universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Verification of the Lightning Potential Index (LPI) in the COSMO-D2 ensemble prediction system

verfasst von / submitted by

Michele Salmi

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

Master of Science (MSc)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 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

Abstract

The COSMO-D2 EPS is the very high resolution, limited-area ensemble prediction system (L-EPS) maintained at the German Weather Service (DWD) and has an horizontal resolution of 2.2 km. At such spatial scales, which lie at the lower end of the mesoscale, deep convection does not need to be parametrized and can instead be resolved directly in the model. At the same time, the development of innovative parameters which combine synoptic scale forcings and intra-cloud physics, like the Lightning Potential Index (LPI), significantly increased the potential accuracy when forecasting heavy showers and thunderstorms. However, such improvements in spatial resolution and modeling also need a proper verification approach in order to put into perspective grid-point related issues such as the double-penalty effect. The probabilistic approach of an EPS applied to high resolution models could nonetheless help increasing the accuracy and the predictability also in case of very localized convective phenomena. The first part of this work is dedicated to the analysis of the two datasets used (the LPI from the COSMO-D2 EPS and the observed lightning activity from the LINET observation network). A preliminary verification based on a conventional measure such as the Symmetric Extremal Dependence Index (SEDI) has also been conducted. In the second part, fuzzy and object based verification methods such as the dispersion Fractions Skill Score (dFSS) and the ensemble-SAL (eSAL) has been used to analyze the COSMO-D2 EPS forecasts of the LPI. This second part is focused on better understanding the spread-errpr relationship in the model, thus investigating possible positive effects on the predictability of convection. In general, the COSMO-D2 EPS tends to generate too little dispersion in its members if compared to the actual model error. Specifically, the ensemble mean generates useful lightning activity forecasts at a spatial scale of around 200 km for the afternoon hours, while the spatial spread of the ensemble members lies at more or less 100 km.

Kurzfassung

Das COSMO-D2 EPS ist das operationelle, kilometerskalige Ensemble Vorhersage System (L-EPS) des Deutschen Wetterdienstes (DWD) und hat eine horizontale Auflösung von etwa 2.2 km. Dieser Gitterpunktabstand erlaubt es, großräumige, hochreichende konvektive Prozesse wie Gewitter oder kräftige Schauer explizit und ohne physikalische Parametrisierung zu modellieren. Spezielle Indizes, die sowohl die Mikrophysik der Wolken als auch die für den Auftrieb vorhandene Energie einbeziehen - wie z. B. der Lightning Potential Index (LPI) - wurden ebenfalls entwickelt, um die Vorhersage hochreichender Konvektion und damit auch der Blitzaktivität auf eine neue Ebene der räumlichen Genauigkeit zu bringen. Mit solch hochaufgelösten Vorhersagen geht jedoch auch ein höheres Fehlerpotential einher, zumindest bei der Gitterpunktverifikation. Die Verwendung eines sehr hochaufgelösten Gitters in einem Ensemble-Vorhersagesystem könnte jedoch enorme Vorteile in Bezug auf Genauigkeit und Vorhersagbarkeit bringen. Der erste Teil dieser Arbeit ist der Analyse der beiden verwendeten Datensätze gewidmet (der LPI aus dem COSMO-D2 EPS und die beobachteten Blitze aus dem LINET-Blitzortungssystem). Eine erste Verifikation mit Hilfe des Symmetric Extremal Dependence Index (SEDI) wurde ebenfalls durchgeführt. Im zweiten Teil wurden innovative Verifikationsansätze wie der Dispersion Fractions Skill Score (dFSS) und der Ensemble-SAL (eSAL) auf den LPI im COSMO-D2 EPS angewendet. Das Hauptziel dieses zweiten Teils ist es, die Beziehung zwischen dem Prognosefehler und dem Ensemble-Spread auf verschiedenen räumlichen Skalen zu bewerten. Für die Sommermonate 2019 zeigt das COSMO-D2 EPS eine allgemeine Tendenz zur Überschätzung der Vorhersagbarkeit der Blitzaktivität. Die Spread-Error Beziehung für verschiedene Vorhersagezeiten variiert aber stark. Mit Hilfe des dFSS kann man zudem untersuchen, wie sich diese Beziehung für sich ändernde räumliche Skalen entwickelt. Im Durchschnitt liefert das System in den Nachmittagsstunden eine brauchbare Blitz-Vorhersage für horizontale Skalen von etwa 200 km. Anhand der Analyse der Ensemble-Streuung kann man aber zeigen, dass das System im Schnitt schon bei rund 100 km die Prognose als "brauchbar" bewerten würde.

Contents

1 Introduction

Forecasting showers and thunderstorms with a high level of accuracy is still a challenge even for very high resolution models. Convection involves various scales in both time and space and is also significantly influenced by smaller scale phenomena such as turbulence and intra-cloud processes. Furthermore, the showery and thundery activity is often triggered and driven by synoptic scale features, such as fronts or troughs and can in return modify the future path and development of such bigger scale processes. This very broad spectrum of scales involved and the complex interactions and energy exchanges between them is the main source of unpredictability when it comes to deep convection in general and lightning activity in particular. Thunderstorms are typically classified as mesoscale phenomena, but are influenced by – and can also significantly influence – both the microscale and the synoptic scale. The uncertainties involved can therefore lead to significant errors in the forecast at various scales.

On the other hand, convective processes can often lead to significant damages for both people and properties. Large hail, high rain rates and strong convective wind gusts are the main source of such human and economic losses. Moreover, the process that separate the electrical charges inside the cumulonimbus clouds (Saunders, 2008), which involves the impacts of liquid water and ice particles in the up- and downdrafts of the convective cell, is responsible for lightning strikes. The lightning discharge is the defining phenomenon when studying convection and is itself a significant source of risk for objects, animals and of course also human beings. Hence, both the society and the economy would greatly benefit from weather models being able to forecast such types of severe weather with a good level of accuracy. Nonetheless, for a very long time the spatial resolution of most weather models has not been high enough to resolve convection explicitly and various types of parametrization were needed. Only towards the end of the 20^{th} century and especially during the last couple of decades mesoscale-resolving, local area models with horizontal resolutions in the kilometer range began to flourish. Thus, at least larger-scale convective processes started to be modeled without the usage of predefined parametrization modules.

Following the increasing spatial resolution, during the 21^{st} century the first algorithms and parameters specifically dedicated to the forecast of lightning activity emerged (McCaul et al., 2009). The Lightning Potential Index (LPI) (Lynn and Yair, 2010; Lynn et al., 2012) has been among the first ones, being developed and tuned using one of the very first high resolution local area models, the Weather Research and Forecasting model (WRF). At the same time, with further improvements in the capacity of supercomputers, the first attempts at creating very high resolution, convection resolving ensemble prediction systems (EPS) were also made. For what

concerns Europe, one of the earliest examples in this field was the COSMO-LEPS system (Montani et al., 2003), the high resolution EPS developed in the framework of the COSMO-Consortium (`https://www.cosmo-model.org/`) already at the beginning of the 21st century.

During the last few years, high resolution EPS with the ability to explicitly resolve convective processes have emerged in many countries. One of the first examples of this new approach has been the German COSMO-DE EPS (Gebhardt et al., 2011), later called COSMO-D2 EPS and nowadays finally named ICON-D2 EPS. With 20 ensemble members and an average horizontal grid spacing of around 2.2 km, the system has been included in the operative processes of the German Weather Service (DWD) for more than a decade now. Moreover, since 2015 the COSMO-D2-EPS calculates an adapted version of the LPI (Blahak, 2015) which tries to provide a high resolution, probabilistic forecast of the potential for lightning strikes. The adapted LPI in use at the DWD interestingly puts together parameters at very different scales. Intra-cloud properties related to the state of water in the updraft are filtered with larger scales convective indices in an effort to deliver a very precise outlook of the risk for electrical discharges in the atmosphere. The fact that this approach is being applied to an high resolution EPS leads to a very promising framework in terms of skill and predictability, which however needs to be verified.

1.1 Motivation

Despite such innovations in forecasting convection with high resolution ensemble systems, until now only limited efforts have been made to assess the performance of this approach. This work, conducted in cooperation with the DWD, is a preliminary verification of the forecast of lightning flashes in an high resolution ensemble model. This is done with a special focus on the spread-error relationship of the system, aiming at investigating also the predictability of this very localized atmospheric phenomenon. Therefore, the study applies both a conventional, grid-point based verification index — the Symmetric Extremal Dependence Index (SEDI) (Ferro and Stephenson, 2011) — and also some measures which are designed to minimize possible double penalty effects using object-based and neighborhood verification approaches such as the Fractions Skill Score, or FSS (Roberts and Lean, 2008) and the Structure-Amplitude-Location, or SAL (Wernli et al., 2008). These last two methods can also be slightly transformed and applied to the ensemble in a more probabilistic form, in order to evaluate the relationship between the ensemble dispersion and the ensemble error. The adapted, probabilistic forms of the two scores are referred to as the dispersion FSS, or dFSS (Dey et al., 2014) and the ensemble SAL, or eSAL (Radanovics et al., 2018). This analysis aims at investigating two key scientific questions:

(a) Quantifying the benefits — if any — brought by the high resolution, probabilistic approach of a L-EPS in forecasting convection in general and lightning activity in particular;

(b) Verifying if the proposed verification metrics — especially the dFSS and the eSAL — are able to give detailed insights and information about the quality of the forecast for lightning activity;

2 Theoretical background

This analysis makes use of the COSMO-D2-EPS LPI output fields, verified against observed lightning activity from the Lightning detection NETwork (LINET). The verification methods applied to the data are the Symmetric Extremal Dependence Index (SEDI), the Fractions Skill Score (FSS) and the Structure-Amplitude-Location (SAL). A brief theoretical description of all the systems and methods is provided below.

2.1 COSMO-D2 Ensemble prediction system

The COSMO-D2 EPS datasets being analyzed in this study originate from the DWD model framework of 2019. A deterministic (ICON) and probabilistic (ICON EPS) model generates outputs at 13 and 40 km, respectively, horizontal resolution for the whole globe. For the European continent, which is object of this study, a finer grid spacing of 6.5 and 20 km, respectively is used (ICON-EU and ICON-EU EPS). Furthermore, a very high resolution model named COSMO-D2 is nested in the European domain and covers the whole of Germany and the alpine region as well as their surroundings with a 2.2 km horizontal grid spacing. Further insights in the COSMO model and the COSMO Consortium can be found at: `https://www.cosmo-model.org/`.

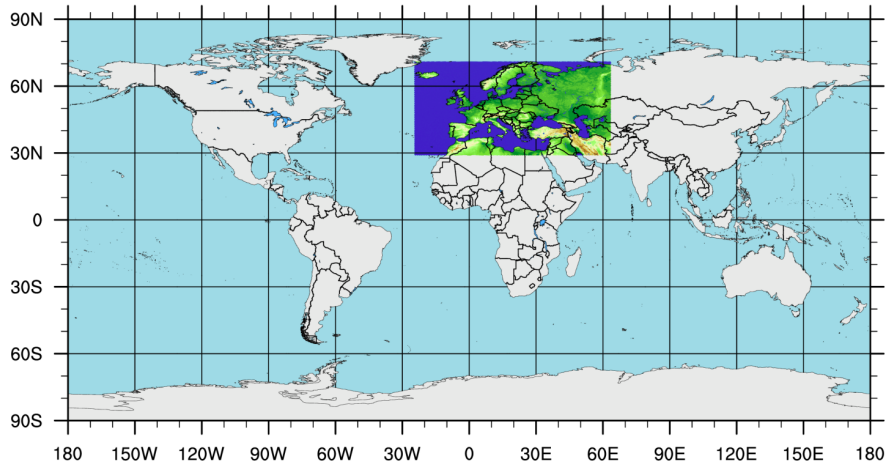

Figure 2.1: ICON-EU and ICON-EU-EPS Model domain and orography. Source: (Reinert et al., 2022)

The very high resolution model, COSMO-D2, can directly resolve larger scale, deep convective processes using 65 terrain following vertical levels. However, phenomena taking place at smaller scales such as shallow convection of intra-cloud processes still need sub-grid parametrization. The cloud microphysics scheme used in COSMO-D2 for example uses a 6-classes closure that comprises the class "graupel". Both the ICON model chain and the COSMO-D2 are thoroughly described on the DWD's website (Baldauf et al., 2018; Reinert et al., 2022) as well as on the website of the COSMO Consortium at `https://www.cosmo-model.org/content/model/documentation/core/default.htm`.

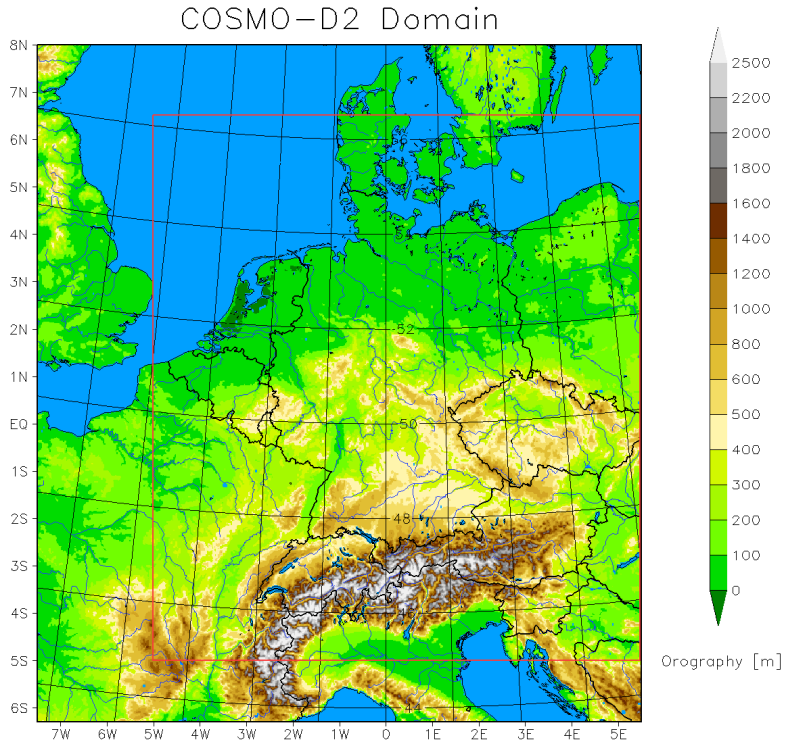

Figure 2.2: COSMO-D2 Model domain and orography in Summer 2019. The domain indicated with the red line is that of the former COSMO-DE model, discontinued in 2017. Source: `https://www.cosmo-model.org/`

The deterministic COSMO-D2 model has also a probabilistic twin, named COSMO-D2 EPS. The latter has the same horizontal and vertical resolution of the COSMO-D2 and it does not differ in the physical properties from its deterministic twin. The boundary conditions come from the european probabilistic model, ICON-EU EPS. The ICON-EU EPS provides 40 different perturbed members and the first 20 members are used to feed the correspondent 20 members of the COSMO-D2 EPS. The Kilometre-scale ENsemble Data Assimilation (KENDA) system, which takes its origins from the Local ensemble Transformed Kalman Filter (LTKF) and has been optimized in the framework of the COSMO Consortium (Schraff et al., 2016; Hunt et al., 2007) provides the initial conditions to the 20 different members of the

ensemble. On top of this, random perturbation methods bring more dispersion to some basic physical modules of the model, including deep convection. COSMO-D2-EPS provides 8 runs per day at main UTC times (00, 03, 06, 09, 12, 15, 18, 21) with 27 hours forecast. The output fields are made available to the users with 15 minutes steps. Table 2.1 summarizes the most important characteristics of the probabilistic model chain over Europe.

