, but remain statistically less significant with respect to precipitation, as shown in an earlier study (Wittmann et al., 2010). Experiences concerning the pure impact of different vertical resolutions on the forecast quality are few. However, it is known that an increase of vertical resolution and, hence, enhanced possibilities to simulate convection-related, micro-physical, and boundary-layer processes, does not necessarily result in an improvement of precipitation forecasts. It is rather related to increased overprediction of precipitation amounts (Aligo et al., 2009).

A further characteristic evident in Fig. 5 is that the precipitation amounts in AROME-EPS develop independently of those in the driving ALADIN-LAEF members, which is indicated by the ensemble spread. In ALADIN-LAEF, the ensemble spread is quite large for certain lead times, ranging from a larger overestimation of the observed precipitation amounts to a large underestimation. This contrasts with

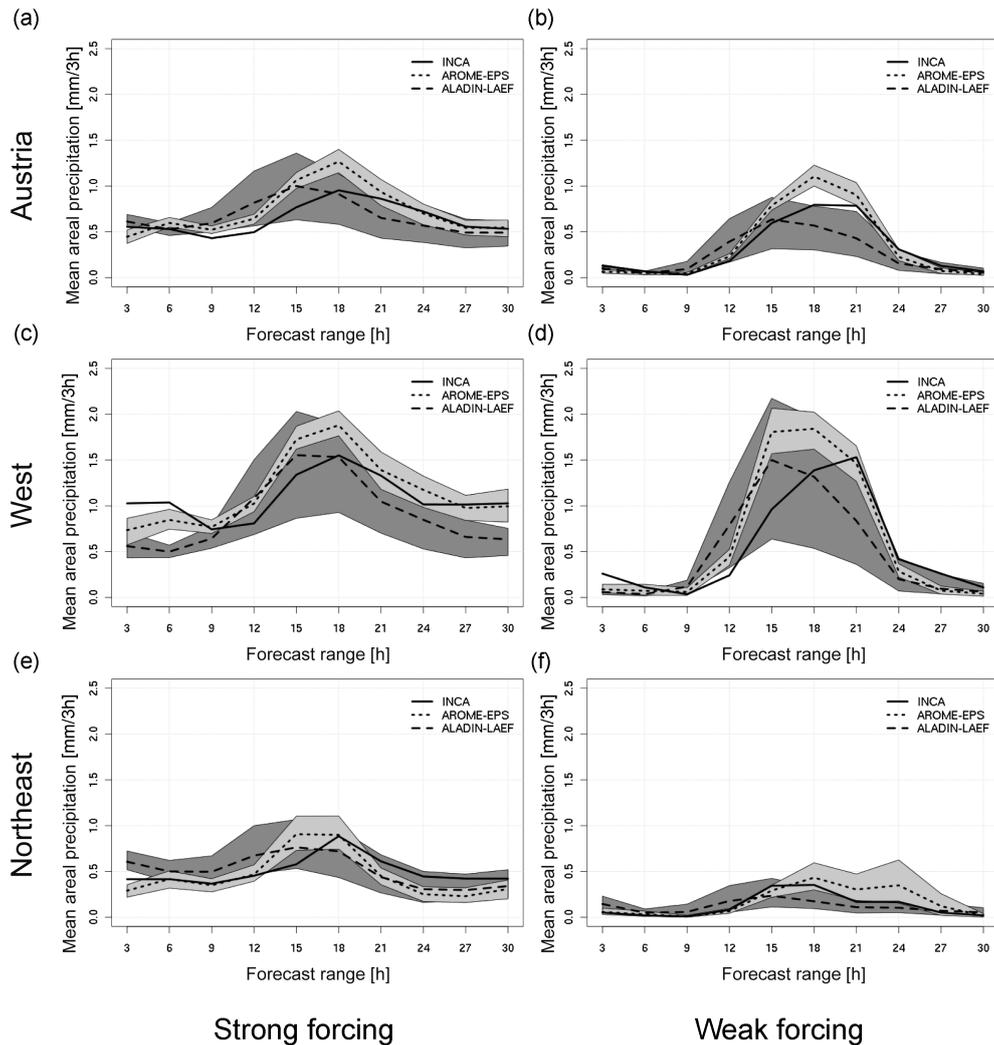


Figure 5. Time evolution of 3-hourly accumulated precipitation forecast for INCA (solid line), ALADIN-LAEF ensemble mean (dashed line), and AROME-EPS ensemble mean (dotted line) for regions Austria (top), West (middle), and Northeast (bottom). Left panels show results for the days with strong synoptic forcing, right panels for weak synoptic forcing. The shaded areas denote the range of individual ensemble member forecasts for ALADIN-LAEF (dark grey) and AROME-EPS (light grey), respectively.

AROME-EPS, which shows a much smaller range of precipitation amounts. This difference in the spread is very likely due to the large influence of the multi-physics configuration in ALADIN-LAEF, compared with the single physics configuration of AROME-EPS.

In order to summarize the findings of Fig. 5, we can state that the ability of the models to forecast the daily precipitation cycle is influenced by both the topography and the type of synoptic forcing. Additionally, there is a general tendency of the finer model, AROME-EPS, to forecast higher precipitation amounts with a temporal maximum later in the day than ALADIN-LAEF. The latter, on the other hand, exhibits a larger variety of simulations, visible through the larger spread, especially over mountainous terrain. In the following, we will discuss several scores (Brier score, SAL scores, and FSS) to demonstrate in which ways the differences in

the diurnal precipitation cycle have an influence on forecast quality.

4.2.1 Brier score components

Figure 6 shows the differences of the components of BS, reliability, resolution, and uncertainty for strong and weak synoptic forcing with different precipitation thresholds for region Austria. BS measures the accuracy of probability forecasts, which is equivalent to the MSE for deterministic forecasts. The value for perfect forecasts is zero. BS has largest values for the lowest precipitation threshold of 0.1 mm/3 h, and decreases for larger thresholds. This is also true for the differences of BS between AROME-EPS and ALADIN-LAEF. However, BS is dominated by the uncertainty component, which is independent of the forecast system but only

