The evolution of forecast uncertainty in a large ensemble

Matjaž Puh

Munich 2024

The evolution of forecast uncertainty in a large ensemble

Matjaž Puh

Dissertation
an der Fakultät für Physik
der Ludwig-Maximilians-Universität
München

vorgelegt von
Matjaž Puh
aus Koper, Slowenien

München, den 27. Mai 2024

Erstgutachter: Prof. Dr. George C. Craig
Zweitgutachter: Prof. Dr. Hella Garny
Tag der mündlichen Prüfung: 08.07.2024

Parts of this thesis are included in:

Puh, M., Tempest, K., Keil, C., Craig, G. C. (2024) Flow dependence of forecast uncertainty in a large convection-permitting ensemble. *Quarterly Journal of the Royal Meteorological Society*, (submitted)

Puh, M., Keil, C., Gebhardt, C., Marsigli, C., Hirt, M., Jakub, F., Craig, G. C. (2023) Physically based stochastic perturbations improve a high-resolution forecast of convection. *Quarterly Journal of the Royal Meteorological Society*, 149(757), 3582–3592. Available from: https://doi.org/10.1002/qj.4574

Craig, G.C., **Puh, M.**, Keil, C., Tempest, K., Necker, T., Ruiz, J.et al. (2022) Distributions and convergence of forecast variables in a 1,000-member convection-permitting ensemble. *Quarterly Journal of the Royal Meteorological Society*, 148(746), 2325–2343. Available from: https://doi.org/10.1002/qj.4305

Abstract

The accurate prediction of atmospheric phenomena is hindered by the inherent chaos of the Earth's atmosphere, coupled with a multitude of uncertainties stemming from observational limitations and model imperfections. These challenges necessitate sophisticated forecasting methodologies capable of quantifying and addressing uncertainties to provide reliable meteorological predictions. Ensemble Prediction Systems (EPS) have emerged as indispensable tools in this regard, offering a probabilistic framework that accommodates the inherent variability of atmospheric processes. However, there are several limitations to ensembles too: from the inadequate representation of physical processes within the model, to sampling errors because of the limited ensemble size. Moreover, the significance of different uncertainty sources varies across weather regimes, which requires a flow-dependent assessment of the evolution of forecast uncertainty, often limited by the high computational cost of running ensemble experiments several times.

This thesis addresses these challenges by performing three different convection-permitting ensemble experiments using the ICON Limited Area Model (LAM). Firstly, the Physically Based Stochastic Perturbation Scheme (PSP) is included as a representation of model error originating from the subgrid scale in the boundary layer, but affecting the smallest resolved scales. The first experiment spans a whole summer season, which allows for a systematic analysis of the impact of the scheme in different synoptic forcing conditions. This shows that PSP efficiently increases ensemble spread of precipitation in weak synoptic forcing, while producing realistic convective structures and without spoiling the forecast skill. During strong forcing, the effect of the scheme is negligible, as expected by design.

The three-month period analysed in the first part offered a unique opportunity to select two representative cases for the weak and strong forcing regimes to be analysed in more detail in the second experiment: the flow-dependence of forecast distributions is studied using a 120 member Icosahedral Nonhydrostatic (ICON) LAM ensemble. The bootstrapping technique is used to investigate the convergence of sampling error for a range of surface and mid-troposphere variables. Convergence is generally observed for the mean and the standard deviation, but not for the 95^{th} percentile, especially in strong forcing. Additionally, maps of uncertainty are introduced, which allow for a more detailed analysis of the spatial pattern of uncertainty and facilitate the interpretation of the sources and evolution of forecast uncertainty in different synoptic forcing conditions. Overall, the main result of the second part of the thesis is the strong connection between uncertainty of forecast variables and convection, with synoptic forcing being crucial in determining the

spatial distribution and uncertainty evolution within 24 hours.

In the last part of the thesis, the longer-lasting impact given by the memory effects of soil moisture and atmospheric stability on forecast uncertainty at the convective scale beyond the first 24 hours of the simulation is studied. Additionally, the impact of these mechanisms is compared with that of altering the initialization time, where ongoing convection is assimilated into the forecast and modifies the evolution of the forecast beyond the first day. The flow-dependent analysis shows that all the studied mechanisms have a larger impact in weak forcing conditions. Although the uncertainty on the convective scale quickly grows and the predictability left comes mostly from the larger-scale flow, there are mechanisms on the smaller scale that can still influence the forecast beyond the usual influence time.

Addressing the challenges of limited ensemble size and uncertainty representation, as well as its flow-dependence, is essential for advancing the capabilities of EPS in providing reliable probabilistic forecasts crucial for mitigating weather-related risks in a changing climate.

Zusammenfassung

Eine genaue Vorhersage atmosphärischer Phänomene wird durch die chaotische Natur der Atmosphäre in Verbindung mit einer Vielzahl weiterer Unsicherheiten erschwert, die sich aus der begrenzten Anzahl an Beobachtungen als auch Näherungen in numerischen Modellen ergeben. Diese Herausforderungen erfordern hochentwickelte Vorhersagemethoden, die in der Lage sind, Unsicherheiten zu reduzieren und zu quantifizieren, um zuverlässige meteorologische Vorhersagen zu liefern. Ensemble-Vorhersagesysteme (EPS) haben sich in dieser Hinsicht als unverzichtbare Instrumente erwiesen, da sie einen probabilistischen Rahmen bieten, der der inhärenten Variabilität atmosphärischer Prozesse Rechnung trägt. Doch auch Ensembles unterliegen verschiedenen Einschränkungen: von der unzureichenden Darstellung physikalischer Prozesse innerhalb des Modells bis hin zu Stichprobenfehlern aufgrund der begrenzten Ensemblegröße. Die Ensemblegröße wird insbesondere durch die verfügbare Rechnerkapazität eingeschränkt, da die Anzahl der Ensemblemember mit den Rechenkosten skaliert. Darüber hinaus variiert die Bedeutung der verschiedenen Unsicherheitsquellen je nach Wetterlage, was eine strömungsabhängige Bewertung der Entwicklung der Vorhersageunsicherheit erfordert.

In dieser Arbeit werden diese Herausforderungen durch die Durchführung von drei verschiedenen konvektionserlaubenden Ensemble-Experimenten unter Verwendung des ICON Modells angegangen. Der Modellfehler wird mit dem physikalisch basierten stochastischen Störungsschema PSP beschrieben, das die Turbulenz innerhalb der Grenzschicht stört. Das PSP Schema repräsentiert Unsicherheiten, die sich aus nicht aufgelösten Grenzschichtprozessen aufgrund endlicher Gittergröße ergeben. Das erste Experiment erstreckt sich über eine ganze Sommersaison, was eine systematische Analyse der Auswirkungen des PSP Schemas unter verschiedenen synoptischen Bedingungen ermöglicht. Dabei zeigt sich, dass PSP bei schwachem synoptischem Antrieb die Ensemblevariabilität des Niederschlags effizient erhöht und dabei realistische konvektive Strukturen erzeugt, ohne die Vorhersagefähigkeit zu beeinträchtigen. Erwartungsgemäß ist die Auswirkung des PSP Schemas bei starkem Antrieb vernachlässigbar.

Der im ersten Teil analysierte Dreimonatszeitraum bietet die einmalige Gelegenheit, zwei repräsentative Fälle für das schwache und das starke Antriebsregime auszuwählen, die im zweiten Abschnitt eingehender analysiert werden. Die Strömungsabhängigkeit der Vorhersageverteilungen wird anhand eines 120 Member umfassenden ICON Ensembles untersucht. Die Bootstrapping-Technik wird verwendet, um die Konvergenz des Stichprobenfehlers sowohl für verschiedene meteorologische Variablen Nahe der Erdoberfläche als auch in der freien Troposphäre zu untersuchen. Konvergenz wird im Allgemeinen für den Mittelwert und die Standardabweichung beobachtet, jedoch nicht für das 95. Perzentil, insbesondere bei starkem Antrieb. Zusätzlich werden Karten der Unsicherheit eingeführt, die eine detailliertere Analyse des räumlichen Musters der Unsicherheit ermöglichen und die Interpretation der Quellen und der Entwicklung der Vorhersageunsicherheit bei verschiedenen synoptischen Antriebsbedingungen erleichtern. Insgesamt ist das Hauptergebnis des zweiten Teils der Arbeit der starke Zusammenhang zwischen der Unsicherheit der Vorhersagevariablen und der Konvektion, wobei der synoptische Antrieb für die Bestimmung der räumlichen Verteilung und der Entwicklung der Unsicherheit innerhalb von 24 Stunden entscheidend ist.

Im letzten Teil der Arbeit werden die längerfristigen Auswirkungen der Speichereffekte von Bodenfeuchte und atmosphärischer Stabilität auf die Vorhersageunsicherheit auf der konvektiven Skala über die ersten 24 Stunden der Simulation hinaus untersucht. Außerdem werden die Auswirkungen dieser Mechanismen mit denen einer Änderung der Initialisierungszeit verglichen, bei der die laufende Konvektion in die Vorhersage aufgenommen wird und die Entwicklung der Vorhersage über den ersten Tag hinaus verändert. Die strömungsabhängige Analyse zeigt, dass alle untersuchten Mechanismen bei schwachem Konvektionsantrieb einen größeren Einfluss haben. Obwohl die Unsicherheit auf der konvektiven Skala schnell zunimmt und die verbleibende Vorhersagbarkeit hauptsächlich von der großskaligen Strömung stammt, gibt es Mechanismen auf der kleineren Skala, die die Vorhersage auch über die übliche Einflusszeit hinaus beeinflussen können.

Die Erforschung der Herausforderungen, die sich aus der begrenzten Ensemblegröße und der Darstellung der Unsicherheit sowie der Abhängigkeit von der Strömung ergeben, ist von entscheidender Bedeutung für die Verbesserung der Fähigkeiten von EPS bei der Bereitstellung zuverlässiger probabilistischer Vorhersagen, die für die Minderung wetterbedingter Risiken in einem sich ändernden Klima entscheidend sind.

Contents

Abstract **v**

Zusammenfassung **vii**

1 Introduction **1**
1.1 Evolution of numerical weather prediction 1
1.2 Probabilistic forecasting 2
1.3 Challenges in ensemble prediction 5
1.3.1 Limited ensemble size 5
1.3.2 Model error representation 8
1.3.3 Flow dependence of convection 9
1.4 Uncertainty and convergence in a big ensemble: a preliminary study 9
1.5 Research questions and outline 14

2 Model and methods **17**
2.1 The ICON model 17
2.2 PSP scheme 19
2.3 Bootstrapping 21
2.4 Neighborhood method 22

3 Ensemble simulations **25**
3.1 Trial run with PSP scheme 25
3.2 Case studies of weak and strong forcing - large ensemble 27
3.3 Case studies of weak and strong forcing - extended forecast 28

4 Improving probabilistic forecasts of convection with a SP scheme **31**
4.1 Background 31
4.2 Results and discussion 33
4.2.1 Synoptic forcing regime classification 33
4.2.2 Diurnal cycle of precipitation 34
4.2.3 Spatial uncertainty of precipitation 35
4.2.4 Probabilistic verification of near-surface variables 37
4.3 Conclusions 39

5 Flow dependence of forecast uncertainty in a large ensemble 41
5.1 Background 41
5.2 Results and discussion 43
5.2.1 Convergence of uncertainty with ensemble size 43
5.2.2 Maps of uncertainty 46
5.2.3 Flow-dependent evolution of forecast uncertainty 50
5.3 Conclusions 54

6 Forecast uncertainty beyond 24 hours 57
6.1 Background 57
6.2 Results 59
6.2.1 Time evolution of precipitation 59
6.2.2 Spatial distribution of precipitation 62
6.3 Summary and conclusions 69

7 Conclusions 71

A Additional figures for chapter 5 77

B Additional figures for chapter 6 85

List of Figures 95

List of Tables 97

List of Abbreviations 99

Bibliography 101

Acknowledgements 111

Chapter 1

Introduction

1.1 Evolution of numerical weather prediction

Over a century ago, Abbe (1901) and Bjerknes (1904) proposed a fascinating idea: predicting the weather by using the laws of physics. They acknowledged that weather prediction could be treated as an initial value problem of mathematical physics. In this context, the future state of the weather is determined by integrating the governing partial differential equations, starting from the currently observed weather conditions. The transformation from theoretical propositions to the present-day practice of weather prediction delineates a path marked by technological advancements, scientific breakthroughs, and an enduring pursuit of accuracy.