Characteristic	COSMO-D2-EPS	ICON-EU-EPS
Horizontal grid spacing	2.2 km	20 km
Vertical levels	65	90
Ensemble members	20	40
Model runs per day	8	8
Forecast range (main runs)	27 h	120 h
Boundary conditions	ICON-EU-EPS	ICON-EPS
Initial conditions	KENDA	LTKF

Table 2.1: Some of the key characteristics of COSMO-D2-EPS and its parent model, ICON-EU-EPS.

2.2 COSMO-D2 Lightning Potential Index - LPI

The Lightning Potential Index or LPI (Lynn and Yair, 2010; Lynn et al., 2012; Salmi et al., 2022) estimates the fraction of energy of the convective updraft which is potentially useful for charge separation inside the cumulonimbus cloud. Accordingly to this definition, the LPI is provided in $J \cdot kg^{-1}$. The equation 2.1 that describes the general version of the LPI for a model unit volume V is based on the convective updraft velocity w and the liquid water to ice water ratio in the most important layer of the cumulonimbus cloud where charge separation and therefore also electrification occurs, i.e. between the height of the 0 °C isotherm $H_{(0°C)}$ and the height of the -20 °C isotherm $H_{(-20°C)}$.

$$LPI = \frac{1}{V} \iiint_{H_{(0°C)}}^{H_{(-20°C)}} \epsilon \; w^2 \; dz \; dy \; dx \tag{2.1}$$

ϵ is a dimensionless function that varies from 0 to 1 and is defined as follows:

$$\epsilon = \frac{2 \cdot (Q_l \cdot Q_s)^{0.5}}{Q_l + Q_s} \tag{2.2}$$

In equation 2.2, Q_l and Q_s are the total liquid water mass mixing ratio and the ice fractional mixing ratio, respectively. Bot are expressed in $kg \cdot kg^{-1}$ and Q_s takes into account also snow, hail and graupel. Following equation 2.2, ϵ approaches 1 only in those cloud layers where the liquid water and the ice water mixing ratios are very close in absolute terms. If one of the two prevails, then the function will tend to

0. The mixture of supercooled liquid water droplets, snow, hail and graupel particles creates the best environment for electrification to occur, thanks to charge separation processes. Through ϵ, such microphysical characteristics of the cumulonimbus cloud enters the LPI algorithm and helps estimating the risk of electrical discharges in the analyzed convective cloud.

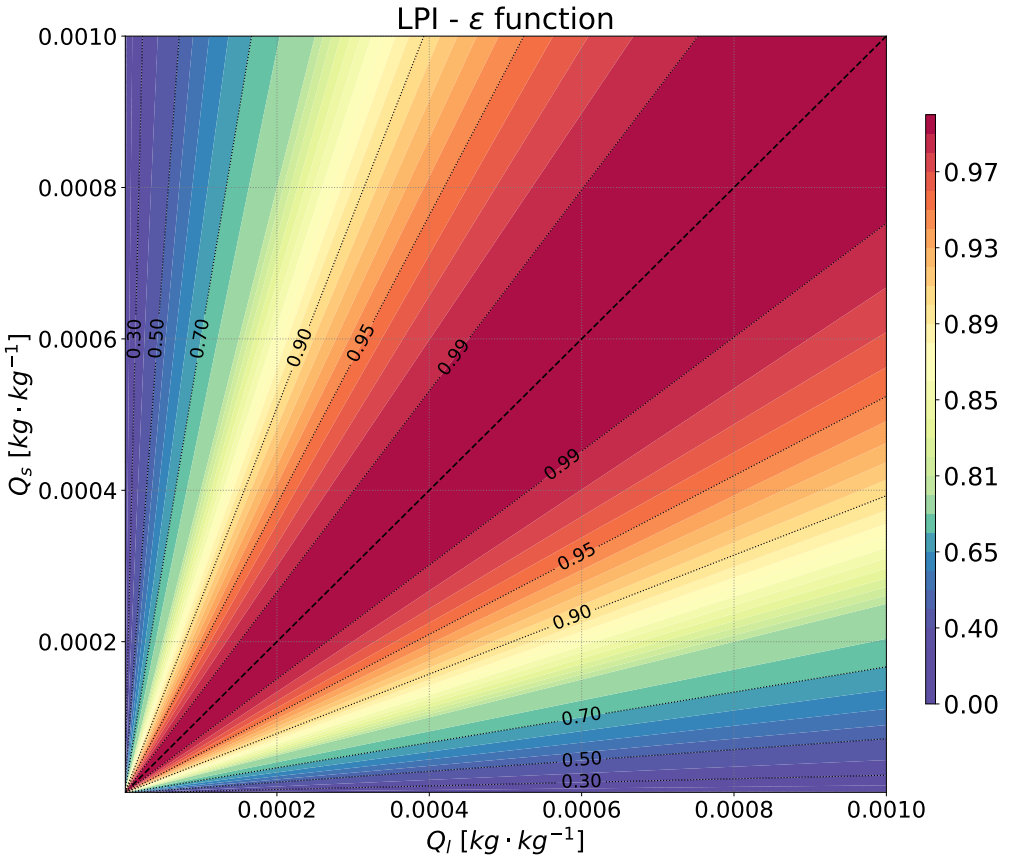

Figure 2.3: Visualization of the ϵ function as described in Eq. 2.2.

The DWD has slightly modified the original LPI formula shown in equation 2.1 and in 2015 developed an improved algorithm which has been included in the COSMO-D2 EPS ever since. The adapted LPI formula is shown in equation 2.3 (Blahak, 2015) and comprises three additional boolean filtering functions f_1, f_2 and $g_{(w)}$, described below after equation 2.3.

$$LPI = f_1 \cdot f_2 \cdot \frac{1}{H_{(-20°C)} - H_{(0°C)}} \cdot \int_{H_{(0°C)}}^{H_{(-20°C)}} \epsilon \cdot w^2 \cdot g_{(w)} \cdot dz \tag{2.3}$$

Specifically, the boolean function f_1 analyzes the highest updraft velocity w, determined from all the vertical layers in the model. f_1 equals 0 if the maximum updraft velocity is less than $w_{max,0} = 1.1\ m \cdot s^{-1}$ for more than half of the model grid points in a square of 10 km around a specific point. This prevents false alarms

in case of relatively weak convection over a determined area. When the updrafts are strong enough to support significant convection, then f_1 is set to 1.

$$f_1 = \begin{cases} 1 & : \ a \geqslant 0.5 \\ 0 & : \ a < 0.5 \end{cases} \tag{2.4}$$

with

$$a = \frac{\iint \begin{cases} 1 & : \ max[w(z)] \geqslant w_{max,0} \\ 0 & : \ max[w(z)] < w_{max,0} \end{cases} \ dx \ dy}{\iint dx \ dy} \quad , \quad w_{max,0} = 1.1 \ m \cdot s^{-1} \tag{2.5}$$

In a similar way, f_2 investigates the mean available convective energy determined from parcel theory — with virtual temperature T_v and surface pressure p_s — over a 20 km square surrounding a grid point. The boolean function takes the value 0 if the calculated, mean buoyancy term B_{ML} fails to get to a predetermined threshold B_0 and equals 1 if the forecasted buoyancy is strong and widespread enough.

$$f_2 = \begin{cases} 1 & : \ B_{ML} \geqslant B_0 \\ 0 & : \ B_{ML} < B_0 \end{cases} \quad , \quad B_0 = -1500 \ J \cdot kg^{-1} \tag{2.6}$$

with

$$B_{ML} = \frac{\iint \int_{p_s - 50hPa}^{p_s - 550hPa} R_d (T_{v,parcel} - T_{v,env}) \ \ d(\ln p) \ dx \ dy}{\iint dx \ dy} \tag{2.7}$$

Both f_1 and f_2 boolean functions have been included with the purpose of minimizing false alarms in the LPI. If convection is modeled to be weak, isolated and/or shallow, then only showers or single-celled, short-lived thunderstorms are usually expected. This brings a much lower risk for electrical discharges to occur.

Finally, $g_{(w)}$ is a relatively simple boolean function that excludes (i.e. takes the value 0) all the vertical layers in the model where the updraft velocity w is less than 0.5 $m \cdot s^{-1}$. In all other cases the function equals 1 and does not interfere with the overall algorithm.

$$g = \begin{cases} 1 & : \ w \geqslant 0.5 \ m \cdot s^{-1} \\ 0 & : \ w < 0.5 \ m \cdot s^{-1} \end{cases} \tag{2.8}$$

This retains only the vertical levels that provide significant lifting, which paired with the right mixing of ice and liquid water particles leads to ideal charge separation conditions.

2.3 Lightning detection system - LINET

The LPI forecasts will be compared to the observed lightning flashes detected by the Lightning detection NETwork (LINET). LINET is a ground based Low to Very

Low Frequency (VLF/LF) lightning detection system that has been developed at the University of Munich during the late 2000s (Betz et al., 2004, 2009).

Figure 2.4: Distribution of the LINET sensors across Europe during Summer 2019. Grey squares are active sensors, red squares are temporarily unavailable sensors. Property of nowcast GmbH.

The LINET sensors detect the magnetic emissions of a stroke with extremely high precision in time (ns) and locate its position through an optimised Time Of Arrival (TOA) algorithm. Working in the VLF/LF range ($\sim$ 10 kHz), the distance between each sensor can be up to 200 km without degrading the precision of the detection algorithm. Compared to conventional VLF detections systems, the LINET network is able to detect weaker lightning below 5 kA, can better distinguish between cloud-to-ground lightning (CG) and intra-cloud lightning (IC) and has an horizontal precision of around 150 m (Karagiannidis et al., 2019), provided of course a sufficiently high density of sensors. This is one order of magnitude higher than the COSMO-D2 horizontal grid spacing and is therefore more than suitable for the purpose of this study. The network also provides an estimation of the height of the discharged electricity, giving a true third dimension to the observed data. However, this feature is not relevant for the analysis conducted in this work. LINET mostly covered just a portion of central Europe in the early stages, but expanded significantly during the last decade, reaching a truly global scale. For the purposes

of this work, it is important to state that the LINET network has its highest sensor density in and around Germany (see Figure 2.4) and that soon after its release it became the official lightning detection network of the DWD.

2.4 Predictability analysis and verification

Verifying lightning activity in the framework of a high resolution EPS comes with two main problematic aspects from a statistical point of view.

The first and basic issue is that lightning strikes are rare phenomena. The datasets are therefore going to show a very low base rate. This makes common verification methods based on contingency tables — such as the hit rate or the false alarm rate — not adequate to express the quality of the forecast, as they degrade very fast to trivial values when it comes to rare events (Ferro, 2007). In order to solve this issue, some new verification methods designed for extreme and/or rare events have emerged during the last decades (Ferro and Stephenson, 2011). The Extremal Dependence Index (EDI) and its complement symmetric twin, the Symmetric Extremal Dependence Index (SEDI) specifically address this problematic.

The second problem is known as double-penalty effect (Mass et al., 2002; Rossa et al., 2008) and affects grid-point verification when the horizontal resolution of a weather model is very high. Suppose the COSMO-D2-EPS is showing very high probabilities for lightning strikes at a certain grid point in the domain at 14:15 UTC on a certain day. At 13:30 UTC on the same day, a thunderstorm causes lightning activity in the grid cell next to the forecasted one. This would not only lead to a false alarm in the contingency table, but would also count as a miss, penalizing the forecast twice. Furthermore, from an operational point of view, a spacial error of around 5 km and a time offset of 45 minutes are mostly considered a very good performance in case of convective activity. As this issue arose with the first high resolution models a couple of decades ago, there have been already several attempts at addressing this problematic. One possible way is adding another dimension to the verification scheme: the spatial scale. This is the concept behind so–called fuzzy verification methods, which verify the skill of the forecasts for different spatial scales. One of the most used scores in this field is the Fractions Skill Score (FSS)(Roberts and Lean, 2008). A different, but equally effective approach is the one applied in object–based verification methods such as the Structure-Amplitude-Location (SAL) (Wernli et al., 2008). In this case, the areas with contiguous lightning activity are considered as one single object and the verification is done by comparing the position and the intensity of the forecasted and the observed objects.

All three approaches (SEDI, FSS and SAL) have been included in this study in order to asses the performance of the COSMO-D2 EPS LPI forecasts and will be briefly described in the following sections. Furthermore, for the FSS and the SAL a so called "probabilistic" version of the verification method is discussed, with the goal of analyzing the spread-error relationship of the high resolution ensemble. This way, one can investigate the ability of the ensemble in modeling the predictability of

a specific variable (in this case the lightning activity) and if the system is capable of correctly assessing the actual uncertainty in the forecast.

2.4.1 Symmetric Extremal Dependence Index – SEDI

The SEDI has been developed by Ferro and Stephenson (2011). It is based on the categories obtained from a usual contingency table such as the one shown in Table 2.2 for binary events. A binary event is defined applying a specific threshold q to a field, as show in Eq. (2.17). The SEDI is a refinement of other verification measures such as the Symmetric Extreme Dependency Score (SEDS) or the Extremal Dependence Index (EDI). These methods came up during the 2000s and are focused on providing a statistically significant and stable verification analysis even when it comes to extremely rare phenomena.

	Observed	Not observed	
Forecasted	a	b	a+b
Not forecasted	c	d	c+d
	a+c	b+d	n

Table 2.2: Generalized contingency table for observed and forecasted events.

In order to make the score base–rate–independent, which is crucial for rare events as the base rate $f_0 = \frac{a+c}{n}$ would tend to zero, the SEDI is only dependent from the Hit rate H and the False alarm rate F and is defined as follows:

$$SEDI = \frac{\log(F) - \log(H) - \log(1-F) + \log(1-H)}{\log(F) + \log(H) + \log(1-F) + \log(1-H)} \tag{2.9}$$

with H being the Hit rate and F being the False alarm rate:

$$H = \frac{a}{a+c} \tag{2.10}$$

$$F = \frac{b}{b+d} \tag{2.11}$$

The SEDI can take values between -1 and +1. When H tends to 0 and F tends to 1, the SEDI gets close to -1 and the verified forecast has the worst skill possible. For a very good forecast, the SEDI tends to +1 as H tends to 1 and F tends to zero. A SEDI value of 0 is equal to the skill of a random forecast.