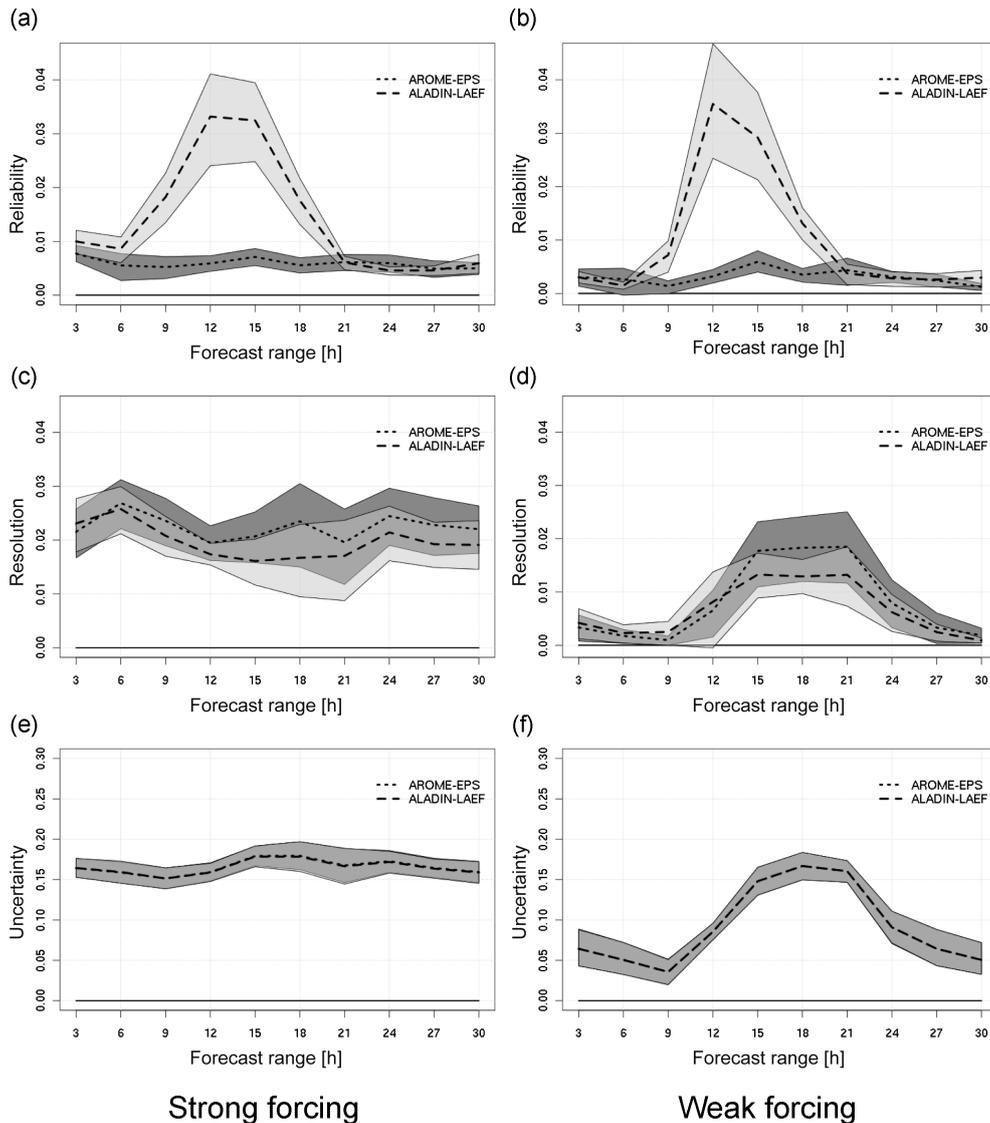


Figure 6. Time evolution of the Brier score components, reliability (top), resolution (center), and uncertainty (bottom), with confidence intervals (shades) for region Austria, AROME-EPS (dotted line), and ALADIN-LAEF (dashed line). The results are shown for a precipitation threshold of 0.1 mm/3 h. Left panels depict results for days with strong synoptic forcing, right panels results for days with weak synoptic forcing.

dependent on the observations. Therefore, the components are shown in Fig. 6, as they provide a more detailed insight into forecast performance than the overall quantity BS.

The unequal diurnal variations of uncertainty for days with strong synoptic forcing and days with weak synoptic forcing are clearly visible in panels e and f, respectively, in Fig. 6. The relatively constant values of uncertainty for strong synoptic forcing and the differences between afternoon (+12 to +24 h forecast range) and early nighttime and morning hours (+3 to +9 h and +27 to +30 h forecast range) for weak synoptic forcing reflect the mean precipitation intensities in Fig. 5a and b. They state that the uncertainty is high whenever there is some possibility of rainfall. In cases of strong syn-

optic forcing, this circumstance persists for the whole day, while there is a period with relatively stable conditions and low probability of rainfall during the morning hours for days with weak synoptic forcing.

The results of the resolution component depicted in panels c and d show very similar daily variations compared to uncertainty. Generally, larger-resolution values are preferable for any forecast system. However, this does not necessarily mean that the forecasts are generally wrong as during the morning hours of days with weak synoptic forcing (panel d) in Fig. 6. It reveals, moreover, that the models keep forecasting low values of precipitation probability regardless of if there is no rain or a little rain reported. However, if the observation sam-

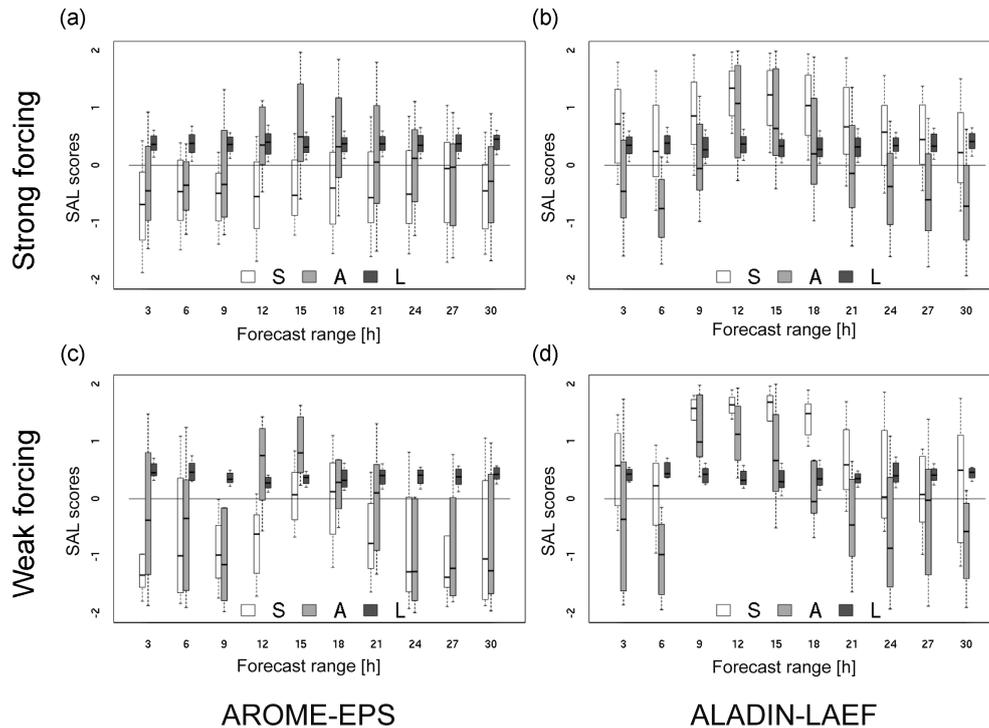


Figure 7. Time evolution of SAL scores for AROME-EPS (left) and ALADIN-LAEF (right) for different forecast ranges in region West. Upper panels (a) and (b) show results for days with strong synoptic forcing, lower panels (c) and (d) for weak synoptic forcing. The boxes are created based on the scores of all individual ensemble members.

ple itself contains values of no rain, results of resolution are less meaningful than for situations with a more balanced distribution of observations. This is the case between noon and early night hours for days with weak synoptic forcing and for the whole day for days with strong synoptic forcing. For these periods, we can observe mostly higher resolution for the forecasts of AROME-EPS than for ALADIN-LAEF, at which the differences are not significant, though. The lower-resolution values for ALADIN-LAEF are presumably due to the smoother precipitation fields compared to AROME-EPS. The smoothness leads to rather medium precipitation probabilities in large areas, which is a disadvantage with regard to resolution compared to sharper forecasts near 0 and 1 (i.e., very low and very high probabilities for rainfall).

The most obvious differences between ALADIN-LAEF and AROME-EPS can be observed for the reliability component (Fig. 6a and b). They can, for the most part, be explained by the time shift between forecast and observation, i.e., by the fact that the precipitation generally starts too early in ALADIN-LAEF forecasts (see Fig. 5a and b). Both models show good (i.e., low values of) reliability during the nighttime and the morning hours (+3 to +6 h and +21 to +30 h forecast range). However, during daytime (starting at +9 h forecast range) ALADIN-LAEF shows significantly higher values of reliability than AROME-EPS with a peak at +12 h of the forecast range. It is the same point in time at which the largest differences between ALADIN-LAEF and INCA

are reported in Fig. 5a and b. The fact that there are also large differences between ALADIN-LAEF and INCA at a longer forecast range (e.g., +21 h) is, however, not reflected in the score. An explanation for this fact is that both the forecasts and INCA reported larger amounts of rain. In this situation, it is easier for the models to differ between no rain and rain. For this reason, the bias in the precipitation intensities of AROME-EPS is also not reflected in the reliability.

4.2.2 SAL scores

The variability of SAL scores with lead time gives insight into the performance of AROME-EPS and ALADIN-LAEF in terms of the structure, amplitude, and location of the predicted precipitation events. Figures 7 and 8 show the SAL scores for the mountainous region West and the lowland region Northeast, respectively. The distributions of SAL values are sampled for the individual ensemble members and classified into days with strong (panels a and b) and weak synoptic forcing (panels c and d). These values differ from those based on the ensemble mean and median forecasts as the averaging produces more smoothed precipitation events, and hence has an influence on the properties described by the SAL method.

