universität wien

MASTERARBEIT/MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Improvement of air traffic management through probabilistic thunderstorm forecasts at Schwechat Airport

verfasst von / submitted by

Maximilian Weissinger, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2022 / Vienna, 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 614
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Meteorologie
Betreut von / Supervisor:	Ass.-Prof. Mag. Dr. Manfred Dorninger

Contents

List of Tables

List of Figures

Abstract

At present, information from Air Traffic Management (ATM) and the forecasters (MET) is processed separately. The thresholds for warnings of potentially dangerous events are set centrally, but without knowing the requirements of the individual work areas that rely on these warnings. The project will go beyond the most commonly used METARs (Airport Weather Analysis) and TAFs (Airport Weather Forecasts) and will provide a more detailed probability of occurrence and alternative events in addition to the weather information. This will be done by means of ensemble forecasts. While first such forecast concepts for fog already exist, a more concrete forecast for thunderstorms will be developed within this project. The aim is to develop automated thunderstorm warnings for the flight sectors of Vienna Airport.

Three different ensemble prediction models are available for the investigation, which differ in model physics, resolution, data assimilation and other properties. On the one hand, the convection parameterizing model ALADIN-LAEF of ZAMG, which has been in operation for years, will be investigated, on the other hand, models of the relatively new genus of convection-resolving models will be investigated. Here C-LAEF of ZAMG and COSMO-D2-EPS of DWD are available. The model data is to be checked for the times of day at which the greatest probability of thunderstorms occurs and then parameters for the effective forecast of thunderstorm events are to be found. These parameters are either calculated directly by the model, such as the simulated radar reflectivity of COSMO-D2-EPS, or are generated via threshold values of CAPE, precipitation and cloud base. These parameters are used to calculate predictions for the Ensemble models and thus generate probability predictions.

The resulting probability forecasts are verified by means of a novel cell tracking system, which uses radar reflectivities of surrounding radars as well as lightning data of the Austrian Lightning Detection System (ALDIS). As verification tools, reliability diagram, ROC curve and the PEV (potential economic value) diagram are used. Subsequently, the prediction concepts are to be validated with a temporally more extensive data set. Further work steps beyond the scope of this study involve the calibration of the probability predictions and the creation of products specifically tailored for ATM and MET.

The results showed a significantly higher prediction skill for high-resolution ensemble prediction models such as C-LAEF and COSMO-D2-EPS and also pointed to a significant advantage over conventional deterministic prediction models. However, it also clearly showed that thunderstorm forecasts have to be made over sufficiently large areas and with a sufficiently large number of events to give the results statistical robustness. Furthermore, it was shown that combining the forecast models into a multi-model ensemble significantly increases the potential forecast skill. The use of flight sectors had the disadvantage that regional differences in forecast quality and occurrence of convection were superimposed by the effect of different sector sizes.

Zusammenfassung

Derzeit werden Informationen von Air Traffic Management (ATM) und den Forecastern (MET) getrennt voneinander verwertet. Die Schwellwerte für die Warnung vor potentiell gefährlichen Ereignissen werden zentral erstellt, ohne jedoch die Anforderungen der einzelnen Arbeitsbereiche, die auf diese Warnungen angewiesen sind, zu kennen. Das Projekt soll über die meist verwendeten METARs (Wetter-Analyse am Flughafen) und TAFs (Wettervorhersage am Flughafen) hinausgehen und zusätzlich zu den Wetterinformationen eine detailiertere Eintrittswarscheinlichkeit sowie alternative Ereignisse mitliefern. Dies soll mittels Ensemble-Vorhersagen geschehen. Während für Nebel bereits erste Vorhersagekonzepte vorliegen, soll in diesem Projekt eine konkretere Vorhersage für Gewitter ausgearbeitet werden. Ziel ist es, automatisierte Gewitterwarnungen für die Flugsektoren des Flughafens Wien zu entwickeln.

Für die Untersuchung stehen drei verschiedene Ensemblevorhersagemodelle zur Verfügung, die sich in Modellphysik, Auflösung, Datenassimilation und anderen Eigenschaften unterscheiden. Einerseits wird das operationell bereits seit Jahren im Einsatz stehende konvektionsparameterisierende Modell ALADIN-LAEF der ZAMG untersucht, andererseits sollen Modelle der relativ neuen Gattung konvektionsauflösender Modelle untersucht werden. Hier stehen C-LAEF der ZAMG und COSMO-D2-EPS des DWD zur Verfügung. Die Modelldaten sollen daraufhin überprüft werden, zu welchen Tageszeiten die größte Eintrittswahrscheinlichkeit für Gewitter vorliegt und anschließend Parameter zur effektiven Prognose von Gewitterereignissen gefunden werden. Diese Parameter werden entweder vom Modell direkt gerechnet, wie die simulierte Radarreflektivität von COSMO-D2-EPS oder werden über Schwellwerte von CAPE, Niederschlag und der Wolkenbasis bestimmt. Mithilfe dieser Parameter werden Vorhersagen für die Ensemble-Modelle gerechnet und somit Wahrscheinlichkeitsprognosen generiert.

Die resultierenden Wahrscheinlichkeitsprognosen werden mittels eines neuartigen Zell-Trackingsystems verifiziert, das sowohl Radarreflektivitäten umliegender Radars, aber auch Blitzdaten des österreichischen Blitzortungssystems (ALDIS) nutzt. Als Verifikationstools werden Reliability-Diagramm, ROC-Kurve und das PEV-Diagram (Potential economic value) eingesetzt. Anschließend sollen die Vorhersagekonzepte mit einem zeitlich umfangreicheren Datensatz validiert werden. Weitere Arbeitsschritte, die über den Rahmen dieser Studie hinausgehen, sind die Kalibrierung der Wahrscheinlichkeitsvorhersagen und die Erstellung von Produkten, die speziell für ATM und MET zugeschnitten sind.

Die Ergebnisse zeigten einen deutlich höheren Vorhersageskill für hochauflösende Ensemblevorhersagemodelle wie C-LAEF und COSMO-D2-EPS und wiesen auch auf einen signifikanten Vorteil gegenüber herkömmlichen deterministischen Vorhersagemodellen hin. Es zeigte aber auch deutlich, dass Gewitterprognosen über ausreichend großen Gebieten und mit einer ausreichend großen Zahl von Ereignissen gemacht werden müssen, um den Ergebnissen eine statistische Robustheit zu geben. Weiters wurde gezeigt, dass die Kombination der Vorhersagemodelle zu einem Multi-Modell-Ensemble den potentiellen Vorhersageskill deutlich erhöht. Die Verwendung von Flugsektoren brachte den Nachteil mit sich, dass räumliche Unterschiede in Vorhersagequalität und Auftreten von Konvektion durch den Effekt der unterschiedlichen Sektorgrößen überlagert wurde.

1. Introduction

1.1 Problem Statement

Notice: Although the coronavirus pandemic has caused a massive decline in international air traffic volume in the past two years, passenger traffic is already recovering as can be seen from ICAO (2022), which means that the issue of a further increase in air traffic volume remains relevant.

In the course of globalisation, the volume of air traffic is growing significantly. According to ICAO (2018), 4.3 billion passengers worldwide were carried on scheduled flights in 2018, an increase of 6.4 percent over the previous year. During the same period, revenue passenger kilometres rose by 7.1 percent to 8.3 trillion kilometres. The transported cargo went up by 2.4 percent, reaching 58 million tonnes by 2018 (ICAO, 2018).

A higher volume of air traffic results in a more compressed time schedule at the airport terminals, but also on the flight routes. Therefore, planning concepts must also become more precise in order to cope with the ever increasing air traffic volume. Thunderstorms, in particular, can be a massive hindrance to tightly scheduled air traffic (Weber et al., 2005).

Flights must avoid thunderstorms en-route and airports must stop all activities - from passenger boarding to apron work and refuelling - as soon as lightning activity (cloud-to-ground as well as cloud-to-cloud lightning) is detected within a certain radius around the airport, usually 3 miles (5km) (Blom, 2008). This can result in delays, rerouting to other airports or even cancellation of services. Especially thunderstorms organised in lines, so-called squall lines, can be a difficult obstacle to fly around and particularly restrict air traffic. Figure 1 shows an extensive field of thunderstorms over the eastern edge of the Austrian Alps, which must be flown around by airplanes on a large scale, as can be seen from the flight trajectories (grey). Based on Federal Aviation Administration (FAA) delay statistics, Weber et al. (2005) estimate direct operational costs of $2 billion (USD) annually for the US commercial airline industry due to thunderstorm related flight delays. According to the FAA (2018) during the years 2012 to 2017 flight delays and cancellations in the United States cost airlines and passengers annually around 20 billion dollars, of which at least in the years 2016 and 2017 over 60% account to adverse weather like low visibility, low ceiling, turbulence and wind shear, snow and icing, as well as thunderstorms.

Weber et al. (2005) suggest the following elements to be considered when developing a more effective air traffic management (ATM) approach specialised on thunderstorm appearance:

- **Forecasts** of thunderstorm position, intensity, movement and height and its further development in fine time steps are needed for the relevant time span of the next six hours

- **Capacity models** convert weather forecasts into time-varying estimates of capacity reductions for all affected areas of ATM, including uncertainty bounds for those events

- **Strategy generation tools** automatically provide optimal ATM strategies that account for the current weather situation by means of rerouting traffic or minimising the delay on airborne and/or ground

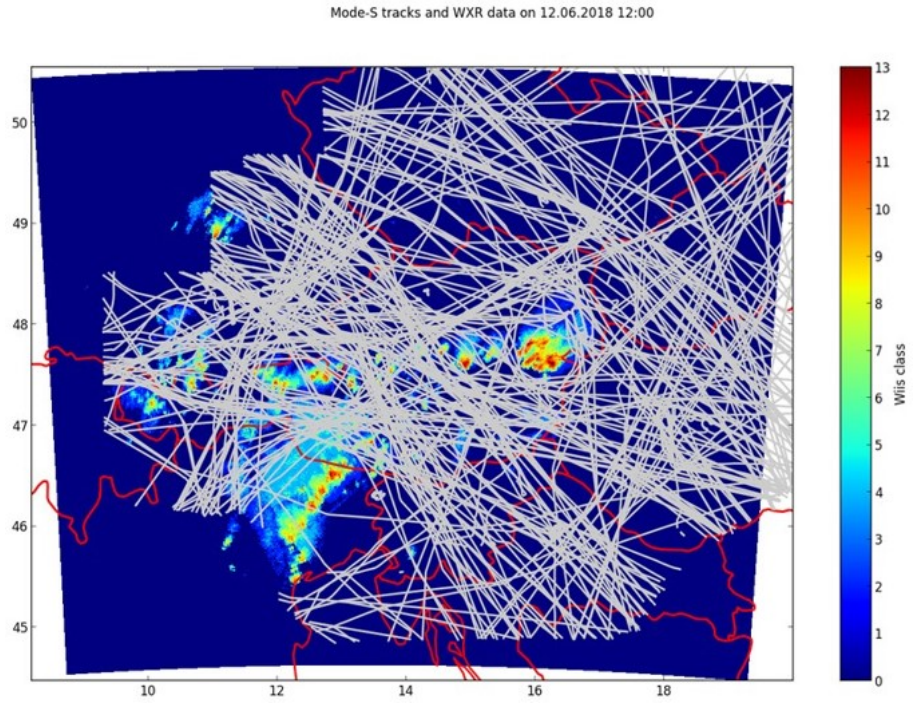

Figure 1: Flight trajectories around areas of pronounced thunderstorm activity show that large detours are necessary to avoid the hazards posed by thunderstorms. Numerous flights are affected. The WXR classes, shown in colour, represent the reflectivities measured by the weather radar (WXR). Here, red represents the highest radar reflectivities, an indication of significant thunderstorm activity. Source: Johannes Sachsperger, MeteoServe Wetterdienst GmbH

– **Airspace capacity enhancements** allow flexible planning of aircraft rerouting without exceeding workload or capacity limits of aircraft or controller to maintain safety in complex situations and high traffic density

Based on the points mentioned above, the objectives of this thesis can be derived. The core task is the automation of thunderstorm forecasts including uncertainties, which will subsequently be used as input for capacity models and for the generation of strategies for complex convective weather situations. To improve forecast quality, high-resolution ensemble weather prediction models are employed. The reason for using such models is carried out in the next sections.

1.2 Application of high-resolution ensemble forecasts

Weather affects many processes in aviation and can have a broad range of consequences, from flight diversions, to congestion at airports, to serious incidents involving aircraft. Due to the different phenomena, the challenges for ATM and meteorologists are manifold. Hazardous conditions like low visibility, low ceiling, wind gusts and turbulence occur frequently in conjunction with convective weather, often even before lightning or hail occurs at all. Many meteorological phenomena that are most disruptive to air traffic are also highly dependent on small-scale processes and even microphysics and therefore require consideration and implementation in forecast models with high spatial resolution. While many high resolution models are able to process various aspects of convection on grid scale, microphysical processes still need to be parameterised.

Schellander-Gorgas et al. (2017) compared the convection parameterising ensemble model ALADIN-LAEF (11km horizontal resolution) with the convection-permitting limited area model AROME-EPS (2.5km horizontal resolution) and concluded that AROME-EPS outperformed ALADIN-LAEF by providing more accurate precipitation forecasts. The study also noted that the advantage of AROME-EPS was especially greater in mountainous areas than in lowlands and that the model offered a better representation of the diurnal convective precipitation cycle. For high precipitation thresholds, the high-resolution model also showed an advantage (Schellander-Gorgas et al., 2017). Similar results were obtained by Weusthoff et al. (2010), who examined three convection parameterising models and their high-resolution counterparts and concluded that the high-resolution models performed better or at least equally well and that the advantages lay particularly in the prediction of convective small-scale precipitation events.

Wastl et al. (2021) compared the convection-permitting limited area model C-LAEF (2.5km horizontal resolution) of ZAMG with the coarser model ALADIN-LAEF (11km horizontal resolution) and ECMWF's global ensemble forecast model IFS-ENS (18km horizontal resolution). The paper concludes that C-LAEF outperforms ALADIN-LAEF and IFS-ENS in all surface variables examined and performs better than ALADIN-LAEF in all upper-air variables assessed. According to the authors, the advantages lie in the higher resolution and an explicit computation of deep convection. Regarding precipitation forecasts, all three ensemble forecast models showed similiar forecast skill for lower precipitation thresholds ($\leq$ 0.5mm/h), while C-LAEF outperforms the other models for higher thresholds ($\geq$ 1.0mm/h) for most leadtimes (Wastl et al., 2021). The study of Schellander-Gorgas et al. (2017) comparing AROME-EPS with ALADIN-LAEF showed that ALADIN-LAEF triggered precipitation too early in the daily precipitation cycle.

The fact that high resolution models outperform models with convection parameterisation is of great interest, especially in the light of the challenge of the complex Austrian topography. This work therefore aims to emphasise the advantage of high resolution forecast models over coarse or even global forecast models.

However, as discussed in Kain et al. (2008), the further increase of grid resolution towards 1km and higher brings up the question whether there is still sufficient added value in doing so. Two configurations of the Weather Research and Forecasting model (WRF) with different resolution were compared, namely WRF4 (4km horiztonal resolution, 35 vertical levels) and WRF2 (2km horizontal resolution, 51 vertical levels), but despite WRF2 showing more detailed features, Kain et al. (2008) conclude that it may be questionable, whether these results justify a tenfold computational effort.

With the computational power rising, the quality of forecasts can also be improved by employing ensemble forecasts. Since thunderstorm forecasts involve considerable uncertainties despite the high-resolution forecasts, an ensemble forecast approach helps to visualise these uncertainties. This allows decision makers in ATM to assess the forecast quality and to plan alternative scenarios in order to increase safety and efficiency. Further information on ensemble forecasts is provided in section 3.1.

1.3 Contribution of this thesis

The present work is intended to facilitate the work of ATM by supporting weather-related decisions with a probabilistic weather forecast. The main questions are the following:

- Deterministic forecasts are extended by a probability of occurrence to indicate possible alternative events. Do probability forecasts have higher forecast skill than deterministic forecasts?
- Ensemble Prediction models of different resolution are compared. Do high resolution convection-resolving models show an significant advantage over coarser resolved convection parameterising forecast models?
- Which of the compared models has the highest forecast skill? Can a multi-model ensemble consisting of different models further improve forecast quality?
- How skillful are forecasts for events that lie one day in advance?
- Airports and the surrounding air space are organised in flight sectors. Therefore, it is of particular interest to ATM to obtain forecasts for the individual sectors. However, these have a different size. How useful are forecasts based on these sectors?
- Can thunderstorm forecasts be automated in order to provide ATM a qualitative first-guess on which they can base their decisions?

This work relates to the meteorological tasks for the implementation of automated capacity forecasts and decision making tools and serves as a preliminary work for further research activities, which subsequently also deal with details from aviation and the activities of ATM. The final goal of this work is to select a suitable NWP model, to evaluate the forecast quality and to highlight probabilistic methods for forecasting thunderstorm events.

1.4 Related Work

This master thesis was realised in the framework of the project "Probabilistic MET information for the Capacity Optimisation in Arrival and Departure Management" (PROB4LOWW), which aims to preprocess meteorological data for ATM to harmonise workflow and decisions in the case of adverse weather like reduced visibility, icing or thunderstorms. The potential of probabilistic weather forecasts for application at Vienna International Airport (LOWW) is to be investigated. The prediction uncertainty input is intended to provide additional support for ATM decisions.

Probabilistic fog forecasts have been conducted at Vienna International Airport by Kneringer et al. (2018) with the result that the forecasts of an ordered logistic regression model (OLR) could keep up with the predictions of a human forecaster. The Meteorological Development Laboratory (MDL) of NOAA has developed a probabilistic thunderstorm guidance Localized Aviation MOS Program (LAMP) which produces forecasts for lead times from 1 to 25 hours with MOS data obtained from the 3km convection permitting High Resolution Rapid Refreshing Modell (HRRR) which is updated on a hourly basis with observations from lightning data and radar reflectivity (Charba et al., 2019).

Kicinger et al. (2012) conducted a feasibility study for an Integrated Airport Capacity Model (IACM) at Hartsfield-Jackson Atlanta International Airport, in which deterministic weather forecasts were replaced by ensemble forecasts. The integration of probabilistic information showed advantages in the feasibility study and attested the airport a 5% higher estimated capacity than benchmark values (Kicinger et al., 2012). Further calibration and validation studies showed that the IACM predictions for departure and arrival rates under different weather conditions could be considered accurate, provided that the traffic demand is properly scaled and the runway configuration was correctly set (Kicinger et al., 2016).

2. Description of the data

The performance of three different ensemble NWP models is examined in this thesis. A short description of data quality and availability will follow. For the verification of model data, grid observational data is required. Therefore, the output of an automatic cell detection based on lightning and radar data, CONVOY cell tracking, is used.

2.1 Observational data

CONVOY (CONVection detection and nOwcasting sYstem) is an automatic state-of-the-art cell tracking system that uses weather radar data and lightning data. Developed by MeteoServe Wetterdienst GmbH, it combines the properties of precise lightning data from the Austrian Lightning Detection and Information System (ALDIS) network and the radar reflectivity from the radar network covering Austria and its neighbouring countries. For the CONVOY cell tracking, the maximum column values of radar reflectivity on a two-dimensional field of equal elevation (maxCAPPI, Constant Altitude Plan Position Indicator) were used. Cells are detected using a threshold of 38dBZ on radar reflectivity or by the presence of lightning activity. Each 5 minutes the cell tracking system generates a polygon for each cell, defined by geographical coordinates and including meta information such as maximum radar reflectivity and lightning activity. A bright band filter reduces false detections in bright bands and intense stratiform precipitation. An example for CONVOY cell tracking is seen on Figure 2.

The combination of these different data sources enables the user to evaluate flashing and non-flashing cells. An important advantage over plain lightning data is that cells can be detected before they trigger lightning. This also meets the requirements of air traffic. While the detection of lightning in the vicinity is primarily relevant for the safety of ground personnel, other accompanying phenomena such as turbulence, gusts, hail and icing, which can already be present before lightning strikes, are also important for the processing of take-offs and landings. Furthermore, the horizontal dimension of thunderstorm cells is also easier to identify using radar reflectivity. Since many of the Area Control Centre (ACC) sectors cover neighbouring countries, the Austrian radar network is not able to cover all of these sectors, as can be seen in Figure 3 (left panel). For this reason, radar stations of the surrounding countries were included into the composite, see Figure 3 (right panel).

Radar data can occasionally be influenced by phenomena such as shadowing effects, which poses a downside of using radar data in Austria, as it lies within the complex topography of the European Alps. Especially in the area of Salzburg and Eastern Tyrol shadowing effects clearly stand out in the cell statistics. Fortunately, this shadowing effect mainly affects few particular airspace sectors that can be taken out of statistics to minimise the effect. The addition of lightning data provides further detection capabilities, however, it can not completely clear out the shadowing effects, as non-flashing cells in these areas are still missing.

Many other weather services have been developing their own cell tracking systems. A similar concept, CONRAD (CONvection in RADar products), is currently in use by the DWD (Deutscher Wetterdienst), where the cells are derived from 2D radar composite data on a 5 minute interval from the German radar network (Lang, 2001; Wapler et al., 2012). For the primary cells a threshold of 46dBZ is applied and several parameters acquired that describe the cell and its temporal development (Wapler et al., 2012). Another project of DWD is the so called CellMOS, which integrates radar data and lightning data into model output statistics (Wapler et al., 2012).