Today, this approach involves tackling a complex system of nonlinear differential equations on a daily basis, spanning approximately half a billion points per time step, projecting weeks to months ahead. It encompasses the intricate interplay of dynamic, thermodynamic, radiative, and chemical processes occurring across scales ranging from hundreds of meters to thousands of kilometers, operating within time frames stretching from seconds to weeks.

The utility of accurate weather forecasts extends far beyond theoretical realms. Beyond saving lives and aiding emergency management, these forecasts mitigate the socioeconomic impact of high-impact weather events, bolstering sectors such as energy, agriculture, transport, and recreation. The tangible benefits outweigh the considerable investments required in scientific research, cutting-edge supercomputing, and observational infrastructure vital for generating such forecasts (Lazo et al., 2009).

At the heart of atmospheric forecasting lie the Navier–Stokes and mass continuity equations, together with the first law of thermodynamics and the ideal gas law, describing the evolution of wind, pressure, density, and temperature across spatial and temporal dimensions (Kalnay, 2003). However, the integration of physical processes operating at unresolved scales, down to the molecular level, involves intricate source terms for mass, momentum, and heat, introduced through friction, condensation, evaporation, and radiative heat exchange (Bauer et al., 2015). Nevertheless, an inherent and crucial attribute of the atmosphere is that it follows particular rules, which makes it a dynamical system.

On large scales, the atmospheric evolution is deterministic, governed by specific physical laws. This means that if we know the atmospheric state at a given moment, we can predict its future conditions through the application of these natural laws. This deterministic characteristic forms the fundamental basis for atmospheric prediction (Toth and Buizza, 2019a).

Two decades after Abbe's and Bjerknes' idea, Lewis Fry Richardson (1922) recognized the atmosphere's deterministic nature and undertook a pioneering effort in numerical weather prediction. Richardson, a British mathematician and physicist, aimed to use a simplified set of equations to forecast future weather conditions through mathematical calculations and early computing methods. Although his forecast was unsuccessful due to the imperfect simplifications, it represents a remarkable achievement considering the calculations were done by hand.

The inaugural utilization of the first electronic computer for weather forecasting occurred in Princeton in 1950 (Charney et al., 1950). This marked a significant milestone achieved by incorporating approximations that effectively characterized the largest scales of motion in the atmosphere. While the simulations conducted in Princeton focused on retrospectively simulating past weather conditions (hindcasts), it was in Stockholm (Bolin, 1955) that the first real-time weather forecasts were made.

Since then, advancements in weather forecasting have been driven by various factors. Improvements in representing unresolved atmospheric processes within global models, the advent of ensemble methods for estimating forecast uncertainty, and the introduction of objective analysis techniques to establish the initial state of the atmosphere have collectively elevated our predictive capabilities. Additionally, the evolution of computing has been crucial in enhancing numerical weather prediction, with computational power increasing substantially, roughly by tenfold every five years since the 1980s (Vitart and Robertson, 2019). This growth in computational capability has played an essential role in refining predictive skill, allowing us to retain accuracy in forecasting one additional day into the future for every decade of dedicated research and development (Bauer et al., 2015). Moreover, fluctuations in predictive accuracy are often a consequence of the varying levels of predictability exhibited by the atmosphere. Some weather regimes exhibit higher predictability, enabling more accurate forecasts over longer periods, while others prove more challenging. Our ongoing understanding of these atmospheric patterns continues to advance, facilitating a more nuanced and refined assessment of our predictive capabilities.

1.2 Probabilistic forecasting

When numerical weather prediction became an everyday task, forecasters realized that sometimes the forecast was particularly "bad" compared to the average. What they did not realize yet was that their assumption of the atmosphere being a deterministic system was actually wrong.

In 1961, Edward Lorenz, a meteorologist, introduced the idea that small changes in initial conditions could lead to vastly different outcomes in complex systems, a phenomenon popularly known as the "butterfly effect" (Lorenz, 1969). Lorenz's chaos theory revolutionized our understanding of complex systems, illustrated through the chaotic behavior of the atmosphere. His theory highlighted that even small fluctuations in data could lead to drastically different outcomes, challenging the traditional deterministic view of predictability in natural phenomena. This insight transformed meteorology, emphasizing the inherent complexity and nonlinearity of atmospheric dynamics, prompting a shift towards probabilistic forecasting methods.

In the early 1990s, several groups of meteorologists embraced the concept of employing ensemble forecasts (e.g. ECMWF, NCEP). The ensemble approach boils down to a straightforward concept: generate N modified forecasts, called ensemble members, each mimicking potential uncertainties in the main (control) forecast. These forecasts are then used to determine the possible outcomes, the most likely values, and the probability of a future variable exceeding or falling below a certain threshold.

Fig. 1.1 shows an example of an ensemble forecast for temperature at a specific location. Initially, all the ensemble members are close together and to the control forecast. The resulting Probability Distribution Function (PDF) is Gaussian, with the most likely value being close to reality. As time proceeds, the nonlinear, chaotic nature of the atmosphere causes the ensemble forecasts to diverge and form 3 clusters, resulting in a multi-modal PDF. The deterministic, control forecast fell into one of the clusters that diverged from reality, which demonstrates the advantage of a probabilistic approach, where the most likely value is close to reality, despite the large uncertainty.

Ensemble techniques in weather prediction have undergone a significant evolution. Initially, methods like time lagged forecasts and random initial condition perturbations, though attempted, yielded limited success. However, the 1990s brought a turning point with the emergence of more promising and sophisticated methodologies. The first iteration of the European Center for Medium-range Weather Forecasting (ECMWF) global ensemble employed Singular Vectors (SV) to simulate initial uncertainties. SV represent the perturbations that exhibit the most rapid growth within a defined time frame (Buizza and Palmer, 1995). In 2008, perturbations from multiple data assimilation cycles, known as Ensembles of Data Assimilations (EDAs), were also incorporated in addition to SV (Buizza et al., 2008), which are still used today.

Initial condition uncertainty, however, is not the only source of uncertainty to be accounted for in an ensemble forecast. An important contributor to uncertainty is the model itself, arising from our restricted understanding of atmospheric phenomena and the inherent limitations of finite grid size. Tackling the former involves perturbing particular elements of the model formulation, such as parameters of simulated microphysical processes. Meanwhile, addressing the latter necessitates meticulous parameterization of subgrid processes that are unresolved by the model. Model uncertainties were first included in the Ensemble Prediction System (EPS) of the Meteorological Service of Canada (Houtekamer et al., 1996). A few years later, a stochastic model perturbation scheme designed to simulate model uncertainties was introduced in the ECMWF ensemble (Buizza et al., 1999).

The implementations of ensemble techniques marked an important shift in operational NWP methodologies. This shift transitioned NWP from a deterministic approach reliant

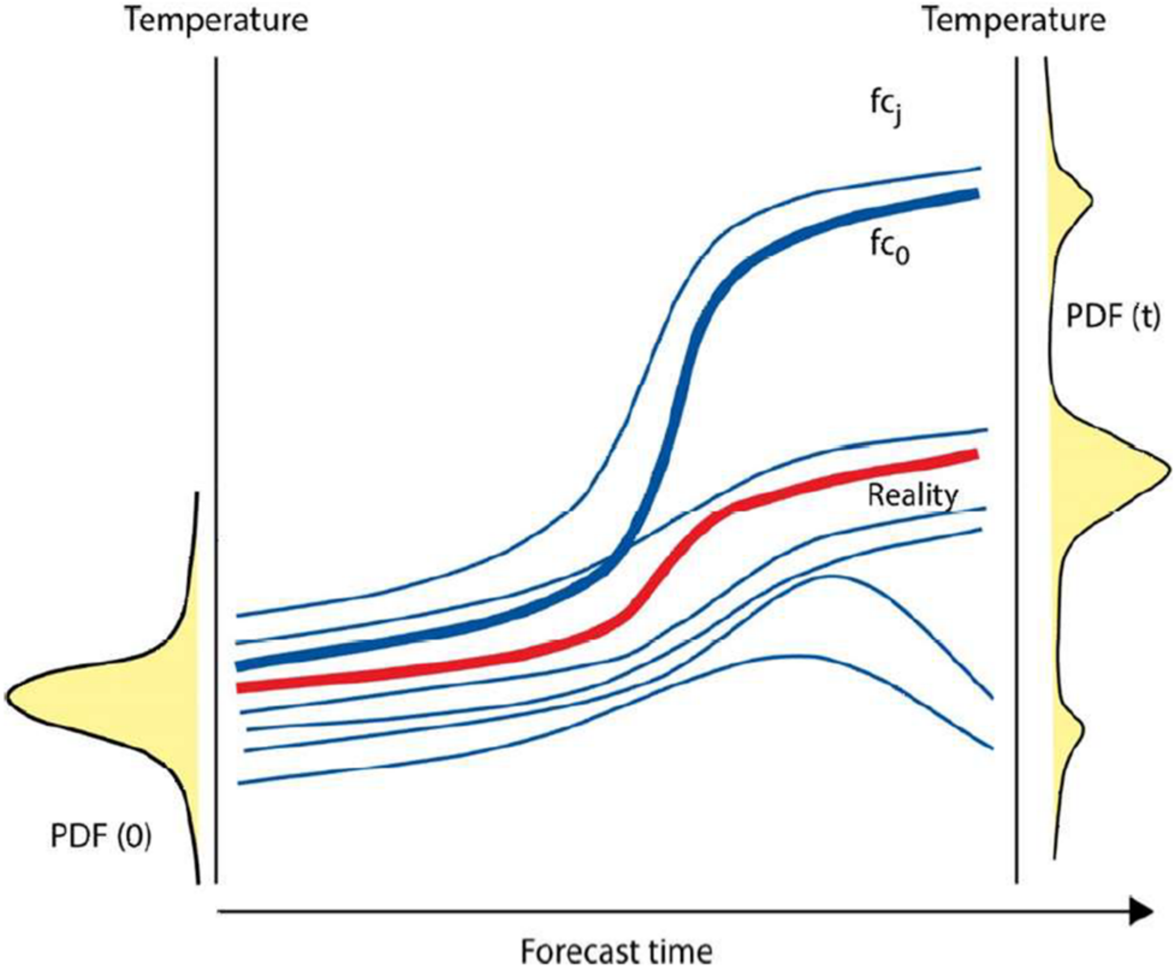

Figure 1.1: Conceptual representation of an ensemble forecast for temperature. Thin blue lines represent the ensemble members (fc_j), while the think blue line is the control forecast (fc_0) and the red line is the reality. The PDF at time 0 is Gaussian, while after a time t it becomes multi-modal due to nonlinear dynamics. Adapted from Toth and Buizza (2019b).

on a single forecast to a probabilistic approach. Nowadays, this paradigm shift is widely acknowledged as a necessity in forecasting, emphasizing the incorporation of uncertainty estimates (Toth and Buizza, 2019a). Whether for short-term, medium-range, monthly, seasonal forecasts, or even the more extended outlooks like decadal and climate projections, ensembles have become indispensable. They not only present the most probable scenarios but also offer critical insights into associated uncertainties. As the predictability diminishes over time, especially with longer forecasts, adopting a probabilistic approach becomes imperative. This evolution underscores the significance of ensembles in modern forecasting, enabling meteorologists to not just predict but also understand the inherent uncertainties within the complex weather systems.