2.4.2 Dispersions Fractions Skill Score – dFSS

The benefits of high resolution weather forecasts are undoubted, especially if the focus lies on convective processes. High resolution prediction systems are a much more accurate representation of reality compared to coarse ones, which plays a role in many different ways. From a better representation of topography and land use to

the importance of being able to handle smaller-scale processes by resolving their correspondent physical equations rather than using predetermined parametrization schemes. Nonetheless, this improved resolution does not necessarily lead to a correspondent increase in accuracy, at least for grid-point verification. As previously discussed, double-penalty effects and possible forecast errors in global models being passed over to nested systems usually affect high resolution models, lowering the accuracy especially when it comes to very localized and rare events, such as lightning strikes. In recent years, such issues have been addressed in a number of ways, with so called fuzzy verification methods probably being the most effective and intuitive solution. One of the most used and studied example of fuzzy verification is the Fractions Skill Score, or FSS (Roberts and Lean, 2008). The FSS expresses the ratio between the actual Mean Square Error (MSE) and the highest possible MSE ($MSE_{(ref)}$) for a specific subset of the model domain with dimension n. In order to neutralize possible double penalty effects, n can increase from 1 gridpoint to several hundreds gridpoints. Rather than verifying the forecast for a specific location, the FSS is therefore able to verify the fields at different spatial scales as n increases.

$$FSS_n = 1 - \frac{MSE_n}{MSE_{n(ref)}} \tag{2.12}$$

and for each sub-domain n with a $N_x \cdot N_y$ grid:

$$MSE_{(n)} = \frac{1}{N_x \cdot N_y} \cdot \sum_{i=1}^{N_y} \sum_{j=1}^{N_x} (A_{(n)ij} - B_{(n)ij})^2 \tag{2.13}$$

$$MSE_{(n)ref} = \frac{1}{N_x \cdot N_y} \cdot \sum_{i=1}^{N_y} \sum_{j=1}^{N_x} (A^2_{(n)ij} - B^2_{(n)ij}) \tag{2.14}$$

The fields $A_{(n)ij}$ and $B_{(n)ij}$ are the derived fractions of the original, binary fields A and B for a specific square — or rectangular, but in our case $N_x = N_y$ — window n and are defined as follows:

$$A_{(n)ij} = \frac{1}{N_x \cdot N_y} \cdot \sum_{k=1}^{N_y} \sum_{l=1}^{N_x} A \left[i + k - 1 - \frac{(n-1)}{2}, j + l - 1 - \frac{(n-1)}{2} \right] \tag{2.15}$$

$$B_{(n)ij} = \frac{1}{N_x \cdot N_y} \cdot \sum_{k=1}^{N_y} \sum_{l=1}^{N_x} B \left[i + k - 1 - \frac{(n-1)}{2}, j + l - 1 - \frac{(n-1)}{2} \right] \tag{2.16}$$

As the datasets used in this work are considerably big in size, the two fractions fields $A_{(n)ij}$ and $B_{(n)ij}$ have been calculated using the summed area table or integral image method. This approach, very common in the field of digital image processing,

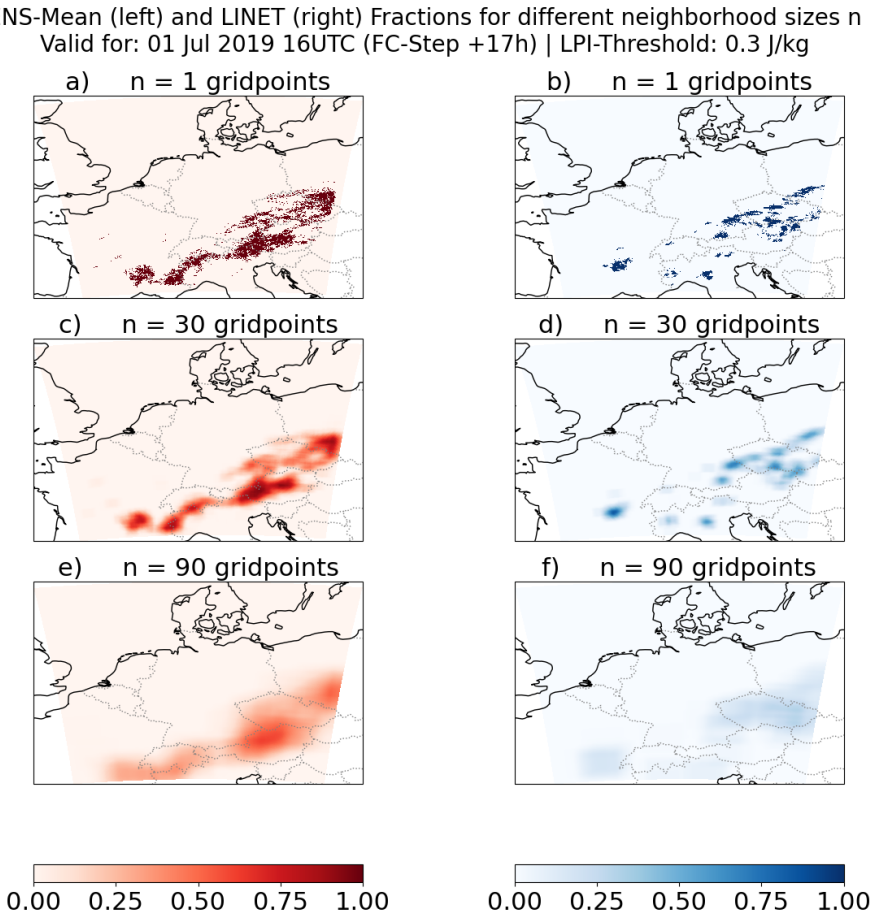

Figure 2.5: Example of calculated fractions for the COSMO-D2 EPS ensemble mean (a), c), e), threshold 0.3 J/kg) and the LINET observed lightning flashes (b), d), f), threshold 1 flash) for different values of the neighborhood size n.

has also been thoroughly described specifically for the FSS case (Faggian et al., 2015).

The two data fields A and B that are used for the calculation of the FSS has to be binary fields. Typically, A and B express a forecasted dataset and an observed dataset which exceed a fixed threshold q, as shown below.

$$A = \begin{cases} 1 & : \ A_{i,j} \geqslant q \\ 0 & : \ A_{i,j} < q \end{cases} \quad and \ B = \begin{cases} 1 & : \ B_{i,j} \geqslant q \\ 0 & : \ B_{i,j} < q \end{cases} \tag{2.17}$$

When the forecast being verified has a very good skill, the FSS approaches +1, while a completely useless forecast would lead to FSS values close to 0. In general, the verified forecast is considered skillful at a specific spatial scale n when $FSS_n \geq 0.5 + \frac{f_0}{2}$, with f_0 being the base rate (Roberts and Lean, 2008).

Shifting the focus to probabilistic forecasts, there are two different ways of making

use of the FSS. First, there is the more conventional one, where the FSS is obtained by analyzing observational data A with a significant product of the ensemble — such as the ensemble mean — B. This study refers to this version of the FSS as "error" FSS or eFSS. The eFSS can be seen as a measure that defines the spatial skill of the EPS. When working with large datasets, the average eFSS over a season will be the mean of all the eFSS values calculated for each timestamp. The second way of interpreting the FSS for EPSs is to compare two members of the same ensemble system rather than observations and forecasts. In this case, A and B are two forecast fields coming from two different members of the ensemble. In other words, this method gives a spatial measure of how much the two ensemble members diverge one from the other. If the whole set of ensemble members is analyzed with this approach and then averaged, the resulting measure can be used as a mean for expressing the ensemble dispersion in space (or spatial spread). For this reason, this study refers to it as "dispersion" FSS or dFSS (Dey et al., 2014). Using the same approach as for the eFSS, for larger datasets the seasonal dFSS is the averaged dFSS computed over several timestamps for each couple of ensemble members. In this study, the single components of the FSS are summed before performing the averages, even if newer studies state that this might not always be the best solution depending on each specific case (Mittermaier, 2021). Given the fact that the spatial spread (dFSS) is calculated and obtained in an analogous way as for the spatial skill (eFSS), the two measures can be directly compared for different spatial scales n. Therefore, thanks to the comparison between the eFSS and the dFSS at different scales, the conventional spread-error relationship gains an additional dimension: the spatial scale.

2.4.3 Ensemble Structure-Amplitude-Location – eSAL

A second approach that targets the problematic effects of grid-point verification for high resolution forecasts is the Structure-Amplitude-Location, or SAL (Wernli et al., 2008). The SAL is classified as an object-based verification method and it is based on the recognition of contiguous features or "objects" in a forecast field, for example large precipitation bands or, considering this study, areas with widespread potential for lightning activity. These objects are then verified against an analogous observed field in terms of their intensity and positioning within the domain. As the name already suggests, the SAL is made of three different components: the Structure S, the Amplitude A and the Location L. S compares the volumetric structure of the detected features from two different fields. A investigates the overall intensity or amplitude of the two fields, regardless of the object-based analysis. Finally, L performs a center of mass analysis over the whole domain as well as a center of mass analysis for each single pair of features within the domain and is therefore a measure for spatial skill.

The S and A components are constrained between -2 and +2, with 0 meaning a perfect match between the two fields for what concerns the shape of the identified

features and the domain-wide intensity of the parameter being verified. If S tends to positive values, then the system is forecasting features that are either too stretched horizontally and/or have much higher peaks in magnitude compared to the observed objects. A on the other side can be seen as a bias measure of the whole field: if A is positive then the forecast has a positive bias compared to the observations. The Location component L is obtained by summing two different parts L_1 and L_2. L_1 expresses the difference between the center of masses of the two fields being compared, calculated over the whole domain regardless of the single objects. L_2 compares the two fields in terms of the average distance of the center of masses of all the identified features from the center of mass of the whole domain. L_1 and L_2 can range from 0 to +1 and L can therefore vary from 0 to +2. If all the L components are 0, then the fields being compared show absolutely no discrepancies for what concerns the center of masses of all the identified features and the detected domain-wide center of mass.

The single components A, L_1, L_2 and S are defined as follows (Wernli et al., 2008):

$$A = \frac{\overline{C} - \overline{D}}{0.5 \cdot (\overline{C} + \overline{D})} \tag{2.18}$$

with $\overline{C}$ and $\overline{D}$ being the domain-wide average of two fields (either observation against forecast or forecast against forecast).

$$L = L_1 + L_2 \tag{2.19}$$

with

$$L_1 = \frac{|\mathbf{x}(C) - \mathbf{x}(D)|}{d} \tag{2.20}$$

$$L_2 = 2 \cdot \left[\left| \sum_M \frac{\sum_{obj} (C \cdot |\mathbf{x}(C) - \mathbf{x}_{obj}(C)|)}{\sum_{obj} C} - \sum_M \frac{\sum_{obj} (D \cdot |\mathbf{x}(D) - \mathbf{x}_{obj}(D)|)}{\sum_{obj} D} \right| \cdot \frac{1}{d} \right] \tag{2.21}$$

d is the largest distance reachable in the considered domain. $\mathbf{x}()$ is the center of mass of the whole field, while $\mathbf{x_{obj}}()$ is the center of mass of each identified object in the domain. $\sum_M$ is the sum over the list of M detected objects in the domain, while $\sum_{obj}$ is the sum of the values in each grid point inside a single object.

$$S = \left(\sum_M \frac{\sum_{obj} \left(C \cdot \frac{C}{C_{max}} \right)}{\sum_{obj} C} - \sum_M \frac{\sum_{obj} \left(D \cdot \frac{D}{D_{max}} \right)}{\sum_{obj} D} \right) \cdot \left[0.5 \cdot \left(\sum_M \frac{\sum_{obj} \left(C \cdot \frac{C}{C_{max}} \right)}{\sum_{obj} C} + \sum_M \frac{\sum_{obj} \left(D \cdot \frac{D}{D_{max}} \right)}{\sum_{obj} D} \right) \right]^{-1} \tag{2.22}$$

Again, $\sum_M$ is the sum over the whole list of M detected objects in the domain, while $\sum_{obj}$ is the sum of the values for each grid point inside each object. C_{max} and D_{max} are the maximum values inside a single object for the two fields.

During the last decade, SAL has already been applied in a number of studies, mostly related to high resolution precipitation forecasts (Wernli et al., 2009; Wittmann et al., 2010; Schneider et al., 2019; Zhaoye et al., 2022), but also to radar reflectivity fields (Lawson and Gallus Jr, 2016) or volcanic ash forecasts (Wilkins et al., 2016). The object-based approach presented in SAL is therefore very effective when it comes to precipitation fields. Given the fact that lightning activity often matches areas with convective precipitation, SAL could deliver insightful information also in the verification of lightning flashes. Furthermore, in recent years also the first papers focusing on converting SAL to be used on ensemble forecasts for precipitation fields already emerged (Radanovics et al., 2018; Marsigli et al., 2019). Nevertheless, in this study two inherently different fields with different units — the observed lightning activity and the LPI — have to be compared. This fact leads to necessary considerations and adjustments to be made prior to the verification process. These are thoroughly investigated in Chapter 4, Sections 4.2 and 4.4. Otherwise, the approach used for the probabilistic version of the SAL is similar to the one described in the previous Section for the dFSS. On the one hand, in a more conventional way, the fields C and D can be for example the ensemble mean forecast verified against the observed dataset. On the other hand, C and D can also represent two forecast fields of two different ensemble members which can be compared with the goal of measuring the spread of the EPS. In the following chapters, this probabilistic version of the SAL will be referenced as eSAL. The total eSAL will then be the average SAL calculated over all the possible pairs of members of the ensemble. An example of eSAL for precipitation fields is documented in Radanovics et al. (2018). Compared to this example, in this study only the Structure component S is obtained in a different way. When considering two ensemble members C and D, the conventional version of Eq. (2.22) has been applied and the results for each pair of the ensemble have then been averaged over the whole EPS. Apart from this, for calculating the eSAL Eq. (2.18), (2.19), (2.20), (2.21) and (2.22) are modified only for the fact that an ensemble average SAL — i.e. the mean of 190 SAL values resulting from the comparison of all the possible couples of the 20 members of the EPS — for each model run is being calculated, as described in Radanovics et al. (2018).

3 Datasets

Both the COSMO-D2 EPS fields and the LINET observed lightning flashes have been made available from April 2019 until September 2019. As discussed in Chapter 4, Section 4.1, this analysis has been conducted using only data from the summer season (June, July and August, or JJA).

3.1 Observations - LINET

The LINET data have been extracted from the DWD database for central Europe from April 2019 until September 2019. They are provided as a list of observed flashes with latitude (deg), longitude (deg), time of observation (s), altitude (m) and amplitude (A), though the last two characteristics are not relevant for this study.

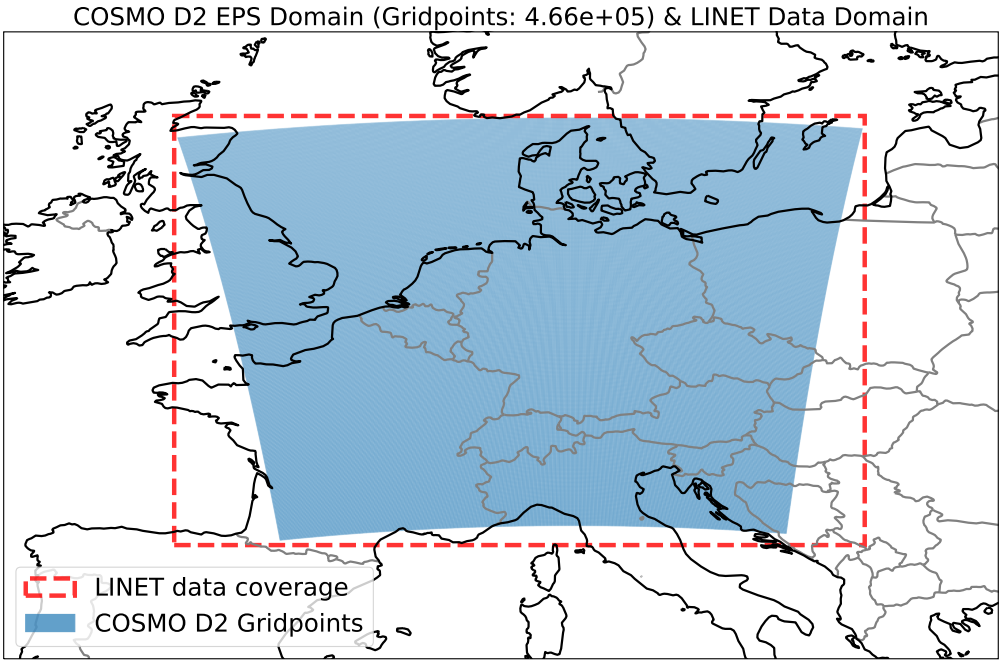

Figure 3.1: Coverage of the LINET data (red dashed line) compared to the COSMO-D2 EPS domain (blue area). From Salmi et al.(2022).

The raw dataset covers large parts of Europe, including of course the whole COSMO-D2 EPS domain. As the data have been provided as validated, no further quality check is applied.

3.2 Forecasts - COSMO-D2 EPS LPI

The COSMO-D2 EPS LPI forecast fields are provided as NetCDF files for the same time window of 2019 with the highest space and time resolution possible, i.e. around 2.2 km grid spacing on average, one forecast step every 15 minutes. The dataset includes all 20 members of the ensemble. Given the significant amount of data involved, only the 00 UTC run for each day with forecasts up to 24 hours is considered in the study. Figure 3.2 shows an example of the EPS product, with overnight convection and significant potential for lightning activity expected in northern Germany at different locations from the EPS members.