A case study on the application of a cell tracking system based on radar data and lightning density has been carried out by Steinacker et al. (2000) and shows clear correlations between the trajectories of the convective cells and the mountain ranges and valleys in the Alpine region.

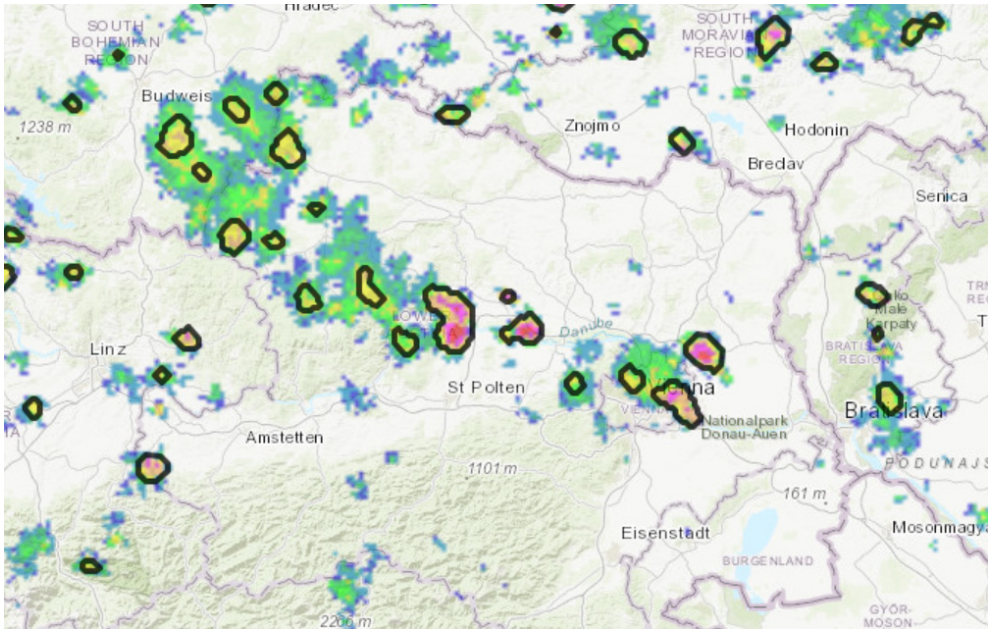

Figure 2: An example of a CONVOY cell overlay over a field of radar reflectivity. The image is from 16.08.2019 at 11UTC, a day with pronounced convection over Lower Austria. Quelle: MeteoServe Wetterdienst GmbH

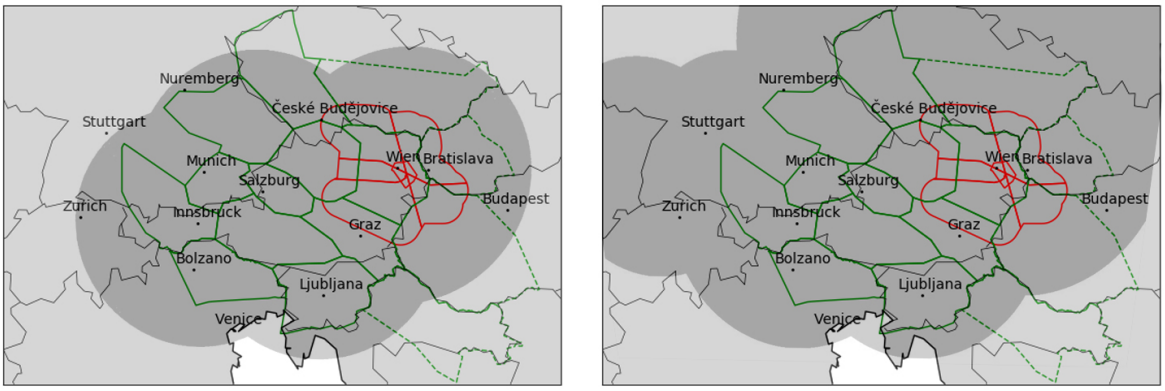

Figure 3: Overview of the areas covered by maxCAPPI radars (left) and extended maxCAPPI (right), with the sectors of Area Control Centre (ACC, green dotted lines) and the sectors of Approach (APP, red solid lines) displayed. Source: MeteoServe Wetterdienst GmbH

2.2 Forecast data

This section presents the models used and highlights the differences in model features.

2.2.1 ALADIN-LAEF model

The ALADIN-LAEF (Aire Limitée Adaption dynamique Développement InterNational – Limited Area Ensemble Forecasting) model is an meso-scale ensemble prediction model put into operation in 2011 by ZAMG (Central Institute for Meteorology and Geodynamics) as part of the Regional Cooperation for Limited-Area modelling in Central Europe (RC LACE) (Belluš et al., 2019). Based on the spectral limited-area model ALADIN, it is hydrostatic and uses hybrid vertical coordinates, semi-implicit timestepping, semi-Lagrangian two-time-level advection scheme, and a set of advanced multiphysics schemes for radiation, turbulent transport, deep and shallow convection and cloud physics among others (Wang et al., 2010, 2011). ALADIN-LAEF offers a horizontal resolution of 11km and 45 vertical levels, it is operationally run twice a day (00/12 UTC) up to a lead time of 72 hours and it has got one unperturbed control member and 16 perturbed ensemble members (Belluš et al., 2019). Further configuration details can be found in Table 1.

According to Wang et al. (2011), the model was initially used purely for the dynamic downscaling of the global ECMWF (European Centre for Medium-Range Weather Forecasts) ensemble forecast, but in 2009 it has been extended by additional features (Wang et al., 2011): The large-scale uncertainties that come with the ECMWF-EPS singular vectors are blended with small-scale perturbations from ALADIN breeding. Secondly, a multi-physics approach was chosen and ensemble members were grouped together with different physical approaches. Finally, a so-called non-cycling surface breeding (NCSB) was implemented to generate the surface initial conditions (IC), because other methods for generating initial conditions where not disturbing surface parameters like surface temperature or soil moisture and therefore were under-dispersive (Wang et al., 2011, 2010). NCSB created the initial conditions by running a short-range forecast which was driven by atmospheric forcing (Belluš et al., 2019).

The current version of ALADIN-LAEF employs an ensemble surface data assimilation (ESDA) which creates the inital conditions by randomly perturbing observations consistent to the observation error statistics and in addition also disturbing the underlying model physics within the data assimilation system (Belluš et al., 2016). ESDA uses the ALADIN surface data assimilation method CANARI (Code d'Analyse Nécessaire à ARPEGE pour ses Rejets et son Initialisation, (Giard and Bazile, 2000)), which is based on the Optimal Interpolation (OI) method (Belluš et al., 2016, 2019).

According to Weidle et al. (2015) CANARI is embedded in the breeding-CANARI-blending cycle, a data assimilation cycle that processes perturbations and data assimilation for the model. It breeds perturbations that have been adapted from its coupling model ECMWF-EPS on scales of ALADIN-LAEF and performs a surface data assimilation creating an surface analysis ensemble using randomly perturbed observations (Belluš et al., 2016). To obtain upper-air perturbations of the initial conditions, a spectral blending is conducted, combining ALADIN-LAEF-scale perturbations and the global-model-scale perturbations of ECMWF (Belluš et al., 2016). The ALADIN-LAEF ensemble members are coupled with the first 16 ECMWF-ENS members and are processed using different combinations of the aforementioned ALARO/ALADIN physics settings (Belluš et al., 2016).

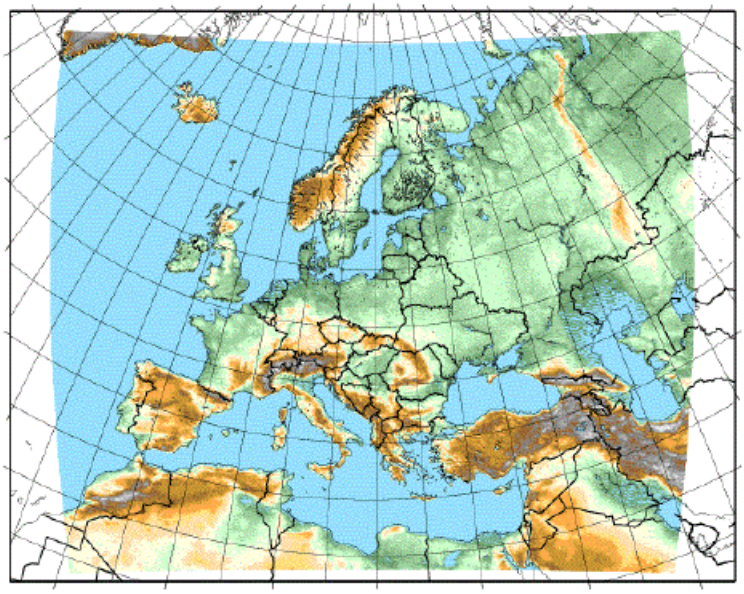

Figure 4: The ALADIN-LAEF forecast domain covers large parts of europe, Source: WMO (2017)

Availability

The ALADIN-LAEF run at 0UTC was kindly provided by ZAMG for the months of May to September for the years 2018 and 2019. The domain of ALADIN-LAEF is depicted in Figure 4.

2.2.2 COSMO-D2-EPS model

COSMO-D2-EPS is an operational ensemble prediction system of the DWD, which was developed in the COSMO consortium (COnsortium for Small-Scale MOdelling), an association of European weather services (Baldauf et al., 2017). According to the authors, it is designed for short range forecasts with a maximum forecast lead time of 45h for the 3UTC run and maximum lead times of 27h for every other run (0,6,12,15,18,21UTC) and has got 65 vertical layers and a terrain-following hybrid coordinate system. COSMO-D2-EPS offers a horizontal grid resolution of about 2.2km and is therefore a convection-permitting non-hydrostatic model that is able to compute deep convection on grid (Baldauf et al., 2017). Shallow convection is, however, a sub-scale process that still needs parameterisation (Baldauf et al., 2017). Further model specifications can be found in Table 1.

Baldauf et al. (2017) state that the generation of the 20 ensemble members in COSMO-D2-EPS is achieved by variation of :

- Lateral boundary conditions
- Initial Conditions
- Soil moisture and sea surface temperature
- Model Physics

The initial conditions for the computation of COSMO-D2-EPS are provided by the local ensemble-based data assimilation system KENDA (Kilometre-scale Ensemble Data Assimilation) which was developed for COSMO (Baldauf et al., 2017). The data assimilation system is based on the concept of

local ensemble transform kalman filter (LETKF) and includes a separate deterministic analysis (Baldauf et al., 2017). This analysis uses the Kalman Gain matrix of the ensemble mean, as the ensemble mean itself is not a well balanced initial state for a deterministic forecast, namely when the ensemble members are not Gauss distributed and this would subsequently lead to an increased spin-up time (Schraff et al., 2016; Baldauf et al., 2017). KENDA creates analyses by comparing observation data and information from 1-hour short forecasts which are run as an ensemble of 40 members and provide the most likely state of the atmosphere as well an estimation of the analysis error (Baldauf et al., 2017). The most likely state is the ensemble mean of the analysis forecasts, the estimation of the analysis error contributes for the spread of the actual ensemble forecast (Baldauf et al., 2017). The first 20 members of the analysis ensemble are used as initial conditions for the COSMO-D2-EPS ensemble (Baldauf et al., 2017). In addition, explicit random perturbations are applied to sea surface temperatures and soil moisture, corresponding to a spread of 1K and 15% relative soil moisture, respectively, to prevent under-dispersion of the ensemble, as is explained in more detail in Baldauf et al. (2017).

Referring to Baldauf et al. (2017), the lateral boundary conditions are passed from ICON-EU (deterministic COSMO-D2) and from ICON-EPS (COSMO-D2-EPS), whereby the ICON-EPS is refined for this purpose, but not to the same resolution as the deterministic ICON-EU. Furthermore, for covering the full COSMO-D2-EPS ensemble the first 20 members of ICON-EPS are chosen randomly and to keep up the supply of boundary conditions within the assimilation cycle, ICON-EU/ICON-EPS must therefore be computed every three hours. It is important to note that COSMO-D2-EPS uses the ICON predictions of three hours ago (Baldauf et al., 2017). A comparison of all the models used is given in Table 1.

The perturbations in model physics are set in such a way that 12 parameters related to parameterized processes, e.g. flat convection, turbulence or microphysics, are varied in such a way that the predictions thereafter receive a qualitative added value (Baldauf et al., 2017). This added value should consist in providing forecasts other than the standard configuration while keeping the forecast quality constant, i.e. avoiding outliers. Which parameters are disturbed is decided by a random procedure, which should nevertheless maintain a certain proximity to the standard configuration (Baldauf et al., 2017).

Availability

COSMO-D2-EPS data is available free of charge on the DWD's opendata server [1]. Since according to Baldauf et al. (2017) a spin-up of 3 to 6 hours is to be expected due to the interpolation of the driving model ICON and most thunderstorm activity is usually expected in the afternoon hours, the 3 UTC run was chosen. This run also offers a longer forecast duration of 45 hours, which would also allow a forecast for the following day. For this project, the data was collected in a sub-area of the domain that is sufficient for the project, from June 20, 2019 to September 30, 2019.

A disadvantage of the model is its choice of domain, because the eastern domain border of COSMO-D2-EPS is only about 100km away from the Austrian border and therefore some flight sectors which should be evaluated are cut off, see Figure 5. The adjustments to this downside are described in Section 3.3.

[1] Open data server access and further legal information can be found here: `https://www.dwd.de/EN/ourservices/opendata/opendata.html;jsessionid=F4A0E2B1043217A4EED9C7C5C33A726D.live11042`, last visited: 17.05.2020, 11:07

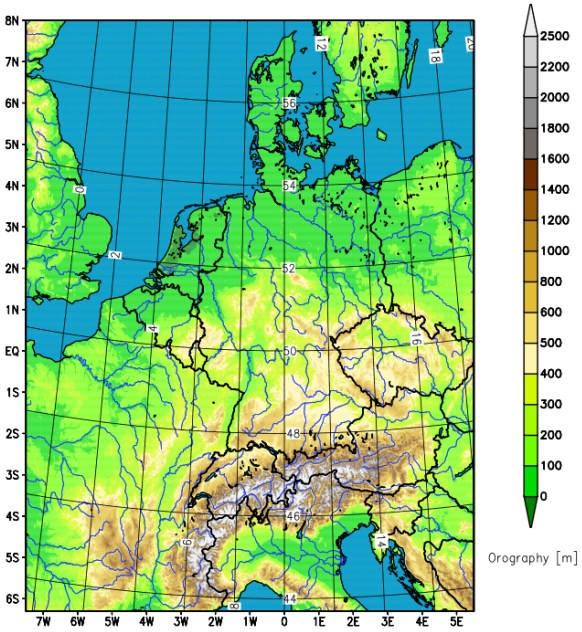

Figure 5: The COSMO-D2(-EPS) forecast domain covers mainly Germany and Austria as well as parts of neighbouring states. Source: Baldauf et al. (2017)

2.2.3 C-LAEF model

C-LAEF (Convection permitting Limited Area Ensemble Forecasting) is a convection-permitting, non-hydrostatic model operationally running at ZAMG since November 2019 (Wastl et al., 2020). The model is based on the spectral limited area model AROME (Seity et al., 2011), which was developed in the ALADIN consortium and is still being enhanced (Wastl et al., 2019). Four forecast runs per day are computed, with the 0UTC and 12UTC runs providing a forecast duration of 60 and 48 hours respectively, the runs at 6UTC and 18UTC are used to maintain the 6-hour assimilation cycle (Wastl et al., 2021). C-LAEF offers a horizontal resolution of 2.5km, has got 90 vertical levels and consists of 16 perturbed members and one control run (Wastl et al., 2019). The code for the adiabatic part was mostly adapted from ALADIN/AROME, while the physics package was implemented from the Mesoscale Non-Hydrostatic research model (MESO-NH, (Lafore et al., 1998)) (Keresturi et al., 2019; Wastl et al., 2019).

For the perturbations Keresturi et al. (2019) attempt to combine small-scale perturbations of the limited area model (LAM) with the large-scale perturbations of the host model ECMWF-IFS, namely its first 16 ensemble members. Since inconsistencies of the initial conditions and the lateral boundary conditions can lead to disturbing gravity waves, the jk-blending method used at ZAMG is applied (Keresturi et al., 2019). In addition to the upper-air parameters assimilated using a 3D-VAR based ensemble data assimilation, the surface parameters are also adjusted and disturbed by an ensemble surface data assimilation, ESDA (Wastl et al., 2021). See Wastl et al. (2019) and Wastl et al. (2021) for further information.

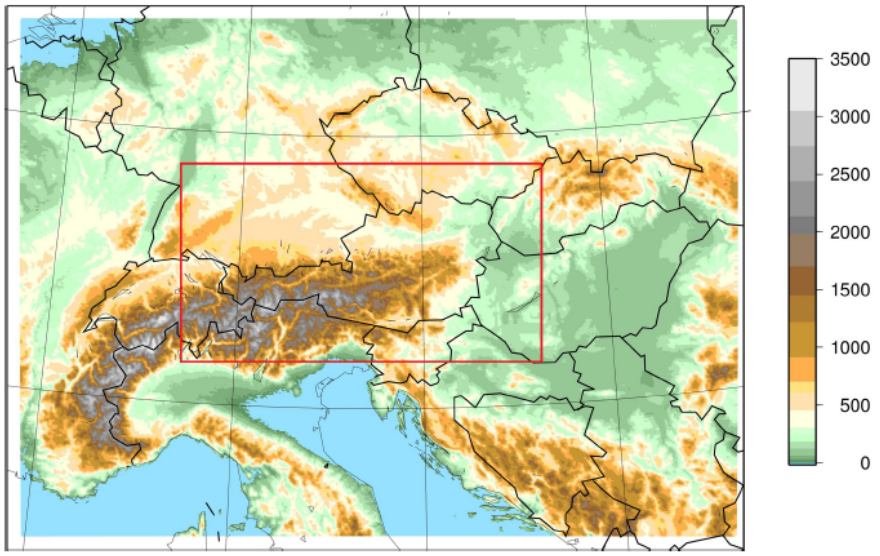

Figure 6: The C-LAEF domain mainly covers the alpine regions and is sufficient for the projects requirements. The red frame is not relevant to the study. Source: Wastl et al. (2019)

Model	ALADIN-LAEF(*1)	COSMO-D2-EPS(*3)	C-LAEF(*4)
Ensemble Size	16+1	20+1	16+1
Horizontal Resolution	11km	2,2km	2,5km
Vertical Resolution	45 layers	65 layers	90 layers
Runs/Day	2 (00,12 UTC)	8 (00,03,06,09,12,15,18,21)	4 (00,06,12,18 UTC)
Forecast Range	+72h	+45h (03 UTC) +27h (00,06,09,12,15,18,21)	+48h (00,06 UTC) +6h (12,18 UTC)(**)
Output Frequency	1h	1h (some variables 15min)	1h
Coupling-Model	time-lagged ECMWF-EPS (Singular Vectors, First 16 Members)	ICON-EU	ECMWF-EPS (lagged)
Coupling Update	6h	1h	3h
Model time step	450s	20s	60s
Assimilation	Ensemble Surface Data Assimilation (ALADIN's CANARI based on OI) (*2)	KENDA (including analysis ensemble)	OI (Surface) 3DVAR (Upper-air)
Perturbation	Breeding-CANARI-blending method, Multi-Physics Scheme Upper-air blending ECMWF SV	KENDA (based on regional LETKF) LBC	ESDA, 3DVAR, LBC, Model Physics

*Table 1: A general comparison of the model properties. Sources: (*1): Weidle et al. (2015), (*2): Belluš et al. (2019), (*3): Baldauf et al. (2017), (*4): Wastl et al. (2019), (**): Configuration of C-LAEF at time of investigation. Regular runs according to Wastl et al. (2019) are: +60h (0 UTC), +48h (12 UTC), +6h (6,18 UTC)*

Availability

The C-LAEF run at 6UTC was kindly provided by ZAMG from 9.July until 30.September 2019. As it was not in operational use then, the 6UTC run also provided 48 hours of forecast lead time. The domain of C-LAEF is depicted in Figure 6.

This study evaluates the forecasts for the time period in which all three models were available, see Figure 11.

3. Methods

3.1 Probabilistic Forecasts

Over many decades, numerical weather forecast models have been continuously improved and extended by new functionalities. The forecast quality increased significantly, however, even the best forecast models are always subject to errors. This is because the models need to represent the initial state of the atmosphere as accurately as possible. However, it is obviously impossible to determine the atmosphere in all its details and this will remain so in the future. Since numerical prediction models calculate numerous non-linear processes, under certain circumstances forecast errors from initial or boundary conditions can grow rapidly and the predictions can sometimes change considerably as the time integration progresses.

The basis for such findings was already founded by Edward Lorenz in 1963 (Lorenz, 1963). He computed a non-linear system of equations, the Lorenz model, for an ensemble of randomly varied initial values. The result of Lorenz's experiment showed the importance of small changes in the initial conditions, which under proper circumstances can increase significantly with time integration. The Lorenz model represented a simplified version of the atmosphere and showed that some initial conditions gave very well predictable results, while others led to rather unlikely solutions (Palmer et al., 2006).