In both geographic regions and for both types of synoptic forcing, the structure score is lower for AROME-EPS than for ALADIN-LAEF, which is, inter alia, a consequence of the model resolution (Wittmann et al., 2010). AROME-

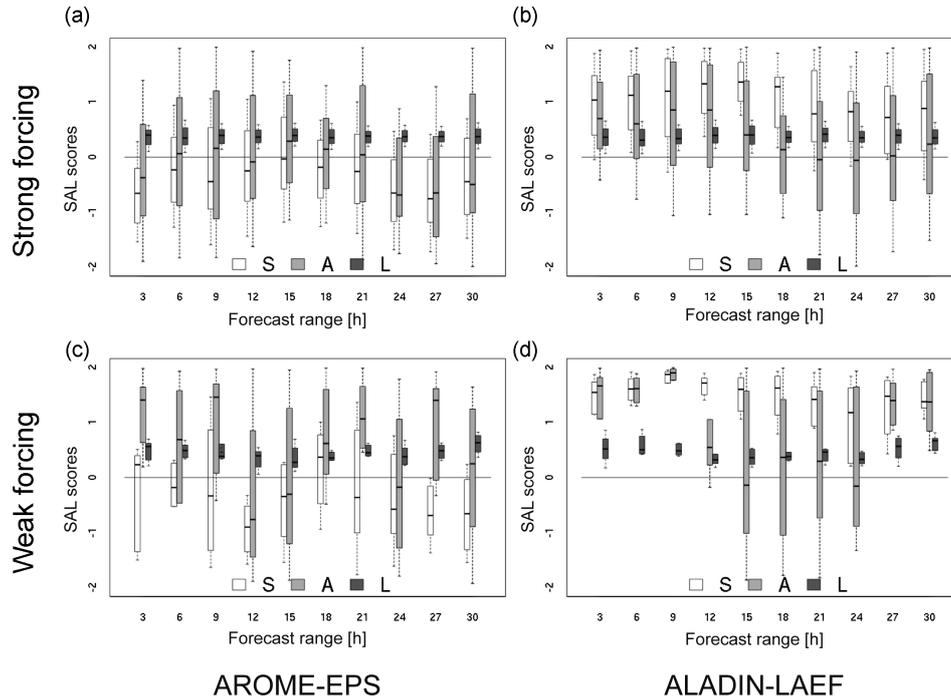


Figure 8. Same as in Fig. 7, but for region Northeast.

EPS produces precipitation events, which are mostly too small and/or too peaked, whereas precipitation objects in ALADIN-LAEF are too large and flat. This is particularly true for days with strong synoptic forcing and for flat terrain. The structure score for ALADIN-LAEF further shows a pronounced diurnal variation for region West, where precipitation events are too large during the day (09:00–15:00 UTC), but more realistic during evening and nighttime. In region Northeast and in weak synoptic forcing, on the contrary, there is a rather damped diurnal variation. This is a sign that precipitation events emerge too early and grow too large over the mountains, whereas over flat land, they are too flat and too widespread during the whole day. AROME-EPS generally shows better agreement with the observed precipitation structures than ALADIN-LAEF during noon (12:00–15:00 UTC) while objects are much too small during the rest of the day. Only on days with strong synoptic forcing and over mountainous terrain does AROME-EPS mostly underestimate the dimension of precipitation events. Also over flat land, structure scores are variable for AROME-EPS, but do not show a perfect daily cycle as for the mountainous areas.

In most instances, the amplitude component reflects the findings shown in Fig. 5, being more apparent for days with weak than for days with strong synoptic forcing. For both EPS models, an overestimation occurs during noon over mountainous terrain (region West; Fig. 7), which is associated with the early onset of convection for ALADIN-LAEF and with the overestimation of precipitation amounts in AROME-EPS. In region Northeast (Fig. 8), the agreement

seems to be much better for days with strong synoptic forcing than with weak synoptic forcing. However, the amplitude score measures the agreement in terms of the percentage share of precipitation amounts. Hence, if the amounts are on a much lower level, as in the case of weak synoptic forcing, amplitude scores appear worse. The large amplitude errors in Fig. 8c and d are, therefore, more dependent on the time shift between simulated and observed peaks of precipitation intensities than on the absolute amount of maximum precipitation intensities, which are fairly well captured.

The location score in both regions provided by the SAL shows not as much variability as the other two components. Nevertheless, an investigation of the distances of observed and forecast centers of mass for the precipitation events can provide useful information. Figure 9a and b show the mean distances for objects pertaining to precipitation thresholds of 0.1 mm/3 h and of 2 mm/3 h for days with strong synoptic forcing, respectively. In general, it can be stated that the distances get shorter with increasing thresholds. This indicates that both ALADIN-LAEF and AROME-EPS are more successful for more intense precipitation events. On the other hand, precipitation objects with very low intensities can be either very small and randomly distributed, which is difficult to predict, or very large, which is easier to predict or detect.

For higher thresholds, Fig. 9b shows that the distances have more variability with time. Although distances are short for earlier hours of the forecast (and the first half of the day), they increase for later forecast hours and reach a maximum at +21 h (21:00 UTC). This effect is much greater

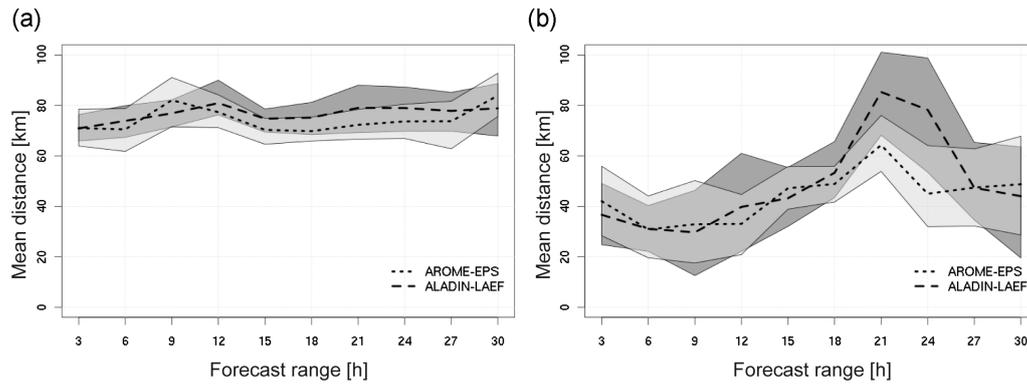


Figure 9. Distances (km) between the centers of mass of observed and forecast precipitation objects for AROME-EPS (dotted) and ALADIN-LAEF (dashed) for thresholds of (a) 0.1 mm/3 h, and (b) 2 mm/3 h. The shades indicate the confidence intervals for AROME-EPS (light grey) and ALADIN-LAEF (dark grey).

in ALADIN-LAEF than in AROME-EPS and it is remarkable that it happens very late in the day, much later than the main peak of precipitation shown in Fig. 5. The reason could be that the precipitation cells are captured well when they are in a mature and well-developed state. Their further development or collapse seems to be better simulated in AROME-EPS. This should be connected to the prognostic (and explicit) treatment of the atmospheric variables describing the evolution of convective activity in AROME. A convection parameterization (in particular, a diagnostic convection scheme, as it is used for some members of ALADIN-LAEF) has more deficiencies in simulating the life cycle of convective objects properly than AROME. In addition, the non-hydrostatic dynamics, higher resolution, and better representation of turbulence and microphysical interactions in the model physics might lead to a more realistic decay of convection in AROME-EPS.

4.2.3 Fractions skill score

The fractions skill score (FSS) indicates how well the ensemble systems predict precipitation at different spatial scales. The grid box widths (1–21 km, corresponding to areas of 1–441 km²) have been selected to investigate the performance of models at very fine scales, near the resolution of the analyzed observations of INCA. At these scales, models have difficulties to reach the level of usefulness (i.e., the target skill as defined in Roberts and Lean, 2008), which can be expected at larger scales. Nevertheless, it is interesting to examine how FSS values change with increasing precipitation thresholds.

Figure 10a and b compare the FSSs for days with strong synoptic forcing and days with weak forcing. FSS values are greater (\sim factor of 2) for strong synoptic forcing than for weak synoptic forcing, since for the latter, precipitation events are generally less structured and lead to the lower level of skill.