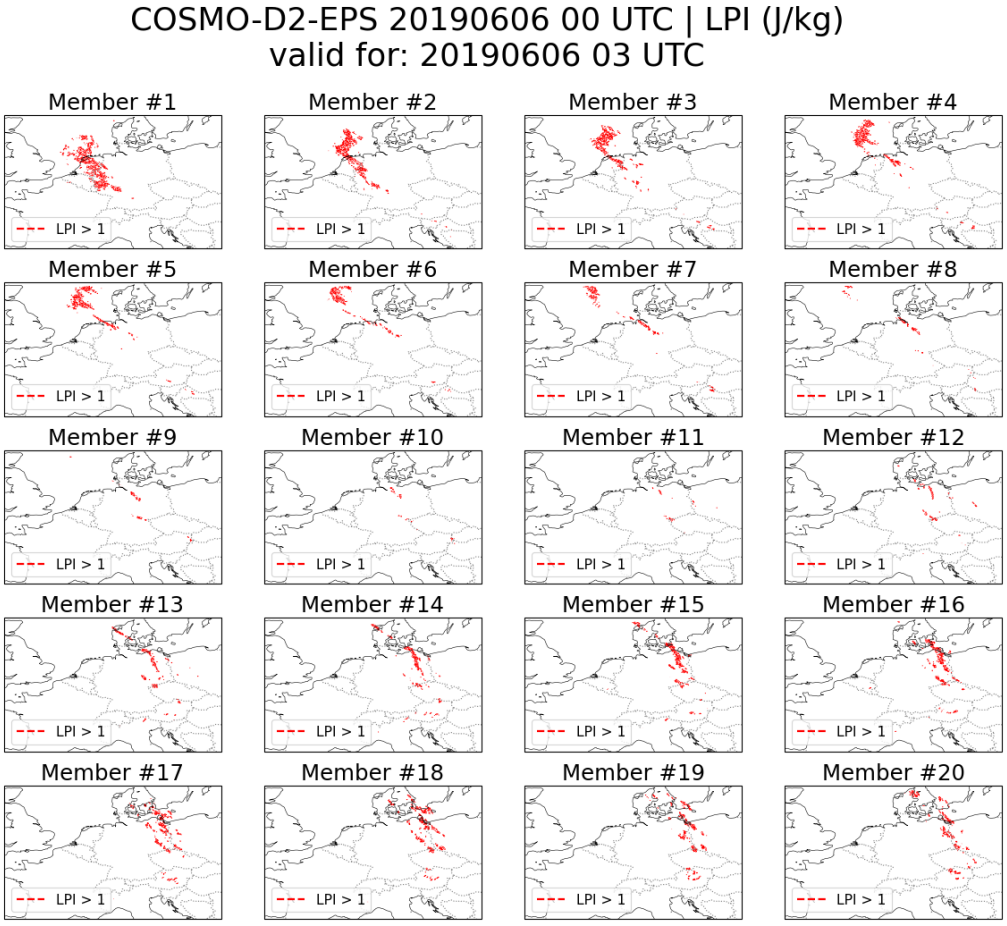

Figure 3.2: Example of LPI output fields from all the 20 members of the high resolution ensemble (contours for LPI $> 1\ J \cdot kg^{-1}$). COSMO-D2 EPS 00 UTC run from June, 6th 2019, valid for June, 6th 2019 at 03 UTC.

3.3 Data homogenization

As shown in Figure 3.1, the LINET dataset is covering a slightly bigger area compared to the COSMO-D2 EPS domain. Furthermore, the lightning data are sparse geographical points in space, while the LPI forecasts are linked to a regular, 2–D mesh grid. Thus, in order to perform the verification, the LINET dataset needs to be trimmed and brought onto the COSMO-D2 EPS mesh grid. This homogenization process in space has been done using a linear nearest neighbor algorithm. The observed lightning flashes have been assigned to the nearest COSMO-D2 EPS grid point, while LINET data falling outside of the COSMO-D2 EPS domain have been rejected.

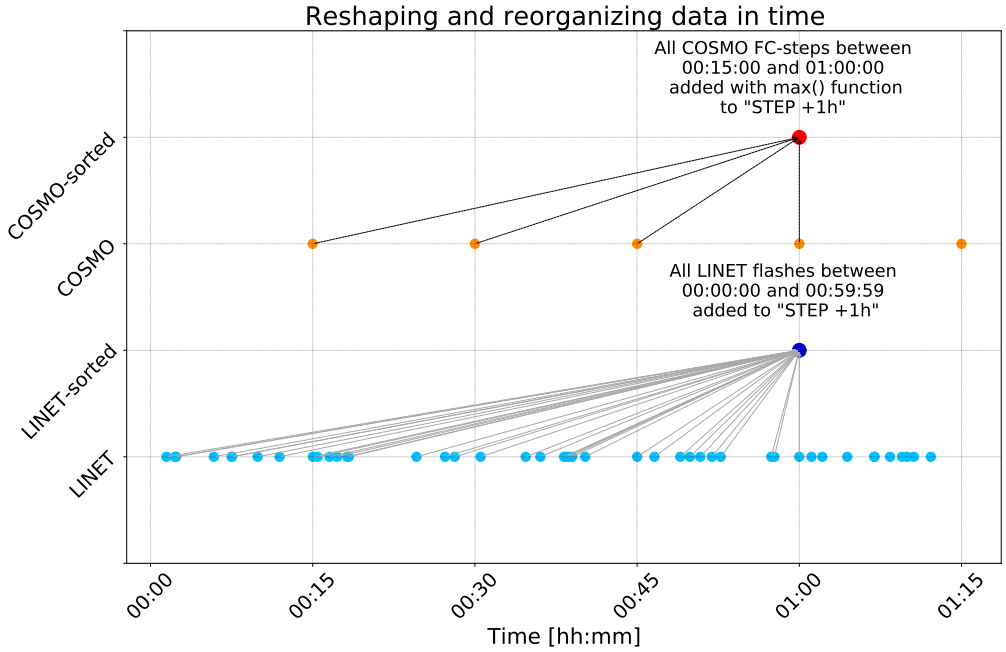

Figure 3.3: Schematic view of the applied procedure to aggregate both datasets in time.

A similar homogenization process is also necessary for the time dimension. The lightning data are sparse points along the time axis, while the LPI dataset has regular time steps every 15 minutes. However, verifying the forecast with such an high frequency would have required a significantly higher computational capacity. Another positive aspect of integrating in time is also the fact that rapid variations in the fields will be filtered out. Therefore, as this work is intended as a preliminary study of the skillfull scale and predictability of convective forecasts, all data have been aggregated using an hourly frequency. The LINET observed flashes have been summed up on an hourly basis, while the maximum LPI value for each hour has been computed. The process is schematized in Figure 3.3. This approach can be considered as an upscaling in time, thus partly neutralizing time offsets in the

model's forecast of convection. This aspect needs to be taken into account when proceeding to the actual verification process.

The resulting, homogenized datasets that will be used for the analysis are therefore gridded fields of number of observed lightning flashes and forecasted LPI values in J/kg, available with an hourly frequency and a spatial resolution of 2.2 km, on average.

4 Methodology

As described in the previous Section, this study has been conducted using hourly fields of observed LINET flashes and forecasted LPI fields. Given the fact that convection is often strongly correlated to daytime heating and that model performances tend to degrade with increasing forecast lead time, an analysis that differentiates between hourly forecast steps is the most appropriate choice. Therefore, all the verification scores shown in the following chapters are aggregated based on the forecast lead time from the start of the model run (00 UTC). Furthermore, as only the 00 UTC COSMO-D2 EPS model runs are taken into account with a maximum forecast lead time of +24 hours, the forecast steps in UTC being verified coincide with the solar time in Central Europe, +1 hour (solar time = CEST -1 hour = UTC +1 hour). This plays a central role when looking at the verification outputs, as most of the convective activity will be concentrated in the afternoon hours. Finally, to asses the skill of the EPS, the ensemble mean has been taken as the reference forecast for the SEDI, the eFSS and the SAL.

4.1 Data filtering

The raw data are available for a period from April to September 2019. Table 4.1 shows the distribution of the monthly observed lightning activity and the corresponding percentage of total lightning in the whole time window. Convection is known to be a summer topic for mid-latitudes and therefore there is no surprise in counting around 92% of the observed lightning flashes in June, July and August. As April, May and September accounts for just 8% of the total lightning activity, this study focuses only on the three summer months.

Month	Number of observed flashes	Percent of total
April 2019	0.4 x 10^6	2%
May 2019	0.8 x 10^6	4%
June 2019	6.4 x 10^6	34%
July 2019	6.5 x 10^6	35%
August 2019	4.2 x 10^6	23%
September 2019	0.4 x 10^6	2%
Total	18.7 x 10^6	100%

Table 4.1: Distribution of LINET observed lightning flashes in dataset per month.

Furthermore, another filtering method has been applied in order to focus the study only on truly active convective days. For achieving this, a simple formula based on the fraction of domain with observed lightning flashes for each hourly forecast step has been used. If the fraction of domain with lightning activity — i.e. the number of gridpoints with at least one observed lightning divided by the total number of gridpoints in the domain — is less than one third of the summer average fraction for that specific forecast step, then the day is discarded. The one third threshold has been chosen in order to maximize the ratio between retained lightning activity and discarded days in the dataset. By doing this, the number of days retained in the study are on average 60% of the total 92 days for the summer season 2019 and at the same time most of the observed lightning activity is also still included in the datasets.

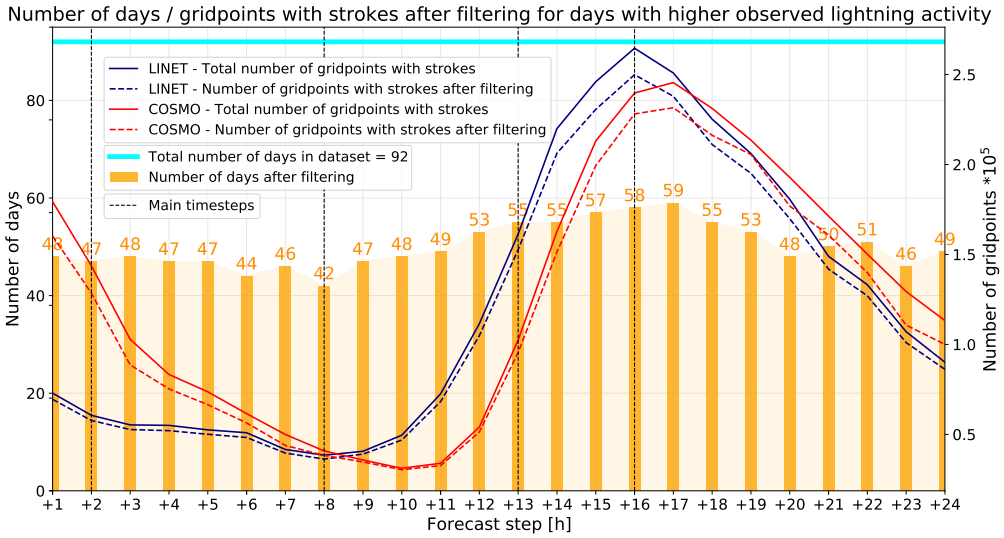

Figure 4.1: Amount of days (bar plot) in dataset and singular gridpoints (lines) being retained after the filtering process compared to the unfiltered datasets. The "main timesteps" are the forecast steps that have been investigated in details during the verification process (see Chapter 5).

Of course, by applying an observations-based filtering method the risk of underestimating the false alarms and therefore of adding a bias to the study is very much given. However, as after the filtering process more than 90% of both the observed and the forecasted lightning activity is retained — as shown in Figure 4.1 — this method is not affecting the analysis in a decisive way, especially during the main convective window in the afternoon hours. This hypothesis is supported by the statistics of the False Alarm Ratio (FAR) for the filtered and unfiltered datasets. Based on the same contingency table described in Table 2.2, the FAR is defined as follows:

$$FAR = \frac{b}{a+b} \tag{4.1}$$

As shown in Figure 4.2, the FAR is very high in both the filtered and the unfiltered dataset, with changes in the range 0.1 to 1% of the FAR. For better reference, also FAR values for simple upscalings in space are also shown.

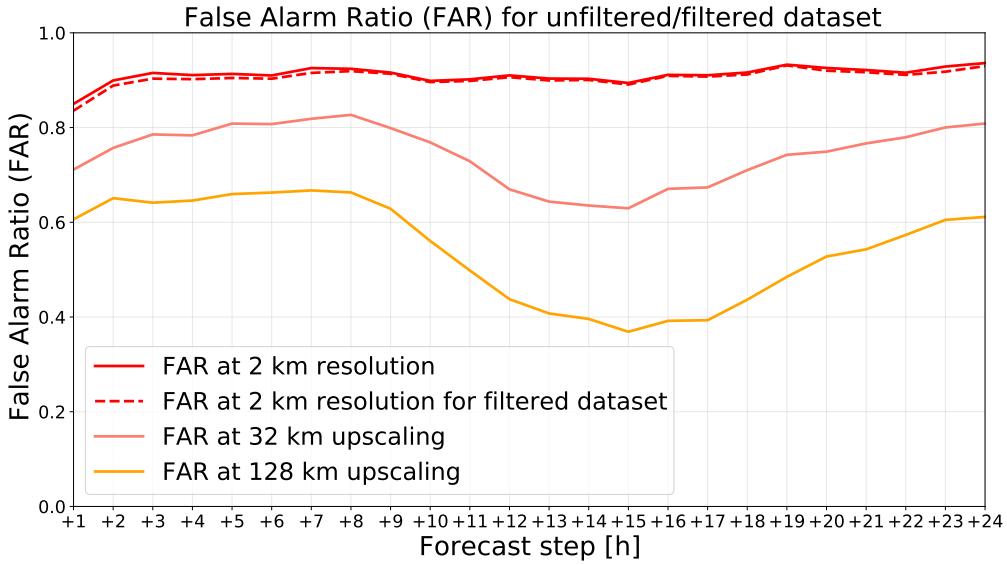

Figure 4.2: False Alarm Ratio (FAR) for the filtered and unfiltered datasets.

Finally, a third filtering approach has been introduced only for the SAL analysis. As the SAL is inherently an object-oriented measure, large portions of the domain without lightning activity might distort the final results. Therefore, on top of the previously described filtering approaches also a geographical zooming algorithm has been applied prior to the SAL analysis. For each day, the maximum and minimum point values (i,j) of the mesh grid with observed lightning flashes have been identified, defining a rectangular area where the lightning activity is occurring. In order to discard large inactive regions and after allowing for a safety buffer of 20 gridpoints (around 45 km) in every direction, the SAL analysis has been conducted only on this zoomed rectangular area. This way, the domain of the analysis changes for each day in the dataset, but as the discarded area does not include lightning activity this has no influence on the quality of the verification.

4.2 Choosing thresholds

Some of the verification measures discussed so far (the SEDI and the FSS) need binary fields to compare the LINET observed flashes to the COSMO-D2 EPS LPI forecasts. For the SAL, a more complex process is needed, as described in Section 4.4. In order to obtain binary fields from the original datasets, specific thresholds need to be defined. The choice has been made by looking at the FSS score for different combinations of thresholds for both the observations and the forecasts. The results are shown in Figure 4.3. For the observed flashes the solution that

maximizes the FSS for most of the forecast steps is obtained by considering more than one flash per hour in one grid point (around 2 km^2) as a proper threshold to set the binary field to 1. For the LPI the decision is not so straight forward, but the threshold that performs best throughout the investigated forecast steps is the one set at more than 1 J/kg.