As Figure 7 depicts, different sets of initial conditions can undergo different growth of errors. While in a.) the development is rather clear to understand, the development in b.) at some point splits up between the two possible solutions. In c.) the split takes place even earlier and leads to a rather unforeseeable behaviour within a few steps of temporal integration. According to Palmer et al. (2006), the radius of the circles is some measure of expected initial error.

Since any prediction with just one set of initial and boundary conditions corresponds to only one possible atmospheric state among many, it is not possible to give any information about the probability of a prediction coming true. Such deterministic prediction models are thus fixed on a possible course of events which would not change even when recalculated. If, however, the initial conditions are changed in each run, as was also the case in the experiment of Edward Lorenz, an ensemble of possible prediction courses is obtained (Lorenz, 1963).

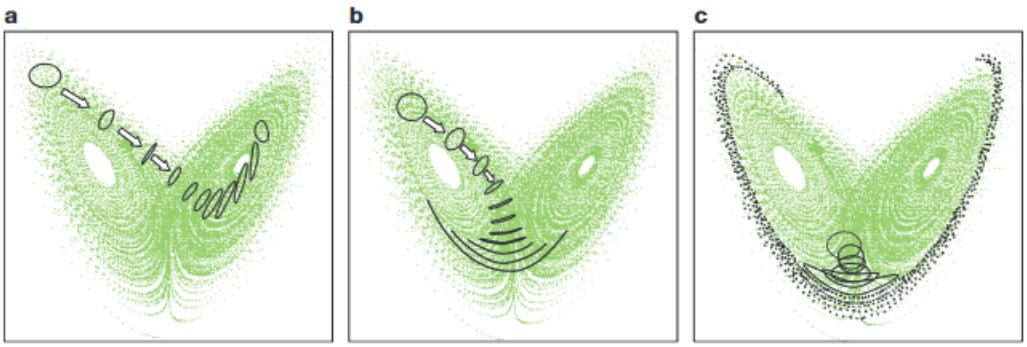

Figure 7: Flow dependent growth of initial uncertainties in a nonlinear system, the Lorenz model. Source: Palmer et al. (2006)

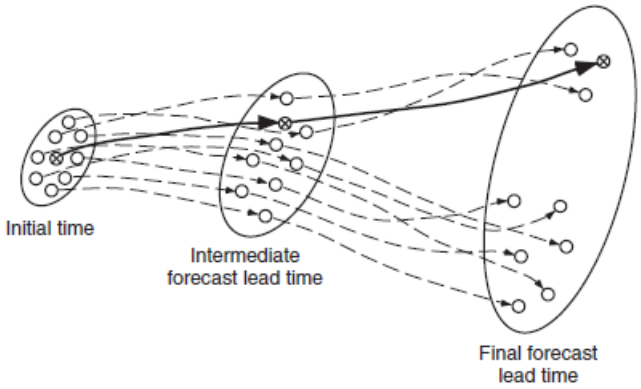

Figure 8: Schematic representation of the most realistic analysis (bold line) and its development through forecast lead time among other ensemble members (dashed lines). The ellipses correspond to the probability distribution of the ensemble at the given points in time. While in the intermediate forecast lead time step the probability distribution tends to get broader, in the final forecast lead time step the development splits up into two sets of realisations similar to each other. Source: Wilks (2011b)

3.1.1 Disturbances of Initial Conditions

Initially, so-called "Poor man's ensembles" were preferred due to the cost-effective use of computing capacity. These ensembles consist of several independent (deterministic) forecasting systems of different weather services, which are jointly evaluated as an ensemble. According to Ebert (2001), a major advantage is the fact that the ensemble obtains its uncertainties through initial conditions and model formulation as well, because the models use different input data, analysis and forecast methodologies and model components to produce forecasts. This reduces the vulnerability of the ensemble forecast for model bias and errors that can cause underdispersion, i.e. the variability of a forecast is lower than that of empirical observations, which is a common problem in ensemble prediction models (Ebert, 2001).

In order to understand how an ensemble of initial conditions is created, one should imagine the following concept described by Wilks (2011b): Assume the atmospheric initial state and its uncertainty as a finite sample in a probability distribution. A number of these points are forming a cloud distributed around the mean estimated state of the atmosphere which form our ensemble of initial conditions. These points were picked randomly and each one represents a plausible state of the atmosphere. Instead of predicting the temporal evolution of the whole initial probability distribution, only the trajectories of those selected initial states are calculated (Wilks, 2011b). Each representative state is dynamically integrated separately and the distribution of the trajectories at each point in time serves as an approximation to the integration of the whole probability distribution, which is the concept of the Monte Carlo method (Wilks, 2011b; Bott, 2016).

Figure 8 shows the development of the ensemble of forecasts when integrated. Like in the Lorenz Model, some forecasts in the ensemble bifurcate in their development and represent an alternative state of the atmosphere. Having understood that uncertainties are dealt with by integrating some slightly different initial states over time, it is important to know how to best select these initial states. Therefore, a closer look at the generation of the meteorological fields on which a prediction is based is necessary. To achieve the number of data points needed to feed a model, the available observational data are far from sufficient and unevenly distributed over the Earth. Observational data hence is

weighed against a background forecast ("first-guess") in order to fill areas with sparse observational data, which is performed by data assimilation.

3.1.2 Data assimilation

In the basic principle of data assimilation a NWP forecast is interpolated to the locations of available observations in order to obtain a model "first-guess" or also called background field x_b, with $H(x_b)$ representing the observation operator responsible for the interpolations between model and observation space (Kalnay, 2002). The difference between the original observations and the model "first guess" is called "observational increment" or "innovation". By adding these innovations depending on certain weights W to the background field leads to the definition of the analysis field (Kalnay, 2002):

$$x_a = x_b + W[y - H(x_b)] \tag{1}$$

The last years have seen significant developments in numerical data assimilation. The most commonly used approaches are statistical analysis and variational analysis, but recently also methods using Kalman filters become of interest in atmospheric sciences. The individual concepts are explained in brief below.

A type of statistical analysis is the optimal interpolation method (OI), which in principle applies a weighted least squares method to fit the first guess fields to the observations (Daley, 1991; Raith, 2008; Kalnay, 2002). The "optimal" weights are dependent on the error covariance matrix of observations and the first guess of the forecast model (Daley, 1991; Kalnay, 2002). However, statistical methods exhibit the problem that they still show small-scale inconsistencies in the fields, the so-called meteorological noise, which is the reason why analysis fields require some filtering in advance (Raith, 2008).

With increasing computing capacity the method of variational analysis becomes of greater interest. In this analysis method the observations and background field are assimilated by iteratively minimising a cost function and thereby approximating the solution (Bouttier and Courtier, 2002). 3D-VAR includes the forecast as well as all observations at one point in time, while the more recent 4D-VAR analysis extends the assimilation process to include all observations within a time period (Bouttier and Courtier, 2002). Thus the temporal development of observations and forecasts is taken into account, which better reflects the dynamic development of the model (Thepaut et al., 1996). However, this is at the expense of the computing capacities and it must be noted that the observations need to be waited for before the 4D-VAR analysis can be started, which may delay the generation of analysis fields (Bouttier and Courtier, 2002; Valur Hólm, 2018).

Another data assimilation method uses a Kalman filter to sequentially assimilate background forecasts and observation data (Evensen, 1994). The Kalman filter creates the analysis in two steps, as described in Vetra-Carvalho (2018): The forecast step creates a forecast from the previous analysis and evolves the mean state of the forecast as well as its covariance. Then the Kalman Gain is calculated in order to incorporate the observations into the new analysis. Finally in the update step the Kalman Gain Matrix acts as weight function to control the influence of the new observations on the background forecast. In the update step both the mean state and the covariance are being updated with the Kalman Gain. However, this only works for linear processes and the atmosphere is highly non-linear (Vetra-Carvalho, 2018).

The Extended Kalman filter was developed to process non-linear models, but it did not solve the problem around costly matrix operations regarding the forward integration of the forecast background covariance matrix (Vetra-Carvalho, 2018; Evensen, 1994). Instead of explicitly evolving mean state and covariance matrix, an ensemble of states is employed in the frame of Monte Carlo method and the mean and covariance is taken from the evolved sample (Vetra-Carvalho, 2018; Evensen, 1994). This Ensemble Kalman Filter (EnKF) still has a significant disadvantage as it is susceptible to undersampling due to a low number of ensemble members (Petrie, 2008). According to the paper, using localisation increases the rank of the forecast error covariance and hence the effective ensemble size. Further developments lead to the local ensemble transform Kalman filter (LETKF), which is used in a specialised version called KENDA-LETKF at the DWD to assimilate data for the COSMO-D2-EPS model (Baldauf et al., 2017). For further information on KENDA-LETKF please refer to Schraff et al. (2016).

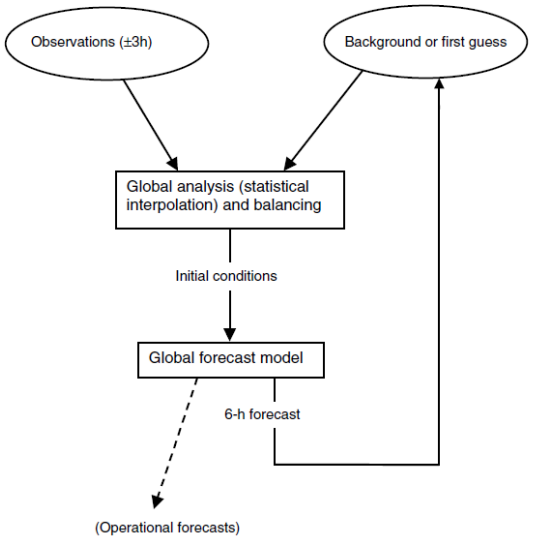

Figure 9: Data assimilation cycle as depicted in Kalnay (2002). Observations are weighed against a background forecast in order to fill areas with sparse observational data. The resulting analysis field serves as the basis for forecasts.

Data assimilation is usually carried out at regular intervals and takes place in so-called assimilation cycles, as shown in Figure 9. These regular assimilations maintain the update cycle to provide recent background forecasts for the next cycle.

3.2 Basics of convection physics

3.2.1 Physics of convective phenomena

Convection, or more precisely atmospheric convection, is the transport of properties by means of some fluid. Meteorologists use the term convection only for the buoyancy related vertical transport of fluid and denote non-buoyant mass transport as advection (Emanuel, 1994; Doswell III, 2001). A more detailed definition can be found in Emanuel (1994).

If the phase changing processes of water play an important role in the convection process, it is referred to as moist convection, which leads to the buildup of convective cumulus clouds. As con-

vection reaches far into the atmosphere, the process is called deep moist convection (DMC), which often results in thunderstorms (Doswell III, 2001). A brief definition of thunderstorms can be found in section 3.2.2

Convective motion is established, when an air parcel is brought out of its thermodynamic balance and gets lifted from the level of free convection (LFC), that is the level, from where buoyant forces allow a further rise of the air parcel. This motion continues up to a certain point, where its density equals that of the surrounding atmosphere, the equilibrium level (EL). The difference between the virtual temperatures of the air parcel and that of the atmosphere integrated from LFC to EL is defined as the convective available potential energy (CAPE) (Markowski and Richardson, 2011):

$$CAPE = \int_{LFC}^{EL} B dz \approx g \int_{LFC}^{EL} \frac{T'_v}{\overline{T_v}} dz \tag{2}$$

B denotes the mass specific buoyancy force (N/kg), LFC and EL are the level of free convection and the equilibrium level respectively. T'_v is the deviation of the virtual temperature of the rising air parcel from the virtual temperature of the surrounding environment $\overline{T_v}$.

CAPE is proportional to the area between the atmospheric temperature profile and that of the air parcel and is defined as positive value with the unit J/kg. The strength and depth of this vertical motion is determined by the distribution of CAPE in the atmosphere. CAPE is necessary for the development of convection and may be used as an indicator to where thunderstorms may form, but it is not sufficient to their development. In order for the air parcel to reach the LFC at all, it often needs to overcome a stable layer of air close to the ground. The energy needed to reach this area is called convective inhibition (CIN) with the unit (J/kg) and is computed similarly to CAPE, but with a negative sign before the integral as this energy needs to be expended (Markowski and Richardson, 2011):

$$CIN = -\int_{0}^{LFC} B dz \approx -g \int_{0}^{LFC} \frac{T'_v}{\overline{T_v}} dz \tag{3}$$

For the convection initiation, three factors are crucial (Doswell III, 1987):

- moisture
- instability
- lifting processes

With the presence of moisture and the atmosphere being unstable and both accounting for CAPE, a certain cause of lifting is needed in order to overcome CIN. Especially in Austria, the orography plays an important role in triggering convection, as results of Bertram and Mayr (2004) suggest. Lifting can also be initiated by phenomena that change the temperature profile of the atmosphere like cold fronts or even outflow boundaries, which are produced by the cold outflow from thunderstorms in the vicinity (Doswell III, 2001). Lifting can further be provided by a convergence line, where the air is forced to rise due to mass conservation. If the differential heating is strong enough, convection can also appear without dynamic forcings, as the stable atmospheric layer near ground is diminished. A more detailed examination of lifting mechanisms is presented in Doswell III (2001).

3.2.2 Prediction of thunderstorms

Thunderstorms are high-reaching and intense atmospheric overturnings that can reach up into the lower stratosphere. They are always accompanied by lightning and a number of meteorological phenomena such as gusty winds and heavy precipitation as well as hail (Bott, 2016). In this study, thunderstorms are identified by their characteristics, namely the occurrence of lightning and the high radar reflectivity inherent in thunderstorms.

Thunderstorms are small-scale events and are strongly connected to microphysical processes that need high resolution representation. Therefore high resolution models are necessary. The great sensitivity to small-scale phenomena makes them difficult to predict and forecasts hence tend to have low probabilities to be correct for a specific location. Deterministic predictions cannot address these uncertainties, which is why ensemble forecast models are used, which can also depict rare events.

For the use of convection parameterizing models like ALADIN-LAEF, over years convection indices were used to determine thunderstorm events and the convective available potential energy (CAPE) was considered. However, convective indices have the disadvantage of taking into account only a few atmospheric levels and therefore lacking a detailed representation of the atmospheric profile. A possible alternative is CAPE, which is an integral over the atmosphere. But further information of the model is needed to whether thunderstorms are triggered or not as CAPE only represents a potential for convection.

A simple concept was developed for ALADIN-LAEF, which was compared with convective indices, in order to find areas of likely thunderstorm formation and triggering within the model: Three conditions were linked to each other, which should indicate the occurrence of thunderstorms. First, the radius of 5 grid points was checked for the presence of at least 200J/kg CAPE at one of the grid points. This includes a radius of about 55km, which takes into account the fact that thunderstorms consume the CAPE in their direct vicinity and therefore the CAPE is often very small at the direct location of the storm. Although thunderstorms require CAPE to form, the amount of CAPE is not directly related to the probability of the formation of thunderstorms, which is why the threshold value was chosen low. Secondly, the presence of precipitation was assumed via the parameter "total precipitation". The threshold value was again chosen low with 1mm/h. Since ALADIN-LAEF with a coarse resolution can reflect geographical factors as well as meteorological details only to a limited extent, the precipitation amounts for convective events are comparatively low and a too high threshold value would leave out too many thunderstorm events. Thirdly, a lower cloud base was chosen as an indication of the occurrence of thunderstorms. Based on temperature and dew point at 2 meters above ground, the spread * 125m was calculated, which is a comparatively good approximation for the cloud base and is considered sufficient for these purposes. In this case a cloud base of 1000m or less was chosen as the threshold.

Due to the high resolution of DWD's COSMO-D2-EPS and ZAMG's C-LAEF, calculated parameters were not used in these models and instead built-in parameters were used, which are directly available as model output. This saves additional calculations and instead shows where thunderstorms trigger in the model. In COSMO-D2(-EPS) the parameter "dbz_cmax" is available for this purpose, namely the simulated maximum radar reflectivity (measured in dBZ) in the entire model column, for which a threshold value is then also selected (Baldauf et al., 2017). C-LAEF provides a parameter called "surfflashdiag" which represents the simulated flash density in flashes/km^2.

Regarding the comparability, using the method applied in ALADIN-LAEF for all three models had no significant advantage, as CAPE in particular is calculated differently within each model and thus

comparability would be limited anyway. For the sake of a simpler calculation, built-in parameters were used.

3.3 Verification of probabilistic forecast models

Air space sectors

Due to their small-scale structure and the large prediction uncertainty arising from the many micro-physical processes involved, the predictive skill of a point thunderstorm forecast tends to be low even for a few hours of forecast lead time (Bouttier and Marchal, 2020). To reduce double-penalty issues caused by location errors, for certain applications it proves useful to verify forecasts of such rare events on defined areas rather than at a single point.

In air traffic, thunderstorms affect not only activities on the apron, but also aircraft on approach to the airport. Therefore, the verification in this work is not carried out by means of spatial verification but by treating forecast and observed thunderstorms within airspace sectors as binary events. However, the airspace and approach sectors have different sizes, which makes comparability difficult. These sectors can be seen in Figure 3. Of these sectors, the easternmost sectors are excluded due to the domain boundary of the COSMO-D2-EPS model over western Hungary. Out of the remaining sectors, several groups were assembled based on their size. The composition of these groups can be seen in Figure 10.

Time period of the studies

The studied time period is determined by the availability of the forecast models. The C-LAEF model was still under development at the time of the investigations and could not be delivered until the beginning of July. Technical issues also resulted in occasional gaps in the time series, which is why only specific days were used on which all models and the observational data were available. The time period for a common study ranges from 9th July 2019 until 30th September 2019 (see Figure 11). To conserve server memory, only certain daytimes were stored and therefore the forecast times used are from 10UTC to 21UTC. During this time period, the highest load at airports and the statistically highest thunderstorm activity overlap during the course of the day. As a study by Neuwirth (2012) shows for an area around the city of Salzburg, the highest lightning activity in terms of cloud-to-ground lightning flashes appears mostly within the daytimes selected for this study, as can be seen in Figure 12. The study area in Neuwirth (2012) includes alpine area as well as alpine foothills on the northern edge of the alps, which is similar to the setting of the flight sectors investigated and therefore seen as representative for this investigation.

If the conditions for thunderstorms were met, e.g. the thresholds exceeded on a certain grid point, the forecasts were given the value 1, otherwise 0. If thunderstorm conditions were met in the observation data within a grid point, the observations were treated likewise. The sum of thunderstorm grid points forecasted or observed were stored along with the total sum of grid points of every sector (e.g the total sector size as a number of grid points).

In the next step, binary values for forecasts and observations are to be derived. If one or more grid points within a sector show forecasted/observed thunderstorms, then the forecast/observation is True for the whole sector. This is conducted for each ensemble member separately and finally the

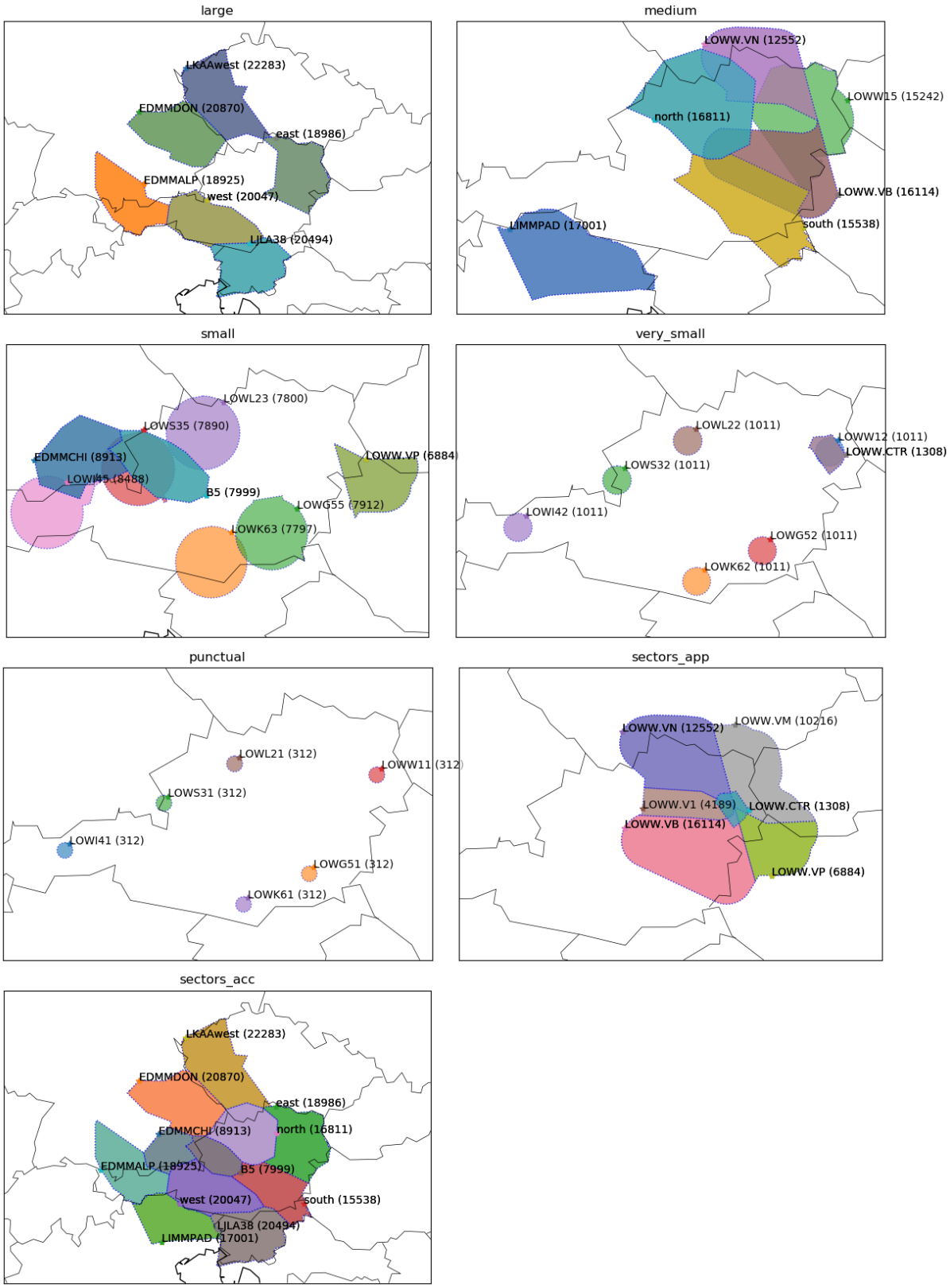

Figure 10: The sectors are grouped by their area size, which is given in the brackets in the unit km^2. In addition, the two sector groups "ACC" and "APP" at Schwechat Airport are listed, which have very different area sizes.