1.3 Challenges in ensemble prediction

Although ensemble forecasting has become very useful and indispensable in weather prediction, it still faces certain challenges. In this section, three challenges will be discussed. Firstly, the high computational cost of running an ensemble limits the number of members in the ensemble forecast, which is an important factor in determining how well a probability distribution of a weather-related variable can be estimated. Therefore, it is crucial to investigate the question of how big an ensemble should be to study and understand forecast uncertainty in a probabilistic framework.

Secondly, it is important to include all the relevant sources of uncertainty when designing an ensemble. Early EPS methodologies, such as the Monte-Carlo approach introduced by Leith (1974), primarily focused on addressing uncertainty in the Initial Conditions (IC). However, over time, this scope has significantly broadened to encompass uncertainties in all aspects of a modeling system, including atmospheric initial states, model physics, numerical methods, lower boundary forcing such as land or sea surface conditions, Lateral Boundary Conditions (LBC), and additional coupling mechanisms such as air-sea interaction (Du et al., 2019). IC have been extensively studied so far and advanced methods to account for the uncertainty in the initial state have been developed. Therefore we focus on model uncertainty, in particular on that arising from unresolved processes due to the limited resolution of NWP models. This has become a particularly relevant issue more recently, as the resolution of operational models reached the kilometer scale, which required a revision of the way physical processes are represented in the models.

Finally, another important aspect to consider in probabilistic forecasting is the flow-dependence of forecast uncertainty on the convective scale. In this context, the influence of the larger-scale flow on convection is of key importance for the forecast evolution and its uncertainty, with important implications for predictability at the convective scale.

In the following subsections, the three issues described here are further discussed: the limited ensemble size, the model error representation and the flow dependence.

1.3.1 Limited ensemble size

In practical terms, the ensemble size, which is the number of member forecasts utilized to formulate the anticipated distribution of a forecast variable, stands out as a critical factor influencing the quality of probabilistic forecasts. The size of the ensemble is particularly crucial for accurately capturing the nuances of distributions, especially when dealing with rare outlier occurrences and non-Gaussian behaviors like multi-modality or heavy tails (Bannister et al., 2017). Nevertheless, owing to computational constraints, operational ensembles within NWP centers typically incorporate fewer than 50 members for global models, and even fewer for limited area models (e.g., Gebhardt et al., 2011; Bouttier et al., 2012; Hagelin et al., 2017; Schwartz et al., 2017; Frogner et al., 2019; Keil et al., 2020). This is not enough to accurately represent the nonlinear evolution of the forecast distributions (Leutbecher, 2019). In an insightful study, Kondo and Miyoshi (2019) conducted experiments using an intermediate atmospheric general circulation model, revealing that approximately 1000 ensemble members were necessary to accurately represent crucial distribution features such as multi-modality and the probability of extreme events.

On the other hand, limited understanding exists concerning the challenge of undersampling in convective-scale NWP. Research by Harnisch and Keil (2015) revealed that augmenting the ensemble size from 20 to 40 members resulted in a more accurate analysis and improved 3-hour forecasts. Similarly, Hagelin et al. (2017) demonstrated a substantial enhancement in precipitation forecast skill by doubling the ensemble size from 12 to 24, utilizing the Met Office convective-scale ensemble (MOGREPS-UK). In a comparative study, Raynaud and Bouttier (2017) evaluated the advantages of increasing ensemble size from 12 to 34 members against increasing horizontal resolution from 2.5 to 1.3 km. Their findings indicated that enlarging the ensemble size proves more advantageous than resolution augmentation for lead times exceeding approximately 12 hours, particularly when dealing with larger uncertainties. This aligns with the observations of Legrand et al. (2016), who identified a growing need for larger samples due to increasing non-Gaussianity with extended forecast lead times. A consistent trend emerges from various studies deploying substantial ensembles on global and regional scales, emphasizing that the most significant non-Gaussianity originates from highly nonlinear processes within deep convective clouds (Miyoshi et al., 2014; Jacques and Zawadzki, 2015). Recent investigations employing data assimilation in 1000-member convective-scale ensembles have underscored the rapid development of non-Gaussianity in less than an hour during deep moist convection, originating in the vicinity of convective updrafts (Kawabata and Ueno, 2020; Ruiz et al., 2021)

The preceding findings collectively indicate that the currently employed ensembles are likely too small at least for some purposes. However, determining the optimal size of ensembles remains a question. Generally, the required number of samples to estimate the distribution of a forecast variable depends on the distribution's shape, which can vary considerably in weather prediction. Figure 1.2 illustrates a conceptual model depicting how the distribution of a forecast variable might change over time. Initially, the uncertainty in the initial conditions is relatively small, assumed to be Gaussian in the data assimilation system. As time progresses, the distribution may broaden, developing asymmetric tails due to factors like non-negativity constraints on quantities such as humidity (left panel). With the influence of nonlinear processes over time, the distribution may assume a complex form with heavy tails indicating higher probabilities of extreme events or even multi-modal distributions associated with preferred regimes (center panel). Eventually, the forecast ensemble loses memory of the initial conditions, and the forecast distribution converges to a broad, smooth, climatological distribution (right panel). In convective-scale forecasting for a limited-area model, the "climatological" distribution represents possible weather within the model when small-scale errors have saturated, subject to synoptic-scale conditions from the driving global ensemble (Selz, 2019). While this "climatological" distribution may be reached within a day or two (Hohenegger and Schär, 2007), it continues to evolve on synoptic timescales rather than being time-independent.

In addition to evolving over time, the shape of the forecast distribution is also subject to the specific forecasting scenario under consideration. It is influenced by factors such as the prevailing weather regime (e.g., convective or clear) and the nature of the forecasted

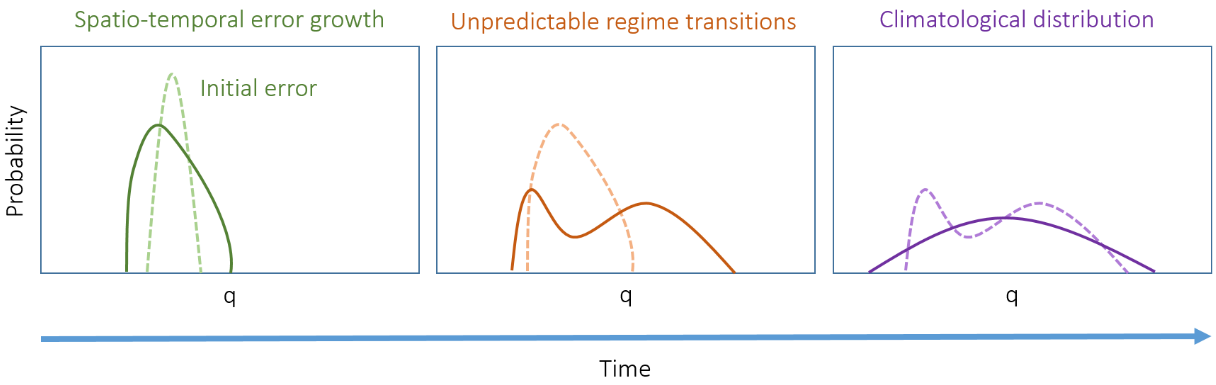
Spatio-temporal error growth
Unpredictable regime transitions
Climatological distribution
Initial error
Probability
q
q
q
Time

Figure 1.2: Conceptual model showing the time evolution of the distribution of a hypothetical forecast variable q. In each panel the dashed line represents the distribution at an earlier time, which evolves into the distribution shown by the solid line. In the centre and right panels, the dashed line is identical to the solid line in the previous panel. See text for further details.

quantity. Variables like precipitation, constrained by non-negativity, exhibit skewed distributions, whereas aggregated measures such as time- or area-averages may display more Gaussian characteristics compared to point values. While it might be feasible to accurately estimate the ensemble mean with a relatively small ensemble, capturing higher moments of the distribution or predicting the probability of extreme events could necessitate significantly larger ensembles (Leutbecher, 2019). The question of determining the optimal ensemble size encompasses a series of inquiries, and the challenges associated with experimenting with large ensembles in NWP make it challenging to provide definitive answers.

Due to concerns that computationally feasible ensemble sizes may not be sufficiently large to accurately represent forecast uncertainty, various techniques have been proposed to enhance the representativeness of smaller ensembles. In global ensembles, employing perturbations based on singular vectors or bred vectors ensures the capture of the most rapidly-growing error modes (Palmer et al., 1998; Toth and Kalnay, 1997). In the context of limited area models, selecting lateral boundary conditions becomes crucial to ensure that the complete spread of the global ensemble is adequately represented (Montani et al., 2011). Research by Marsigli et al. (2014) highlights that the lack of diversity in the global ensemble, providing boundary conditions to a limited-area ensemble prediction system, can be a significant limitation, especially when the global ensemble is small or based on a single forecast model. The impact of this limitation is likely to vary with the weather regime, as demonstrated by Keil et al. (2014) in their study of ensemble forecasts in a convection-permitting model. They found that the driving model is the primary source of uncertainty when the synoptic forcing of convection is strong, whereas model physics perturbations influencing the triggering of convection become the main source of uncertainty when the forcing is weak.

In conclusion, the size of probabilistic weather forecasts ensembles is constrained by cost, often limiting them to smaller sizes. Assessing errors stemming from these limited sizes proves challenging because the distribution of a forecast variable, as observed by a larger ensemble, is unknown, making it uncertain how many ensemble members are necessary for accurate sampling.

1.3.2 Model error representation

In regional convection-permitting EPSs, three primary types of uncertainties are included. Firstly, the central aspect lies in initial condition uncertainty within forecast ensembles, usually achieved by conducting multiple simulations that commence with slightly perturbed initial states derived from data assimilation procedures. Secondly, lateral boundary condition uncertainty is typically addressed through an ensemble of global model simulations that offer diverse large-scale flow patterns covering the simulation domain of the regional model. Lastly, model uncertainty arising from unresolved or inadequately represented physical processes stands as another important source of uncertainty and poses a significant challenge in ensemble forecasting.

Various approaches have been conceived to integrate model uncertainty into convection-permitting EPSs. To address unresolved, subgrid-scale physical phenomena, stochastic perturbation techniques are being employed, while uncertainties in the structure of physical processes are commonly handled through methods such as "multiphysics" or perturbed parameter strategies (e.g. Berner et al., 2017; Fleury et al., 2022; Roberts et al., 2023, and references therein). However, current convection-permitting EPSs frequently exhibit underdispersion in near-surface variables (e.g. Bouttier et al., 2012; Raynaud and Bouttier, 2017), and devising ensemble construction methodologies that adequately capture the multitude of uncertainties at play in nature remains a persistent challenge.