For all weather situations, ALADIN-LAEF shows better values for the lowest thresholds of 0.1 and 0.5 mm. The converse result is observed for higher thresholds above 2 mm. For 5 mm/3 h, ALADIN-LAEF has hardly any skill on the very fine scales for days with weak synoptic forcing. This means that small, scattered showers and thunderstorms, which typically occur on these days, cannot be simulated well by the model with coarser model resolution. In AROME-EPS, there is at least a certain skill for small intense precipitation events, although it is not on a level considered to be useful.

These results are comparable to the main outcomes of Le Duc et al. (2013) and Schwartz et al. (2009). Le Duc et al. (2013) also found that the coarser 10 km ensemble showed slightly better results for light rains than the finer 2 km one. Both models had lower skill in predicting heavy rain; however, for the higher precipitation thresholds, the 2 km ensemble performed better than the 10 km one. Schwartz et al. (2009) partially found the same behavior of FSS for coarse 12 km and fine models (2 and 4 km resolution). The coarser model clearly outperformed the finer ones for light rain, whereas the 4 km model showed better skill at a high threshold of 5 mm h⁻¹.

In the previous sections, the discussion provided an overview on the whole 3-month period. In the following section, evaluations focus on a single selected day. This is done in order to show the forecast behavior of the ensembles in a single, concrete weather situation.

4.3 Case study

A typical convective day with weak synoptic forcing is selected to show the evolution of precipitation in AROME-EPS and ALADIN-LAEF in more detail. Here, more emphasis is put on the observation of the numbers, volumes, and distribution of the precipitation objects.

Figure 11 illustrates the precipitation at different times of INCA analyses on 29 April 2014 and the ensemble means of

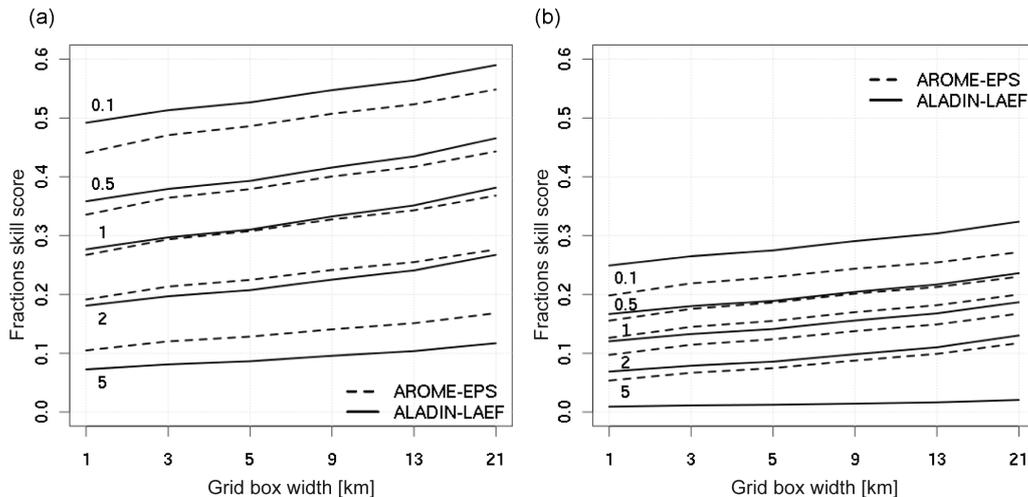


Figure 10. FSSs for (a) strong synoptic forcing, and (b) weak synoptic forcing of AROME-EPS (dashed) and ALADIN-LAEF (solid line) for the region Austria. Numbers denote the precipitation thresholds (mm). The values represent averages for all hours of lead time.

AROME-EPS and ALADIN-LAEF. On this day, continuous light rain was reported in Austria's mountainous terrain, near the main Alpine ridge during the morning hours, as shown in the first row of Fig. 11. At the same time, the lowlands in the east and north were dry. In the lowlands, precipitation activities in terms of small showers started from approximately 11:00 UTC in second row of Fig. 11. Over the course of the day, the focus of precipitation was increasingly shifted to the flat lands in the north, east, and southeast of Austria as well as to Slovenia and northern Italy. The peak rain intensity was around 15:00 UTC, shown at 14:00 UTC in third row of Fig. 11. Rain in the inner Alpine areas had diminished. In contrast, the showers in the flat regions continued until the time of sunset. Then, their activity also weakened, which is visible in the bottom row of Fig. 11.

Figure 12 gives the characteristics of the precipitation forecasts of ALADIN-LAEF and AROME-EPS, such as the temporal evolution of the mean areal precipitation in Fig. 12a, the number of precipitation objects in Fig. 12b, and the temporal evolution of the SAL scores in Fig. 12c. For the selected day, precipitation amounts for the region Austria are slightly underestimated by the both ensemble systems. Further, only a minor fraction of ensemble members reach the observed precipitation intensities at noon. By investigating the structures of the precipitation forecasts, further insight into the behavior of the ensemble systems is provided. The number and volume of precipitation objects describe how models perform in a spatial context. In this respect, AROME-EPS clearly shows more ability to replicate the real spatial structure of precipitation. Although the number of objects in the region Austria is too low during the first forecast hours, the further development as observed by the INCA analysis in Fig. 12b is described well. In the ALADIN-LAEF forecast, the number of precipitation objects is very low and mostly a product of the lower resolution. The volumes of the pre-

cipitation events are in direct connection with their number (not shown). ALADIN-LAEF overestimates the volumes to the same degree as it underestimates their numbers. However, it shows a clear diurnal variation of the volumes with a maximum around noon, which is not indicated by AROME.

The fact that ALADIN-LAEF tends to produce fewer but larger precipitation objects does not lead to worse verification statistics for ALADIN-LAEF. On the contrary, in most regions, the hit rate is higher for ALADIN-LAEF than for AROME-EPS and the number of missed events is lower. AROME-EPS, on the other hand, outperforms ALADIN-LAEF in terms of correct negatives and false alarms (not shown).

These results are also reflected in the temporal evolution of SAL scores in Fig. 12c. As expected, the structure score S is too high for ALADIN-LAEF, due to the overestimation of the volumes of precipitation objects. At the same time, however, AROME-EPS produces a low S score which means that it still produces precipitation objects that are too small and peaked compared to INCA.

Interestingly, there is a late peak in the S score between the 26–28 h lead time in both models, which follows a short minimum at 25 h lead time. This is also slightly reflected in the A score. The sequence of minimum and peak is related to a nightly shower, which was also simulated by the ensembles, but with a delay of approximately 2 h. The location or L score is rather constant in time for both ensemble models. This means that they were able to reproduce the changing spatial focus and distribution of precipitation during the day.

5 Summary and conclusions

In this paper, we investigate the forecast performance of the 2.5 km convection-permitting ensemble AROME-EPS by