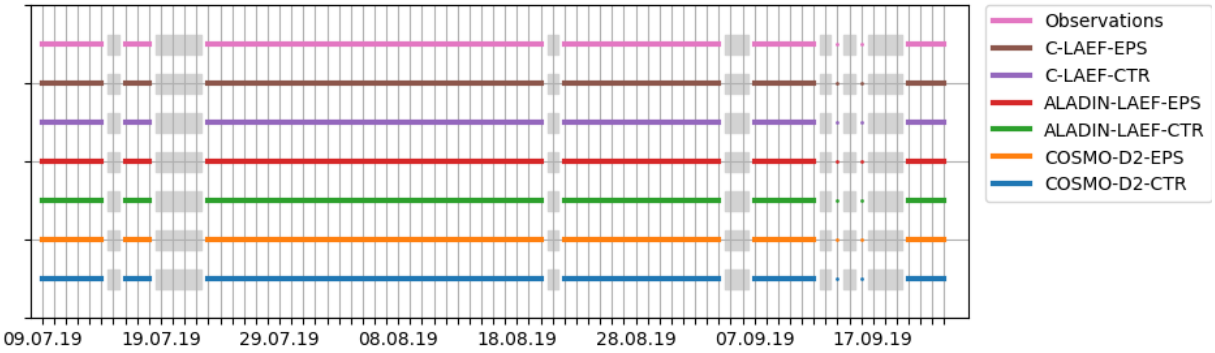

Figure 11: The simultaneous availability of all data (forecasts and observations) is limited to the period from 9.July 2019 to 23.September 2019 (64 days available in total). Due to data issues there are some gaps within the dataset. These days where excluded for all models (grey areas).

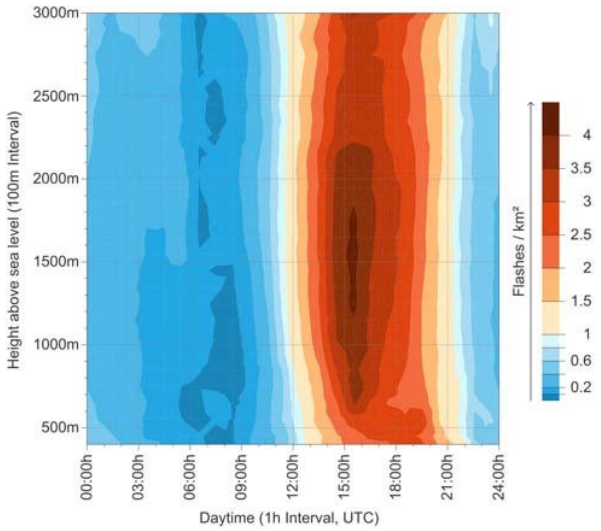

Figure 12: This figure shows the mean lightning activity (cloud-to-ground lightning flashes only) by time of day in the study area situated around Salzburg. Source: Neuwirth (2012)

sum of ensemble members predicting thunderstorms was divided by the sum of all disturbed ensemble members. The deterministic run was excluded from this process. At the end of data processing, observations and the deterministic run are thus available as binary values, while the predictions of the ensembles are available as probabilistic values (between zero and one). These processed values are finally used to set up contingency tables, see Figure 13. Based on these evaluations the forecasts are assessed according to various measures, that will now be explained in a brief overview. For further details please refer to Wilks (2011b) and Jolliffe and Stephenson (2012).

3.3.1 Reliability Diagram

A contingency table as seen in Figure 13 assigns every event to one of four possible outcomes (if the forecasts are binary), namely if an event was forecasted (a,b) or not (c,d) and if an event occurred (a,c) or not (b,d). To get an overview on how well the predicted probabilities fit to the observed frequencies, a reliability diagram can be consulted. An example is depicted in Figure 14.

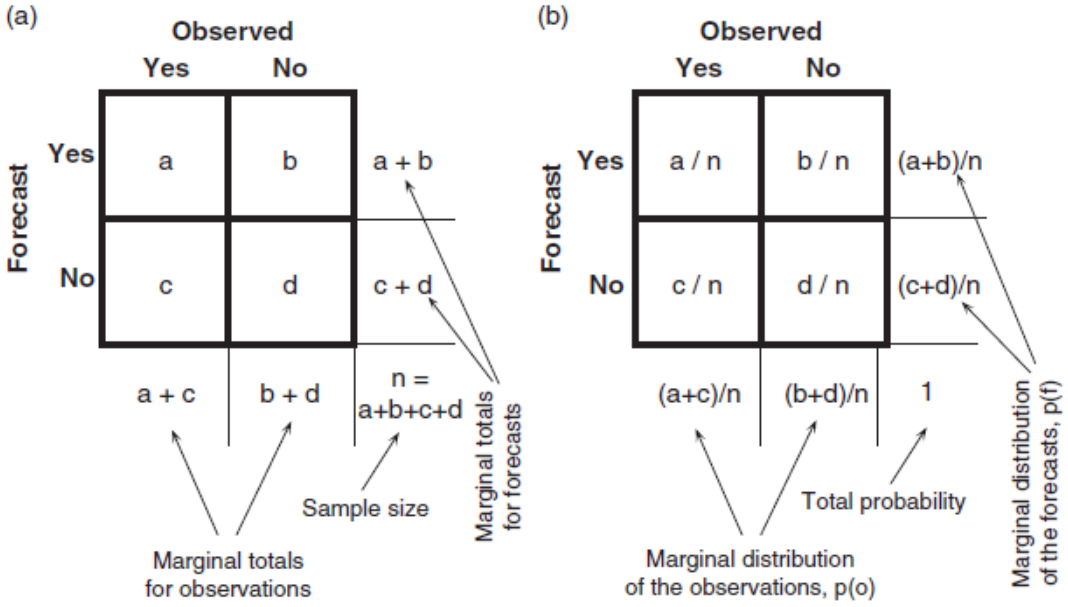

Figure 13: Contingency Table for dichotomous non-probabilistic verification (left) and joint distributions for probabilistic verification (right). Source: Wilks (2011b)

The observation data is supplied as binary data on every time step with 1 being an observed event and 0 being otherwise. The forecast data consists of values between zero and one depending on the number of ensemble members predicting the event. The reliability diagram is basically a plot of predicted probabilities over the relative frequency of observations. As the prediction probabilities are continuously distributed, the values must be arranged in bins B_k in order to allow a classification (Bröcker and Smith, 2007).

All forecast values X_i that fall into a specific bin B_k are stored by their indices in a collection of indices I_k (Bröcker and Smith, 2007):

$$I_k := \{i; X_i \in B_k\} \tag{4}$$

For every bin the mean prediction probability r_k is computed

$$r_k := \frac{\sum_{i \in I_k} X_i}{\#I_k} \tag{5}$$

with $\#I_k$ as the number of elements in a bin B_k. The relative frequency of observations f_k is calculated by dividing the sum of the observed events Y_i by the sum of events for each bin B_k (Bröcker and Smith, 2007):

$$f_k := \frac{\sum_{i \in I_k} Y_i}{\#I_k} \tag{6}$$

With all the points, a curve is constructed that allows various statements to be made about the model and its ability to cover the spectrum of observations. Using the reliability diagram, some important properties of a model can be assessed:

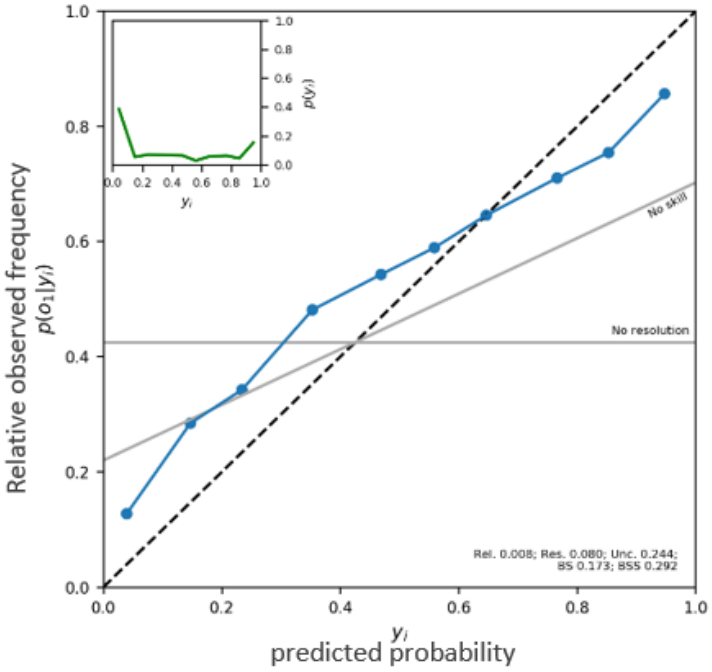

Figure 14: Example of a Reliability diagram with the predicted probability on the X-axis and the relative observed frequency on the Y-axis. The horizontal line presents the climatological base rate of an observed event. The diagonal line represents perfect reliability, the grey line halfway between perfect reliability and climatology-line is called the no-skill-line. The sharpness diagram in the upper left corner shows the distribution of forecasts over the given bins. It gives a clue on how sharp and confident a forecast is, but does not necessarily display the reliability of a forecast.

The **reliability** describes the distance of the curve from the diagonal, the perfect reliability. The closer the curve is to the diagonal, the better the predicted probabilities match the relative observed frequency. Reliability only checks for statistical consistency (Bröcker and Smith, 2007).

The **resolution**, which states whether a prediction can distinguish between forecast situations with different frequencies of occurrence (Wilks, 2011b). The flatter the curve, the lower the resolution of a forecast. A climatology forecast therefore has no resolution (WWRP, 2017).

The **sharpness** can be determined as the tendency of a forecast to predict extreme values (eg. values close to 100% probability or very unlikely events of almost 0%). However, a sharp forecast does not necessarily indicate a reliable forecast, but could also give unrealistic certainty (Owens and Hewson, 2018).

The reliability diagram is often accompanied with rank histograms, discrimination diagrams or sharpness diagrams. As seen in Figure 14 in the upper left corner, a sharpness diagram is used to visualise how many forecasts fall into which bin. If the distribution is U-shaped, the forecast has a high sharpness.

Figure 15 (a) shows an overview over the most common biases that may arise in a forecast. If the forecast indicates higher probabilities for an event than the frequency at which they will actually occur, the forecast has a wet bias (overforecasting). Otherwise, if the forecast probabilities are too low, a dry bias (underforecasting) is present. If the reliability curve is too flat (e.g. has a horizontal alignment) the forecast is said to have poor resolution or tends to be overconfident (also called underdispersion) which means that forecast probabilities more often tend to extreme values than the relative observed frequencies (Wilks, 2011b). This leads to a low ensemble spread, which in a well calibrated model im-

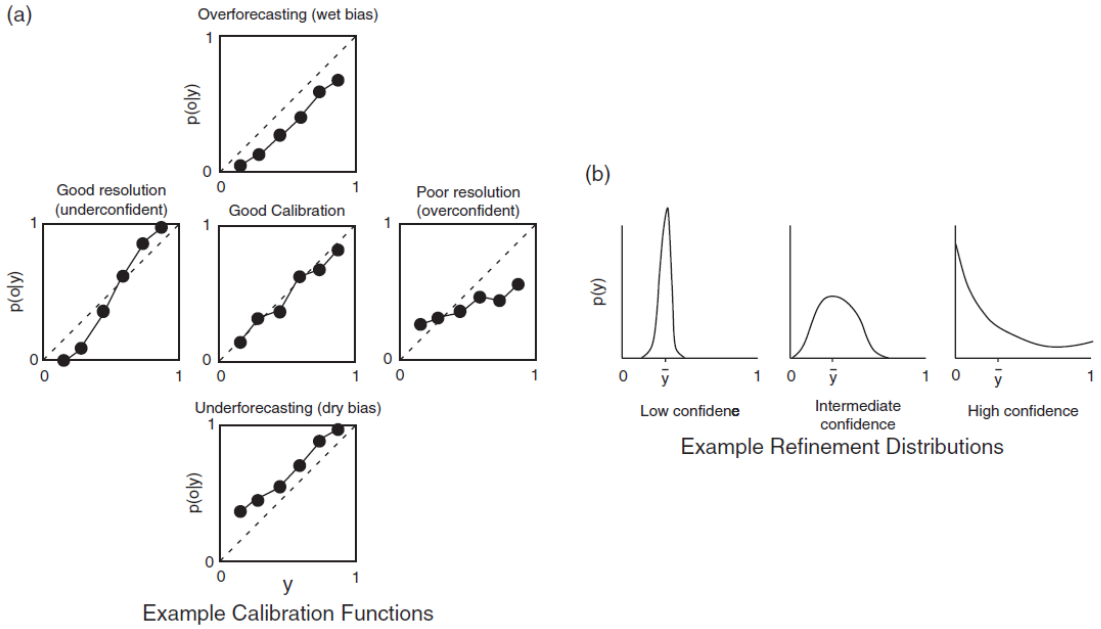

Figure 15: (a) Different types of bias in Reliability diagrams. (b) Forecasts distributed over all bins. Source: Wilks (2011b)

plicates a high confidence (Warner, 2010). However, the forecast reliability is low and thus the forecast gives a wrong certainty. Underdispersion poses a common problem to many ensemble forecast systems.

The opposite case, an underconfident forecast means high resolution, but also unrealistic overdispersion in the way that observations too rarely lie outside the ensemble because of high ensemble spread (Wilks, 2011a,b). This results in a low confidence and probabilities arranged around a certain mean. Different states of confidence can be taken from Figure 15 (b). Whereas low confidence means that forecast probabilities hardly ever reach extreme probabilities like 0 or 1, high confidence forecasts rarely show ambiguous probability values and often get close to 0 or 1 (Wilks, 2011a,b).

However, a reliability diagram is conditioned on the forecasts and only applies statistical measures in order to compare forecast probability and frequency of observations (WWRP, 2017). Therefore the reliability diagram is often used in combination with a relative operating characteristic (ROC) curve that provides information over how well a forecast can distinguish between event and non-event.

If predictive performance is to be determined in the form of a scalar value, the Brier Score (BS) is often employed. It is basically the mean squared error of a probabilistic forecast (Wilks, 2011b),

$$BS = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2, \tag{7}$$

where y_k denotes the forecast probability and o_k represents the binary observation.

As probability forecasts only have a range from zero to one the BS also is bound in this same range with zero being a perfect forecast and erroneous forecasts having higher values of BS (Wilks, 2011b). Murphy (1973) developed an algebraic decomposition that allows the determination of three central properties of a probabilistic forecast:

$$BS = \underbrace{\frac{1}{n}\sum_{i=1}^{I} N_i(y_i - \overline{o_i})^2}_{\text{"Reliability"}} - \underbrace{\frac{1}{n}\sum_{i=1}^{I} N_i(\overline{o_i} - \overline{o})^2}_{\text{"Resolution"}} + \underbrace{\overline{o}(1-\overline{o})}_{\text{"Uncertainty"}} \tag{8}$$

$\overline{o_i}$ represents the subsample relative frequency or average observation and $\overline{o}$ is the overall relative frequency or sample climatology. The number of forecasts within the same probability bin is denoted with N_i and y_i represents the forecast probability of each bin.

The BS may be an easy way to get a scalar result on the mean square forecast error of a forecast, however it has the disadvantage of being sensitive to the climatological frequency (WWRP, 2017): It is easier to get good BS values for events with a rare occurrence without having any real skill which may distort the results.

The Brier Score can also be defined as a Skill Score (Wilks, 2011b):

$$BSS = \frac{BS - BS_{ref}}{BS_{perfect} - BS_{ref}} = 1 - \frac{BS}{BS_{ref}} \tag{9}$$

BS_{ref} is the Brier Score of the climatology and $BS_{perfect}$ is the Brier Score of a perfect forecast (Wilks, 2011b). The Brier Skill Score (BSS) delivers the relative skill of a forecast to predict an event over the climatology. This poses an improvement over the BS as it takes the climatology into account. However, the BSS gets unstable when used upon small sample sets and especially rare events need bigger sample sets (WWRP, 2017).

The BSS is also connected to the reliability diagram. A perfect prognosis is associated with a BSS of 1 or a reliability curve directly at the diagonal line in Figure 16. Every reliability curve between the diagonal and the no-skill-line leads to a positive BSS and therefore poses an improvement over the reference forecast (Weisheimer and Palmer, 2014). Negative values of BSS arise if the forecast has less skill than the reference forecast.

3.3.2 Relative Operating Characteristic (ROC) and Area under Curve (ROC-AUC)

The relative operating characteristic is another graphical method for forecast verification, depicted in Figure 17. The ROC diagram does not contain full information on the joint distribution of forecasts and observations (Wilks, 2011b). Rather it measures the ability of a forecast to distinguish between events and non-events and therefore poses the opportunity to get information on the potential forecast skill of a model (WWRP, 2017).

The ROC diagram compares the hit rate H (or probability of detection, POD), which is the number of all hits by the number of all observed events

$$H = \frac{a}{a+c} \tag{10}$$

and the false alarm rate F (or probability of false detection, POFD), which is the number of all false alarms by the number of all non-observed events (Wilks, 2011b).

$$F = \frac{b}{b+d} \tag{11}$$

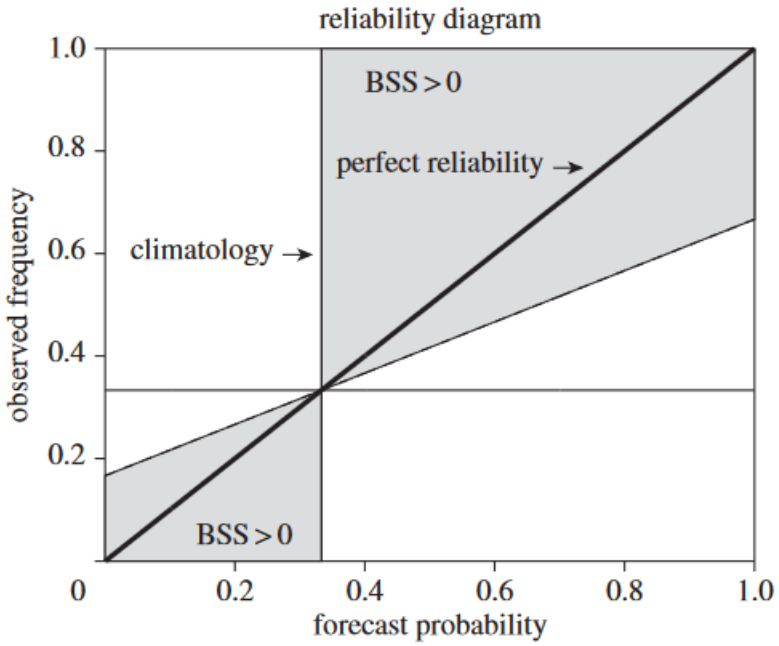

Figure 16: The Brier Skill Score (BSS) in connection with the reliability diagram. A perfect prognosis would be situated directly at the diagonal line. Every forecast better than the reference forecast (e.g. the climatology) would lead to a positive BSS and its curve woud lie within the grey area. Source: Weisheimer and Palmer (2014)

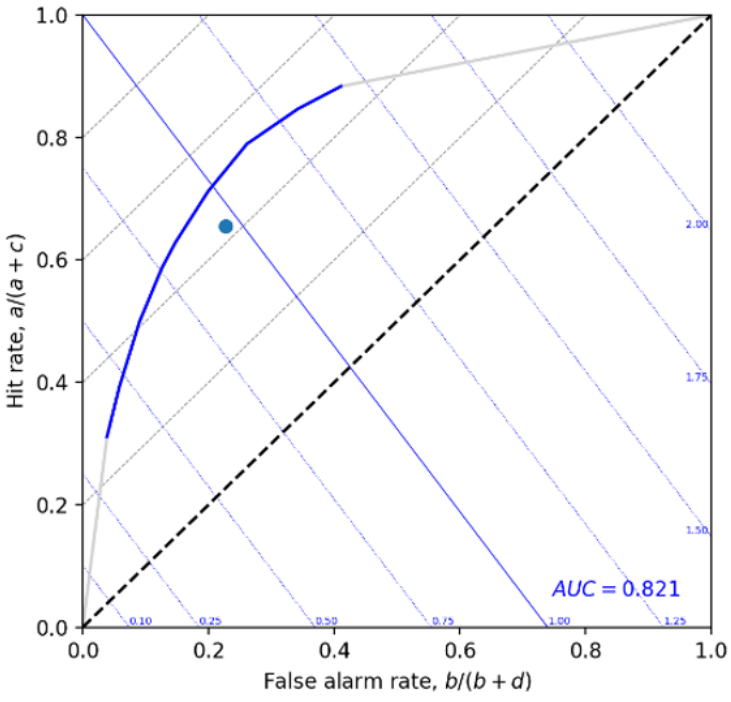

Figure 17: Example of a ROC-Diagram. It visualises the ability of a model to distinguish between events and non-events. This enables the user to measure the potential forecast skill.