Boundary layer turbulence, along with cloud microphysics and their interplay with aerosols, constitutes significant sources of model uncertainty in convection forecasts (Clark et al., 2016). The initiation of convection is closely tied to boundary layer processes, yet these processes remain incompletely resolved due to their inherently small scales. In numerous existing boundary layer schemes, turbulent processes are depicted by a mean state within a grid box. This approach results in inadequate small-scale variability, particularly hindering or postponing the initiation of convection, especially in the absence of convective forcing to trigger convection (Kühnlein et al., 2014). This issue is addressed by Kober and Craig (2016) and Hirt et al. (2019), who developed the Physically Based Stochastic Perturbation Scheme (PSP) for subgrid processes that targets the coupling between subgrid turbulence and resolved convection. The scheme is introduced in more detail in section 2.2.

In summary, the ongoing challenge is to develop ensemble methodologies that effectively capture the multitude of uncertainties in natural systems and combine them to create physically consistent and effective variability in the ensemble forecast.

1.3.3 Flow dependence of convection

The development of convection is influenced by several factors. Firstly, the cooling of the troposphere plays a significant role in generating instability, primarily through dynamically induced ascent. This cooling at higher altitudes is balanced by heating and moisture addition in the boundary layer, either from surface heat fluxes or through advection, which results in the formation of Convective Available Potential Energy (CAPE). However, the mere presence of CAPE doesn't assure convection will take place. Often, triggering mechanisms from mesoscale and local features are necessary to overcome convective inhibition caused by a capping inversion at the upper boundary of the planetary boundary layer. These triggering features can include convergence lines, various boundary-layer formations, and disturbances from previous rounds of convective clouds, like outflow boundaries and gravity waves. Identifying the meteorological characteristics impacting predictability is a challenging task.

The impact of dynamical forcing on convection in mid-latitudes is often characterized as either strong or weak. Typically, this classification relies on the presence or absence of synoptic or mesoscale dynamical features capable of inducing upward motion and the formation of CAPE. However, identifying such features is often subjective, necessitating a more precisely defined measure of the convective environment's influence. To address this need, Done et al. (2006) introduced the convective adjustment timescale, which gauges the extent to which convection aligns with larger-scale forcing. When inhibition of convection is weak and triggering disturbances are abundant, convection occurs whenever instability exists, leading to the rapid consumption of CAPE. Conversely, when inhibition is strong and triggering disturbances are scarce, CAPE can accumulate, indicating a non-equilibrium state. Equilibrium conditions often coincide with strong forcing because dynamical ascent weakens inversions, while widespread convection supplies numerous triggering disturbances.

In conclusion, the distinction between strong and weak dynamical forcing of convection in mid-latitudes and the use of the convective adjustment timescale provide insight into the convective environment and offer a more precise measure of the influence of larger-scale forcing on convection. Understanding these dynamics is crucial, as they dictate the forecast evolution on the convective scale.

1.4 Uncertainty and convergence in a big ensemble: a preliminary study

The motivation for this PhD project was born from a study on properties of distributions of forecast variables in a big convection-permitting ensemble. The 1000 member ensemble data set used by Craig et al. (2022) includes 14-hour forecasts, with 3 km resolution, for eight different days that featured convective weather over Germany. Although the length and number of forecasts is limited, the large ensemble size provides an opportunity to characterize the forecast distributions with exceptional accuracy, and to address the question of how big an ensemble is required for a wide variety of forecast variables drawn from different distributions.

Three distribution types The first goal of this investigation was to inspect histograms of the various forecast quantities for different regions and times, in order to identify the characteristic types of distribution produced by the ensemble. Each forecast variable was found to have a typical shape, and these shapes could be classified into three broad categories: quasi-normal distributions, highly skewed distributions, and mixtures with two or occasionally three peaks. Table 1.1 shows which variables are assigned to each category. A distribution is classified as quasi-normal if it is unimodal, with a relatively small skew. In most cases, these distributions are fitted well by a Gaussian function. Variables with this distribution shape include temperature (see Fig. 1.3), all wind components at 500 hPa, and mean sea-level pressure. The quasi-normal shape was found for all neighborhood widths, averaging regions, and forecast lead times. Note that this subjective description does not take into account outlier values, such as the temperature or vertical velocity anomalies at the core of a convective updraft, which are very rare (around 0.1% of grid points) in the case considered here.

Variables showing highly skewed distributions include precipitation and reflectivity. These quantities are both related to hydrometeor content and hence bounded by zero. The example precipitation distribution shown in Figure 1.3 is closer to lognormal than normal in shape. Note that the full distribution of precipitation rates includes a point mass at zero representing members that have zero precipitation at this location, and would be best described as a mixture that resembles a combination of a lognormal distribution and delta function at zero.

The last group, mixture distributions, includes the specific saturation deficit, q$_{def}$, and other humidity variables. It soon became apparent that an important factor influencing the distributions was the distinction between cloudy (saturated) regions and unsaturated air. To make this distinction more obvious, specific saturation deficit (q$_{def}$) was plotted instead of specific humidity. This quantity is defined as the difference between the saturation water vapor mixing ratio at the grid point temperature and the actual mixing ratio and is related to relative humidity:

$$q_{def} = q_{sat} - q = q\left(RH^{-1} - 1\right), \tag{1.1}$$

where $RH = q/q_{sat}$ is the relative humidity. Figure 1.3 shows that the complex distribution shape arises as a mixture of distinct distributions in cloudy and clear regions.

The time evolution of the shape of the specific saturation deficit distribution shown in Figure 1.4 resembles that of the conceptual model in Figure 1.2. The distribution begins relatively Gaussian at the analysis time but undergoes narrowing and increasing asymmetry as time elapses. At a 6-hour lead time, the histogram concentrates near the zero bound of saturation deficit, indicating widespread cloudiness in nearly all ensemble members. By the end of the 14-hour forecast, the histogram exhibits peaks for both cloudy and clear-sky conditions. The evolution of the distribution beyond this point is uncertain, but this form aligns with expectations for the "climatological" distribution, where predictability of convective cloud locations diminishes. At this stage, the overall humidity is influenced by large-scale conditions, yet there is no skill in predicting the cloudiness of a specific gridpoint.

While the overall distribution behavior aligns with the conceptual model in Fig. 1.2, clear identification of all stages for each forecast variable is challenging. A higher time resolution would be necessary to observe the impact of the initial loss of convection predictability, and a longer simulation time may be required before considering the distributions as climatological, as defined in Fig. 1.2.

Table 1.1: Classification of variables. All 3-dimensional variables are extracted at 500 hPa.

Category	Variable
Quasi-normal	Temperature
	Horizontal wind velocity
	Vertical wind velocity
	Mean sea level pressure
Highly skewed	Precipitation
	Reflectivity
Mixture	Specific humidity
	Specific saturation deficit
	Relative humidity

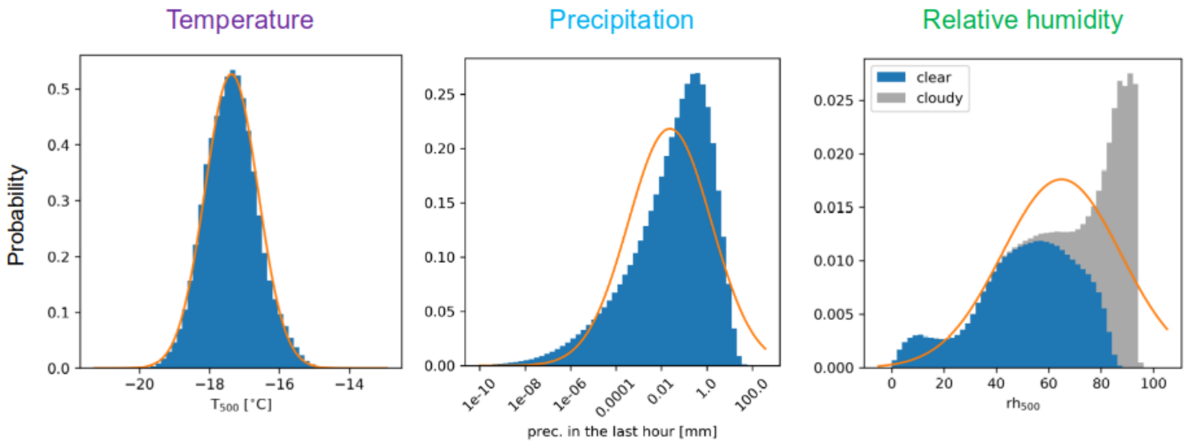

Figure 1.3: Examples of histograms for the three categories: (left) temperature at 500 hPa, quasi-normal; (center) hourly precipitation, highly skewed; (right) relative humidity, mixture. The criterion to distinguish cloudy grid points is a simulated radar reflectivity higher than -19 dBZ at 500 hPa. A Gaussian function with the same mean and standard deviation is shown for comparison (solid lines). Adapted from Craig et al. (2022).

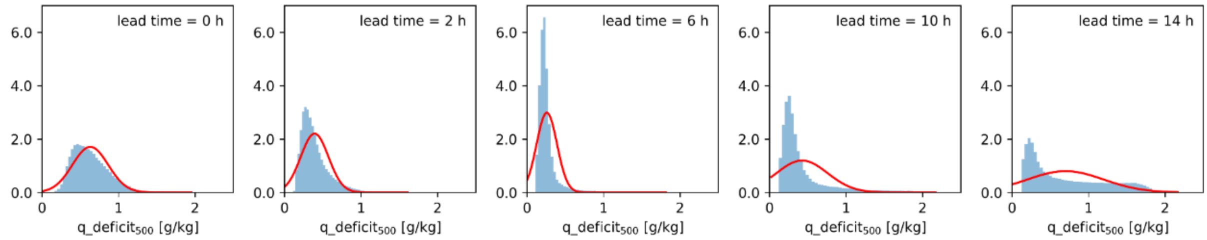

Figure 1.4: Evolution of the histogram of specific saturation deficit. Forecast lead time increases from left to right from the first time step (1200 UTC) to 14 h (0200 UTC). A Gaussian distribution function with the same mean and standard deviation is shown for comparison (solid lines). Adapted from Craig et al. (2022).

Asymptotic convergence of sampling error Estimates of forecast quantities constructed from a small ensemble will suffer from sampling error, but should converge to an accurate value as ensemble size increases. To provide a quantitative measure of this convergence, estimates were made using a range of ensemble sizes, subsampled from the 1,000-member ensemble. Confidence intervals for the estimates were constructed as follows. For each ensemble size, 10,000 test ensembles were created by bootstrapping with replacement (more details on the method in section 2.3). The forecast parameter (e.g., ensemble mean) was computed from each test ensemble to create a distribution of estimates. The 2.5^{th} and 97.5^{th} percentiles of this distribution then define the 95% confidence interval. Figure 1.5 (top row) shows an example of this confidence intervals for the mean, standard deviation and the 95^{th} percentile of precipitation at a specific location and a specific time and how they vary with ensemble size. The bootstrap sample median shows that the estimate of the mean is biased low for small ensemble sizes and all three quantities. The rate of decrease in the width of the confidence interval for smaller ensemble sizes is irregular, although it becomes smoother for larger ensemble sizes.