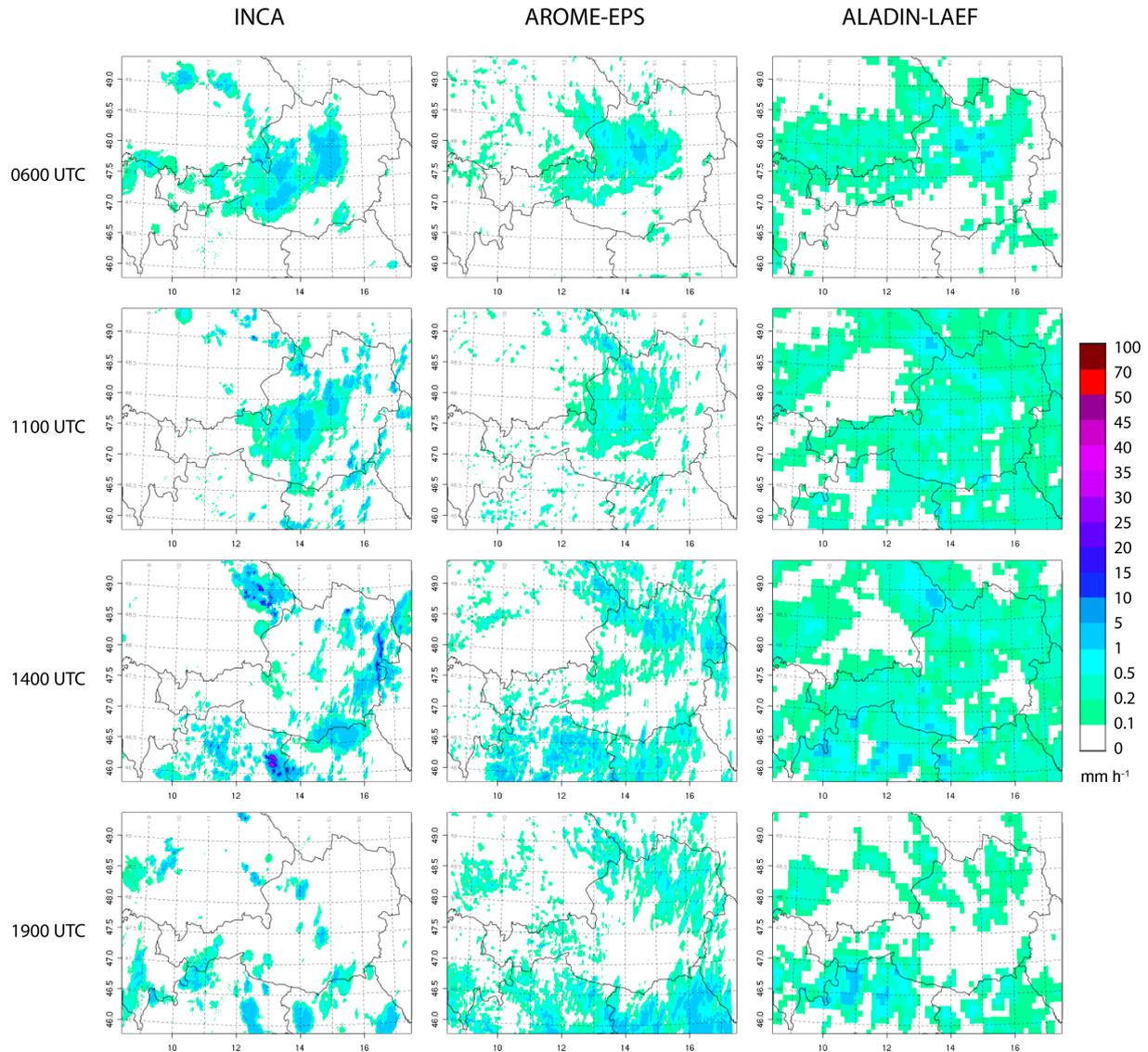


Figure 11. Observed (INCA, first column) and forecast (AROME-EPS and ALADIN-LAEF, second and third columns, respectively) development of precipitation on 29 April 2014 shown for selected times (rows). The panels show 1-hourly accumulated precipitation sums (mm).

comparison with the regional 11 km ensemble ALADIN-LAEF to reveal the benefit provided by a CP EPS. The regional EPS, ALADIN-LAEF, involves several sources of forecast perturbations, such as initial condition perturbations by blending ECMWF-EPS with ALADIN-LAEF breeding vectors and assimilation of perturbed surface observations, and a multi-physics scheme. The high-resolution, convection-permitting AROME-EPS solely performs downscaling of the ALADIN-LAEF forecasts. The performance of the ensembles is evaluated for a 3-month period during the convective season of 2011 and for a typical convective day in April 2014 with a special focus on precipitation events in mountainous terrain and lowland regions. The aim is to show whether the convection-permitting ensemble provides

benefits to the regional ensemble with deep convection parameterization. The evaluation is conducted using a combination of standard deterministic and probabilistic verification scores and selected spatial verification measures. The former are applied on several main forecast parameters for surface and upper levels, and the latter – according to their definition – only for precipitation.

The forecast quality for the main meteorological parameters (except precipitation) for the surface and selected upper levels is strongly dependent on the model bias and is rather balanced, except for diurnal variations near the surface. However, characteristic differences are revealed by the investigation of the precipitation forecasts. A known drawback of models using deep convection schemes proves true, which

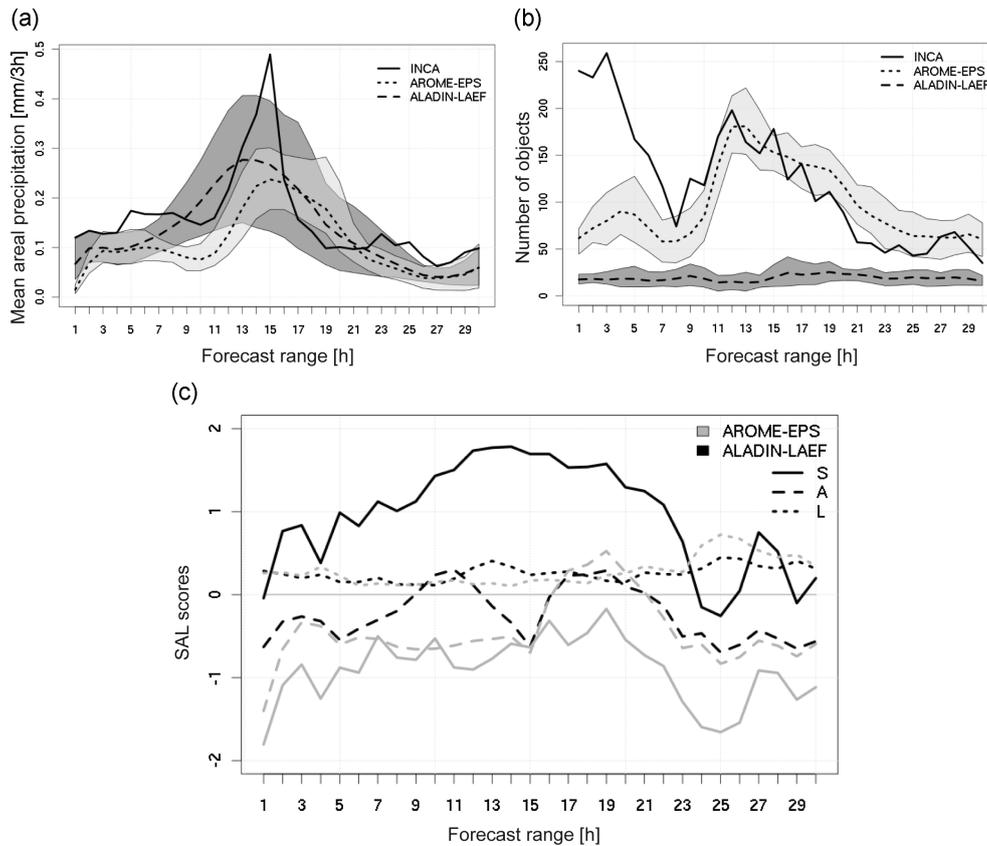


Figure 12. Characteristics of the precipitation forecasts of ALADIN-LAEF and AROME-EPS on 29 April 2014. **(a)** Temporal evolution of the mean areal precipitation compared with INCA and **(b)** temporal evolution of the number of precipitation objects. Dashed and dotted lines in **(a)** and **(b)** represent the ensemble mean and grey shades the ensemble spread. **(c)** Temporal evolution of *S* (structure), *A* (amplitude), and *L* (location) scores of the ensemble means of ALADIN-LAEF (black) and AROME-EPS (grey).

is the premature onset of precipitation in the daily cycle by ALADIN-LAEF (see, e.g., Wittmann et al., 2010; Weusthoff et al., 2010). On the other hand, an overestimation of precipitation intensities at the peak of convection activities by AROME-EPS is also confirmed, which has been assumed in previous validations. Both of these properties are found to be more pronounced in mountainous than in flat regions.

ALADIN-LAEF shows skill in the prediction of probabilities for low precipitation thresholds, i.e., to distinguish between rain and no rain. This is also true for small scales, but it is again dependent on the time of day, as the early onset of precipitation has a negative influence on the verification scores. AROME-EPS, on the other hand, has a better ability to capture the diurnal cycle of convective precipitation, especially over mountainous terrain. At small spatial scales, it further demonstrates better performance for higher precipitation thresholds. The results of the evaluations in this study lead to the conclusion that the convection-permitting ensemble is more skillful in the precipitation forecast than its mesoscale counterpart, the regional ensemble. The positive impact is larger for the mountainous areas than for the lowlands. Nevertheless, the knowledge of which precip-

itation situations can be better modeled by the convection-permitting ensemble is important to have. For many applications, e.g., for large-scale extreme events, such as the central Europe flooding event of 2013, the best solution will be a combination of both systems: the coarser ensembles with longer forecast range for (pre)warnings and the convection-permitting ensemble for the detailed specification of the expected event. Regarding different time and length scales in that way could lead to the generation of seamless forecast products (e.g., Drobinski et al., 2014; Vitart et al., 2008).