A perfect forecast is indicated by a ROC-curve that leads vertically to the upper-left edge of the diagram and from this point horizontally to the upper-right point (WWRP, 2017). This is achieved, when the false alarm rate is zero for all bins and the hit rate is one for all bins, e.g. all forecasts were hits or

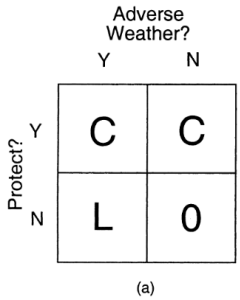

Figure 18: The different outcomes of a cost-loss decision model as an expense matrix. Regardless of whether the event has occurred, the costs (C) must be paid. If no protection is in place, the loss (L) occurs when the event occurs. Source: Wilks (2001)

correct negatives and there were no false alarms or misses. A useless forecast is indicated by a ROC-curve at the diagonal. It is not sensitive to forecast biases and therefore carries no information about the reliability. The diagram is conditioned on the observations and hence poses a good addition to the reliability diagram (WWRP, 2017).

For deterministic forecasts a point can be determined by means of hit rate and false alarm rate. If the point lies below the ROC curve, as in Figure 17, the deterministic forecast shows a lower skill than the ensemble forecast. Otherwise, it would perform better.

For probabilistic forecasts the outcome of a relative operating characteristic can be put into a specific value, the area under the curve (ROC-AUC). The area of a useless forecast is less or equal to 0.5, the area of a perfect forecast at 1. Note that a biased and unreliable forecast may still have resolution and therefore performs well in a ROC-diagram, which why a ROC-diagram is said to only measure the potential skill of a forecast (WWRP, 2017).

3.3.3 Potential Economic Value

With consistency and quality assessed using reliability diagram and ROC diagram respectively, Murphy (1993) defines a third kind of goodness that is important when evaluating the usability of a forecast: The economic value of a forecast is derived from a static cost-loss decision model that assumes that a protective action can prevent or mitigate potential harm from the event that is being predicted (Jolliffe and Stephenson, 2012). The economic value is strongly depending on the balance between the cost that the user pays for forecasts over a period of time and the loss that a user expects to incur if a damaging event occurs during that period when having no forecast product at hand (Murphy, 1993). As can be seen in Figure 18, hits and false alarms count to the costs (C), whereas misses represent the loss (L) that shall be eliminated by using forecasts.

The economic value of a forecast is defined by Jolliffe and Stephenson (2012) as the benefit of using a forecast compared to a reference (mostly the base climatology) for individual users or businesses:

$$V = \frac{E_{climate} - E_{forecast}}{E_{climate} - E_{perfect}} \tag{12}$$

The best possible outcome would require the user to have full knowledge of the weather development and on when to protect. The user would only need to protect if an event of climatological frequency (s) occurs (Jolliffe and Stephenson, 2012):

$$E_{perfect} = s\,C \tag{13}$$

This serves as an upper bound of how much can be saved using a forecast at this cost. The climatological expense as a reference value is defined by the decision on whether to protect or not. The decision maker therefore needs to weigh up which costs are higher for him: Either protecting always or risking the loss, provided that the climatological base rate of the said event is known to him. The aim is to find the best strategy for the user by minimising the expense (Jolliffe and Stephenson, 2012):

$$E_{climate} = min(C, sL) \tag{14}$$

The forecast expense is defined by the outcome of the forecast. Taking into account the contingency table and the expense matrix leads to the following expression:

$$E_{forecast} = C\frac{a}{n} + C\frac{b}{n} + L\frac{c}{n} \tag{15}$$

By transforming the equation using the definitions of hit rate and false alarm rate, the following form can be acquired (Jolliffe and Stephenson, 2012):

$$E_{forecast} = F(1-s)\,C - H\,s\,(L-C) + s\,L \tag{16}$$

Jolliffe and Stephenson (2012) use the expression $\alpha = C/L$ as the user specific cost loss ratio to define the economic value as:

$$V = \frac{min(\alpha, s) - F(1-s)\alpha + H\,s(1-\alpha) - s}{min(\alpha, s) - s\,\alpha} \tag{17}$$

The above equation also highlights the importance of the base rate in relation to the user's decision as there are two solutions depending on whether the cost-loss rate is higher or lower than the base rate:

$$E_{climate} = \begin{cases} C, & if \quad C/L < s \\ sL, & if \quad C/L > s \end{cases} \tag{18}$$

This makes sense in the way that the costs at a climatological frequency s do not exceed the loss value, otherwise the user would act uneconomically. In this context, the prediction system is finally compared to evaluate how high the benefit turns out to be quantitatively. The curves of both functions are plotted together in a so-called potential economic value diagram (PEV diagram), see Figure 19.

The potential economic value of some forecast describes the benefits compared to the reference forecast with a value of 1 representing the perfect forecast. In addition to the value, the width of the distribution is also important when evaluating forecasts. The broader the distribution, the more users can be served with their individual C/L ratios. The red curve in Figure 19 depicts a deterministic forecast. To assess the economic value of a probabilistic forecast, the potential economic value of all ensemble members is plotted (blue curves). The upper enveloping curve (black curve) represents the economic value of an ensemble forecast and is therefore always equal or greater than any ensemble member. The PEV diagram thus shows the range of potential usefulness for different customers.

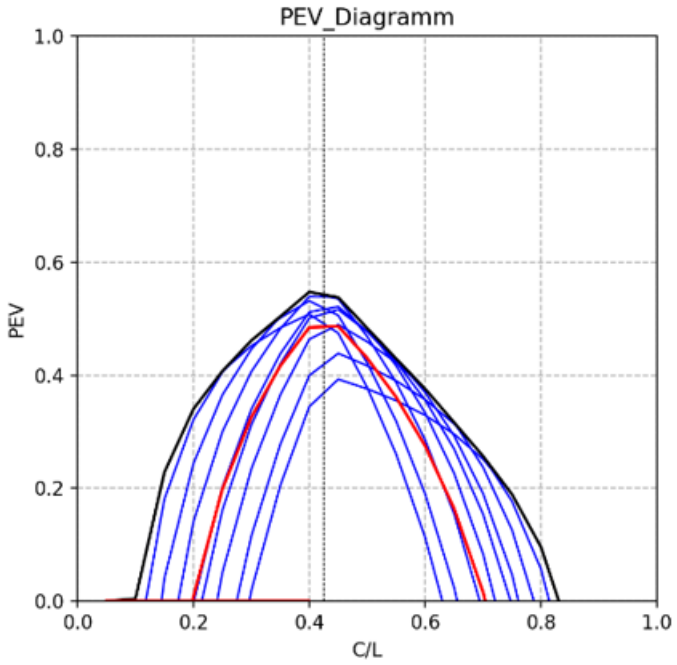

Figure 19: Example for the Potential Economic Value of a probabilistic forecast model. The blue lines display the potential value of single ensemble members, whereas the black enclosing curve shows the overall potential value of the whole ensemble forecast. The red line represents the accompanying deterministic forecast.

The best possible case for consulting a forecast model as protection from potential loss is the personal cost-loss ratio being equal to the base rate (Jolliffe and Stephenson, 2012). Towards smaller or larger C/L ratios it is more difficult to get some advantage over the climatology. C/L is always situated between 0 and 1 as $C > L$ would be useless to the customer and $C < 0$ would mean paying the customer to use the forecast. Figure 19 illustrates a PEV-diagram showing the potential economic value for different cost-loss ratios. Note that the PEV curves can also have negative values for some C/L ratios, which would mean that a potential customer would have a better outcome ignoring the forecast (Wilks, 2001).

As Wilks (2001) states, the ROC diagram and the PEV diagram partly overlap in their mechanics. It is further stated that the PEV diagram however has the advantage of showing conditional biases whereas the ROC curve does not. This circumstance provides a good reason to add further verification methods to the ROC diagram.

4. Results

The models were evaluated using the following combination of parameters and thresholds that were found empirically:

- ALADIN-LAEF-EPS
  - CAPE greater or equal 200 J/kg within a radius of 5 grid points
  - Total precipitation greater than 1mm/h, convective precipitation was not available
  - Cloud base less than 1000m (calculated from dewpoint spread)
- COSMO-D2-EPS
  - Simulated radar reflectivity ("dbz_cmax") greater or equal 30 dBZ
- C-LAEF
  - Flash density ("surfflashdiag") greater than 0.001 flashes/km^2)

The method for ALADIN-LAEF was developed to identify areas where thunderstorms have triggered according to the model, since convective parameters only provide information about the potential for convection. For COSMO-D2-EPS, the threshold for the simulated radar reflectivity was set to 30dBZ, although the threshold in CONVOY is significantly higher at 38dBZ. Comparisons with different thresholds (25, 30 and 35 dBZ) have shown the best forecast reliability for 30 dBZ, which indicates a dry bias in forecasts of simulated radar reflectivity. For C-LAEF, the threshold for the flash density was set to 0.001 flashes/km^2 because lower thresholds lead to background noise and higher values resulted in too many missed events.

4.1 Deterministic versus Probabilistic Forecasts

The question whether an ensemble forecast is more skillful than a deterministic forecast is of great importance as the computational effort is by far greater when running a forecast model several times. For this reason reliability, potential forecast skill and the potential economic value of all models were compared. For this evaluation the forecast data of the first forecast day (lead times 10h to 21h) for the the "ACC" sectors was used, see Figure 10 for an overview over the sectors. The forecasts were created using the above-mentioned parameter combinations.

The deterministic forecast is denoted as point within the ROC-diagram and as can be seen in Figure 20 (upper left panel), it lies within the ROC-curve, because its false alarm rate is higher than the false alarm rate of the corresponding class of the same hit rate in the probabilistic forecast. Hence, the probabilistic forecast shows more potential skill than the deterministic forecast.

The reliability diagram in Figure 20 (upper right panel) shows how well calibrated a forecast model is, that is, how accurately the forecast probability matches the observed frequency of the forecasted event. All examined model forecasts, especially the ALADIN-LAEF model forecast tend to have too small forecast probability compared to the observed frequency. This underforcasting or dry bias poses a conditional bias that may be corrected by calibration of the forecast, but which would go beyond the scope of this thesis.

The PEV-diagram in Figure 20 (lower left panel) reveals a higher potential economic value for the probabilistic forecast for all potential users also covered by the deterministic model and even expands the scope of possible users in terms of higher or lower C/L ratios for their specific needs. Generally speaking, all the ensemble prediction models investigated show higher skill and a higher PEV value than their deterministic counterparts regardless of their resolution. What stands out is the comparable skill of the low resolution ensemble forecast and the deterministic forecast of the high resolution model, as can be seen in the ROC diagram and the PEV diagram as well. However, the low resolution ensemble prediction model still covers a broader span of C/L ratios than the deterministic models.

The use of an ensemble prediction approach in fine-scale forecast models is not only owed to the effort of getting probabilistic information in addition to the forecast. When using high-resolution models one also has to very accurately describe the initial conditions and hence, forecast errors tend to grow faster with increasing lead time. The ensemble approach is hence necessary to address uncertainties and to fully exploit fine scale NWP models (Mittermaier and Csima, 2017).

4.2 Convection Parameterising versus Convection Resolving Models

When one examines Figure 20, one also notices the significant difference between ALADIN-LAEF on the one hand and C-LAEF and COSMO-D2-EPS on the other. This can mainly be attributed to the different resolution of about 11km (ALADIN-LAEF) and about 2km (C-LAEF, COSMO-D2-EPS).

Models of coarser grid scale like ALADIN-LAEF need to use parameterisation to take convection into account, with deep convection in particular being a significant source of error (Schellander-Gorgas et al., 2017). Instead of expecting perfect point forecasts, the focus when using high-resolution models is on a better statistical representation of convection (Weusthoff et al., 2010). When comparing high resolution models with their driving models of a coarser horizontal resolution, Weusthoff et al. (2010) found that the high resolution models outperformed or at least performed equal to the lower resolution models in the majority of spatial scales and different precipitation thresholds. This supports the statement of this paper, namely that the high resolution models produced better or at least equally skillful results on different sized sector classes.

One problem that arises explicitly from the low resolution of models like ALADIN-LAEF is the inability to resolve downdrafts (Davis et al., 2003). In weather conditions with weak synoptic forcing, the formation of cold pools by downdrafts is an important factor for the propagation of mesoscale convective systems (MCS) (Corfidi, 2003). Convective parameterisation schemes (CPS) compute convection in independent model columns, which results in cold pools getting represented poorly and thus the movement of the MCS occurs too slowly, resulting in higher displacement errors (Schellander-Gorgas et al., 2017). For ALADIN-LAEF, it is also suspected that deep convection parameterisation could be the reason for a premature triggering of convection over complex topography (Wittmann et al., 2010). As already mentioned the advantage of high resolution forecast models lies in mountainous regions with complex topography as these regions receive better representation (Schellander-Gorgas et al., 2017).

4.3 On the added value of a multi-model ensemble forecast

Ensemble forecast models require substantial amounts of computation capacity and therefore are expensive in operational use. To keep costs low it is common practice to produce either time-lagged

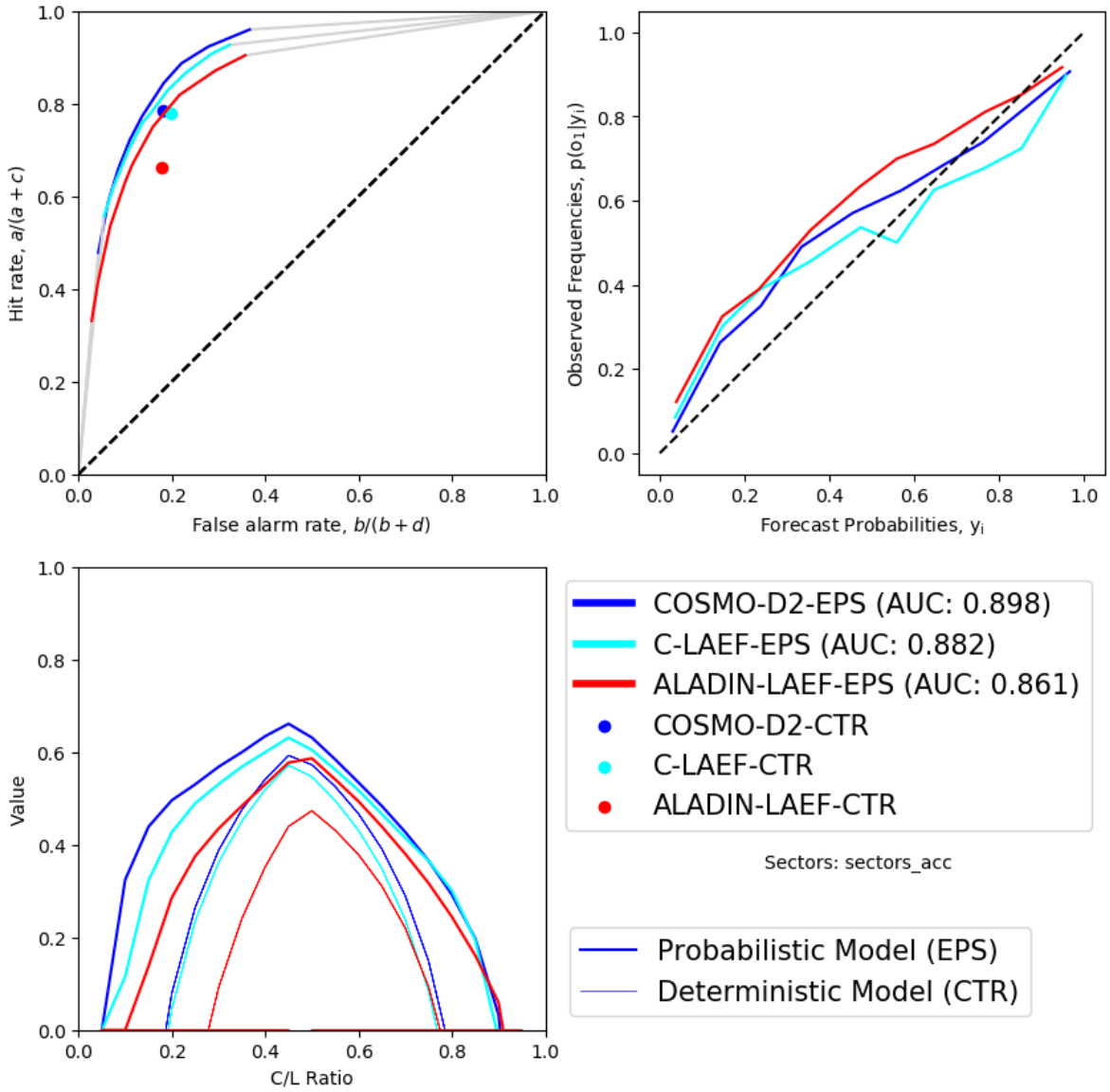

Figure 20: Relative operating characteristic (ROC, top left), reliability diagram (top right) and potential economic value (PEV, bottom left) in comparison for all models for the first forecast day (lead times 10h to 21h). In the ROC diagram, the points represent the deterministic models associated with the ensemble models in the respective colour. In the PEV diagram, the associated deterministic models are shown by thinner curves of the same colour.

ensembles by taking into account previous model runs or employing multi-model ensembles by incorporating ensemble forecasts from different models. The latter was examined in the context of this work. In this section, the forecast data for the "ACC" sectors on the first forecast day (lead times 10h to 21h) was used. The forecast methods for each model are presented at the beginning of chapter 4.

When comparing the ROC curves in Figure 21 (upper left panel), all combinations of ensemble forecasts examined in this study show comparable or higher skill than the high resolution single model ensemble forecasts. Except for the combination "ALADIN-LAEF-EPS + C-LAEF-EPS", which is outperformed by the COSMO-D2-EPS model at some thresholds, as can be seen on the ROC diagram and the PEV diagram. The combination of the two high-resolution models ("C-LAEF-EPS + COSMO-D2-EPS") outperforms the combinations of one high-resolution model with a model of coarser resolution. The best result, however, was obtained with the combination of all three models, which shows

slight leads in forecast skill and potential value for most bins in ROC-diagram and PEV-diagram respectively. Worth mentioning is the insight of Bouttier and Marchal (2020) that using an ensemble of significantly lower skill as part of a multi-model-ensemble does not necessarily lead to a higher overall skill. It was also stated that results improved notably when conducting calibration and post processing to the ensemble.

The reliability diagram in Figure 21 (upper right panel) reveals a significant dry bias of the ALADIN-LAEF-EPS forecasts for most probability bins. This also seems to affect the multimodel forecasts that include the ALADIN-LAEF-EPS ensemble members. The models C-LAEF and COSMO-D2-EPS also show a dry bias for low probability bins but to a lesser extent. The PEV diagram in Figure 21 (lower left panel) for $C/L < s$ shows a significant improvement, when ALADIN-LAEF-EPS ensemble members are joined by members of the high resolution models. Otherwise, there are no significant differences in economic value. The combination "C-LAEF-EPS + COSMO-D2-EPS" and especially the combination of all three models stand out with the highest values for potential economic value. For $C/L > s$, however, the combination "ALADIN-LAEF-EPS + C-LAEF-EPS" has a comparable economic value to "COSMO-D2-EPS + C-LAEF-EPS". This could be explained by the fact that for C/L ratios greater than the base rate, the economic value of the high resolution models is not significantly greater than that of ALADIN-LAEF-EPS.

4.4 Temporal development of Forecast Skill

For many applications, the forecast of the current day's weather development is of primary interest. However, it should also be reviewed how well NWP models can represent the following day, that is, forecasts with a lead time of more than 24 hours. In general, the prediction skill can be expected to decrease with increasing lead time, since errors such as in the initial and boundary conditions and in the model formulation propagate, grow progressively and subsequently reach ever larger scales. According to Mittermaier and Csima (2017) rapid error growth at or near grid scale has significant contribution to the higher rate of forecast decay in high resolution models. Therefore, high resolution NWP model benefit particularly from employing an ensemble approach.

Figure 22 shows the temporal development of the ROC-AUC value for the three forecast models investigated on the "ACC"-sectors. The forecasts were processed using the combination of parameters and thresholds mentioned at the beginning of chapter 4. The ROC-AUC values were computed for time intervals of three hours, e.g. the ROC-AUC value at a lead time of 12 hours was calculated using the forecasts from 10 hours lead time up to 12 hours to keep up the sample size. Clearly, the probabilistic forecasts of all three ensemble forecast models outperform their deterministic counterparts at all lead times investigated.

The ROC-AUC curves also show a diurnal variation in forecast skill for both days and for all models. Especially the values in the late evening feature a considerable decrease of forecast skill, while the forecast skill rises in the afternoon hours. Brunet et al. (2019) attributes this behavior to poor representation of nocturnal convection.