The rate of convergence of the forecast estimates can be examined by plotting the width of the confidence interval as a function of the ensemble size N. If the ensemble size is large enough, and the underlying distribution is well behaved (e.g., has finite moments), the Central Limit Theorem (CLT) states that for a large number of independent and identically distributed random variables, the sampling distribution of the normalized sum will tend towards a normal distribution without dependence upon the underlying distribution's shape (Dekking et al., 2005). The standard error of the mean of this sampling distribution will then be proportional to N-1/2. This normality of the sampling distribution can also be extended to a wide range of other statistics, including those of the standard deviation and of quantiles (Walker, 1968). The width of the 95% confidence interval is a multiple of the sampling distribution's standard error, and can also be expected to converge as N-1/2. However, this behavior is only expected in the limit of large ensemble size, and it is not certain whether it can be observed for the available meteorological distributions and ensemble sizes.

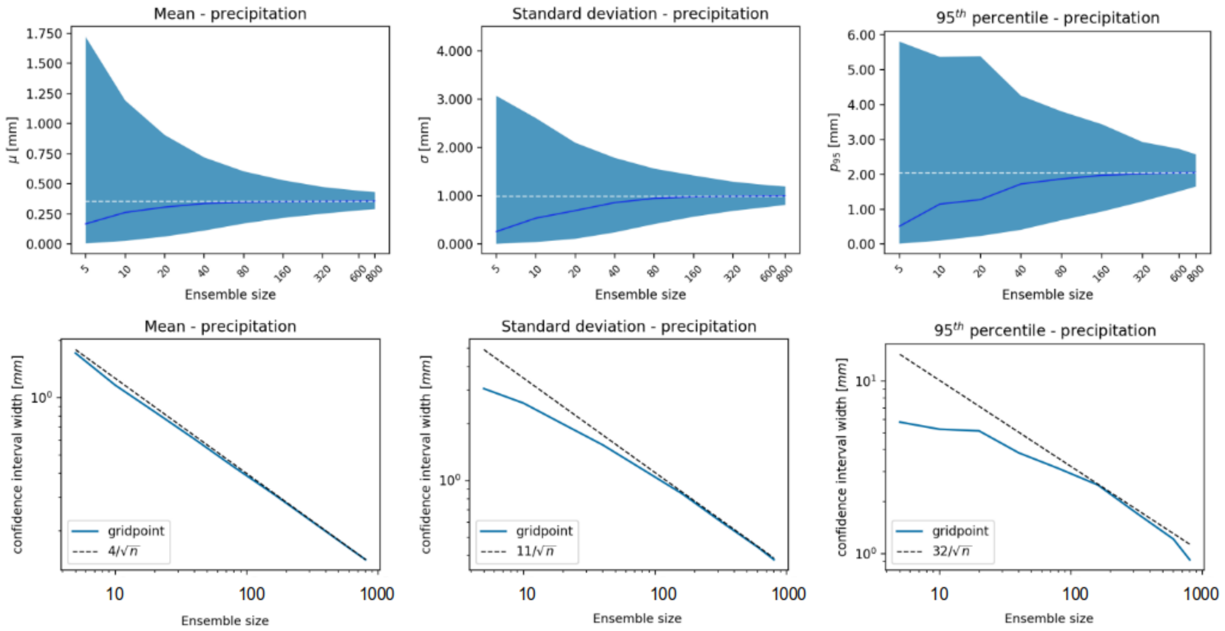

Figure 1.5: (top row) mean, standard deviation and 95th percentile of hourly precipitation. The bands show the 95% confidence interval determined for the 10,000 bootstrapped samples (bounded by 2.5% and 97.5% quantiles), while the solid lines show the respective median values. The dashed white horizontal line indicates the median of the distribution computed using all 1000 members. (bottom row) Width of the 95% confidence intervals shown above. Dashed lines show reference curves with slope $\mathrm{N}^{-1/2}$, fitted by eye. Adapted from Craig et al. (2022).

The bottom row in Figure 1.5 shows the width of the confidence interval of hourly precipitation as a function of the ensemble size. The resulting straight line in the left panel means that for the ensemble mean of hourly precipitation, the width of the confidence interval decreases proportional to $N^{-1/2}$. It is quite striking that the convergence follows the asymptotic law even for ensemble sizes of less than 10 members. For the estimates of standard deviation of temperature and saturation deficit, the $N^{-1/2}$ scaling appears to hold for ensemble sizes of 100 or greater, but the deviations for smaller ensemble sizes are often significant. For the 95^{th} percentile, the $N^{-1/2}$ scaling is not observed even with up to 1000 members.

The results of this motivational study suggest that in deciding what size of ensemble is needed for a particular forecasting problem, it is important to consider whether the ensemble is large enough for the asymptotic convergence behavior to be established. It is significant that for the forecast problems considered here, the ensemble sizes of 40-100 currently used in operational weather forecasting are more than adequate to show convergence of the ensemble mean, and in most cases sufficient for the standard deviation, although clearly inadequate for more extreme events such as the 95^{th} percentile. It is tempting to speculate that, since the operational ensembles are often evaluated in terms of their standard deviation (RMS spread), relative to the mean error, the current ensemble sizes have been chosen as the minimum necessary to give a useful estimate of spread.

Although the presented study offered a significant insight into the field of forecast uncertainty, a number of limitations are to be addressed in additional work. Foremost, the limited number of cases did not allow for a systematic flow-dependent analysis. Several studies (e.g. Keil et al., 2014, 2019; Bachmann et al., 2020) have shown that predictability of convective precipitation is strongly related to the larger-scale forcing. Owing to the interaction between larger-scale flow and convection, strong forcing is typically more predictable than weak forcing. In other words, the area-averaged intensity of convection can be predicted with greater accuracy over an extended period. Predictability is often assessed by examining the spread of the forecast ensemble's distribution for a specific variable, such as precipitation, but we show above that distributions can have a complex shape and therefore more accurate measures of uncertainty are needed. Other limitations of the preliminary study are the lack of surface variables analysis, which typically experience more variability, especially at convective spatial- and time-scales, and the focus on specific locations to study forecast uncertainty and sampling error. These issues will be addressed in the present thesis, to give a more complete insight into the flow-dependent evolution of forecast uncertainty in probabilistic forecasting.

1.5 Research questions and outline

This dissertation addresses the majority of the limitations in work done so far. A period of three months is covered, allowing a systematic convective weather regime classification. A stochastic scheme is used in a state-of-the-art NWP model, to enlarge the ensemble by including a new representation of model uncertainty. The resulting ensemble of 120 members is one of the largest convection-permitting ensembles run so far. Moreover, the analysis of sampling error convergence introduced above is extended to surface variables and the spatial distribution of uncertainty is studied, focusing on the influence of the larger-scale flow on the convective scale. Finally, the evolution of forecast uncertainty beyond 24 hours is investigated.

The dissertation is divided into three parts. Firstly, it assesses a physically-based stochastic perturbation (PSP) scheme with a focus on its flow-dependent impact. Secondly, it explores the flow-dependent representation and evaluation of forecast uncertainty, as well as of sampling error. Lastly, it extends the analysis to longer lead times, up to 48 hours, with a focus on memory effects that extend the predictability of convection on the second day of the simulation.

The **research questions** that are posed in the thesis are:

1. Does the PSP scheme systematically improve the probabilistic skill of convection-permitting ensemble forecasts over Germany?

2. Is a 120-member ensemble sufficiently large to observe convergence of sampling error with a fully-fledged NWP ensemble?

3. How does the convective weather regime affect the evolution of uncertainty of forecast variables and how does it influence its spatial distribution?

The first part of the thesis evaluates the impact of the PSP scheme (Kober and Craig, 2016; Hirt et al., 2019), which has been implemented in the convection-permitting ICON-D2 ensemble prediction system at Deutscher Wetterdienst (DWD) and run for a three-month trial experiment in summer 2021. The scheme mimics the impact of boundary layer turbulence on the smallest resolved scales and impacts in particular convective precipitation. A weather regime-dependent systematic evaluation is carried out, including the verification against observations, with a particular focus on ensemble spread and the spread to skill ratio. As the scheme increases the ensemble spread without significantly affecting the forecast skill, it is suitable for increasing the ensemble size of the ICON-D2 ensemble forecast.

In the second part of the thesis, the flow-dependence of forecast distributions is studied using a 120 member ICON-D2 ensemble with PSP for two representative case studies of weak and strong synoptic forcing of convection, chosen from the three-month period examined in the first part. The bootstrapping technique is used to investigate the convergence of sampling error for a range of surface and mid-tropospheric variables. Additionally, maps of uncertainty are introduced, which allow for a more detailed analysis of the spatial pattern of uncertainty and facilitate the interpretation of the sources and evolution of forecast uncertainty up to 24 hours in different synoptic forcing conditions.

The analysis of the evolution of forecast uncertainty is extended to 48 hours in the third part of the thesis. Another pair of case studies is used in the investigation of flow dependence of the impact of the PSP scheme on the evolution of the forecast beyond one day. This reveals interesting mechanisms that transfer the uncertainty arising from the first day of weak forcing to the second day, as well as the impact that a different initialization time can have on the evolution of forecast uncertainty.

Outline The outline of the thesis is the following: **Chapter 2** introduces the limited-area version of the ICON model, as well as the PSP scheme and the used methods, such as bootstrapping and the neighborhood method. In **Chapter 3**, the ensemble simulations are presented and their setup is described, along with the weather situation of the case studies. **Chapter 4** shows the results of the first part of the investigations, the trial run with the PSP scheme, as described above. **Chapter 5** is the main part of the thesis and shows the flow-dependent evaluation of forecast uncertainty using a big ensemble, which is extended to 48 hours in **Chapter 6**. Finally, **Chapter 7** summarizes the results and presents the conclusions of this work.

Chapter 2

Model and methods

In this chapter, the limited-area NWP ICON model will be described first. It is used operationally at DWD, both as a deterministic model and in an EPS. We use this model for all our experiments, since our group actively contributes to its development and it offers a state-of-the-art framework for our investigations on forecast uncertainty. Additionally, the Physically Based Stochastic Perturbation Scheme (PSP) scheme will be introduced. The scheme is used in all our experiments as an additional source of variability originating from the subgrid turbulence in the Planetary Boundary Layer. It also allows to increase ensemble size, which is needed for the sampling error analysis. The resampling method used for this purpose is bootstrapping, which is introduced thereafter. Finally, the neighborhood method is described, which can reduce the sampling error associated with small ensemble size.

2.1 The ICON model

All numerical simulations are performed with the ICON model in its limited-area mode ICON-D2, which is used in operational weather forecasting at DWD since February 2021 (Reinert et al., 2021). ICON employs an unstructured icosahedral-triangular Arakawa-C grid in the horizontal direction, formed by spherical triangular cells that seamlessly cover a simulation domain. The ICON-D2 domain covers Central Europe (see Figure 2.1) with a grid spacing of 2 km (542,040 grid cells roughly encompassing 1400 km x 1600 km) and 65 vertically discretized layers from the ground up to 22 km above mean sea level. As described in Zängl et al. (2015), its dynamical core is based on the non-hydrostatic equations for fully compressible fluids. The prognostic variables are the edge horizontal wind speed, vertical wind speed, air density, virtual potential temperature, mixing ratios and, when using the two-moment microphysics scheme (Seifert and Beheng, 2006a), the number density of hydrometers. Time integration is performed using a two-time level predictor-corrector scheme. Hourly ICON-D2 output data is interpolated onto a uniform, rotated pole coordinate consisting of 651 x 716 grid points (466,116 in total) with a grid spacing of 2.2 km.