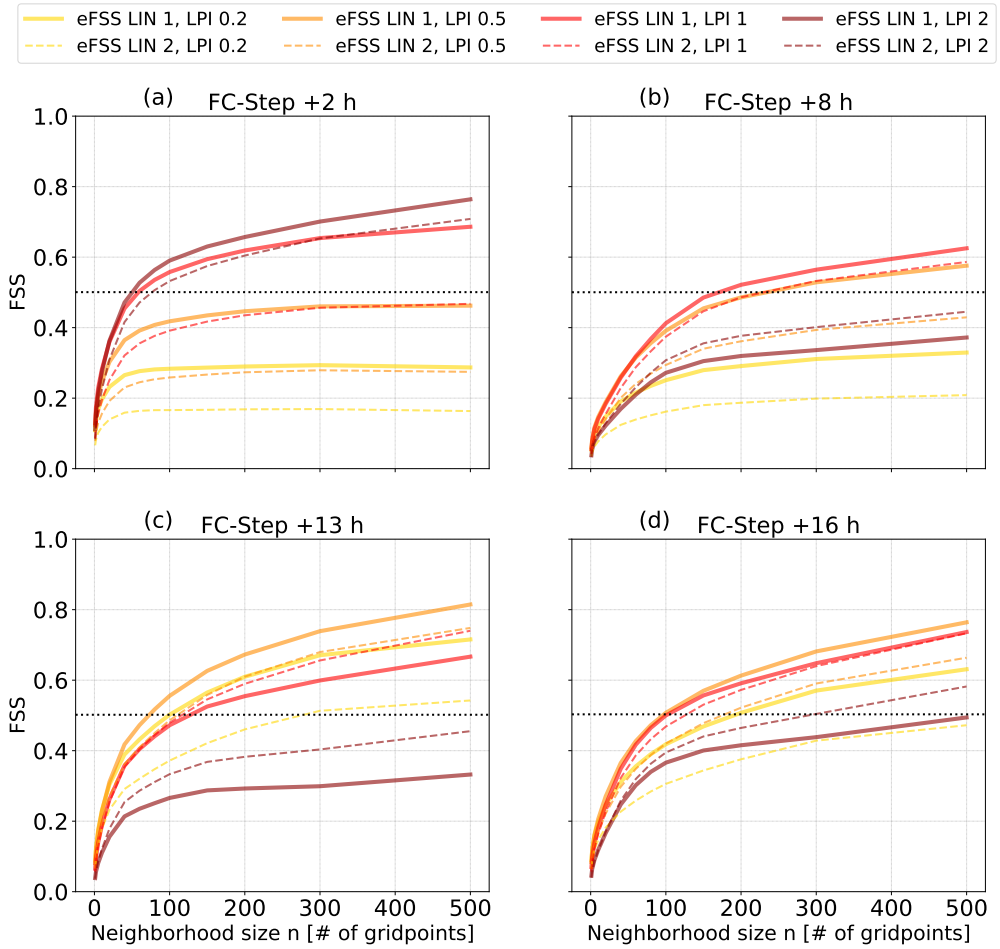

Figure 4.3: FSS scores obtained by applying different thresholds to the LPI fields (0.2, 0.5, 1.0, 2.0 J/kg) and the LINET observed fields (1 and 2 lightning flashes in gridpoint) for the selected forecast steps +2h (a), +8h (b), +13h (c) and +16h (d).

Therefore, the thresholds used for creating binary fields for the SEDI and the FSS as well as for the identification of the objects in the SAL method in this study are set to:

$$LPI > 1 \;\; J \cdot kg^{-1} \quad and \quad LINET \; flashes > 1 \tag{4.2}$$

4.3 SEDI spatial upscale

The FSS is known for allowing users to investigate the skillful scale of a forecast, i.e. the spatial scale at which the forecast becomes useful. In order to introduce a spatial scale to the verification process already from the SEDI analysis, a simple upscale process has been applied to both the observations and the forecasts prior to the SEDI calculation. Besides the original grid spacing (around 2 km), significant upscales at 4, 8, 16, 32, 64 and 128 km have been chosen. The LINET observed flashes have been summed up for all the gridpoints included in the upscale size, while for the LPI the maximum value in the upscaling window has been taken. The result is a simplified version of the fuzzy verification approach that can be evaluated at best with the FSS analysis.

4.4 LPI - LINET Statistical relationship

As partially discussed in previous chapters, the two datasets being compared in this study are similar from an operational point of view in a forecasting environment, but are two clearly separate physical quantities that are measured with different units. In this section, a brief description of the statistical characteristics of the datasets is shown, with a particular focus on the relationship between the two in order to provide homogeneous fields for the SAL analysis.

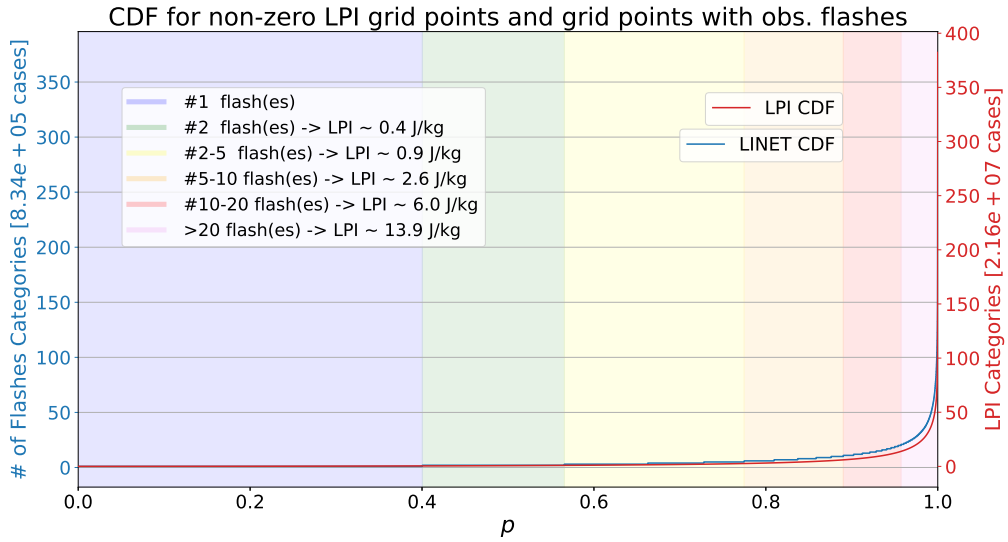

Figure 4.4: Cumulative Distribution Function (CDF) for the LINET observed flashes and the LPI fields. Significant LINET thresholds (number of flashes in gridpoint) and the corresponding are also shown for better reference.

The cumulative distribution functions (or CDF) for the LPI dataset using the ensemble mean and the LINET dataset is shown in Figure 4.4. As both datasets are known to be positive if lightning activity is observed or expected, only non-zero LPI

and LINET gridpoints have been included in the analysis. Several considerations can be made by looking at the two distributions, but the most important one is that both functions are very similar. This is expected, as the LPI is intended to be a forecasting parameter for the occurrence of lightning flashes. Lightning flashes are relatively rare phenomena and even if the CDF plot shown in Figure 4.4 only considers non-zero values in the distribution, this is reflected in the function. In fact, the majority of the gridpoints in the database (60 to 70% of the non-zero gridpoints) take values below 5 flashes per hour or 2.5 J/kg maximum LPI in an area of around 2 km^2. However, there are also extreme events with up to several hundreds flashes per hour with similar LPI maximum values. It is interesting to note that by comparing the two distributions, the threshold $LINET > 2\ flashes$ corresponds to $LINET > 0.9\ J \cdot kg^{-1}$, which is not exactly the threshold which maximizes the FSS as discussed in Section 4.2. Also, the total amount of gridpoints in the whole dataset (i.e. including all the gridpoints in the domain for all the available days and forecast steps) is approximately 10^9. As shown on the y-axes in Figure 4.4, the gridpoints with at least one observed flash are around 10^6, while the ones with non-zero LPI values around 10^7. This leads to two considerations: first, that the vast majority of the gridpoints in the datasets ($> 99\%$) prior to the filtering processes described in section 4.1 show no lightning activity and second that the LPI is non-zero for a significantly larger number of gridpoints compared to the observed lightning flashes.

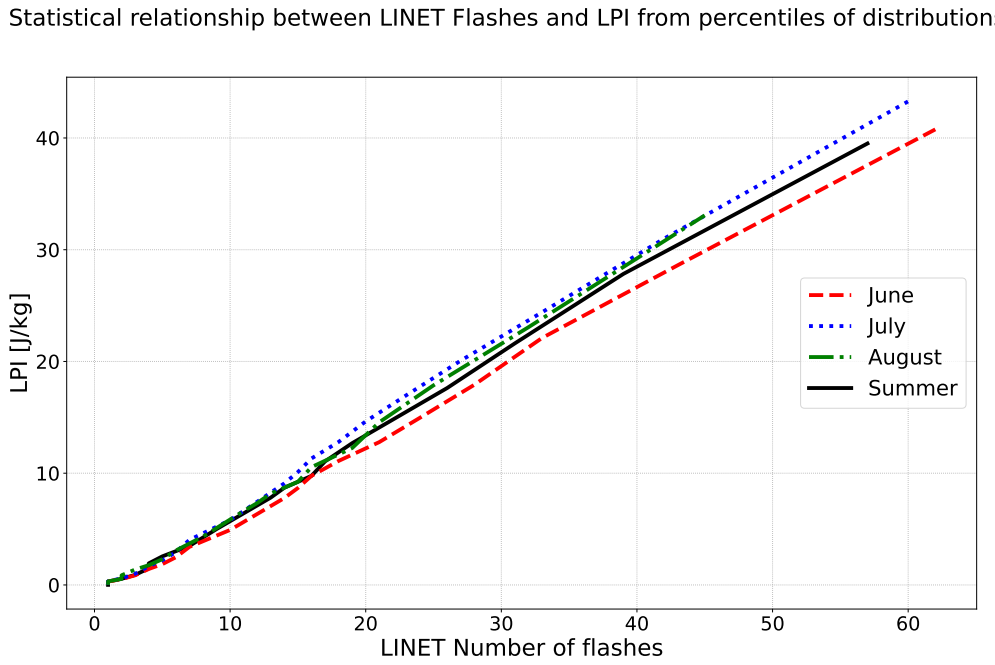

Figure 4.5: Statistical relation between the LPI fields and the LINET dataset based on significant percentiles of both distribution for the whole Summer season as well as for the single summer months.

While for the SEDI and the FSS the usage of binary fields with fixed thresholds automatically remove the issue of comparing two quantities having different units, for the calculation of the SAL components one more step is needed. In fact, the magnitude of the fields does play a role when it comes to the SAL analysis and this implies that the two fields being compared must have the same unit. Therefore, a fitting model needs to be applied to the LPI fields in order to express them in terms of number of lightning flashes. By looking at the percentiles of the two distributions, shown in Figure 4.5, the relation between the two quantities does seem almost linear, though not symmetric. Furthermore, the differences between each month are negligible and the Summer season can be treated as a whole. However, if the analysis is performed for each single forecast step in the dataset — as shown in Figure 4.6 — then some significant discrepancies emerge and the relation is not always linear. As this study is being performed for each forecast step separately, different fitting parameters will be applied for different forecast lead time.

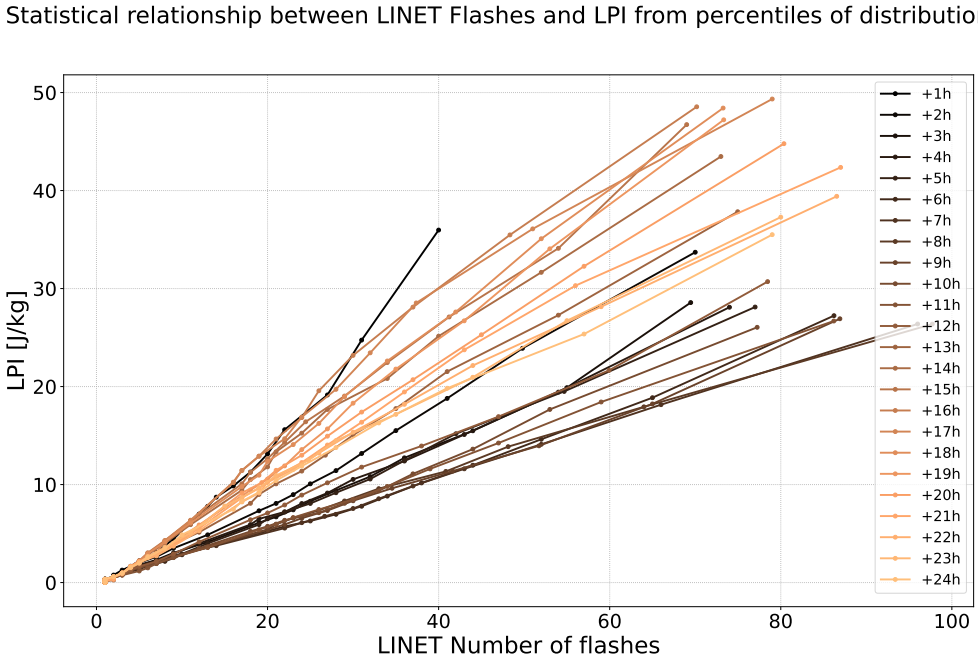

Figure 4.6: Relation between the LPI fields and the LINET dataset based on significant percentiles of both distribution for the whole Summer season focused on each of the 24 forecast steps.

In order to find the best fitting for all the 24 forecast steps while using the same process, an exponential function of the form:

$$a \cdot x \cdot exp(b \cdot x) + c \tag{4.3}$$

has been used. The parameters a, b and c leading to the best curve fitting results have been found for each forecast step by applying a non-linear least squares

regression analysis. Ideally, a different dataset should be used to train the fitting model. However, as no further data were available for previous Summer seasons, the same dataset being analyzed has been used for training purposes. By looking at the intra-month and intra-season variability of the relation as already shown in Figure 4.5, the statistical relationship between the two datasets does seem to be pretty consistent throughout the whole season.