This study is considered as the initial point for further investigations and improvement of the convection-permitting ensemble AROME-EPS. The low spread of the prevailing AROME-EPS version is a clear drawback compared to ALADIN-LAEF. Therefore, future enhancements of AROME-EPS will involve components which will presumably increase ensemble spread. Among those upgrades will be ensemble data assimilation and physics perturbations (multimodel and stochastic). The expectation with these components is that forecast errors will be reduced, and that a more realistic simulation of forecast uncertainties will be achieved.

6 Code and/or data availability

The ALADIN-LAEF and AROME codes including all related intellectual property rights, are owned by the members of the LACE consortium and ALADIN consortium. Access to the ALADIN-LAEF and AROME systems, or elements thereof, can be granted upon request and for research purposes only. INCA code and INCA data are only available subject to a licence agreement with ZAMG.

Acknowledgements. We gratefully acknowledge all the LACE/ALADIN/HIRLAM colleagues who have contributed to the development of AROME. ECMWF has provided the computer facilities and technical help implementing ALADIN-LAEF and AROME-EPS on the ECMWF HPCF.

Edited by: A. Kerkweg

Reviewed by: two anonymous referees

References

- Ahijevych, D., Gilleland, E., Brown, B., and Ebert, E.: Application of spatial forecast verification methods to gridded precipitation forecasts, *Weather Forecast.*, 24, 1485–1497, 2009.
- Aligo, A. E., Gallus Jr., W. A., and Segal, M.: On the Impact of WRF Model Vertical Grid Resolution on Midwest Summer Rainfall Forecasts, *Weather Forecast.*, 24, 575–594, 2009.
- Barthlott, C., Burton, R., Kirshbaum, D., Hanley, K., Richard, R., Chaboreau, J. P., Trentmann, J., Kern, B., Bauer, H.-S., Schwitalla, T., Keil, C., Seity, Y., Gadian, A., Blyth, A. M., Mobbs, S., Flamant, C., and Handwerker, J.: Initiation of deep convection at marginal instability in an ensemble of mesoscale models: A case-study from COPS, *Q. J. Roy. Meteor. Soc.*, 137, 118–136, 2011.
- Bauer, H. S., Weusthoff, T., Dorninger, M., Wulfmeyer, V., Schwitalla, T., Gorgas, T., Arpagaus, M., and Warrach-Sagi, K.: Predictive skill of a subset of models participating in D-PHASE in the COPS region, *Q. J. Roy. Meteor. Soc.*, 137, 287–305, 2011.
- Bechtold, P., Semane, N., Lopez, P., and Chaboureau, J.-P., Beljaars, A., and Bormann, N.: Breakthrough in forecasting equilibrium and non-equilibrium convection, *ECMWF Newsletter*, 136, 15–22, 2013.
- Ben Bouallégue, Z., Theis, S. E., and Gebhardt, C.: Enhancing COSMO-DE ensemble forecasts by inexpensive techniques, *Meteorol. Z.*, 22, 49–59, 2013.
- Bougeault, P.: A simple parameterization of the large-scale effects of cumulus convection, *Mon. Weather Rev.*, 113, 2108–2121, 1985.
- Bouttier, F., Vié, B., Nuissier, O., and Raynaud, L.: Impact of Stochastic Physics in a Convection-Permitting Ensemble, *Mon. Weather Rev.*, 140, 3706–3721, 2012.
- Bowler, N. E., Arribas, A., Mylne, K. R., Robertson, K. B., and Beare, S. E.: The MOGREPS short-range ensemble prediction system, *Q. J. Roy. Meteor. Soc.*, 134, 703–722, 2008.
- Brier, G. W.: Verification of forecasts expressed in terms of probability, *Mon. Weather Rev.*, 78, 1–3, 1950.
- Bryan, G. H., Wyngaard, J. C., and Fritsch, J. M.: Resolution requirements for the simulation of deep moist convection, *Mon. Weather Rev.*, 131, 2394–2416, 2003.
- Bukovsky, M. S., Kain, J. S., and Baldwin, M. E.: Bowing convective systems in a popular operational model: Are they for real?, *Weather Forecast.*, 21, 307–324, 2006.
- Caron, J.: Mismatching perturbations at the lateral boundaries in limited-area ensemble forecasting: A case study, *Mon. Weather Rev.*, 141, 356–374, 2013.
- Casati, B. L., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocerlich, M., Damrath, U., Ebert, E. E., Brown, B. G., and Mason, S.: Review forecast verification: current status and future directions, *Meteor. Appl.*, 15, 3–18, 2008.
- Clark, A. J., Gallus Jr., W. A., and Chen, T.-C.: Comparison of the Diurnal Precipitation Cycle in Convection-Resolving and Non-Convection-Resolving Mesoscale Models, *Mon. Weather Rev.*, 135, 3456–3473, 2007.
- Clark, A. J., Gallus Jr., W. A., Xue, M., and Kong, F.: A Comparison of Precipitation Forecast Skill between Small Convection-Allowing and Large Convection-Parameterizing Ensembles, *Weather Forecast.*, 24, 1121–1140, 2009.
- Clark, A. J., Kain, J. S., Stensrud, D. J., Xue, M., Kong, F., Coniglio, M. C., Thomas, K. W., Wang, Y., Brewster, K., Gao, J., Wang, X., Weiss, S. J., and Du, J.: Probabilistic Precipitation Forecast Skill as a Function of Ensemble Size and Spatial Scale in a Convection-Allowing Ensemble, *Mon. Weather Rev.*, 139, 1410–1418, 2011.
- Davis, C. A., Manning, K. W., Carbone, R. E., Trier, S. B., and Tuttle, J. D.: Coherence of warm season continental rainfall in numerical weather prediction models, *Mon. Weather Rev.*, 131, 2667–2679, 2003.
- Davison, A. C. and Hinkley, D. V.: *Bootstrap Methods and their applications* – Cambridge University Press, Cambridge, UK, 193 pp, 1997.
- Done, J. M., Craig, G. C., Gray, S. L., Clark, P. A., and Gray, M. E. B.: Mesoscale simulations of organized convection: Importance of convective equilibrium, *Q. J. Roy. Meteor. Soc.*, 132, 737–756, 2006.
- Drobinski, P., Ducrocq, V., Alpert, P., Anagnostou, E., Béranger, K., Borga, M., Braud, I., Chanzy, A., Davolio, S., Delrieu, G., Estournel, C., Boubrahmi, N., Font, J., Grubišić, V., Gualdi, S., Homar, V., Ivančan-Picek, B., Kottmeier, C., Kotroni, V., Lagouvardos, K., Lionello, P., Llasat, M., Ludwig, W., Lutoff, C., Mariotti, A., Richard, E., Romero, R., Rotunno, R., Roussot, O., Ruin, I., Somot, S., Taupier-Letage, I., Tintore, J., Uijlenhoet, R., and Wernli, H.: HyMeX: A 10-Year Multidisciplinary Program on the Mediterranean Water Cycle, *B. Am. Meteorol. Soc.*, 95, 1063–1082, 2014.
- Ferro, C. A. T.: A probability model for verifying deterministic forecasts of extreme events, *Weather Forecast.*, 22, 1089–1100, 2007.
- Gebhardt, C., Theis, S. E., Paulat, M., and Ben Bouallégue, Z.: Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries, *Atmos. Res.*, 100, 168–177, 2011.
- Geleyn, J.-F., Catry, B., Bouteloup, Y., and Brožková, R.: A statistical approach for sedimentation inside a microphysical precipitation scheme, *Tellus*, 60A, 649–662, doi:10.1111/j.1600-0870.2008.00323.x, 2008.