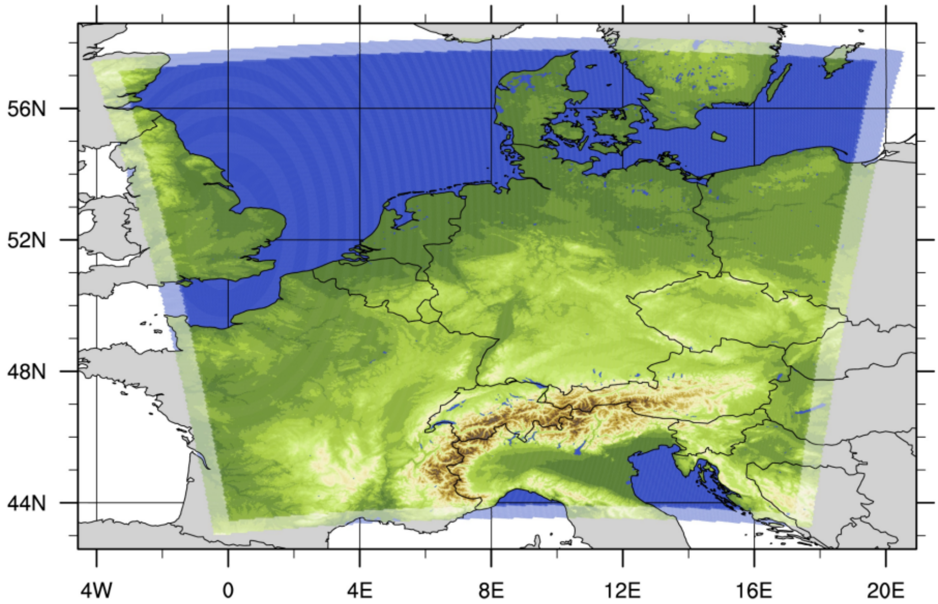

Figure 2.1: Domain of the ICON-D2 limited-area model with orography. From Felger (2022).

Initial Conditions (IC) and Lateral Boundary Conditions (LBC) are the backbone in limited area modelling and represent a major source of forecast uncertainty. In the operational ICON-D2 ensemble prediction system (ICON-D2-EPS) at DWD, IC uncertainty is provided by the ensemble data assimilation system ICON-D2-KENDA (Kilometer Scale Ensemble Data Assimilation, Schraff et al., 2016). In the 40-member ICON-D2-KENDA the model state is updated hourly by assimilating observations and 1-hour first guess forecasts. In 2021 only conventional observations (synoptic stations, radiosondes, wind profilers and aircrafts) were operationally assimilated.

The uncertainty representation in LBC stems from ensemble forecasts generated by the coarser grid model. The global ICON-EPS has a horizontal grid spacing of 40 km (26.5 km since November 2022). An ICON-EU nest is embedded online in the global ICON simulation and covers the entire Euro-Atlantic region with half the grid spacing. The ICON-EU ensemble provides the ICON-D2 LBC. Forecast variability in the ICON-EU-EPS is attained by 40-member IC perturbations generated by the ensemble data assimilation with an assimilation cycle of 3 hours, and by ensemble physics perturbations where a random combination of tuning parameters is set for each of the ensemble members and fixed throughout the forecast horizon. As in DWD's operational setup (Reinert et al., 2021), ICON-EU ensemble forecasts initialized three hours before the initialization time of the ICON-D2 ensemble are used. Therefore, the LBC is updated hourly using the ICON-EU-EPS output at lead times 3—27 hours. Since we primarily focus on the impact of model uncertainties we consider the impact of IC and LBC uncertainty together and call it Initial and Boundary Conditions (IBC) uncertainty.

2.2 PSP scheme

The insufficient representation of physical processes in numerical models causes under-dispersion in probabilistic forecasts, especially for near-surface variables. In other words, the forecast is overconfident, because processes that cause variability in the atmospheric system are not represented properly. Many of these processes play a role in convection, like cloud microphysics, cold pool dynamics and boundary layer turbulence. The latter is of key importance in convection triggering and its missing variability could lead to significant forecast errors: for example, the systematic tendency for convection to be weak and late in convection-permitting NWP models (Trentmann et al., 2009; Kühnlein et al., 2014).

Traditional, deterministic boundary layer turbulence parametrizations were developed for a grid size of tens of kilometers, which implied averaging the effect of many eddies contained in one grid box. However, operational NWP models now have a horizontal resolution of the order of 1 km, which matches the size of the largest eddies in the boundary layer (Fig. 2.2). These eddies should lead to significant variability on the resolved scale, which requires a stochastic representation, since the eddies themselves are not resolved. The presence of such an eddy and its interaction with others can, for example, cause a strong enough updraft to trigger the formation of a convective cell, which is partially resolved when it grows upscale.

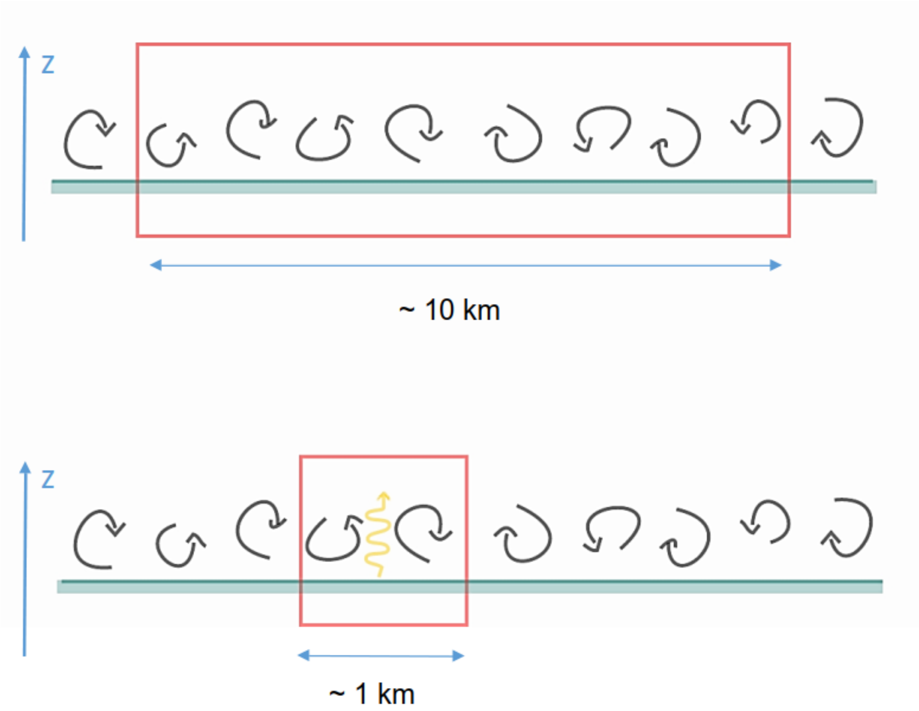

Figure 2.2: Schematic of PBL turbulence and the comparison of the size of PBL turbulent eddies with the gridbox size in a coarse-resolution NWP model (top) and a high-resolution model (bottom).)

In recent years, efforts have been made to develop and test stochastic boundary layer turbulence schemes that reintroduce the missing small-scale variability. Kober and Craig (2016) developed a physically-based stochastic perturbation (PSP) scheme that uses turbulent kinetic energy and flux information from the model's turbulence parameterization to compute the corresponding variances in temperature, moisture and vertical velocity. Spatially and temporally correlated stochastic increments are then added to the model fields to introduce the resolved portion of this turbulent variability. Using the scheme in the Consortium for Small-scale Modeling (COSMO) model, they find that stochastic perturbations lead to triggering of additional convective cells and improve precipitation amounts in simulations of two days with weak synoptic forcing of convection. In a case with strong forcing, the boundary layer perturbations have little impact, as expected, since the amount of precipitation is controlled by the mesoscale and synoptic environment. The PSP scheme has been revised and improved by Hirt et al. (2019), whose version is used in this work and is described in the following.

The PSP scheme reintroduces the variability by means of stochastic perturbations that are scaled according to the turbulence variability. The following stochastic perturbations are added to the temperature, humidity and vertical velocity tendencies with $\Phi \in \{T, q_v, w\}$:

$$\partial_t \Phi|_{\mathrm{PSP}} = f_z \alpha_{\mathrm{tuning}}\ \eta\ \frac{1}{\tau_{\mathrm{eddy}}}\ \frac{l_{\mathrm{eddy}}}{\Delta x_{\mathrm{eff}}} \sqrt{\overline{\Phi'^2}}. \tag{2.1}$$

where

$$f_z(z) = \begin{cases} \frac{z}{z_0} & 0 \leq z < z_0 \\ 1 & z_0 \leq z \leq H_{PBL} \\ 1 - \frac{z - H_{PBL}}{z_0} & H_{PBL} < z < H_{PBL} + z_0 \\ 0 & z \geq H_{PBL} + z_0. \end{cases}, \quad z_0 = 500\,m \tag{2.2}$$

The perturbations are based on a horizontal random field $\eta(x, y, t|\tau_{eddy}, \Delta x_{eff})$. It evolves over time by an autoregressive process with a time correlation corresponding to the characteristic life time of turbulent eddies $\tau_{eddy} = 10\,min$ thereby allowing for memory effects due to missing scale separation (Berner et al., 2017). The random field η also has a spatial correlation of Δx_{eff} (via an approximate Gaussian convolution), which corresponds to the smallest effectively resolved scale, $\Delta x_{eff} = 5\Delta x$ (see e.g. Bierdel et al., 2012). Importantly, the random field is scaled according to the subgrid standard deviation $\sqrt{\overline{\Phi'^2}}$ computed by the deterministic turbulence parameterization (Raschendorfer, 2001). Furthermore, the scheme becomes scale-adaptive by multiplying the perturbations with $\frac{1}{\sqrt{N_{eddy}}}$, since the number of eddies of scale $l_{eddy} = 1000$ m in a grid box, $N_{eddy} = \frac{\Delta x^2}{l_{eddy}^2}$, characterizes the variance of the subgrid scale impact (Craig and Cohen, 2006). In the vertical, we linearly taper the perturbations to zero above the top of the boundary layer (H_{PBL}), as well as in the lowest part of the boundary layer, near the surface, as described by factor f_z. Finally, the parameter α_{tuning} should be of order one and independent of weather regimes or model resolution. Here, α_{tuning} is set to 5. The implementation of PSP into ICON closely follows the version from Hirt et al. (2019) implemented in the COSMO model including the autoregressive process and the tapering of the perturbations at the top of the boundary layer, but excluding the perturbations of the horizontal wind. Fig. 2.3 shows an example of the random perturbation field, the subgrid standard deviation and the resulting combied perturbations of temperature in the COSMO simulations by Hirt et al. (2019).