To assess the diurnal development of convective cells in the observations as well as the representation within the associated forecasts, the forecast and observation frequencies were compared with respect to the time of day. In Figures 23, 24 and 25 the areas represent the range of values reached by the individual sectors. The green areas correspond to the observed frequencies, the red areas to the deterministic forecast frequencies and the blue areas to the probabilistic forecast frequencies. Prob-

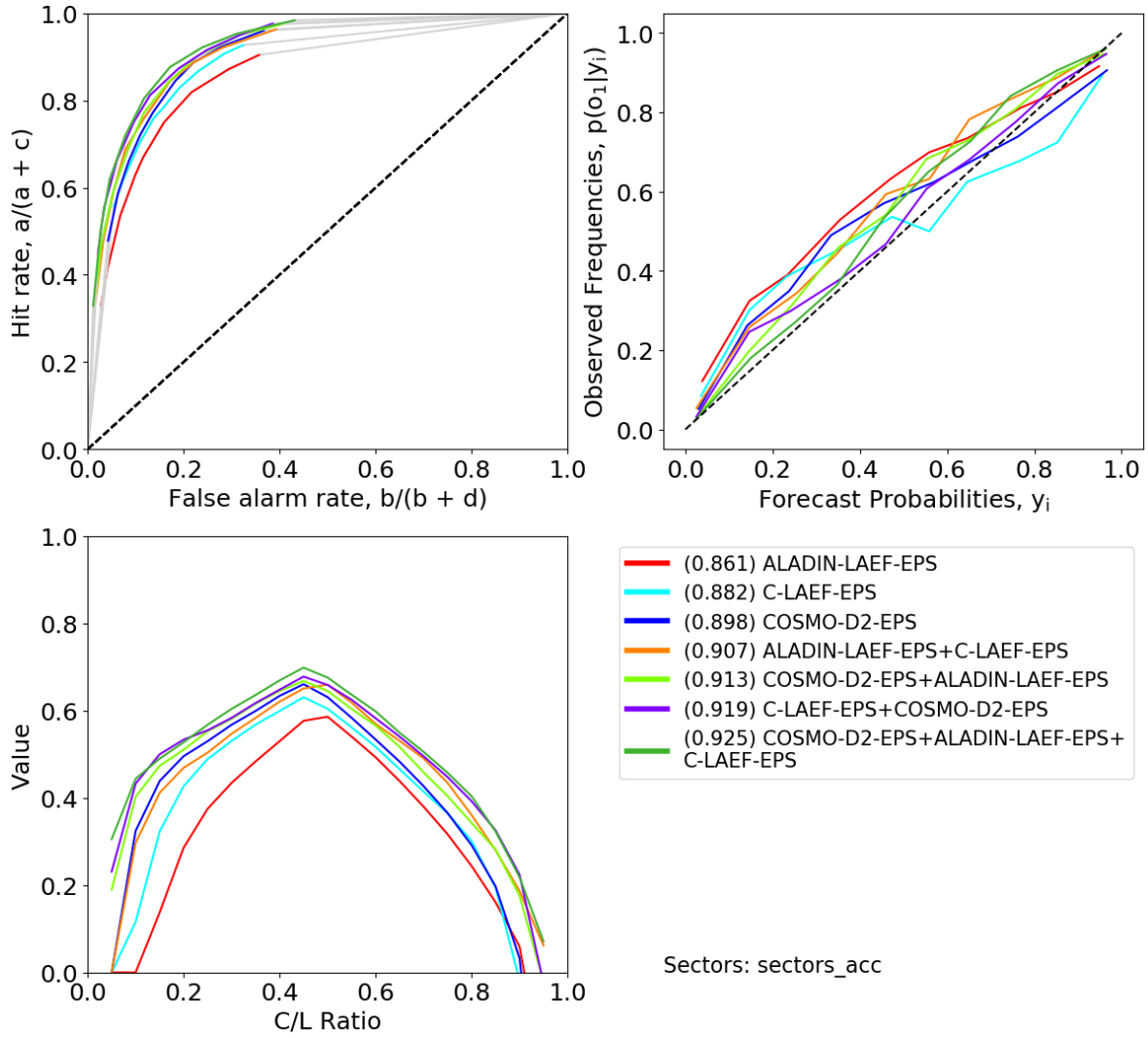

Figure 21: Multi model ensembles were created and compared against the individual model ensembles to highlight potential advantages. Especially the combination of models of high-resolution increased the potential forecast skill significantly. The best results for most probability bins were achieved combining all three ensemble forecast models. This is forecast data for the first forecast day (lead times 10h to 21h).

abilistic forecasts were counted as "Forecasted" if the percentage of ensemble members forecasting a convective cell within a certain sector is higher than the base rate of the event. Generally, forecasts and observations were counted, if more than one percent of the sector was covered by a cell. The dashed lines represent the 25th and 75th percentile of the distribution of the sectors within the colored area. The thick lines represent the mean frequency observed or forecasted.

The aim of the figures is to identify daytime dependent biases in the forecast frequency due to a lack of representation of the convective cycle, e.g. differences in the time of initiation and the diurnal maximum. It must be taken into account that the sample size varies among the sectors as some sectors were more prone to the occurrence of convective events than others. Furthermore, the sectors are of different size which also reflects on the observation and forecast frequency. Attention should therefore be paid specifically to the percentile curves and the mean value when making comparisons.

A unique feature is also the fact that the coloured area of the observation frequencies has a different shape for the individual models, although every model uses the same observational dataset. This is

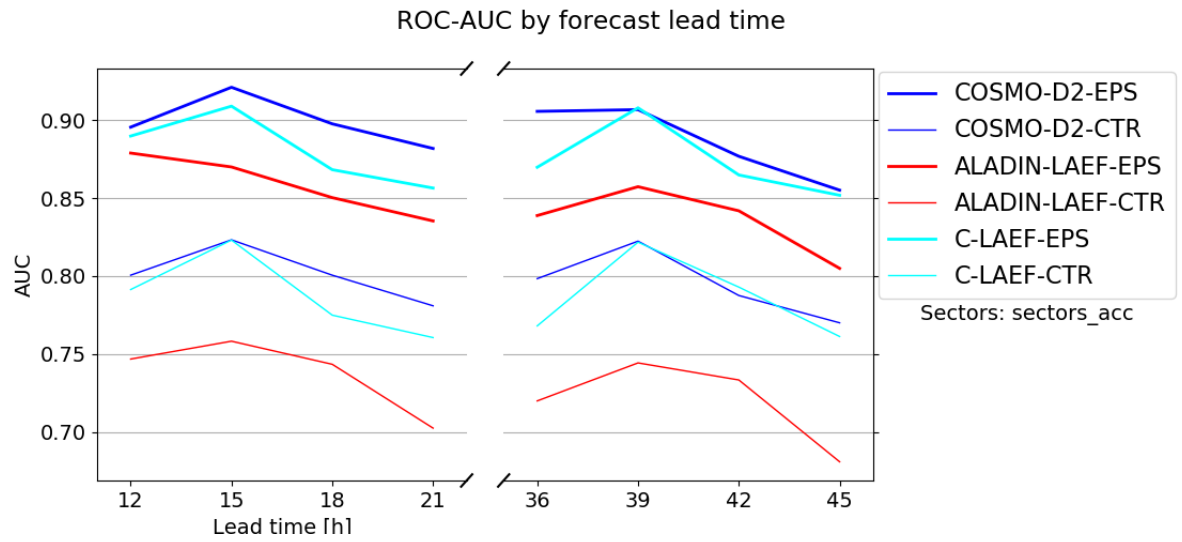

Figure 22: ROC-Area under curve as a measure of model skill calculated for three-hour intervals averaged over the entire period studied (64 days in total). Remarkable is the strong diurnal dependence of the forecast skill for all three models.

attributed to the different grid point resolution, which is considerable especially between ALADIN-LAEF (about 11km) and COSMO-D2-EPS/C-LAEF (about 2km). The models should therefore not be compared directly with each other. More attention should be paid to the differences between observation and forecast frequency and these differences should then be compared.

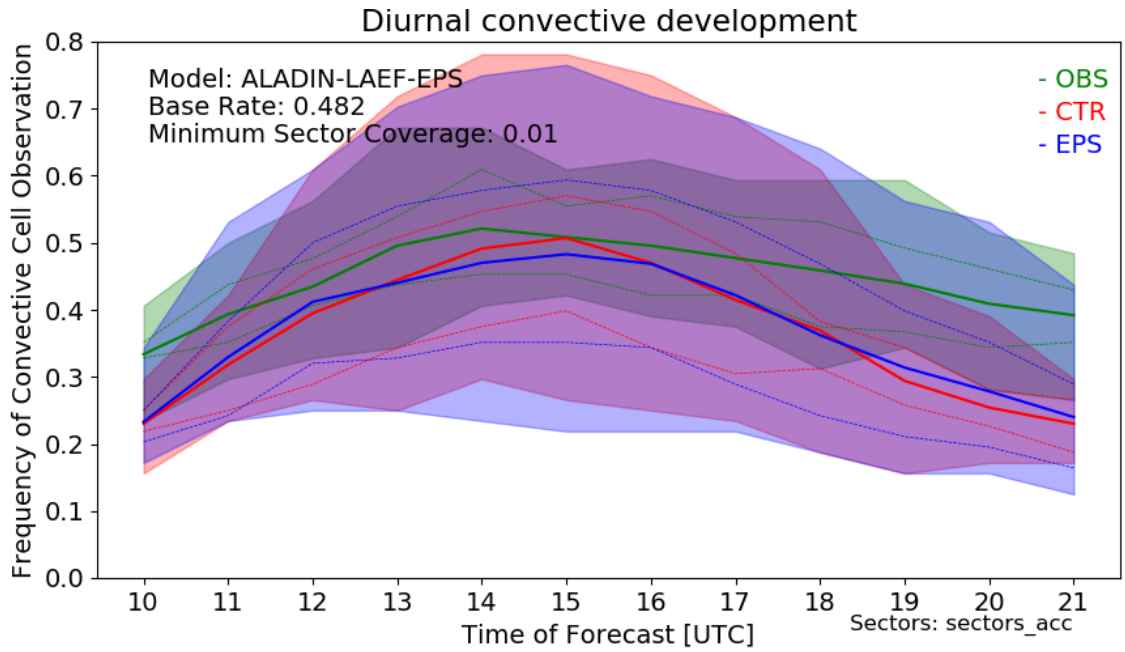

Figure 23: Daytime convective development of ALADIN-LAEF model compared to the observations. The coloured areas represent the observation frequency (OBS) and the the forecast frequency for the deterministic (DET) and the probabilistic (EPS) forecast for all sectors. The dashed lines indicate the 25th and 75th percentile of all sectors evaluated, the bold line shows its mean. All frequencies are mean values over all 64 days examined.

In Figure 23 the ALADIN-LAEF forecasts seem to be shifted to lower forecast frequencies over the whole observed day, which might be attributed to the overall dry bias that has been observed with ALADIN-LAEF convection forecasts. The difference between forecast (deterministic: red line, ensemble: blue line) and observation frequency (green line) is highest in the evening hours from 18UTC, which could be seen as a sign that in ALADIN-LAEF the convection is too much tied to diurnally fluctuating parameters like radiation and therefore convection decreases too early. When comparing this figure with Figure 22, a significant decrease in forecast skill is seen for ALADIN-LAEF especially after 17UTC. The erroneous diurnal convection cycle of ALADIN-LAEF might at least be contributing to this decrease.

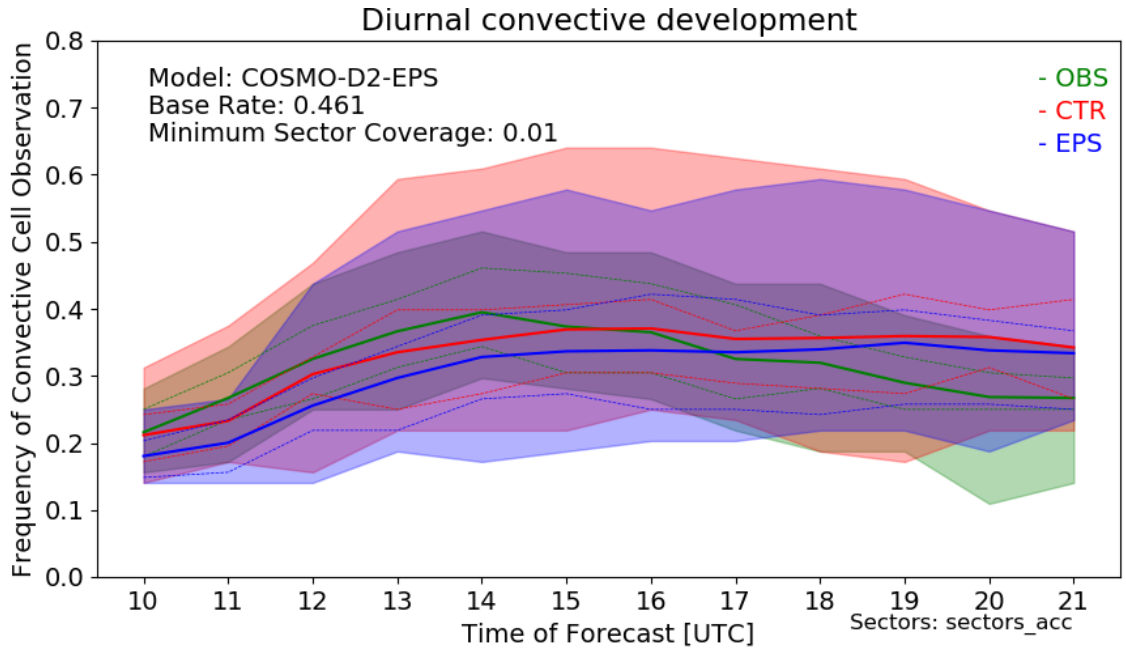

Figure 24: Daytime convective development of COSMO-D2-EPS model compared to the observations. The coloured areas represent the observation frequency (OBS) and the the forecast frequency for the deterministic (DET) and the probabilistic (EPS) forecast for all sectors. The dashed lines indicate the 25th and 75th percentile of all sectors evaluated, the bold line shows its mean. All frequencies are mean values over all 64 days examined.

When comparing COSMO-D2-EPS forecasts with their corresponding observations in Figure 24, the diurnal convection cycle shows the biggest deviations in the evening hours as well as ALADIN-LAEF, but in this case tending to extend convection too far into the night. C-LAEF, as can be seen in Figure 25, shows a similar performance to COSMO-D2-EPS in the morning hours, but tends to have a wet bias in the late afternoon and evening hours, which could at least partially explain the higher decrease in skill than in COSMO-D2-EPS.

4.5 Usability of Forecast Verification on Flight sectors

One of the main goals of the thesis is to evaluate, whether performing NWP forecasts on ATM airspace sectors is feasible and offers any benefit to existing methods. Traditional forecast verifications are carried out using dichotomous, multi-category or continuous values. Especially with regard to ensemble forecasts also probabilistic values are common. High resolution forecasts highlighted the need for a method that considers the double penalty issue in the case of a spatial and/or temporal displacement

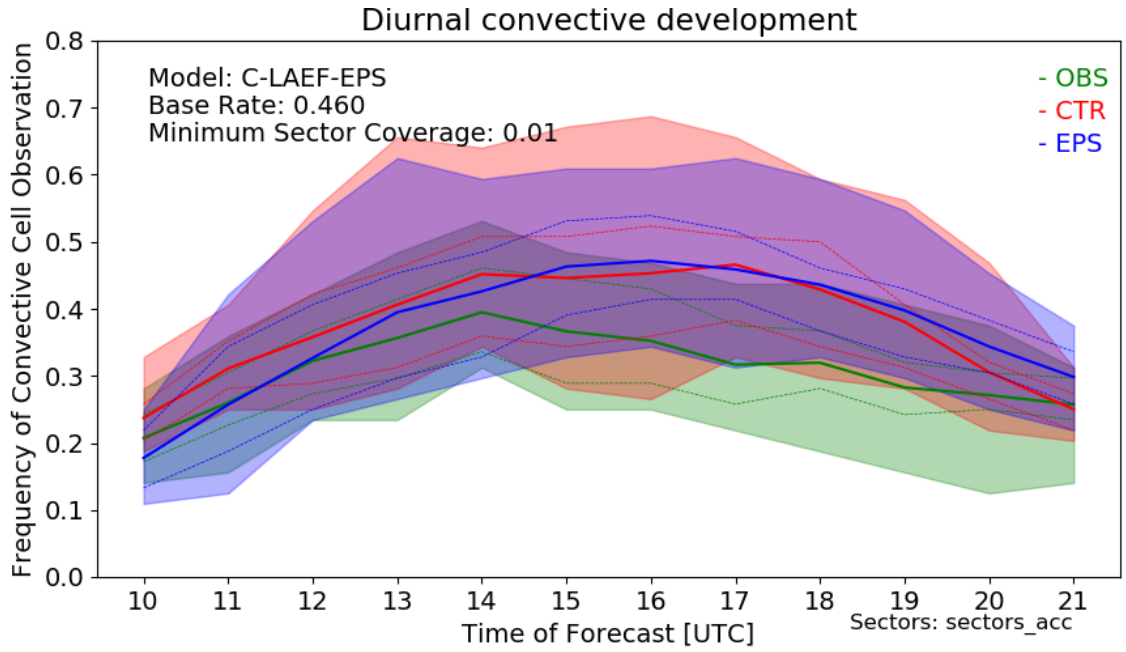

Figure 25: Daytime convective development of C-LAEF model compared to the observations. The coloured areas represent the observation frequency (OBS) and the the forecast frequency for the deterministic (DET) and the probabilistic (EPS) forecast for all sectors. The dashed lines indicate the 25th and 75th percentile of all sectors evaluated, the bold line shows its mean. All frequencies are mean values over all 64 days examined.

and takes into account structures in the "neighbourhood" around the observed location, called spatial verification. The different methods of spatial verification can be found in Jolliffe and Stephenson (2012), a general overview of forecast verification is offered on the webpage of WWRP (2017).

When addressing spatial verification it is also important to mention two basic perspectives, which are explained in more detail in Jolliffe and Stephenson (2012):

- Displacement Methods
- Filtering Methods

Displacement methods use an object-oriented point of view, where features like convective cells as a whole are investigated. The forecast attributes are evaluated based on how much a forecasted cell needs to be displaced or how much a field needs to be deformed in order to match the field of observations (Jolliffe and Stephenson, 2012).

Filtering methods according to Gilleland et al. (2009) address the double penalty effect mainly by filtering fields at increasingly coarser scales in order to get information about the skillful scale of a forecast. The filtered or upscaled field is then verified against the observations on this coarser grid. In principle, forecasting on ATM airspace sectors does some sort of upscaling. However, all sectors are differently sized and shaped, which makes a comparison difficult.

A medium to strong correlation was found for the size and forecast skill of the sectors for all models and sector groups investigated. For the sectors of the group "ACC" the correlation coefficient was 0.60 for ALADIN-LAEF, 0.72 for COSMO-D2-EPS and 0.78 for C-LAEF. Clearly, other effects on forecast

skill were superimposed, like a reduced skill in the easternmost sectors, especially around and east of Vienna.

Figure 26 shows the approach sectors "APP" (right panels) and the Area Control Centre sectors "ACC" (left panels) of Vienna International Airport. The "APP" sectors exhibit significantly lower values, which despite the smaller size compared to "ACC" sectors indicates reduced predictive skill in the easternmost areas.

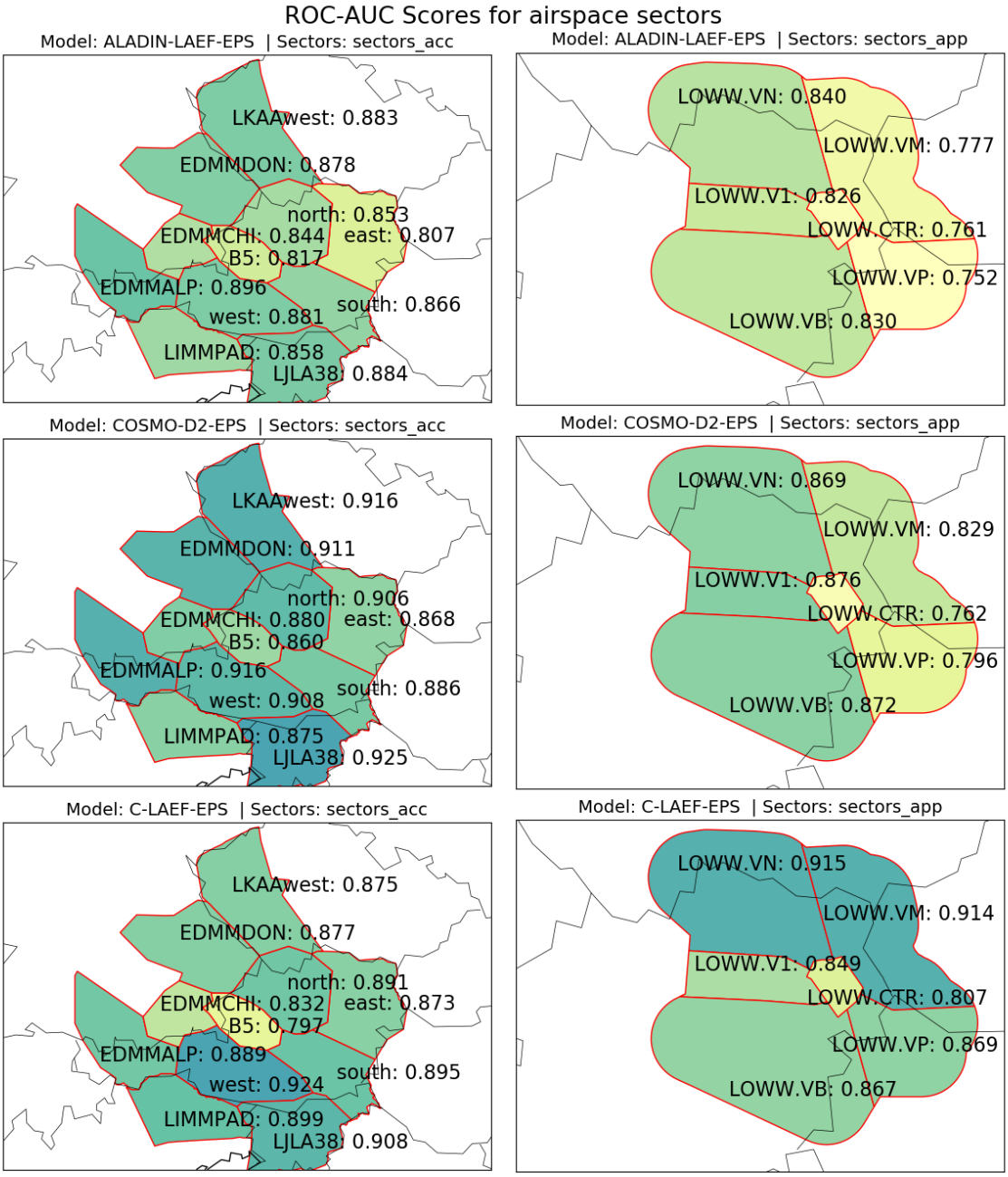

Figure 26: Sector charts displaying the ROC-AUC score for the "ACC" and "APP" sectors for all ensemble forecast models investigated. The scores were computed over the full study period of 64 days.