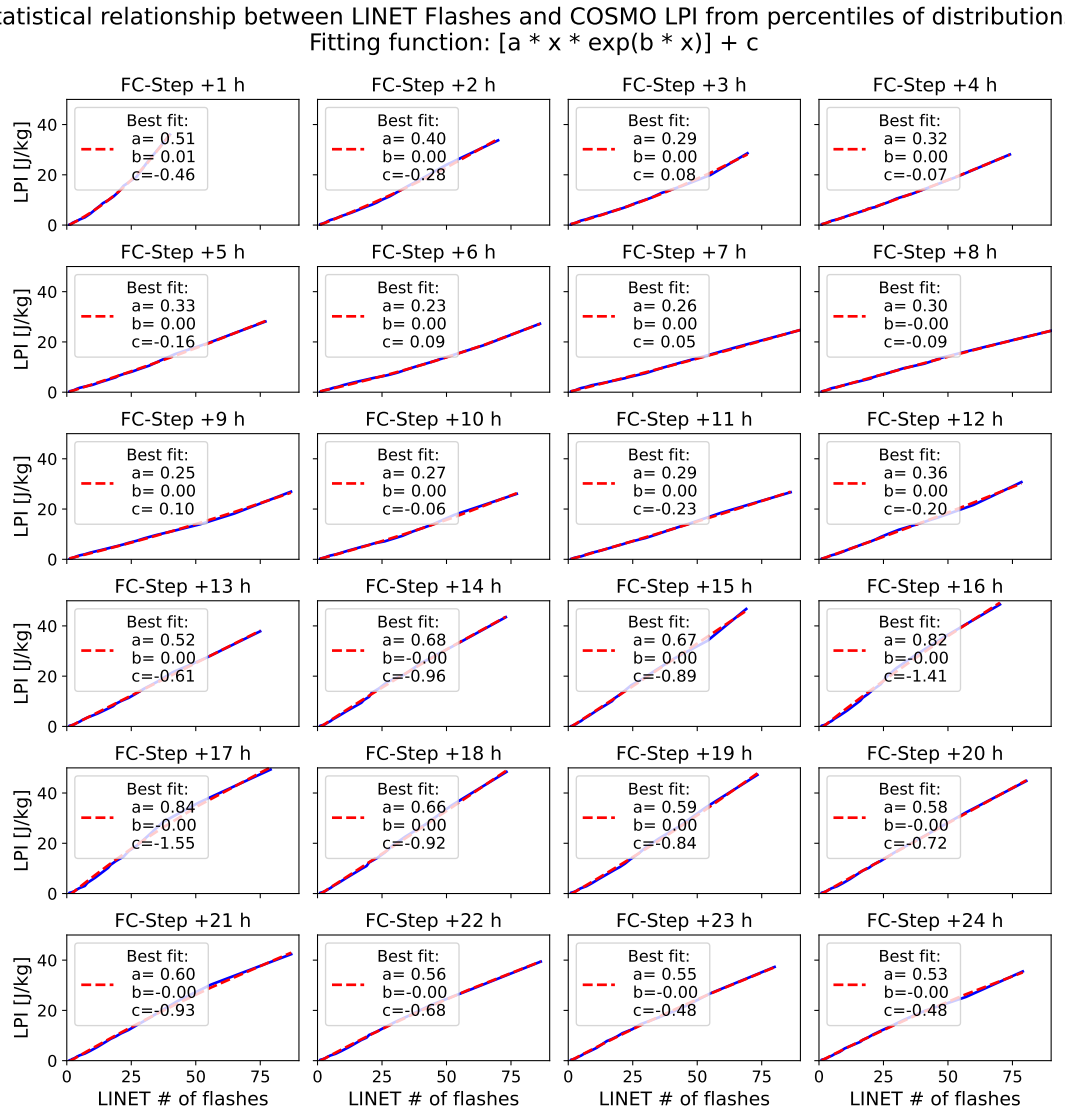

Figure 4.7: Relation between the LPI fields and the LINET dataset based on significant percentiles of both distribution for the whole Summer season focused on each of the 24 forecast steps.

The result of the curve fitting process for each forecast step is shown in Figure 4.7. Using this relation, the LPI fields have been translated into number of lightning flashes and can be used in the SAL analysis. When the LPI values get close to zero,

the relation can significantly vary from time step to time step and might not always be consistent for what concerns the threshold of one flash in gridpoint. For this reason and in order to comply with the same thresholds used for the SEDI and the FSS analysis defined in Equation (4.2), LPI values falling below 1 J/kg have been forced to get a value of 0 lightning flashes in the translated fields, regardless of the fitting process.

5 Verification outcomes

In Chapter 4, the importance of the time of the day in this analysis has already been discussed. Before showing the detailed verification results of this study it is important to briefly analyze the distribution of the lightning activity throughout the 24 considered forecast steps (Figure 5.1). By considering only the 00 UTC model runs up to 24 hours lead time, each forecast step always coincide with the same solar hour of the day +1 hour. This leads to significant differences in the distribution of the lightning activity. The first forecast steps corresponding to the morning hours are 5 to 6 time less active compared to the afternoon hours, when the diurnal convection reaches its peak.

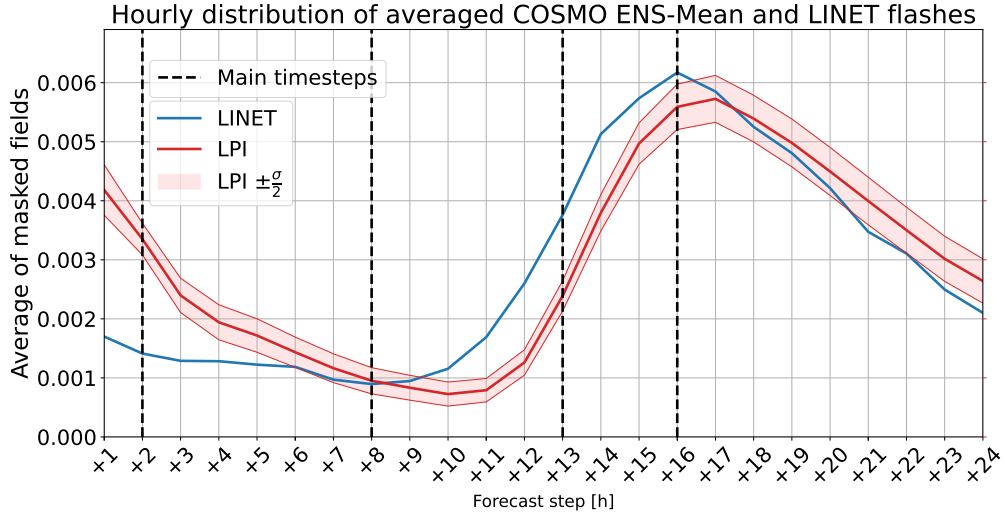

Figure 5.1: Fraction of the domain with observed (LINET) and forecasted (LPI) lightning activity throughout the 24 forecast steps. The calculation is based on the binary (masked) fields that have been then used for the SEDI and the FSS algorithms. For the LPI, half of the standard deviation of the EPS is also shown. Four main forecast steps at +2h, +8h, +13h and +16h have been chosen at significant frames of the daily convective cycle and the model's forecast lead time. For better reference: in this study, the forecast steps coincide with the solar time for Central Europe +1 hour.

Furthermore, Figure 5.1 already shows some interesting peculiarities. First, the LPI distribution seems to show higher potential for lightning activity at the beginning of the model run and in general during nighttime compared to the LINET

distribution. This might be lied to elevated, overnight convection being considered too strong and also to some data assimilation issues for what concerns the convective available potential energy (CAPE) at the beginning of the run. Another interesting feature is the evident offset in time for the start of the convective cycle between +9h and +15h. According to the observation, the lightning activity starts increasing on average at around +10h (11:00 solar time) and reaches the peak at +16h (i.e. in the late afternoon). The LPI distribution does model this behavior very well in terms of magnitude, but there is a clear offset in time, with the curve starting to increase at around +11h and reaching the maximum at +17h. The fact that convection resolving, high resolution models can show a slight delay in triggering diurnal convection is well documented (Ban et al., 2014). Finally, in order to provide some more detailed verification outcomes, 4 main timesteps have been identified: One at the beginning of the model run (+2h), one right before the start of the diurnal convection (+8h), one when daytime convection is rapidly increasing (+13h) and one at the peak of the daily cycle (+16h). For these timesteps, specific charts have been produced in the following Sections.

5.1 Symmetric Extremal Dependence Index - SEDI

In Figure 5.2 the results of the SEDI analysis for the 24 forecast steps and for the spatial upscalings from the original grid spacing of the model (around 2 km) up to 128 km are shown.

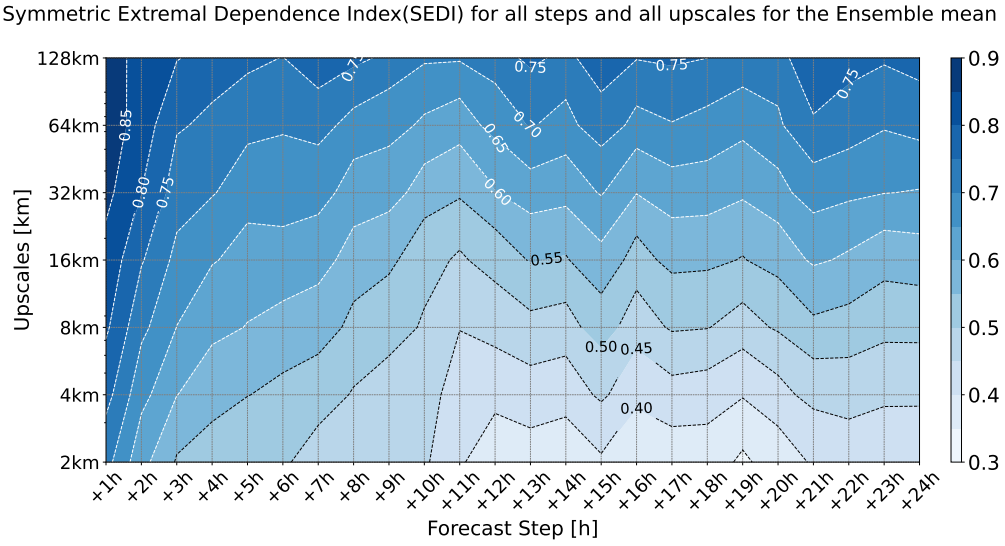

Figure 5.2: SEDI for all the 24 forecast steps and for selected spatial upscales.

The COSMO-D2 EPS produces useful information (SEDI > 0) throughout the 24 hours lead time. By looking at the 2 km resolution, the forecast skill is very good

at the beginning of the model run (SEDI values between 0.6 and 0.7) and drops to values around 0.35 during the main convective window in the afternoon hours. This is somehow expected as two main effects come into play. On the one hand there is the typical skill degradation with increasing forecast lead time: small errors in the initial conditions and the intrinsic chaotic nature of the atmosphere lead to a gradual decay of the model performance as the forecast evolves. This aspect is however not predominant in this study and would be better analyzed if the datasets used would comprise all the available model runs, in addition to the one starting at 00 UTC. Most importantly, on the other hand the diurnal convection itself is a source of uncertainty as it transfers significant amounts of energy to different scales, some of which are not resolved explicitly in the model and rely on approximated parametrization schemes. Nevertheless, the SEDI score at the original resolution of 2 km can already be considered useful in forecasting the potential for lightning activity. For larger spatial scales, the SEDI score gets significantly better and reaches values around 0.75 for the 128 km upscale. At this scale, the LPI forecast has a very good ratio between the hit rate and false alarm rate and provides reliable information on the risk of lightning activity. It is interesting to note that, for many of the upscales considered in this study, the worst skill according to the SEDI is reached at +11h from the beginning of the model run. This is again linked to the time offset of the COSMO-D2 EPS in triggering the diurnal convective cycle that has already been discussed in the opening Section of this Chapter. This aspect will be investigated further with the help of the FSS in the following Section.

5.2 Dispersion Fractions Skill Score - dFSS

Figure 5.3 shows the results of the analysis for the eFSS (b) and the dFSS (c) at varying spatial scales for the 24 forecast steps included in the study. The different spatial scales are defined by the neighborhood square window of side n, which goes from 1 gridpoint (correspondent to the original resolution of 2.2 km) up to 500 gridpoints (which is equal to a square window of approximately $1100 \cdot 1100\ km^2$). As discussed in Section 2.4.2, a forecast is considered skillful if $FSS \geq 0.5 + \frac{f_0}{2}$. However, the base rate f_0 of the datasets used in this study is very low and close to zero for all the forecast steps that are being analyzed. For this reason, a white line has been added in Figure 5.3 where the FSS is equal to the value of 0.5, as in this case this is approximately the threshold at which the LPI forecast reaches a useful skill. The study of the relationship between the eFSS and dFSS provides many in depth information. At many scales, the dFSS is showing higher values compared to those of the eFSS. At this stage it is worth remembering the fact that the dFSS is a measure of the ensemble spread and that the higher the dFSS, the lower the spread between the members. So a dFSS/eFSS ratio larger than 1 implies that the ensemble members are not diverging enough in order to fully cover the actual uncertainty that is being observed in the forecast. For better clarity, the dFSS/eFSS ratio is shown in Figure 5.4. Interestingly, a clear pattern can be recognized. At

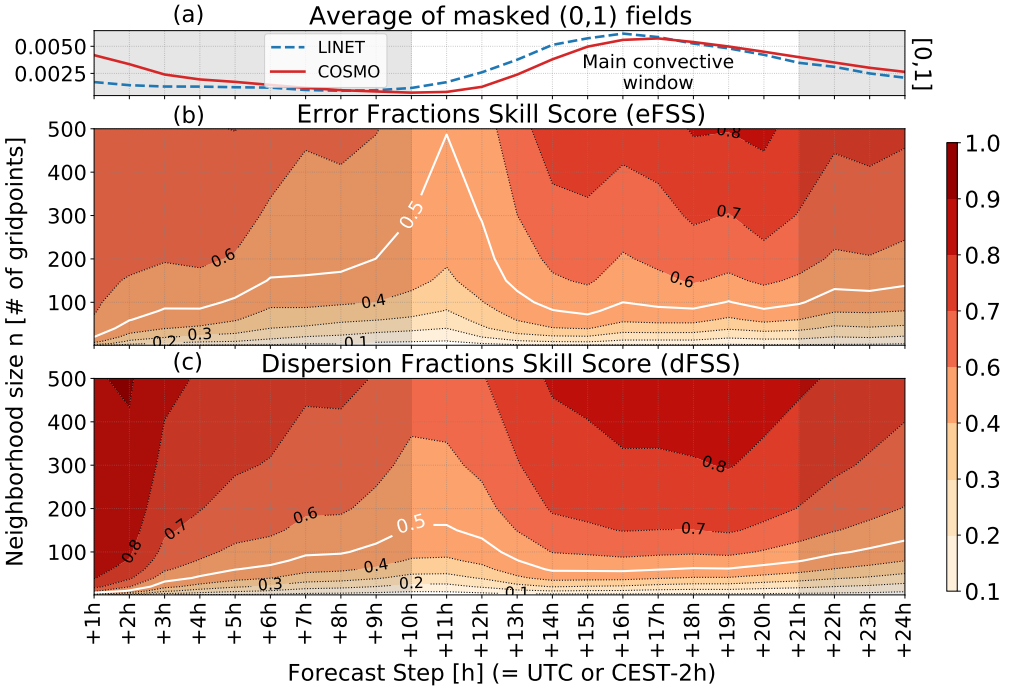

Figure 5.3: (b) Error Fractions Skill Score (eFSS) for 24 forecast steps and neighborhood sizes up to 500 grid points. (c) Dispersion Fractions Skill Score (dFSS) for 24 hourly forecast steps and neighborhood sizes up to 500 grid points. For COSMO-D2, 1 grid mesh equals $2.2 \cdot 2.2\ km^2$. For better reference, (a) is showing the average of all the masked fields used for the study (i.e., the average fraction of the domain with observed/forecasted lightning activity), with a clear maximum of the convective activity during the afternoon and evening hours. From (Salmi et al., 2022).

smaller scales the EPS is showing an overdispersive behavior, while the ratio is inverted at larger scales. This implies that — at least for single convective cells or single convective systems — the members of the ensemble are providing solutions that are too different from member to member compared to the performance of the ensemble average (which of course is relatively poor at this scale). On the contrary, when it comes to large scale convective features the members are diverging too little compared to the actual performance of the ensemble mean. In general, the discrepancies are larger at the beginning of the model run and tend to diminish during the main convective window. As every ensemble does need some spin up time in order to let the disturbances in the different members grow, this is an expected behavior from the COSMO-D2 EPS.