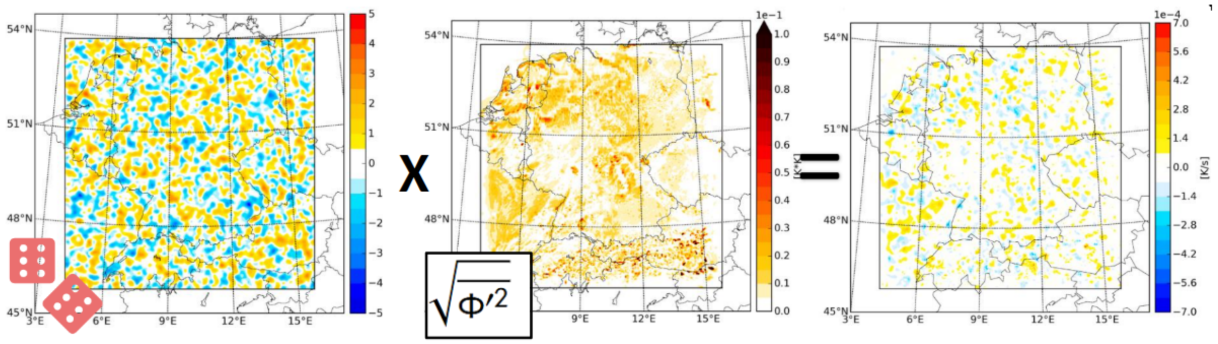

Figure 2.3: Example of the perturbation field created by the PSP scheme in the COSMO model: random field (left), subgrid standard deviation of 2m temperature (center) and resulting tendency perturbations (right). Adapted from Hirt and Craig (2018).

2.3 Bootstrapping

Because of the many degrees of freedom within the atmosphere, ensembles of operational size inherently contain a discernible sampling uncertainty. This arises from the inability of an ensemble to perfectly replicate the true distribution when the number of ensemble members is fewer than the degrees of freedom. Consequently, this sampling uncertainty contributes to inaccuracies in forecast predictions. To estimate the extent of this inaccuracy, bootstrapping, a technique introduced by Davison and Hinkley (1997), can be employed. Bootstrapping involves sampling from a distribution with replacement to generate a statistically equivalent new distribution. Specifically, non-parametric bootstrapping is utilized when the underlying distribution is unknown. This method enables the inference of statistical properties about the underlying distribution without making any assumptions about it. By sampling with replacement from an empirical Cumulative Distribution Function (CDF), bootstrapped distributions are created, which can then be utilized to construct Confidence Intervals (CIs). These CIs provide a probability indicating the likelihood that a statistic falls within the bounds of the chosen interval, e.g. 95% (Jolliffe, 2007). This approach proves valuable in estimating the actual sampling uncertainty.

Following Craig et al. (2022) and Tempest et al. (2023), we also use this approach: different ensemble sizes are used to generate estimates by subsampling from the full, 120-member ensemble. To quantify this convergence, confidence intervals are calculated by creating 1,000 test ensembles for several ensemble sizes using bootstrap sampling with replacement. Various forecast parameters, like the ensemble mean of temperature at 2m, are then computed for each test ensemble to create a distribution of estimates. The 95% confidence interval is defined by the 2.5^{th} and 97.5^{th} percentiles of this distribution.

2.4 Neighborhood method

In numerous forecasting problems, especially those involving convective or subseasonal to seasonal scales, predictions often pertain to averaged quantities. This approach is adopted because a substantial portion of the ensemble's variability may be linked to rapidly changing weather systems, which can obscure predictable variations on larger scales (Toth and Buizza, 2019a). If the small-scale variations are uncorrelated among ensemble members, increasing the averaging region results in a decrease in variability. Additionally, for probabilistic predictions of cumulus convection, enhancing the effective ensemble size can be achieved by sampling statistics within neighborhoods rather than individual grid points (Ebert, 2009; Ben Bouallègue et al., 2013; Hagelin et al., 2017). While these methods can help mitigate sampling errors arising from a small ensemble size, their applicability hinges on specific assumptions about the variability of the weather being predicted.

Several neighborhood approaches have been shown to effectively improve probabilistic forecasts (e.g. Theis et al., 2005; Schwartz et al., 2010; Hitchens et al., 2013). Here a neighborhood method is applied following Craig et al. (2022) to effectively increase the size of the ensemble, with the difference that in this work, the neighborhood is circular and is defined by the radius of this circle, as in Blake et al. (2018). By treating each individual grid point in the selected area as an independent member of the ensemble, the effective size of the ensemble is increased. However, this approach only adds new information if the different grid points are not correlated with one another. For this method to work, the neighborhood size must be large enough for the convection at different grid points within the area to be uncorrelated. Moreover, the statistical properties must be homogeneous, which means that they are not affected by factors such as significant changes in orography or synoptic weather conditions.

Craig et al. (2022) showed that the neighborhood method can reduce the sampling error associated with small ensemble size. The success of the method depends on the small-scale variability of the forecast quantity being uncorrelated in space. The additional points in the neighborhood then provide independent realizations of the variability, leading to a larger effective ensemble size. A quantitative estimate of this increase for ensemble mean precipitation or humidity showed that increase in effective ensemble size corresponds to a correlation length of about 10 gridpoints, which is larger than the effective resolution of the model and may be evidence of some degree of convective organization.

Since the neighborhood method relies on random variability in space, its success for the convection forecasts considered here cannot be generalized to other phenomena such as fog or synoptic weather systems which have smoother spatial structures. Even for 500 hPa temperature the method brought no benefit. This suggests that convective-scale ensemble forecasting systems may be able to use smaller ensemble sizes than the global systems used for medium-range forecasting. It is also possible that sub-seasonal to seasonal forecasts, where the synoptic weather systems can sometimes be regarded as small-scale noise, would again benefit from neighborhood methods.

Chapter 3

Ensemble simulations

This chapter introduces and compares the ensemble simulations used in this work. The three research questions posed in the introduction are addressed by performing three different experiments with the ICON limited-area model, as summarized in Table 3.1. Firstly, a trial run (TR) was performed with the PSP scheme for 3 summer months to assess the improvement of the probabilistic skill given by the scheme, as described in section 3.1. Secondly, section 3.2 describes the two case studies with a 120-member ensemble (CS24) that were used to study the convergence of sampling error, as well as the evolution and spatial distribution of uncertainty in weak and strong convective forcing regime. Lastly, another pair of case studies (CS48), performed to extend this analysis beyond 24 hours, is described in section 3.3

Table 3.1: Description of the ensemble simulations. The IBC and PSP columns indicate the number of different realizations of the respective perturbations.

id	number of days	number of members	IBC	PSP
TR	92	20	20	20
CS24	2	120	40	3
CS48	4	40	40	40

3.1 Trial run with PSP scheme

The trial run of the PSP scheme was made possible by a collaboration with DWD, who showed interest in the newly-implemented scheme in ICON. It consists of 92 24-hour forecasts in summer 2021. The ICON ensemble had 20 members and each member had its own instance of initial and boundary conditions (IBC) as well as its own random seed of the PSP scheme. The simulations were entirely run on DWD machines by Christoph Gebhardt and Chiara Marsigli, while most of the post-processing and analysis was done at the Meteorological Institute in Munich.

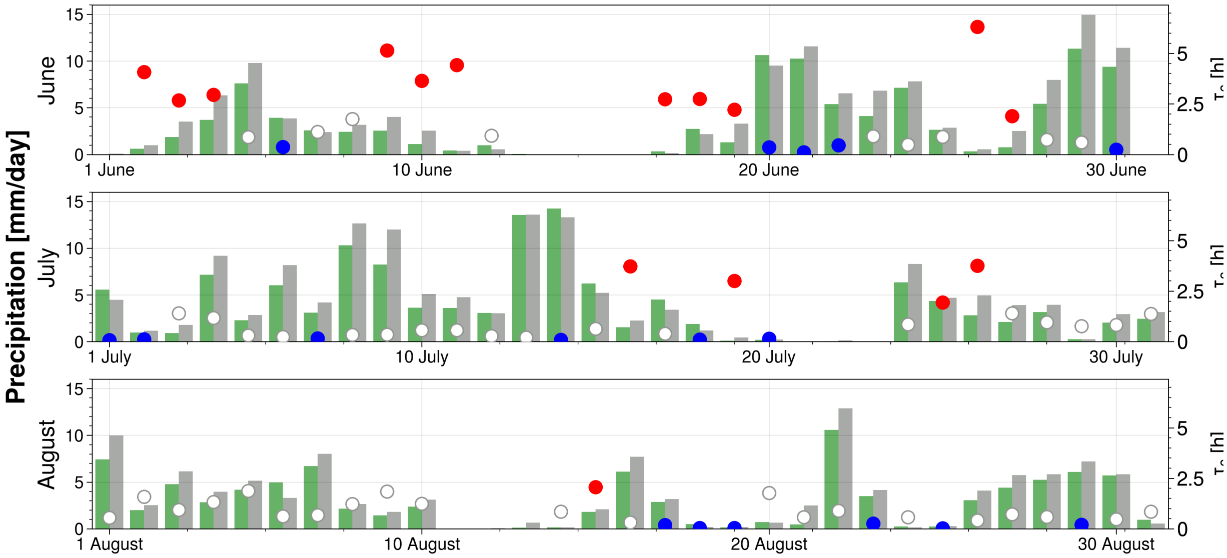

Figure 3.1: Timeseries of June, July and August 2021 illustrating the daily 24-h accumulated precipitation (bars) and convective adjustment timescale τ_c (dots). The colors of the dots represent weak (red), intermediate (white) and strong (blue) forcing regimes (see chapter 4 for details). Green bars depict the reference run ensemble mean and grey bars the radar-observed daily area-averaged rainfall. From Matsunobu et al. (2024).

The ICON model was used for the experiment, more specifically the ICON-D2-EPS of DWD (DWD, Reinert et al., 2021), operational since February 2021, having a horizontal grid spacing of approximately 2.2 km, 65 vertical levels, 20 ensemble members with initial conditions from the KENDA data assimilation system (Schraff et al., 2016) and lateral boundary conditions from the operational ICON-EU ensemble. The model runs for the 24-hour forecasts were initialized at 00 UTC, using the 21 UTC runs of ICON-EU-EPS of the day before for the lateral boundary conditions. The output was saved hourly for the main prognostic variables. The trial period spans 3 months of summer 2021, from 26 May to 31 August 2021, which means 98 days of 24-hour forecasts in total.

The trial simulation consists of two separate experiments with slightly different setups: the reference run and the stochastic run with the PSP scheme turned on. The only representation of model uncertainty in the reference run, which mirrors the operational setup, are the parameter perturbations, which are constant in forecast lead time and space, but vary among the ensemble members and between forecast runs. In the "psp" run, the PSP scheme is applied with a different random seed to each ensemble member, on top of parameter perturbations. The scheme is described now.

Figure 3.1 shows the daily precipitation and convective forcing classification for the whole period. In summer 2021 the weather was characterized by abundant precipitation, with the largest accumulations in the last 10 years on average over Germany (DWD, 2022). Several high impact weather events occurred, including the floods in western Germany (13-14 July) and the hailstorms in southern Germany in the last third of June, including a squall line with widespread severe winds on 29 June.