- Gerard, L., Piriou, J.-M., Brožkova, R., Geleyn, J.-F., and Banciu, D.: Cloud and precipitation parameterization in a meso-gamma scale operational weather prediction model, *Mon. Weather Rev.*, 137, 3960–3977, 2009.
- Gilleland, E., Ahijevych, D. A., Brown, B. G., and Ebert, E. E.: Verifying forecasts spatially, *B. Am. Meteorol. Soc.*, 47, 1365–1373, 2010.
- Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction and Estimation, *J. Am. Stat. Assoc.*, 102, 359–378, 2007.
- Haiden, T. and Pistotnik, G.: Intensity-dependent parameterization of elevation effects in precipitation analysis, *Adv. Geosci.*, 20, 33–38, doi:10.5194/adgeo-20-33-2009, 2009.
- Haiden, T., Kann, A., Wittmann, C., Pistotnik, G., Bica, B., and Gruber, C.: The Integrated Nowcasting through Comprehensive Analysis (INCA) System and Its Validation over the Eastern Alpine Region, *Weather Forecast.*, 26, 166–183, 2011.
- Haiden, T., Magnusson, L., Tsonevsky, I., Wetterhall, F., Alfieri, L., Pappenberger, F., de Rosnay, P., Muñoz-Sabater, J., Balsamo, G., Albergel, C., Forbes, R., Hewson, T., Malardel, S., and Richardson, D.: ECMWF forecast performance during the June 2013 flood in Central Europe, ECMWF – Technical Memorandum, 723, available at: http://old.ecmwf.int/publications/library/ecpublications/_pdf/tm/701-800/tm723.pdf, last access: 2 September 2014.
- Hall, P., Horowitz, J. L., and Jing, B.-Y.: On blocking rules for the bootstrap with dependent data, *Biometrika*, 82, 561–574, 1995.
- Hanley, K. E., Kirshbaum, D. J., Roberts, N. M., and Leoncini, G.: Sensitivities of a Squall Line over Central Europe in a Convective-Scale Ensemble, *Mon. Weather Rev.*, 141, 112–133, 2013.
- Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, *Weather Forecast.*, 15, 559–570, 2000.
- Jolliffe, I.: Uncertainty and inference for verification measures, *Weather Forecast.*, 22, 637–650, 2007.
- Kühnlein, C., Keil, C., Craig, G. C., and Gebhardt, C.: The impact of downscaled initial condition perturbations on convective-scale ensemble forecasts of precipitation, *Q. J. Roy. Meteor. Soc.*, 140, 1552–1562, 2014.
- Le Duc, L., Saito, K., and Seko, H.: Spatial-temporal fractions verification for high-resolution ensemble forecasts, *Tellus A*, 65, 18171, doi:10.3402/tellusa.v65i0.18171, 2013.
- Liu, C., Moncrieff, M. W., Tuttle, J. D., and Carbone, R. E.: Explicit an Parameterized Episodes of Warm-Season Precipitation over the Continental United States, *Adv. Atmos. Sci.*, 23, 91–105, 2006.
- Masson, V.: A physically-based scheme for the urban energy budget in atmospheric models, *Bound.-Lay. Meteorol.*, 94, 357–397, 2000.
- Mlawer, E. J., Taubman, S. J., Brown, P. D., Iacono, M. J., and Clough, S. A.: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave, *J. Geophys. Res.*, 102D, 16663–16682, 1997.
- Morcrette, J.-J.: Radiation and cloud radiative properties in the ECMWF operational weather forecast model, *J. Geophys. Res.*, 96D, 9121–9132, 1991.
- Murphy, A. H.: A new vector partition of the probability score, *J. Appl. Meteor.*, 12, 595–600, 1973.
- Nurmi, P.: Recommendations on the verification of local weather forecasts, ECMWF Technical Memoranda, 430, 19 pp., available at: <http://old.ecmwf.int/publications/library/do/references/show?id=86094> (last access: 20 October 2014), 2003.
- Peralta, C., Ben Bouallègue, Z., Theis, S. E., Gebhardt, C., and Buchhold, M.: Accounting for initial condition uncertainties in COSMO-DE-EPS, *J. Geophys. Res.*, 117, 1–13, doi:10.1029/2011JGD016581, 2012.
- Pergaud, J., Masson, V., and Malardel, S.: A parameterization of dry thermals and shallow cumuli for mesoscale numerical weather prediction, *Bound.-Lay. Meteorol.*, 132, 83–106, 2009.
- Pinty, J. P. and Jabouille, P.: A mixed phase cloud parameterization for use in a mesoscale nonhydrostatic model: Simulations of a squall line and of orographic precipitation, *Preprints, Conf. on Cloud Physics*, Everett, WA, Amer. Meteor. Soc., 217–220, 1998.
- Richard, E., Buzzi, A., and Zängl, G.: Quantitative precipitation forecasting in the Alps: The advances achieved by the Mesoscale Alpine Programme, *Q. J. Roy. Meteor. Soc.* 133, 831–846, 2007.
- Ritter, B. and Geleyn, J.-F.: A comprehensive radiation scheme for numerical weather prediction models with potential applications in climate simulations, *Mon. Weather Rev.*, 120, 303–325, 1992.
- Roberts, N. M. and Lean, H. W.: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events, *Mon. Weather Rev.*, 136, 78–97, 2008.
- Romine, G. S., Schwartz, C. S., Berner, J., Fossell, K. R., Snyder, C., Anderson, J. L., and Weisman, M. L.: Representing Forecast Error in a Convection-Permitting Ensemble System, *Mon. Weather Rev.*, 142, 4519–4541, 2014.
- Rotach, M., Ambrosetti, P., Appenzeller, C., Arpagaus, M., Fontanaz, L., Fundel, F., Germann, U., Hering, A., Liniger, M., Stoll, M., Walser, A., Ament, F., Bauer, H., Behrendt, A., Wulfmeyer, V., Bouttier, F., Seity, Y., Buzzi, A., Davolio, S., Corazza, M., Denhard, M., Dorninger, M., Gorgas, T., Frick, J., Hegg, C., Zappa, M., Keil, C., Volkert, H., Marsigli, C., Montani, A., McTaggart-Cowan, R., Mylne, K., Ranzi, R., Richard, E., Rossa, A., Santos-Muñoz, D., Schär, C., Staudinger, M., Wang, Y., and Werhahn, J.: MAP D-PHASE real time demonstration of weather forecast quality in the Alpine region, *B. Am. Meteorol. Soc.*, 90, 1321–1336, 2009.
- Schumacher, R. S., Clark, A. J., Xue, M., and Kong, F.: Factors Influencing the Development and Maintenance of Nocturnal Heavy-Rain-Producing Convective Systems in a Storm-Scale Ensemble, *Mon. Weather Rev.*, 141, 2778–2801, 2013.
- Schumacher, R. and Clark, A.: Evaluation of Ensemble Configurations for the Analysis and Prediction of Heavy-Rain-Producing Mesoscale Convective Systems, *Mon. Weather Rev.*, 142, 4108–4138, doi:10.1175/MWR-D-13-00357.1, 2014.
- Schwartz, C. S., Kain, J. S., Weiss, S. J., Xue, M., Bright, D. R., Kong, F., Thomas, K. W., Levit, J. J., Coniglio, M. C., and Wandishin, M. S.: Toward Improved Convection-Allowing Ensembles: Model Physics Sensitivities and Optimizing Probabilistic Guidance with Small Ensemble Membership, *Weather Forecast.*, 25, 263–280, doi:10.1175/2009WAF2222267.1, 2010.
- Schwartz, C. S., Romine, G. S., Smith, K. R., and Weisman, M. L.: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble kalman filter, *Weather Forecast.*, 29, 1295–1318, 2014.
- Schwartz, C. S., Romine, G. S., Sobash, R. A., Fossell, K. R., and Weisman, M. L.: NCAR's experimental real-time convection-