Another notable feature in Figure 5.3 is the minimum in performance between +10h and +12h, with the eFSS struggling to reach values of 0.5 also for extremely large spatial scales. The EPS is providing forecasts with very low skill also for neighborhood sizes up to 500 gridpoints, which implies that for these forecast

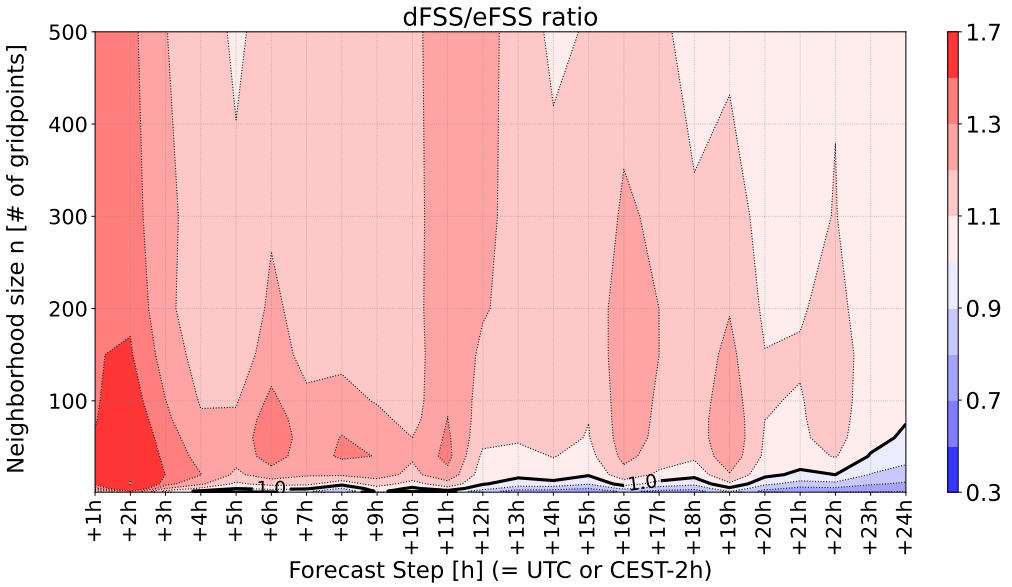

Figure 5.4: dFSS/eFSS ratio for all the forecast steps and the neighborhood sizes. Positive values identify areas where the ensemble is underdispersive, while negative values implies a higher spread than necessary in the EPS. A ratio of one (thick black line in the plot) indicates the perfect ensemble spread.

steps a strong bias is present in the model. This is the same signal that has already been discussed in the opening Section of this Chapter and that emerged also in the SEDI analysis, with an evident, delayed start of the diurnal convective cycle in the COSMO-D2 EPS compared to the observations. The model wrongly delays the triggering of the first convective cells of about one hour, but once the convective processes have been triggered, the forecast skill rapidly returns to more than acceptable eFSS values also at smaller scales. At this point the FSS analysis can provide further in depth details about the behavior of the EPS. In fact, by looking at the dFSS distribution the same feature corresponding to the same forecast step can be recognized. The dFSS minimum is not so pronounced and is less intense compared to the one visible on the eFSS chart, but the signal is evident. This means that the EPS is succeeding in modeling this sudden deterioration in the predictability and that the ensemble members are diverging a little bit more in correspondence of this singularity. In other words, the COSMO-D2 EPS is able to partly cover the average time offset that is present in its mean: some ensemble members correctly signal the possibility of an earlier triggering of the convective cycle. A dedicated, in depth analysis focusing on this specific issue might be needed to confirm such hypotheses.

Finally, for what concerns the scale at which the COSMO-D2 EPS forecast is considered skillful (identified by the white line denoting the 0.5 eFSS values in Figure 5.3), it is important to underline that this lies at around 200 km — or slightly

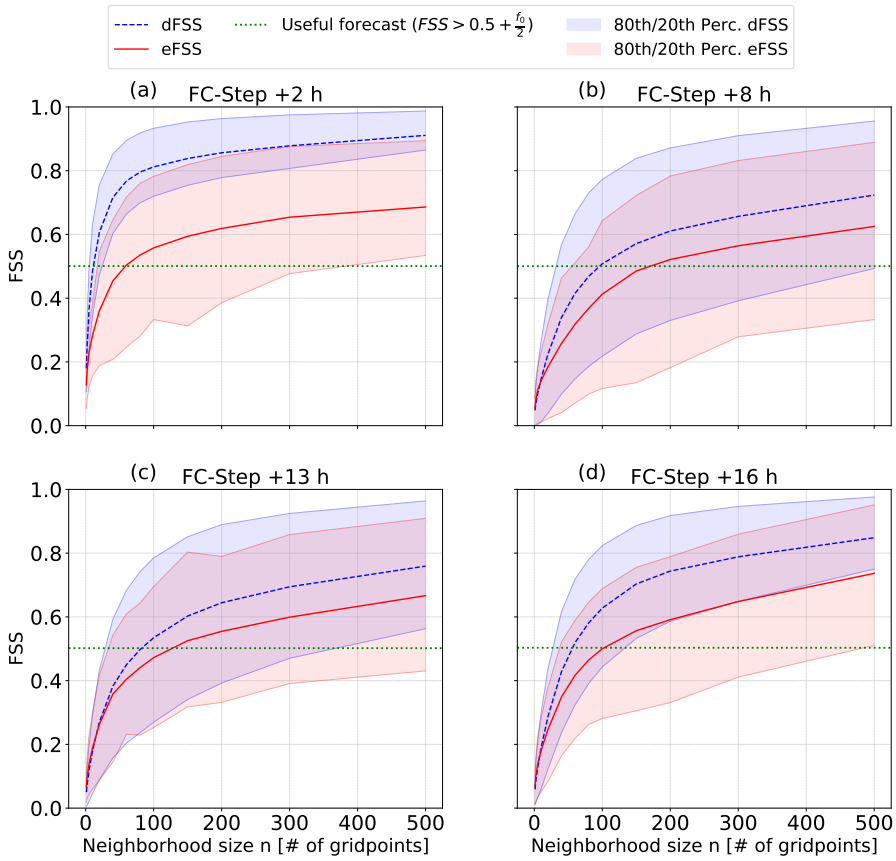

Figure 5.5: Summer average of the eFSS and dFSS for the selected time steps +2h (a), +8h (b), +13h (c) and +16h (d). The skillful threshold $0.5 + \frac{f_0}{2}$ is highlighted. 20^{th} and 80^{th} percentiles of the datasets leading to the eFSS/dFSS are shaded. For the eFSS, the whole dataset is composed by single daily values. For the dFSS a total of 190 daily values — one for each couple of ensemble members — have been processed. (Salmi et al., 2022).

less than 100 gridpoints — during the main convective window, in the afternoon hours. This is almost twice as much if compared with the results in the dFSS chart, where the 0.5 line is located close to 50 gridpoints for the same time window. This leads to the conclusion that the COSMO-D2 EPS is producing a forecast that is on average deemed skillful and useful already at a scale of around 100 km, while the ensemble mean is actually skillful at around 200 km. The ensemble is therefore spatially too optimistic and overconfident.

When looking at the data presented so far, it has to be noted that the eFSS and dFSS scores for each forecast step are mean values for the whole summer season. In fact, for the eFSS, one value per day — for the number of days shown in Figure 4.1 — is actually being calculated and for the dFSS a total of 190 values for each of these days — one for each pair of ensemble members — are available. For this reason, the

internal variability of both datasets is also of interest as this can provide further details regarding the capability of the ensemble spread to cover the full spectrum of daily error values. In order to achieve this, four main timesteps have been chosen, as previously described and for these four forecast steps (+2h, +8h, +13h, +16h) also the 20^{th} and the 80^{th} percentiles of both the eFSS and the dFSS distributions are calculated. Figure 5.5 shows the results of this analysis. Even looking at the whole distribution, it is clear that especially at higher neighborhood sizes the COSMO-D2 EPS produces too little spread if compared with the full spectrum of error values that have been observed. However, from these charts as well it is evident that the ratio between the dFSS and the eFSS gets better at smaller scales.

5.3 Ensemble Structure-Amplitude-Location - eSAL

Figure 5.6 shows the results of the analysis for all the three components of the SAL using the modified (i.e. translated into number of flashes) LPI fields. The conventional SAL calculated by comparing the observations and the ensemble mean forecasts is shown with the blue thick line. The dashed red line is referring to the eSAL, which is intended to be a measure of the ensemble spread. One of the key facts that can be immediately inferred from the chart is that the ensemble spread for the Structure and Amplitude components is almost non existent. In fact the eSAL for the panels (a) and (b) is constantly close to zero. This behavior does not come as a surprise, as the S and A components only investigate the form, volume and magnitude of the areas with lightning activity. These characteristics are strongly tied to the way the model physics treats convective processes and also directly depend on the way the LPI algorithm works and are on the other end much less dependent on larger scale processes. The random perturbations applied to the deep convective processes in the COSMO-D2 EPS does not seem to create significant dispersion between the members when it comes to the intensity and shape of the convective cells or systems leading to lightning activity. For this reason, when computing the average among all the members, any random discrepancies tend to compensate each other. Nevertheless, considering the fact that the adapted SAL for ensembles has only been applied to precipitation fields in the current literature, there is also the hypothesis that the S and A components are simply not ideal quantities to describe the probabilistic spread-error relationship in an EPS when it comes to lightning activity and should be investigated further. After this long premise, the standard SAL analysis for the ensemble mean does convey important details also for the Structure and Amplitude components. By looking at the S plot (a), the COSMO-D2 EPS LPI produces areas with potential for lightning activity that are on average clearly too large and intense if compared with the observations. Of course, this aspect is strongly dependent on the thresholds which can be arbitrarily chosen to define a SAL-object in the LPI fields. Furthermore, the LPI — as the

name suggests — expresses the potential for lightning activity. In this study it has been treated as (or translated to) a direct forecast of the occurrence of lightning flashes only for verification purposes. This leads to the hypothesis that the index itself might have been calibrated this way (i.e. more sensitive, identifying larger areas with the potential for lightning activity) on purpose. For what concerns the Amplitude component shown in panel (b), the ensemble average clearly overdoes the actual lightning activity during the nighttime hours and in the morning (forecast steps from +1h to +10h and from +19h to +24h). This result is coherent with the difference between the general distribution of the observations and the one of the forecasts already shown and discussed in Figure 5.1. As the Amplitude component is a simple measure of the magnitude of the fields over the whole domain and the thresholds applied are the same, the two plots are completely equivalent.

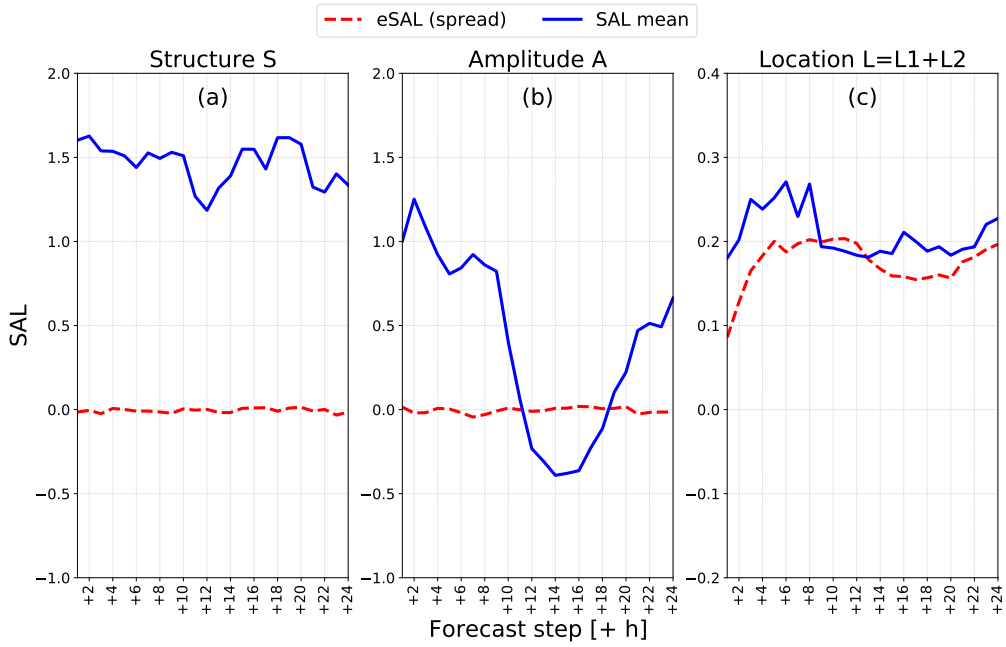

Figure 5.6: (a) Structure, (b) Amplitude and (c) Location components of SAL for 24 forecast steps for the ensemble mean. The eSAL (or spread) has been calculated between all ensemble members. (Salmi et al., 2022).

On the other end, the Location component is for sure the most informative part of the SAL study for what concerns the spread-error relationship as in this case the COSMO-D2 EPS is clearly able to generate a certain amount of dispersion between its members. However, the standard SAL computed with the ensemble average (representing the ensemble error) is slightly higher for most of the forecast steps if compared with the eSAL (which gives a measure of the ensemble spread). Therefore, the SAL analysis for the Location component points at a general, slight lack of ensemble spread if compared with the average predictability of the phenomenon. In other words, the EPS is underestimating the actual ensemble mean error. This is

coherent with the FSS analysis, at least for spatial scales above 10 gridpoints, or 20 km. At this point, it is important to remember the fact that the SAL Location component does not work with fixed spatial scales, but rather averages over several different spatial scales, depending on the scale of each identified object in the domain. Therefore, a direct comparison with the results from the FSS analysis would not be appropriate. During the time of the day with the highest convective activity, both the SAL and the eSAL distributions reach a secondary minimum. This aspect might be lied to the fact that the lightning activity is more sparse and less organized for the morning and the nighttime hours, which could introduce some more sources of error in the forecast. On the contrary, during the main convective window stronger, more organized convective systems might lead to larger objects and therefore also to larger spatial scales. This could result in a better performance in locating them from the model. Overall, the ratio between eSAL and SAL (or the spread-error ratio) is slightly better in comparison to the dFSS/eFSS ratio discussed in the previous section. One hypothesis for this is that the Location component of the SAL is only verifying the displacement of the objects in the two fields, leaving possible magnitude bias to the Structure and the Amplitude components. For what concerns the SAL analysis, in general it can be stated that the COSMO-D2 EPS is delivering a very good performance in terms of locating the areas with lightning activity and assessing the uncertainty connected to this forecast. The performance is not as good when it comes to the intensity and the shape of the single features that are being forecasted.

Finally, the same approach used for the FSS analysis has been applied to the SAL. Therefore, the overall results presented so far are averaged over the whole time period and for the eSAL also over the ensemble members. In a similar way as for Figure 5.5, a detailed analysis about the total variance of both datasets can be performed by plotting all the available SAL and eSAL values. As previously discussed, the only SAL component that shows a significant ensemble spread is the Location, while a further spread-error analysis for S and A would not add any valuable information to this study. Furthermore, as L is actually composed by the sum of $L1$ and $L2$ (see Equations 2.20 and 2.21), a L1/L2 scatterplot for all the available data of SAL and eSAL seems the optimal choice to visualize the results. In Figure 5.7 the resulting scatterplots for $L1$ and $L2$ for the same selected forecast steps as for the FSS analysis are shown. It can be inferred that $L1$ and $L2$ contributes on average with the same amount to the total Location component and also the spread-error relationship of the two component is very similar for all the investigated forecast steps. Another important point to note is that the distributions of the daily SAL (blue dots) and eSAL (red dots) values in the L1/L2 space are different, with the daily ensemble spread not covering the whole spectrum of the verified ensemble mean error. However, when looking at the daily member-to-member variability of the eSAL (orange dots), it can be seen that this covers large parts of the overall mean error dataset. In general, the COSMO-D2 EPS is therefore capable of generating sufficient dispersion between its members when it comes to assess the spatial predictability of the lightning activity.