Especially the easternmost sector "LOWW.VP" stands out in terms of relatively poor forecasting skill compared to the neighboring sectors, as seen in Figure 26. Interestingly, the poor forecasting skill for the easternmost sectors is only present in "ALADIN-LAEF" and "COSMO-D2-EPS" models, but not in "C-LAEF", which shows relatively good forecasting skill for the eastern sectors. More information can be obtained by comparing the convective forecasting behaviour of the models with the observations. Figure 27 reveals several missed or underrepresented convective events especially for "ALADIN-LAEF" and "COSMO-D2-EPS" at the end of August and in early September 2019. Visual verification suggests that "C-LAEF" seems to perform better, although some events where missed too and others were overestimated, as seen at the end of July 2019. Overall, the number of convective events is also low.

Another important detail: The centre sector of the airport stands out in particular with its significantly lower prediction skill. This is clearly due to the size of the sector.

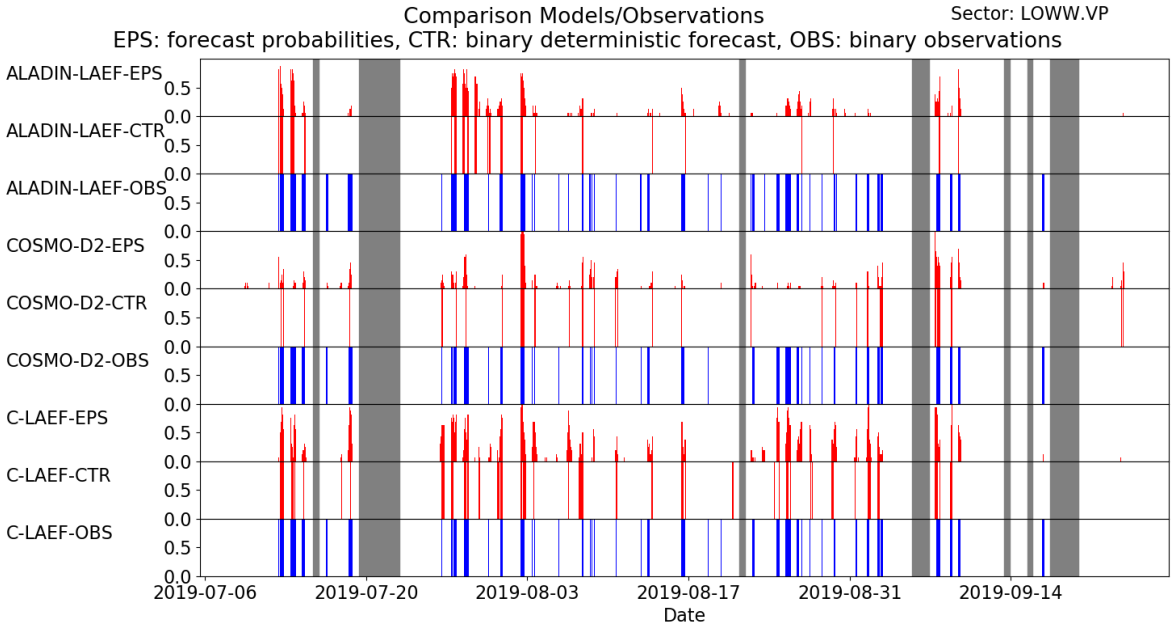

Figure 27: Comparison between convective forecasting behaviour of the investigated models and the observations for the sector "LOWW.VP" of "APP" group. "EPS" is the ensemble forecast model, "CTR" the deterministic forecast and "OBS" the observations. Note that the observations are the same for all models, but are projected on different model grids.

In neighbouring sectors, for example the sector "LOWW.VN", the models uniformly show a significantly higher prediction skill and the number of convective events is also higher than in "LOWW.VP". When studying Figure 28, it becomes apparent that all models seem to correctly identify thunderstorms more often. A remarkable difference can also be noticed when comparing ensemble forecasts with their deterministic counterparts. Especially "ALADIN-LAEF" and "C-LAEF" show some cases, particularly at the end of August, where the deterministic model does not detect convective events at all or only to a very limited extent in time. The visible verification leads to the assumption that the ensemble forecast in many cases tend to reduce timing errors for convection.

Regarding the poor performance of COSMO-D2-EPS in the eastern sectors, it can be argued with the close domain border over Hungary, which could limit the forecast quality, especially for thunderstorms moving into the sectors from the east and southeast. The poor performance of ALADIN-LAEF

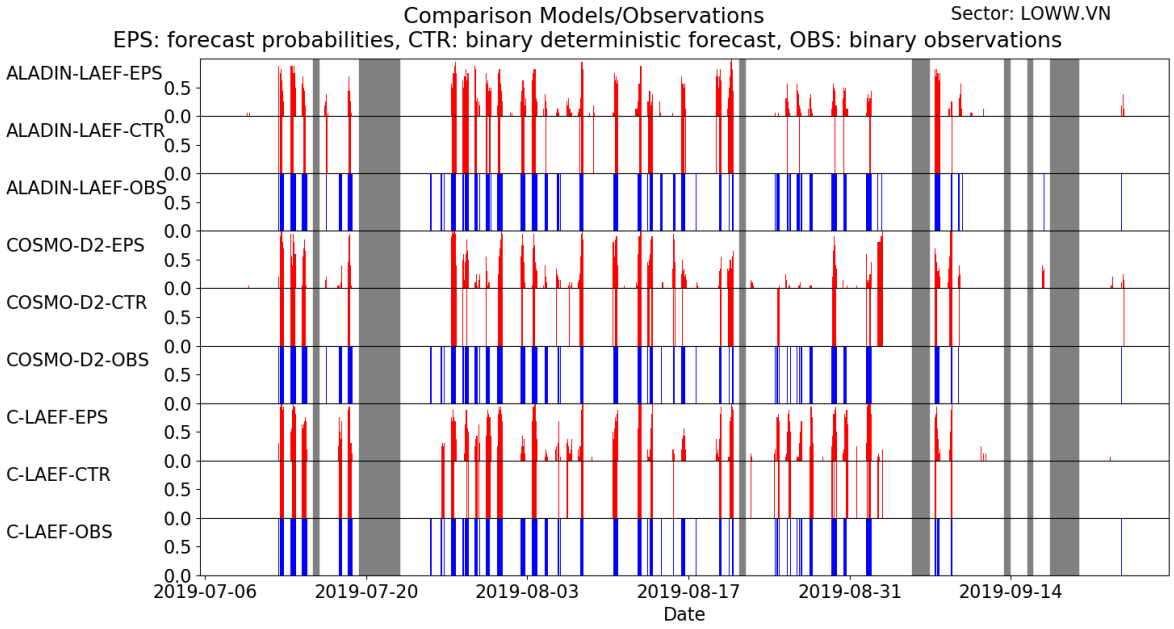

Figure 28: Comparison between convective forecasting behaviour of the investigated models and the observations for the sector "LOWW.VN" of "APP" group. "EPS" is the ensemble forecast model, "CTR" the deterministic forecast and "OBS" the observations. Note that the observations are the same for all models, but are projected on different model grids.

might be connected to the coarse horizontal resolution and the associated lack of representation of the topography.

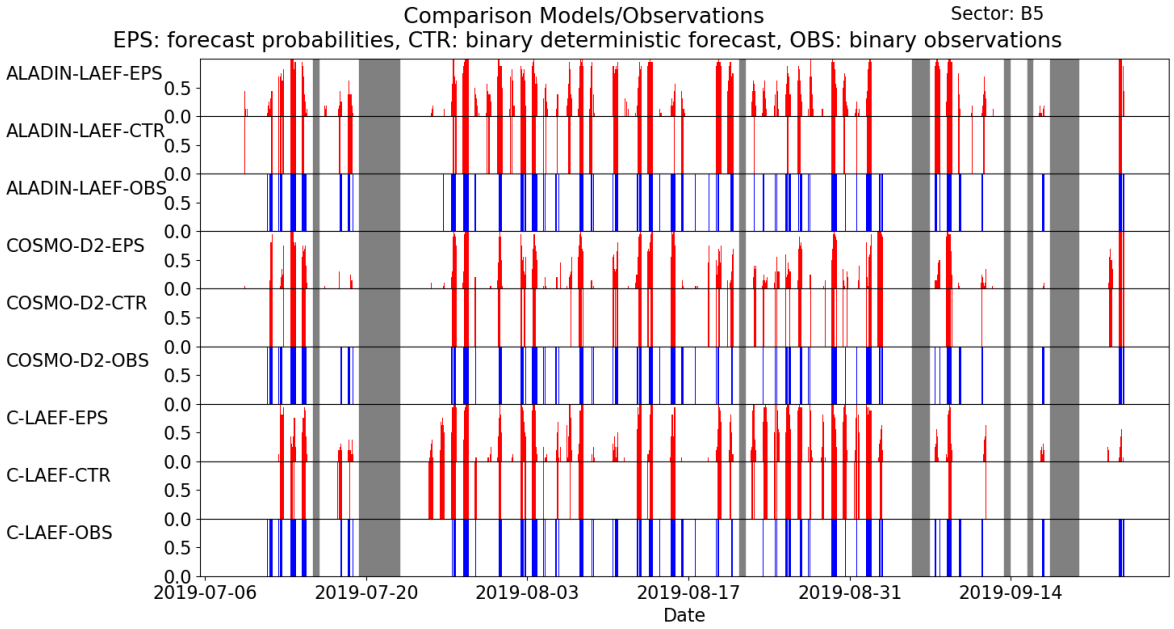

Figure 29: Comparison between convective forecasting behaviour of the investigated models and the observations for the sector "B5" of "ACC" group. "EPS" is the ensemble forecast model, "CTR" the deterministic forecast and "OBS" the observations. Note that the observations are the same for all models, but are projected on different model grids.

In Figure 26 also the sector "B5", that includes the city of Salzburg and is characterised mostly by alpine area, stands out in the models "ALADIN-LAEF" and "C-LAEF", exhibiting poor forecasting skill. This sector coincides to a large extent with an area prone to radar shadowing effects, as already mentioned in section 2.1. Hence, it would be expected, that the observations turn out to be sparse in this sector. Visual comparison of "B5" in Figure 29 with neighbouring sectors suggests, that the effect does not have a significant influence on the observations. However, the forecasts tend to be worse in the alpine area, as there are numerous days with misses and false alarms as well. A shadowing effect would probably rather lead to more false alarms, as the observations are missing. It should also be noted that the sector is smaller in area than the surrounding ones.

The strong dependence of the skill on the sector size makes it considerably more difficult to assess local differences. It is also desirable to define the skillful scale of a forecast to get an idea which scales a forecast is able to reliably resolve. Using airspace sectors makes it difficult to make any assumptions on this scale. However, smaller sectors quickly show a smaller sample size.

4.6 Notes on Observational Data

Before any evaluations could be made at all, the observation data was checked for consistency. The first test was a visual study, where CONVOY convective cells were overlaid on fields of radar reflectivity to check for any inconsistencies. The diurnal development of convection and the movement of the cells showed full agreement with the control data. However, lightning data could not be compared. Despite, the result of the visual control was very satisfying. It is worth to be mentioned that the CONVOY cell data is very close to its underlying raw observation data sources, the ALDIS lightning data and the radar data from the composite. In addition, some corrections were made, e.g. the bright-band correction. Thanks to extensive meta data saved together with the cells, missing data could be detected easily and handled properly.

To get an overview of the thunderstorm acticity, a plot of the thunderstorm days was created. It shows the number of days on which a convective cell was detected at least on one full hour, projected on to the grid of the C-LAEF model. This was done considering all convective cells (Figure 30) and with flashing cells only (Figure 31).

The number of thunderstorm days using the lightning data source of CONVOY alone was too low, considering that July and August have a large contribution to the total number of thunderstorms a year. This is clearly shown in Figure 31. However, the combination of both datasets tends to assume a slightly too high number of thunderstorm days. The threshold in radar reflectivity to identify thunderstorms is commonly used for this purpose, but there are also non-flashing cells that might reach the threshold. When comparing the values of the combined CONVOY to the results in Taszarek et al. (2019), the distribution appears to be realistic. According to their results with ZEUS and EUCLID lightning data, the maximum values of thunderstorm days in August are located mainly around the highest mountain peaks in the Central Alps from Switzerland to western Carinthia, whereas in June, the convection is more widely spread in the flatlands around the Alps. This might be the reason for the outstanding maxima of thunderstorm days in Tyrol, as June is not represented in the study and therefore the distribution of thunderstorm days might be slightly shifted. The comparison between forecasted convection and observations of the sector "EDDMALP" speaks against an erroneous over-representation of convection in Tyrol, as can also be seen on the high values of potential skill for all models in Figure 26.

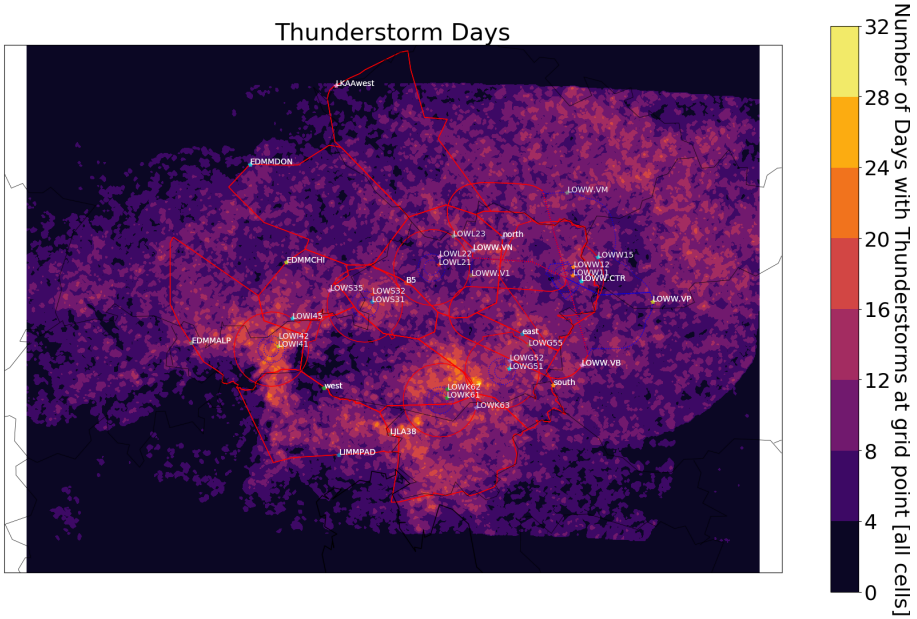

Figure 30: Map of Thunderstorm days, using lightning data and radar data. A day was considered as thunderstorm day if one thunderstorm criteria (lightning flash appearance and/or radar reflectivity threshold) was met for at least one hour on that day. The observational data was evaluated on the grid of C-LAEF.

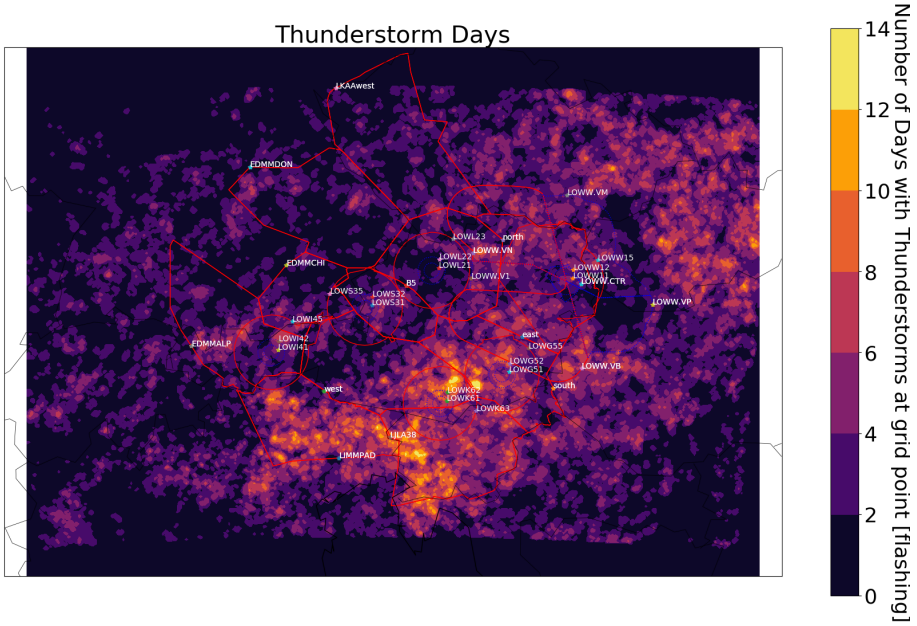

Figure 31: Map of Thunderstorm days, using lightning data only. A day was considered as thunderstorm day if the thunderstorm criteria (lightning flash appearance) was met for at least one hour on that day. The observational data was evaluated on the grid of C-LAEF.

As the further investigation of the observation data is beyond the scope of the Masters Thesis, it is pointed out that comparable cell tracking systems are used by many weather services worldwide.

5. Discussion and Outlook

5.1 Forecast Skill of Ensembles

The results of this thesis have clearly proven that ensemble forecasts outperform deterministic models in forecast skill and potential economic value regardless of model resolution. Since the computation of several forecasts for the purpose of an ensemble is computationally expensive, there is also the option of employing a multi-model ensemble. The combination of different models is capable of increasing forecast skill despite one model of the set performs worse or has a lower model resolution. Clearly, high-resolution models contribute the greatest gains in predictive skill. It was also found that the high-resolution models with an explicit computation of convection perform significantly better than models of lower resolution that rely on convection parameterisation. In the direct model comparison the best performance was attained by the COSMO-D2-EPS of the DWD followed closely by C-LAEF of ZAMG.

When assessing the temporal development of forecast skill, the high-resolution forecast models exhibited only a marginal decrease of forecast skill on the second day. One potential explanation would be constant synoptic conditions in which thunderstorms are largely driven by solar radiation. However, all three models studied showed a loss of performance in the early nighttime hours which implies issues with the diurnal cycle, e.g. timing problems concerning the subsiding of thunderstorms in the evening. Another reason for the low forecasting skill in the evening hours could be frontally driven thunderstorms that superimpose the radiation-driven thunderstorms. The former also occur in the evening and at night and are more difficult for models to forecast. Especially ALADIN-LAEF shows a significant timing error in the early nighttime which might at least in part be attributed to a too early decrease in forecasted convection, as is strengthened by Figure 23.

Verifying convection forecasts on airspace sectors was investigated and showed difficulties with comparing sectors of different size and shape. Regional differences in thunderstorm activity and forecast skill were superimposed by the properties of the sectors and therefore the study specifically aimed to examine the models and the general quality of the predictions instead of investigating local differences. Furthermore, due to the short period of forecasts studied, in some areas the sample size of convective events was low and resulted in further differences in forecast quality. It is therefore recommended to use a bigger time interval of forecasts as thunderstorms tend to occur rarely on a local scale and even some persistent weather conditions may distort studies concerning local differences.

5.2 Further work and Outlook

Spatial verification and post-processing methods

In order to be able to determine local differences in the frequency of occurrence of deep convection and for better comparability in general, the approach of evaluation based on airspace sectors is rather not recommended. Using gridded tiles of equal size seems much more practical in this application, as comparison is easier to conduct. To make the most of the benefits of an ensemble prediction system, the better statistical representation of convection compared to deterministic models has to be exploited. For this purpose, a number of spatial verification methods have been developed. A good overview of the most common methods is provided by Gilleland et al. (2009).