- allowing ensemble prediction system, *Weather Forecast.*, 30, 1645–1654, doi:10.1175/WAF-D-15-0103.1, 2015.
- Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., and Masson, V.: The AROME-France Convective-Scale Operational Model, *Mon. Weather Rev.* 139, 976–991, 2011.
- Taillefer, F.: CANARI – Technical Documentation – Based on ARPEGE cycle CY25T1 (AL25T1 for ALADIN), available at: http://www.cnrm.meteo.fr/gmapdoc/IMG/ps/canari_doc_cy25t1.ps (last access: 14 December 2015), 2002.
- Tennant, W.: Improving initial condition perturbations for MOGREPS-UK, *Q. J. Roy. Meteor. Soc.*, 141, 2324–2336, doi:10.1002/qj.2524, 2015.
- Theis, S. E., Hense, A., and Damrath, U.: Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach, *Meteor. Appl.*, 12, 257–268, doi:10.1017/S1350482705001763, 2005.
- Toth, Z. and Kalnay, E.: Ensemble forecasting at NMC: the generation of perturbation, *B. Am. Meteorol. Soc.*, 74, 2317–2330, 1993.
- UK Met Office: Benefits of high resolution ensemble forecasts, available at: <http://www.metoffice.gov.uk/research/news/2014/high-resolution-ensembles>, last access: 21 October 2014.
- Vana, F., Benard, P., Geleyn, J.-F., Simon, A., and Seity, Y.: Semi-Lagrangian advection scheme with controlled damping: an alternative to nonlinear horizontal diffusion in a numerical weather prediction model, *Q. J. Roy. Meteor. Soc.*, 134, 523–537, 2008.
- Vié, B., Molinié, G., Nuissier, O., Vincendon, B., Ducrocq, V., Bouttier, F., and Richard, E.: Hydro-meteorological evaluation of a convection-permitting ensemble prediction system for Mediterranean heavy precipitating events, *Nat. Hazards Earth Syst. Sci.*, 12, 2631–2645, doi:10.5194/nhess-12-2631-2012, 2012.
- Vitart, F., Buizza, R., Balmaseda, M. A., Balsamo, G., Bidlot, J.-R., Bonet, A., Fuentes, M., Hofstadler, A., Molteni, F., and Palmer, T. N.: The new VarEPS—monthly forecasting system: A first step towards seamless prediction, *Q. J. Roy. Meteor. Soc.*, 134, 1789–1799, 2008.
- Wang, Y., Kann, A., Bellus, M., Pailleux, J., and Wittmann, C.: A strategy for perturbing surface initial conditions in LAMEPS, *Atmos. Sci. Lett.*, 11, 108–113, 2010.
- Wang, Y., Bellus, M., Wittmann, C., Steinheimer, M., Weidle, F., Kann, A., Ivatek-Šahdan, S., Tian, W., Ma, X., Tascu, S., and Bazile, E.: The Central European limited-area ensemble forecasting system: ALADIN-LAEF, *Q. J. Roy. Meteor. Soc.*, 137, 483–502, 2011.
- Wang, Y., Tascu, S., Weidle, F., and Schmeisser, K.: Evaluation of the Added Value of Regional Ensemble Forecasts on Global Ensemble Forecasts, *Weather Forecast.*, 27, 972–987, 2012.
- Wang, Y., Bellus, M., Geleyn, J.-F., Ma, X., Tian, W., and Weidle, F.: A New Method for Generating Initial Condition Perturbations in a Regional Ensemble Prediction System: Blending, *Mon. Weather Rev.*, 142, 2043–2059, 2014.
- Weckwerth, T., Bennett, L., Miller, L., Van Baelen, J., Di Girolamo, P., Blyth, A., and Hertneky, T.: An Observational and Modeling Study of the Processes Leading to Deep, Moist Convection in Complex Terrain, *Mon. Weather Rev.*, 142, 2687–2708, 2014.
- Weidle, F., Wang, Y., Tian, W., and Wang, T.: Validation of Strategies using Clustering Analysis of ECMWF EPS for Initial Perturbations in a Limited Area Model Ensemble Prediction System, *Atmosphere-Ocean*, 51, 284–295, 2013.
- Weisman, M. L., Skamarock, W. C., and Klemp, J. B.: The resolution dependence of explicitly modeled convective systems, *Mon. Weather Rev.*, 125, 527–548, 1997.
- Wernli, H., Paulat, M., Hagen, M., and Frei, C.: SAL – A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts, *Mon. Weather Rev.*, 136, 4470–4487, 2008.
- Weusthoff, T., Ament, F., Arpagaus, M., and Rotach, M. W.: Assessing the Benefits of Convection-Permitting Models by Neighborhood Verification: Examples from MAP D-PHASE, *Mon. Weather Rev.*, 138, 3418–3433, 2010.
- Wilks, D. S.: *Statistical Methods in the Atmospheric Sciences*, Academic Press, 464 pp., chap. 7, 1995.
- Wilks, D. S.: *Statistical Methods in the Atmospheric Sciences*, London, Academic Press, 2nd Edn., 627 pp., 2006.
- Wittmann, C., Haiden, T., and Kann, A.: Evaluating multi-scale precipitation forecasts using high resolution analysis, *Adv. Sci. Res.*, 4, 89–98, doi:10.5194/asr-4-89-2010, 2010.
- Wulfmeyer, V., Behrendt, A., Bauer, H., Kottmeier, C., Corsmeier, U., Blyth, A., Craig, G., Schumann, U., Hagen, M., Crewell, S., Di Girolamo, P., Flamant, C., Miller, M., Montani, A., Mobbs, S., Richard, E., Rotach, M., Arpagaus, M., Russchenberg, H., Schlüssel, P., König, M., Gärtner, V., Steinacker, R., Dorninger, M., Turner, D., Weckwerth, T., Hense, A., and Simmer, C.: The Convective and Orographically induced Precipitation Study: A research and development project of the World Weather Research Program for improving quantitative precipitation forecasting in low-mountain regions, *B. Am. Meteorol. Soc.*, 89, 1477–1486, doi:10.1175/2008BAMS2367.1, 2008.
- Wulfmeyer, V., Behrendt, A., Kottmeier, C., Corsmeier, U., Barthlott, C., Craig, G. C., Hagen, M., Althausen, D., Aoshima, F., Arpagaus, M., Bauer, H.-S., Bennett, L., Blyth, A., Brandau, C., Champollion, C., Crewell, S., Dick, G., Di Girolamo, P., Dorninger, M., Dufournet, Y., Eigenmann, R., Engelmann, R., Flamant, C., Foken, T., Gorgas, T., Grzeschik, M., Handwerker, J., Hauck, C., Höller, H., Junkermann, W., Kalthoff, N., Kiemle, C., Klink, S., König, M., Krauss, L., Long, C. N., Madonna, F., Mobbs, S., Neining, B., Pal, S., Peters, G., Pigeon, G., Rotach, E., Rotach, M. W., Russchenberg, H., Schwitalla, T., Smith, V., Steinacker, R., Trentmann, J., Turner, D. D., van Baelen, J., Vogt, S., Volkert, H., Weckwerth, T., Wernli, H., Wieser, A., and Wirth, M.: The Convective and Orographically-induced Precipitation Study (COPS): the scientific strategy, the field phase, and research highlights, *Q. J. Roy. Meteor. Soc.*, 137, 3–30, 2011.
- Xue, M., Kong, F., Weber, D., Thomas, K. W., Wang, Y., Brewster, K., Droegemeier, K. K., Kain, J., Weiss, S., Bright, D., Wandishin, M., Coniglio, M., and Du, J.: CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2007 spring experiment. Extended Abstracts, 22nd Conference on Weather Analysis and Forecasting/18th Conference on Numerical Weather Prediction, Park City, UT. Amer. Meteor. Soc., available at: <http://ams.confex.com/ams/pdfpapers/124587.pdf> (last access: 19 December 2016), 2007.
- Xue, M., Kong, F., Thomas, K., Gao, J., Wang, Y., Brewster, K., Droegemeier, K., Kain, J. S., Weiss, S. J., Bright, D. R., Coniglio, M. C., and Du, J.: CAPS realtime multi-model convection-allowing ensemble and 1-km convection-resolving forecasts for

the NOAA Hazardous Weather Testbed 2009 spring experiment. Preprints, 23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction, Omaha, NE, Amer. Meteor. Soc., 16A.2, available at: http://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154323.htm (last access: 19 December 2016), 2009.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., and Mylne, K.: The Economic Value Of Ensemble-Based Weather Forecasts, B. Am. Meteorol. Soc., 83, 73–83, 2002.