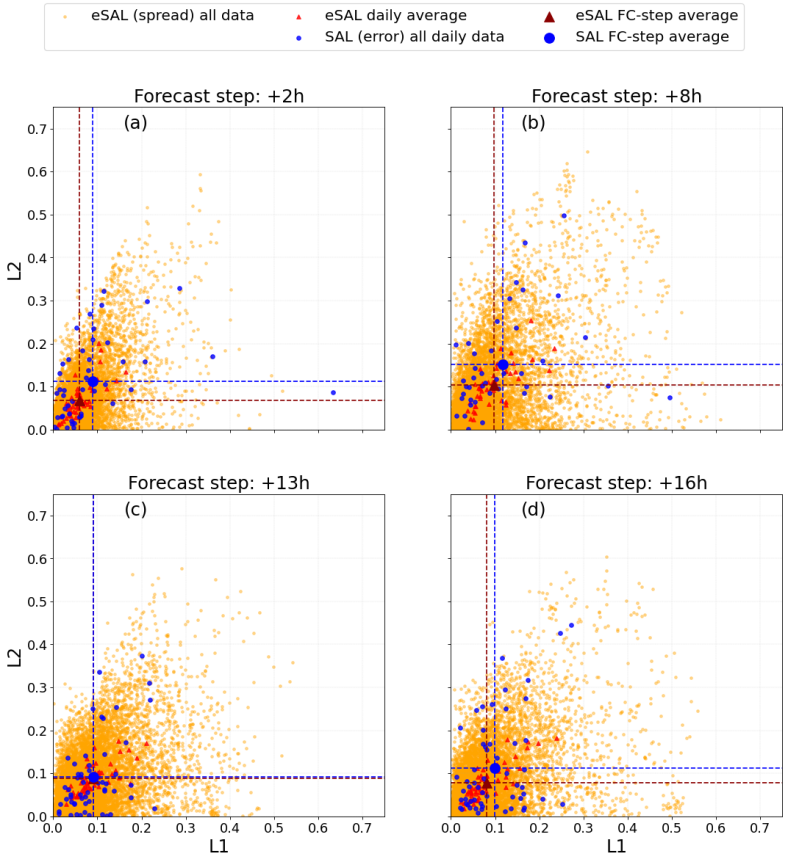

Figure 5.7: Summer average of the $L1$ and $L2$ components for the classical SAL (mean error) and for the ensemble SAL (eSAL or spread) for the selected time steps +2h (a), +8h (b), +13h (c) and +16h (d). The single daily values of the SAL that compose the overall summer averages are plotted as well. For the eSAL (or spread), both the daily averages calculated from the 190 daily values for each couple of ensemble members and every single values available in the dataset are shown.

6 Conclusion and outlook

In this study, some well established verification measures have been adapted and used to asses the performance of the high resolution COSMO-D2 EPS Lightning Potential Index (LPI) forecasts during the summer months of 2019. The Symmetric Extremal Dependence Index (SEDI), the Fractions Skill Score (FSS) and the Structure Amplitude Location (SAL) have been applied to the ensemble mean forecast, showing an overall good performance of the EPS in localizing areas with lightning activity. With the help of a probabilistic version of the FSS and the SAL, the relationship between the ensemble spread and the ensemble mean error has also been investigated, showing an overall lack of dispersion in the ensemble forecasts compared to the mean skill. In general, this means that the EPS suffers a slight underestimation of the low predictability of the lightning activity. Interestingly, when analyzed using the FSS approach, this lack of spread is particularly present at larger spatial scales, while for features located at the lower end of the mesoscale the generated ensemble spread seems to be enough or even slightly too large. It has been also shown that the EPS is at least partly correctly catching some of the uncertainty in the forecast deriving from the often delayed onset of the daily convective cycle in the model. Furthermore, using the FSS, an average skillful scale for the ensemble mean of around 200 km has been determined during the main convective window, in the afternoon hours. On the other hand, the same analysis conducted on the ensemble spread shows that the EPS would see the forecast as skillful already at a spatial scale of 100 km. Finally, most of the study focused on the spatial performance of the LPI forecasts and not on the magnitude of the fields. This is mainly due to the fact that the LPI is in fact a completely different measure with different units compared to the observed lightning flashes. However, using a statistical model, a SAL analysis also on the intensity of the fields has been conducted with controversial results. If the COSMO-D2 EPS performance in locating areas with lightning activity is good, some more concerns are present for possible biases in the magnitude of the signal.

As this was a preliminary verification of a very high resolution, convection-resolving EPS in forecasting lightning activity, there are many branches of this analysis which could be expanded further. First of all, the study has been conducted only on the 00 UTC model runs for 24 forecast steps. This leads to strong discrepancies in the dataset between each forecast step, as convection (and therefore also lightning flashes) mostly occur during the afternoon hours. In order to decouple the forecast lead time from the daily convective cycle, all the available model runs (one every 3 hours) should be included in future analysis. By doing this, a cleaner view on the variation of the spread-error relationship with the advancing forecast lead time would be obtained. Furthermore, this would also enable an in-depth investigation

of possible benefits coming from data assimilation during the morning hours, before the start of the convective cycle. Given the interesting results obtained with the dFSS/eFSS analysis, which shows discrepancies in the spread-error relationship depending on the spatial scale, a further probabilistic investigation focusing on the mesoscale might also be worthy.

In this analysis the fields have been aggregated on an hourly basis, although a maximum frequency of 15 minutes could be possible. For this reason, one possible further step would be to increase the frequency of the analysis as this could for example provide further insights concerning the time delay in the model at the start of the daily convective cycle. Nevertheless, it has already been documented as such a significant increase in the frequency of the verification could lead to an overall worse performance also for lightning activity (Mittermaier et al., 2022a,b). Another possible evolution of this study would be a differentiated verification of sub-domains based on specific topographical characteristics (for example the lowlands in northern Germany, the hilly central Germany and the Alpine region), as topography plays a central role in triggering convection. Also, a comparison between the LPI and other indexes that are relevant for lightning activity (either using parcel theory or considering the cloud microphysics) would be beneficial in order to put this performance in relation to similar and also not so similar forecasting approaches. Finally, also the observation dataset could be different, as the LPI is not a direct forecast of the number of lightning flashes. The usage of innovative pattern recognition algorithms on satellite or radar data to detect deep convection might for example also provide an interesting observational database.

References

Baldauf, M., Gebhardt, C., Theis, S., Ritter, B., and Schraff, C. (2018). Beschreibung des operationellen kürzestfristvorhersagemodells COSMO-D2 und COSMO-D2-EPS und seiner ausgabe in die datenbanken des deutscher wetterdienstes (DWD). Technical Report in German, available at https://www.dwd.de/DE/leistungen/nwv$_{c}$osmo$_{d}$2$_{a}$enderungen/nwv$_{c}$osmo$_{d}$2$_{a}$enderungen.html.

Ban, N., Schmidli, J., and Schär, C. (2014). Evaluation of the convection-resolving regional climate modeling approach in decade-long simulations. *Journal of Geophysical Research: Atmospheres*, 119(13):7889–7907. https://doi.org/10.1002/2014JD021478.

Betz, H.-D., Schmidt, K., Laroche, P., Blanchet, B., Oettinger, W., Defer, E., Dziewit, Z., and Konarski, J. (2009). LINET-an international lightning detection network in europe. *Atmos. Res.*, 91:564–573. https://doi.org/10.1016/j.atmosres.2008.06.012.

Betz, H.-D., Schmidt, K., Oettinger, P., and Wirz, M. (2004). Lightning detection with 3-D discrimination of intracloud and cloud-to-ground discharges. *Geophys. Res. Lett.*, 31(L11108). https://doi.org/10.1029/2004GL019821.

Blahak, U. (2015). LPI (Lightning Potential Index) derived from COSMO-DE fields, COSMO general meeting Wroclaw 2015. Presentation available at https://www.cosmo-model.org/content/consortium/generalMeetings/general2015/parallel/lpi$_{b}$lahak.pdf.

Dey, S., Leoncini, G., Roberts, N., Plant, R., and Migliorini, S. (2014). A spatial view of ensemble spread in convection permitting ensembles. *Mon. Wea. Rev.*, 142:4091–4107. https://doi.org/10.1175/MWR-D-14-00172.1.

Faggian, N., Roux, B., Steinle, P., and Ebert, B. (2015). Fast calculation of the fractions skill score. *Mausam*, 66(3):457–466. https://doi.org/10.54302/mausam.v66i3.

Ferro, C. and Stephenson, D. (2011). Extremal dependence indices: improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting*, 26:699–713. https://doi.org/10.1175/WAF-D-10-05030.1.

Ferro, C. A. T. (2007). A probability model for verifying deterministic forecasts of extreme events. *Weather and Forecasting*, 22(5):1089 – 1100. https://doi.org/10.1175/WAF1036.1.

Gebhardt, C., Theis, S., Paulat, M., and Ben Bouallègue, Z. (2011). Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmos. Res.*, 100:168–177. https://doi.org/10.1016/j.atmosres.2010.12.008.

Hunt, B. R., Kostelich, E. J., and Szunyogh, I. (2007). Efficient data assimilation for spatiotemporal chaos: A local ensemble transform kalman filter. *Physica D: Nonlinear Phenomena*, 230(1):112–126. https://doi.org/10.1016/j.physd.2006.11.008.

Karagiannidis, A., Lagouvardos, K., Lykoudis, S.and Kotroni, V., Giannaros, T., and Betz, H. (2019). Modeling lightning density using cloud top parameters. *Atmos. Res.*, 222:163–171. https://doi.org/10.1016/j.atmosres.2019.02.013.

Lawson, J. R. and Gallus Jr, W. A. (2016). Adapting the sal method to evaluate reflectivity forecasts of summer precipitation in the central united states. *Atmospheric Science Letters*, 17(10):524–530. https://doi.org/10.1002/asl.687.

Lynn, B. and Yair, Y. (2010). Prediction of lightning flash density with the WRF model. *Adv. Geosci.*, 23:11–16. https://doi.org/10.5194/adgeo-23-11-2010.

Lynn, B., Yair, Y., Price, C., Kelman, G., and Clark, A. (2012). Predicting cloud-to-ground and intracloud lightning in weather forecast models. *Weather and Forecasting*, 27:1470–1488. https://doi.org/10.1175/WAF-D-11-00144.1.

Marsigli, C., Alferov, D., Astakhova, E., Duniec, G., Gayfulin, D., Gebhardt, C., Interewicz, W., Loglisci, N., Marcucci, F., Mazur, A., Montani, A., Tsyrulnikov, M., and Walser, A. (2019). Studying perturbations for the representation of modeling uncertainties in ensemble development (SPRED final report). *COSMO Technical Report*, 39. http://www2.cosmo-model.org/content/model/documentation/techReports/cosmo/docs/techReport39.pdf.

Mass, C. F., Ovens, D., Westrick, K., and Colle, B. A. (2002). Does increasing horizontal resolution produce more skillful forecasts?: The results of two years of real-time numerical weather prediction over the pacific northwest. *Bulletin of the American Meteorological Society*, 83(3):407 – 430. https://doi.org/10.1175/1520-0477(2002)083%3C0407:DIHRPM%3E2.3.CO;2.

McCaul, E., Goodman, S., LaCasse, K., and Cecil, D. (2009). Forecasting lightning threat using cloud-resolving model simulations. *Weather and Forecasting*, 24:709–729. https://doi.org/10.1175/2008WAF2222152.1.

Mittermaier, M. P. (2021). A "meta" analysis of the fractions skill score: The limiting case and implications for aggregation. *Mon. Wea. Rev.*, 149(10):3491–3504. https://doi.org/10.1175/MWR-D-18-0106.1.

Mittermaier, M. P., Landman, S., Csima, G., and Goodman, S. (2022a). Convective-scale numerical weather prediction and warnings over Lake Victoria: Part II—can model output support severe weather warning decision-making? *Meteorol. Apps.*, 29(3):e2055. https://doi.org/10.1002/met.2055.

Mittermaier, M. P., Wilkinson, J., Csima, G., Goodman, S., and Virts, K. (2022b). Convective-scale numerical weather prediction and warnings over Lake Victoria: Part I—evaluating a lightning diagnostic. *Meteorol. Apps.*, 29(3):e2038. https://doi.org/10.1002/met.2038.

Montani, A., Capaldo, M., Cesari, D., Marsigli, C., Modigliani, U., Nerozzi, F., Paccagnella, T., and Tibaldi, S. (2003). Operational limited–area ensemble forecasts based on the 'lokal modell'. *ECMWF Newsletter*, 98:2–7. https://www.ecmwf.int/sites/default/files/elibrary/2003/14626-newsletter-no98-summer-2003.pdf.

Radanovics, S., Vidal, J.-P., and Sauquet, E. (2018). Spatial verification of ensemble precipitation: an ensemble version of SAL. *Weather and Forecasting*, 33:1001–1020. https://doi.org/10.1175/WAF-D-17-0162.1.

Reinert, D., Prill, F., Frank, H., Denhard, M., Baldauf, M., Schraff, C., Gebhardt, C., Marsigli, C., and Zängl, G. (2022). DWD database reference for the global and regional ICON and ICON-EPS forecasting system. Technical Report available at https://www.dwd.de/DWD/forschung/nwv/fepub/icon$_{\mathrm{d}}$atabase$_{\mathrm{m}}$ain.pdf.

Roberts, N. and Lean, H. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, 136:78–97. https://doi.org/10.1175/2007mwr2123.1.

Rossa, A., Nurmi, P., and Ebert, E. (2008). *Overview of methods for the verification of quantitative precipitation forecasts*, pages 419–452. Springer Berlin Heidelberg, Berlin, Heidelberg.

Salmi, M., Marsigli, C., and Dorninger, M. (2022). Predictability analysis and skillful scale verification of the lightning potential index (lpi) in the cosmo-d2 high resolution ensemble system. *Advances in Science and Research*, 19:29–38. https://doi.org/10.5194/asr-19-29-2022.

Saunders, C. P. R. (2008). Charge separation mechanisms in clouds. *Space Sci. Rev.*, 137:335–354. https://doi.org/10.1007/978-0-387-87664-1$_2$2.

Schneider, L., Barthlott, C., Hoose, C., and Barrett, A. I. (2019). Relative impact of aerosol, soil moisture, and orography perturbations on deep convection. *Atmospheric Chemistry and Physics*, 19(19):12343–12359. https://doi.org/10.5194/acp-19-12343-2019.

Schraff, C., Reich, H., Rhodin, A., Schomburg, A., Stephan, K., Periáñez, A., and Potthast, R. (2016). Kilometre-scale ensemble data assimilation for the COSMO model (KENDA). *Q.J.R. Meteorol. Soc.*, 142:1453–1472. https://doi.org/10.1002/qj.2748.

Wernli, H., Hofmann, C., and Zimmer, M. (2009). Spatial forecast verification methods intercomparison project: Application of the sal technique. *Weather and Forecasting*, 24(6):1472 – 1484. https://doi.org/10.1175/2009WAF2222271.1.

Wernli, H., Paulat, M., Hagen, M., and Frei, C. (2008). SAL-a novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev.*, 136:4470–4487. https://doi.org/10.1175/2008MWR2415.1.

Wilkins, K. L., Watson, I. M., Kristiansen, N. I., Webster, H. N., Thomson, D. J., Dacre, H. F., and Prata, A. J. (2016). Using data insertion with the name model to simulate the 8 may 2010 eyjafjallajökull volcanic ash cloud. *Journal of Geophysical Research: Atmospheres*, 121(1):306–323. https://doi.org/10.1002/2015JD023895.

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

Zhaoye, P., Yang, K., and Wang, C. (2022). Impacts of cumulus parameterizations on extreme precipitation simulation in semi-arid region: A case study in northwest china. *Atmosphere*, 13(9). https://doi.org/10.3390/atmos13091464.

Acknowledgements

This study has been developed in cooperation with the German Weather Service (DWD) and parts of it have already been published with the decisive support of the University of Vienna (Salmi et al., 2022). The analysis would not have been possible without the support and the data provided by the DWD, especially in the person of Dott. Chiara Marsigli. Dott. Marsigli defined the initial idea, which has then been adapted and applied to the specific EPS and the specific parameter in this work. I also want to thank Professor Manfred Dorninger (University of Vienna, department of Meteorology) for the decisive inputs, motivation and feedbacks he provided throughout this journey.

List of Tables

List of Figures