To further improve the quality of the ensemble forecasts, performing some kind of postprocessing is inevitable. WMO (2021) provides a comprehensive guideline on how to correctly setting up a post-processing for deterministic and ensemble prediction systems.

Prolonged Time Period

Due to the availability of the models examined in this study, the time period investigated was limited to somewhat just over 2 months. In order to obtain statistically more meaningful statements, it is strongly recommended to conduct the study over a longer period of time. The smaller the scale studied, the smaller the sample size. Furthermore, the night time hours should be included into the evaluations in order to investigate the nightly convective behaviour in more detail.

Interdisciplinary work stages

After the successful validation of a convection forecasting methodology, this information has to be converted into information that is directly accessible to air traffic management, as has been described in Weber et al. (2005) and Kicinger et al. (2016). In order to conduct capacity estimates for airport and flight routes, intensive interaction between meteorologists and aviation experts is necessary.

Bibliography

Baldauf, M., C. Gebhardt, S. Theis, B. Ritter, and C. Schraff (2017), Beschreibung des operationellen Kürzestfristvorhersagemodells COSMO-D2 und COSMO-D2-EPS und seiner Ausgabe in die Datenbanken des DWD. Available: https://www.dwd.de/SharedDocs/downloads/DE/modelldokumentationen/nwv/cosmo_de/cosmo_de_dbbeschr_version_0_9_201705.pdf?__blob=publicationFile&v=2.

Belluš, M., Y. Wang, and F. Meier (2016), Perturbing surface initial conditions in a regional ensemble prediction system. *Monthly Weather Review* **144**(9), 3377–3390. URL: https://doi.org/10.1175/MWR-D-16-0038.1, doi:10.1175/MWR-D-16-0038.1.

Belluš, M., F. Weidle, C. Wittmann, Y. Wang, S. Taşku, and M. Tudor (2019), Aire limitée adaptation dynamique développement international – limited area ensemble forecasting (aladin-laef). *Advances in Science and Research* **16**, 63–68. doi:10.5194/asr-16-63-2019.

Bertram, I. and G. Mayr (2004), Lightning in the eastern Alps 1993-1999, part I: Thunderstorm tracks. *Natural Hazards and Earth System Science* **4**. doi:10.5194/nhess-4-501-2004.

Blom, K. (2008), Lightning detection in support of airport authority decision making. aerodrome meteorological observation and forecast study group (amofsg), seventh meeting, montréal, 9 to 12 september 2008, amofsg/7-sn no. 17. https://www.icao.int/safety/meteorology/amofsg/Pages/Study-Notes.aspx?meeting=AMOFSG/7.

Bott, A. (2016), *Synoptische Meteorologie: Methoden der Wetteranalyse und-prognose*. Springer-Verlag, Berlin, Heidelberg.

Bouttier, F. and P. Courtier (2002), Data assimilation concepts and methods. Tech. rep., ECMWF. URL: https://www.ecmwf.int/node/16928.

Bouttier, F. and H. Marchal (2020), Probabilistic thunderstorm forecasting by blending multiple ensembles. *Tellus A: Dynamic Meteorology and Oceanography* **72**(1), 1–19. URL: https://doi.org/10.1080/16000870.2019.1696142, arXiv:https://doi.org/10.1080/16000870.2019.1696142, doi:10.1080/16000870.2019.1696142.

Brunet, D., D. Sills, and N. Driedger (2019), On the Evaluation of Probabilistic Thunderstorm Forecasts and the Automated Generation of Thunderstorm Threat Areas during Environment Canada Pan Am Science Showcase. *Weather and Forecasting* **34**. doi:10.1175/WAF-D-19-0011.1.

Bröcker, J. and L. A. Smith (2007), Increasing the Reliability of Reliability Diagrams. *Weather and Forecasting* **22**(3), 651–661. URL: https://doi.org/10.1175/WAF993.1, arXiv:https://journals.ametsoc.org/waf/article-pdf/22/3/651/4640721/waf993_1.pdf, doi:10.1175/WAF993.1.

Charba, J. P., F. G. Samplatsky, A. J. Kochenash, P. E. Shafer, J. E. Ghirardelli, and C. Huang (2019), LAMP Upgraded Convection and Total Lightning Probability and "Potential" Guidance for the Conterminous United States. *Weather and Forecasting* **34**(5), 1519–1545. URL: https://doi.org/10.1175/WAF-D-19-0015.1, arXiv:https://journals.ametsoc.org/waf/article-pdf/34/5/1519/4883100/waf-d-19-0015_1.pdf, doi:10.1175/WAF-D-19-0015.1.

Corfidi, S. (2003), Cold Pools and MCS Propagation: Forecasting the Motion of Downwind-Developing MCSs. *Weather and Forecasting* **18**. doi:10.1175/1520-0434(2003)018<0997:CPAMPF>2.0.CO;2.

Daley, R. (1991), *Atmospheric Data Analysis*. Cambridge University Press.

Davis, C. A., K. W. Manning, R. E. Carbone, S. B. Trier, and J. D. Tuttle (2003), Coherence of warm-season continental rainfall in numerical weather prediction models. *Monthly Weather Review* **131**(11), 2667 – 2679. URL: https://journals.ametsoc.org/view/journals/mwre/131/11/1520-0493_2003_131_2667_cowcri_2.0.co_2.xml, doi:10.1175/1520-0493(2003)131<2667:COWCRI>2.0.CO;2.

Doswell III, C. (1987), The Distinction between Large-Scale and Mesoscale Contribution to Severe Convection: A Case Study Example. *Weather and Forecasting* **2**. doi:10.1175/1520-0434(1987)002<0003:TDBLSA>2.0.CO;2.

Doswell III, C. (2001), Severe Convective Storms—An Overview. *Meteorological Monographs* **28**. doi:10.1175/0065-9401-28.50.1.

Ebert, E. (2001), Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Monthly Weather Review - MON WEATHER REV* **129**. doi:10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Emanuel, K. A. (1994), *Atmospheric Convection*. Oxford University Press, 580 pp.

Evensen, G. (1994), Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans* **99**(C5), 10143–10162. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94JC00572, arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/94JC00572, doi:https://doi.org/10.1029/94JC00572.

FAA (2018), Air traffic by the numbers 2018. Press material, URL: https://web.archive.org/web/20190120003651/https://www.faa.gov/air_traffic/by_the_numbers/media/Air_Traffic_by_the_Numbers_2018.pdf (retrieved from WebArchive), visited: 2021-09-09.

Giard, D. and E. Bazile (2000), Implementation of a new assimilation scheme for soil and surface variables in a global nwp model. *Monthly Weather Review* **128**(4), 997 – 1015. URL: https://journals.ametsoc.org/view/journals/mwre/128/4/1520-0493_2000_128_0997_ioanas_2.0.co_2.xml, doi:10.1175/1520-0493(2000)128<0997:IOANAS>2.0.CO;2.

Gilleland, E., D. Ahijevch, B. Brown, B. Casati, and E. Ebert (2009), Intercomparison of spatial forecast verification methods. *Weather and Forecasting - WEATHER FORECAST* **24**, 1416–1430. doi:10.1175/2009WAF2222269.1.

ICAO (2018), Presentation of 2018 Air Transport Statistical Results. https://www.icao.int/annual-report-2018/Documents/Annual.Report.2018_Air%20Transport%20Statistics.pdf.

ICAO (2022), Effects of Novel Coronavirus (COVID-19) on Civil Aviation: Economic Impact Analysis. https://www.icao.int/sustainability/Documents/Covid-19/ICAO_coronavirus_Econ_Impact.pdf.

Jolliffe, I. and D. Stephenson (2012), *Forecast Verification: A Practitioner's Guide in Atmospheric Science. 2nd Edition*. Wiley and Sons Ltd.

Kain, J., S. Weiss, D. Bright, M. Baldwin, J. Levit, G. Carbin, C. Schwartz, M. Weisman, K. Droegemeier, D. Weber, and K. Thomas (2008), Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing nwp. *Weather and Forecasting - WEATHER FORECAST* **23**. doi:10.1175/WAF2007106.1.

Kalnay, E. (2002), *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press. doi:10.1017/CBO9780511802270.

Keresturi, E., Y. Wang, F. Meier, F. Weidle, C. Wittmann, and A. Atencia (2019), Improving initial condition perturbations in a convection-permitting ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society* **145**(720), 993–1012. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3473, doi:10.1002/qj.3473.

Kicinger, R., J.-T. Chen, M. Steiner, and J. Pinto (2016), Airport capacity prediction with explicit consideration of weather forecast uncertainty. *Journal of Air Transportation* **24**, 18–28. doi:10.2514/1.D0017.

Kicinger, R., J. Krozel, M. Steiner, and J. Pinto (2012), Airport capacity prediction integrating ensemble weather forecasts. *AIAA Infotech at Aerospace Conference and Exhibit 2012* doi:10.2514/6.2012-2493.

Kneringer, P., S. Dietz, G. Mayr, and A. Zeileis (2018), Probabilistic Nowcasting of Low-Visibility Procedure States at Vienna International Airport During Cold Season. *Pure and Applied Geophysics* **176**. doi:10.1007/s00024-018-1863-4.

Lafore, J. P., J. Stein, N. Asencio, P. Bougeault, V. Ducrocq, J. Duron, C. Fischer, P. Héreil, P. Mascart, V. Masson, J. P. Pinty, J. L. Redelsperger, E. Richard, and J. Vilà-Guerau de Arellano (1998), The meso-nh atmospheric simulation system. part i: adiabatic formulation and control simulations. *Annales Geophysicae* **16**(1), 90–109. URL: https://angeo.copernicus.org/articles/16/90/1998/, doi:10.1007/s00585-997-0090-6.

Lang, P. (2001), P6.7 cell tracking and warning indicators derived from operational radar products. *30th Int. Conf. Radar Met., Munich, Germany* .

Lorenz, E. N. (1963), Deterministic nonperiodic flow. *Journal of Atmospheric Sciences* **20**(2), 130 – 141. URL: https://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469_1963_020_0130_dnf_2_0_co_2.xml, doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Markowski, P. and Y. Richardson (2011), *Mesoscale meteorology in midlatitudes*, vol. 2. John Wiley & Sons.

Mittermaier, M. and G. Csima (2017), Ensemble versus deterministic performance at the kilometer scale. *Weather and Forecasting* **32**. doi:10.1175/WAF-D-16-0164.1.

Murphy, A. H. (1973), A New Vector Partition of the Probability Score. *Journal of Applied Meteorology* **12**(4), 595–600. URL: https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2, arXiv:https://journals.ametsoc.org/jamc/article-pdf/12/4/595/4967197/1520-0450(1973)012_0595_anvpot_2_0_co_2.pdf, doi:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Murphy, A. H. (1993), What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Weather and Forecasting* **8**(2), 281–293. URL: https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2, arXiv:https://journals.ametsoc.org/waf/article-pdf/8/2/281/4650824/1520-0434(1993)008_0281_wiagfa_2_0_co_2.pdf, doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Neuwirth, C. (2012), Lightning density distribution and hazard in an alpine region. *Journal of Lightning Research* **4**, 166–172. doi:10.2174/1652803401204010166.

Owens, R. G. and T. Hewson (2018), ECMWF Forecast User Guide. *ECMWF eLibrary, Online Ressource*. URL: https://www.ecmwf.int/node/16559, doi:10.21957/m1cs7h.

Palmer, T., R. Buizza, R. Hagedorn, A. Lawrence, M. Leutbecher, and L. Smith (2006), Ensemble prediction: A pedagogical perspective. *ECMWF newsletter* pp. 10–17. URL: https://www.ecmwf.int/node/18024, doi:10.21957/ab129056ew.

Petrie, R. E. (2008), *Localization in the ensemble Kalman Filter*. Master's thesis, University of Reading, Department of Meteorology.

Raith, W. (2008), *Erde und Planeten*. De Gruyter, Berlin, Boston. URL: https://www.degruyter.com/view/title/17869, doi:https://doi.org/10.1515/9783110198027.

Schellander-Gorgas, T., Y. Wang, F. Meier, F. Weidle, C. Wittmann, and A. Kann (2017), On the forecast skill of a convection-permitting ensemble. *Geoscientific Model Development* **10**(1), 35–56. URL: https://www.geosci-model-dev.net/10/35/2017/, doi:10.5194/gmd-10-35-2017.

Schraff, C., H. Reich, A. Rhodin, A. Schomburg, K. Stephan, A. Periáñez, and R. Potthast (2016), Kilometre-scale ensemble data assimilation for the COSMO model (KENDA). *Quarterly Journal of the Royal Meteorological Society* **142**(696), 1453–1472. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2748, doi:10.1002/qj.2748.

Seity, Y., P. Brousseau, S. Malardel, G. Hello, P. Bénard, F. Bouttier, C. Lac, and V. Masson (2011), The AROME-France Convective-Scale Operational Model. *Monthly Weather Review* **139**(3), 976 – 991. URL: https://journals.ametsoc.org/view/journals/mwre/139/3/2010mwr3425.1.xml, doi:10.1175/2010MWR3425.1.

Steinacker, R., M. Dorninger, F. Woelfelmaier, and T. Krennert (2000), Automatic Tracking of Convective Cells and Cell Complexes from Lightning and Radar Data. *Meteorology and Atmospheric Physics* **72**, 101–110. doi:10.1007/s007030050009.

Taszarek, M., J. Allen, T. Púčik, P. Groenemeijer, B. Czernecki, L. Kolendowicz, K. Lagouvardos, V. Kotroni, and W. Schulz (2019), A climatology of thunderstorms across europe from a synthesis of multiple data sources. *Journal of Climate* **32**(6), 1813 – 1837. URL: https://journals.ametsoc.org/view/journals/clim/32/6/jcli-d-18-0372.1.xml, doi:10.1175/JCLI-D-18-0372.1.

Thepaut, J., P. Courtier, G. Belaud, and G. Lemaître (1996), Dynamical structure functions in a four-dimensional variational assimilation: A case study. *Quarterly Journal of the Royal Meteorological Society* **122**, 535–561.

Valur Hólm, E. (2018), Lecture notes on assimilation algorithms. European Centre for Medium-Range Weather Forecasts, Reading, UK. URL: "https://www.ecmwf.int/sites/default/files/elibrary/2008/16931-assimilation-algorithms.pdf".

Vetra-Carvalho, S. (2018), The Ensemble Kalman Filter, Part 1: Theory. University of Reading, Lecture notes from Data Assimilation Training in collaboration with the ECMWF, 7-10th March. URL: http://www.met.reading.ac.uk/~darc/training/ecmwf_collaborative_training/EnKF_Lecture1_Sanita.pdf.

Wang, Y., M. Belluš, C. Wittmann, M. Steinheimer, F. Weidle, A. Kann, S. Ivatek-Šahdan, W. Tian, X. Ma, S. Tascu, and E. Bazile (2011), The Central European limited-area ensemble forecasting system: ALADIN-LAEF. *Quarterly Journal of the Royal Meteorological Society* **137**(655), 483–502. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.751, doi:10.1002/qj.751.

Wang, Y., A. Kann, M. Belluš, J. Pailleux, and C. Wittmann (2010), A strategy for perturbing surface initial conditions in LAMEPS. *Atmospheric Science Letters* **11**(2), 108–113. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/asl.260, doi:10.1002/asl.260.

Wapler, K., M. Goeber, and S. Trepte (2012), Comparative verification of different nowcasting systems to support optimisation of thunderstorm warnings. *Adv. Sci. Res.* **8**, 121–127.

Warner, T. T. (2010), *Numerical Weather and Climate Prediction*. Cambridge University Press. doi:10.1017/CBO9780511763243.

Wastl, C., Y. Wang, A. Atencia, F. Weidle, C. Wittmann, C. Zingerle, and E. Keresturi (2021), C-laef: Convection-permitting limited-area ensemble forecasting system. *Quarterly Journal of the Royal Meteorological Society* **147**(735), 1431–1451. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3986, arXiv:https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3986, doi:https://doi.org/10.1002/qj.3986.

Wastl, C., Y. Wang, A. Atencia, and C. Wittmann (2019), Independent perturbations for physics parametrization tendencies in a convection-permitting ensemble (pSPPT). *Geoscientific Model Development* **12**, 261–273. doi:10.5194/gmd-12-261-2019.

Wastl, C., Y. Wang, and C. Wittmann (2020), Application of convection-permitting EPS C-LAEF at ZAMG. In: *Proceedings of the Joint 30th ALADIN Workshop and HIRLAM ASM 2020, 30/03-03/04/2020, visio-conference*, ALADIN Consortium, Ljubljana, conference material. URL: https://www.umr-cnrm.fr/aladin/IMG/pdf/c-laef_ljubljana.pdf.

Weber, M. E., J. J. Evans, M. M. Wolfson, R. DeLaura, B. Moser, B. Martin, J. V. Welch, J. Andrews, and D. Bertsimas (2005), Improving air traffic management during thunderstorms. *24th Digital Avionics Systems Conference* **1**, 3.D.2–31.

Weidle, F., Y. Wang, T. Schellander-Gorgas, M. Belluš, C. Wastl, and C. Wittmann (2015), ALADIN-LAEF and high resolution ensemble systems at ZAMG. Presentation on Ensemble Prediction models at ZAMG. URL: https://www.umr-cnrm.fr/aladin/IMG/pdf/laef_forecastersmeeting_2015.pdf.

Weisheimer, A. and T. Palmer (2014), On the reliability of seasonal climate forecasts. *Journal of the Royal Society, Interface / the Royal Society* **11**, 20131162. doi:10.1098/rsif.2013.1162.

Weusthoff, T., F. Ament, M. Arpagaus, and M. Rotach (2010), Assessing the Benefits of Convection-Permitting Models by Neighborhood Verification: Examples from MAP D-PHASE. *Monthly Weather Review - MON WEATHER REV* **138**, 3418–3433. doi:10.1175/2010MWR3380.1.

Wilks, D. S. (2001), A skill score based on economic value for probability forecasts. *Meteorological Applications* **8**(2), 209–219. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1017/S1350482701002092, doi:10.1017/S1350482701002092.

Wilks, D. S. (2011a), On the Reliability of the Rank Histogram. *Monthly Weather Review* **139**(1), 311–316. URL: https://doi.org/10.1175/2010MWR3446.1, arXiv:https://journals.ametsoc.org/mwr/article-pdf/139/1/311/4257531/2010mwr3446_1.pdf, doi:10.1175/2010MWR3446.1.

Wilks, D. S. (2011b), *Statistical methods in the Atmospheric Sciences*. Elsevier Academic Press, Amsterdam; Boston.

Wittmann, C., T. Haiden, and A. Kann (2010), Evaluating multi-scale precipitation forecasts using high resolution analysis. *Advances in Science and Research* **4**(1), 89–98. URL: https://asr.copernicus.org/articles/4/89/2010/, doi:10.5194/asr-4-89-2010.

WMO (2017), Joint wmo technical progress report on the global data processing and forecasting system and numerical weather prediction research activities for 2017. Tech. rep., WMO, Geneva. URL: https://www.wmo.int/pages/prog/www/DPFS/documents/Austria_2017.pdf.

WMO (2021), Guidelines on ensemble prediction system postprocessing. Tech. Rep. WMO-No. 1254, World Meteorological Organization, Geneva.

WWRP (2017), Forecast verification methods across time and space scales. https://www.cawcr.gov.au/projects/verification/, World Weather Research Programme, 7th International Verification Methods Workshop, Berlin, 2017, Accessed: 2021-07-14.

Danksagung

Ich danke folgenden Personen für die Unterstützung beim Verfassen meiner Masterarbeit:

- Ass.-Prof. Mag. Dr. Manfred Dorninger für die herausragende Betreuung, für die hilfreichen Tipps und das konstruktive Feedback. Des Weiteren bedanke ich mich auch für die unvergesslichen Exkursionen wie das Feldpraktikum in Turnau oder das Messpraktikum in der Schwarzau.
- DI Dr. Lukas Strauss für die großartige Unterstützung und die vielen anregenden Unterhaltungen rund um meine Masterarbeit, von ihm stammt auch die Idee für das Thema dieser Masterarbeit.
- Dr. Johannes Sachsperger, Dr. Martin Steinheimer und DI Dr. Lukas Strauss für die Versorgung mit wertvollen statistischen Skripten für die Datenauswertung und die unzähligen nützlichen Ratschläge.
- Meinen KollegInnen bei MeteoServe für die Unterstützung und besonders Dr. Christina Tavolato-Wötzl, die mir mit ihrem umfangreichen Wissen zu Konvektionsparametern zur Seite stand.
- Dr. Markus Kerschbaum für das konstruktive Feedback bei zahlreichen Meetings und der Möglichkeit, meine Arbeit im Rahmen des PROB4LOWW-Projektes verfassen zu können. Ich danke außerdem der Austro Control für das Zurverfügungstellen der Radardaten.
- Mag. Christoph Wittmann und Dipl.-Met. Florian Weidle von der ZAMG für die Bereitstellung der Modelldaten und die ausführliche Hilfestellung in der Nutzung der Daten.
- dem Deutschen Wetterdienst für die umfassende Versorgung mit Modelldaten über seine Open Data-Plattform.
- meinen StudienkollegInnen und besonders Lukas Kugler für seine Tipps und das Korrekturlesen meiner Arbeit.
- meiner Familie, meinen Eltern und meinem Bruder, dass sie in dieser herausfordernden Zeit stets an mich geglaubt haben, mein Studium ermöglichten und mir immer wieder Mut gegeben haben.