3.2 Case studies of weak and strong forcing - large ensemble

Again, the ICON-D2 model is used for the experiments. It features a horizontal grid spacing of approximately 2.2 km, 65 vertical levels and initial conditions sourced from the KENDA data assimilation system (Schraff et al., 2016). The lateral boundary conditions are based on the operational ICON-EU ensemble, and the 24-hour forecasts commence at 00 UTC using the 21 UTC runs of ICON-EU-EPS from the previous day for the lateral boundary conditions, updated hourly at lead times ranging from 3 to 27 h. Hourly output is saved for the primary prognostic variables. The numerical experiments are conducted with 120 independent ensemble members, consisting of 40 different initial and boundary condition sets combined with three distinct random seeds of the PSP stochastic boundary layer scheme (Kober and Craig, 2016; Hirt et al., 2019). The scheme mimics the impact of boundary layer turbulence on the smallest resolved scales and impacts in particular convective precipitation. It was shown to improve the forecast in terms of the spread to skill ratio (Puh et al., 2023; Matsunobu et al., 2024). As in the operational ensemble, each of the 40 members in the three groups also has parameter perturbations to represent uncertainty in the formulation of the parameterization schemes, with different values chosen randomly for each ensemble member. The simulations use the two-moment microphysics scheme by Seifert and Beheng (2006b).

Two case studies, representative of weak and strong coupling of convection to the synoptic flow, are chosen following the analysis performed by Puh et al. (2023), who used the convective adjustment timescale to classify 32 days in summer 2021, based on the strength of synoptic forcing of convection (see Section 4.2.1). The two chosen days were also used in Matsunobu et al. (2024) as examples of different convective regimes.

On 10 June, the weak forcing case, a weak geopotential gradient resulted in a weak north-westerly flow over Germany (see Figure 3.2). Scattered convection was triggered around noon, reaching its peak intensity at 14 UTC. Atmospheric conditions stabilized towards the end of the day, marking the conclusion of the daily cycle of convection. Among the 16 cases of weakly forced summer convection in 2021, 10 June is one of the most typical, with a significant impact of PSP on precipitation. The second case, 29 June, was selected as a strong forcing day for several reasons. It was characterized by a strong geopotential gradient and the largest accumulated precipitation in the summer, as estimated by the radar network. A Mesoscale Convective System (MCS) along the cold front in southern Germany caused high-impact weather, such as severe winds, hail, and heavy precipitation. Furthermore, 29 June was a suitable choice due to the similar lead time (+14 h) for the maximum convection activity to 10 June, which is beneficial for comparing uncertainty in the initial conditions. This day was also part of an intensive observation period (IOP 5, 28-30 June) of the Swabian MOSES field campaign (Kunz et al., 2022).

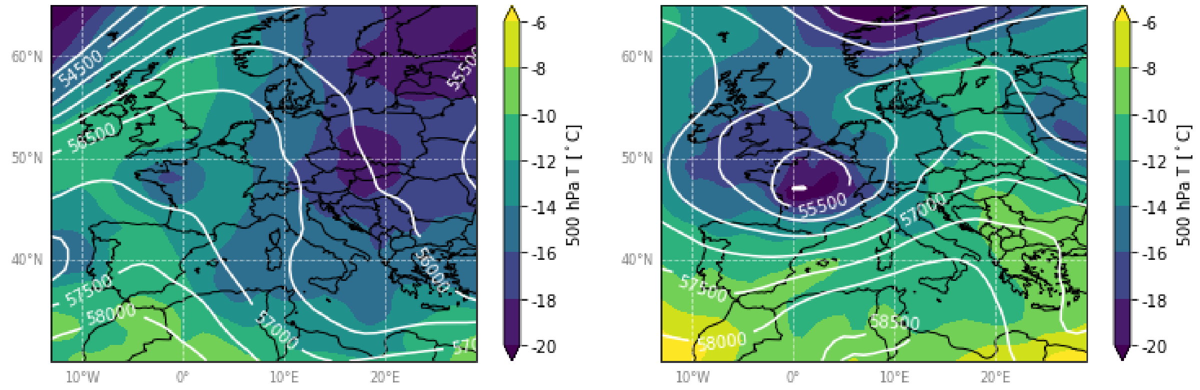

Figure 3.2: ERA5 reanalysis of geopotential (m^2 s^{-2}, white contours) and temperature at 500 hPa (°C, shading) for 10 June 2021, 12 UTC (left, weak forcing) and 29 June 2021, 12 UTC (right, strong forcing).

3.3 Case studies of weak and strong forcing - extended forecast

Lastly, another dataset consisting of 40 ensemble members was produced to study the evolution of forecast uncertainty beyond 24 hours. The setup of the ICON model is similar to the one in the previous section, with the difference that every ensemble member has its own random seed for the PSP scheme, which was found to have a positive impact on the variance in the ensemble. Additionally, simulations are run for up to 48h. Figure 3.3 shows the conceptual flow chart of the four simulations: two with the PSP scheme turned on and two without it (named reference simulations). The initialization time for one pair of experiments is 0 UTC (running for 48h), for the other pair it's 12 UTC (running for 36h). The aim of the different initialization times was to examine how ongoing convection and the updated environmental conditions affect the forecasts through data assimilation.

The second pair of case studies is from August 2022. A "case study" in this section is defined as two consecutive days of the same convective forcing, as we wanted to assess the flow-dependent characteristics of the evolution of forecast uncertainty and a forcing regime transition would have added additional complexity. The cases have been chosen subjectively, based on the observed convection evolution on the visible satellite and radar images.

The first weak forcing day, 26 August 2022, was characterized by a weak pressure and geopotential gradient over most of continental Europe, with a shallow cyclone near Iceland and a high pressure system over eastern Scandinavia and Russia. The atmosphere over Germany was increasingly unstable as colder air aloft was being advected by the weak south-westerly wind at 500 hPa over Germany. This environment was overlapping with a moderate vertical wind shear, since the surface wind was predominantly northerly to north-westerly, favoring convection organization in multi-cell convective storms, whose clouds were covering most of Germany by sunset. Convective activity continued throughout

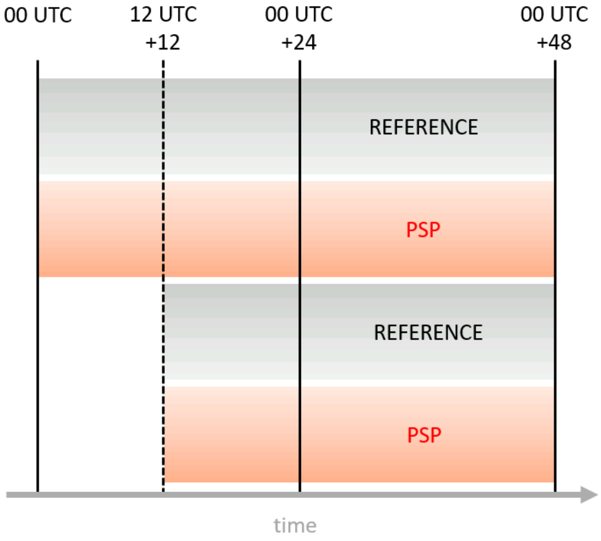

Figure 3.3: Flow chart of the 4 simulations with the extended lead time (CS48).

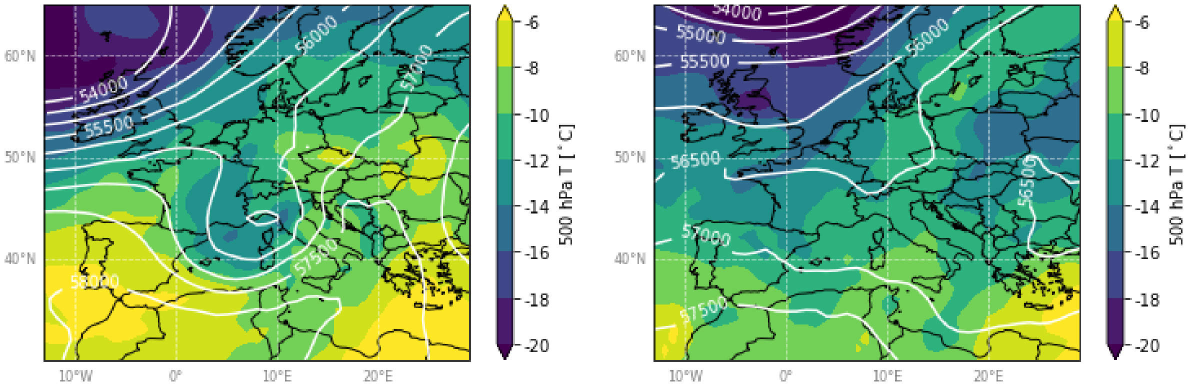

Figure 3.4: ERA5 reanalysis of geopotential (m^2 s^{-2}, white contours) and temperature at 500 hPa (°C, shading) for 19 August 2022, 00 UTC (left, strong forcing) and 27 August 2022, 00 UTC (right, weak forcing).

the night in Eastern Germany and left a widespread cloud layer that persisted until the afternoon on the second day. This made convective initiation on 27 August uncertain, which happened nevertheless. On this day, convection was limited to Southern and Eastern Germany, where a weak cyclonic vortex formed in the lower troposphere. Figure 3.4 shows the temperature and geopotential fields at 500 hPa for 00 UTC on 27 August.

The first day of strong forcing was 18 August 2022. The weather in Europe was affected by a deep low over Iceland and a shallow trough over Western Europe with an associated shallow cyclone over the Northern Mediterranean (see Figure 3.4). In the morning on 18 August, an extremely severe convective system that developed around the Balearic Isles hit Corsica with wind gusts up to 62.2 m/s and continued raging on its path all the way to southern Czechia, causing hail up to 11cm in diameter in Italy and killing 12 people in total, while 106 were injured (ESSL, 2022). The remnants of this exceptional convective system were then advected over Germany, where they merged with locally triggered convection into large coherent precipitating systems, which slowly moved northwards. During the night, convective activity intensified over Southern Germany, with slowly moving convective systems that caused large accumulations of precipitation. This situation persisted throughout the second day, when the precipitation area gradually moved eastward.

Chapter 4

Improving probabilistic forecasts of convection with a stochastic perturbation scheme

4.1 Background

Ensemble prediction systems (EPS) have been implemented to account for the chaotic nature of the atmosphere, imperfect NWP models and uncertain initial conditions. They provide multiple scenarios as an alternative to a single forecast started from the best estimate of the initial state. This is particularly beneficial on the kilometer scale that is now resolved by the high-resolution NWP models, where errors grow rapidly due to nonlinear processes, significantly limiting the predictability (Hohenegger and Schär, 2007).

One of the main advantages of using an ensemble instead of a single deterministic forecast is its ability to represent different sources of uncertainty. A major source of error in limited-area models is the uncertainty in the initial and boundary conditions (Lorenz, 1965), which is commonly introduced by constructing the ensemble using perturbed input fields of global ensembles. Another important source of uncertainty is the model itself, as a consequence of our limited knowledge about atmospheric phenomena and the finite grid size. The former can be addressed by perturbing specific components of the model formulation (e.g. parameters), while the latter requires a careful parameterization of subgrid, unresolved processes. Since convection is mostly resolved at the kilometer scale, the quality of its forecast is limited by our understanding of the key physical processes in convection, like boundary layer turbulence, cloud microphysics (e.g. Thompson et al., 2021; Matsunobu et al., 2022), and cold pool dynamics (e.g. Hirt and Craig, 2021). There is still much fundamental research required in this field, including detailed observations of these processes (Clark et al., 2016). On the other hand, convective initiation is often driven by unresolved boundary layer processes, especially when synoptic forcing is weak and local mechanisms are the main factor in overcoming convection inhibition. In such circumstances, high-resolution models have shown insufficient convective initiation (see e.g.

Clark et al., 2016).

In the context of a convection-permitting EPS, the insufficient representation of physical processes in the model is likely a cause for the underdispersion of the ensemble, especially for near-surface variables. This may be mitigated by stochastic schemes: Bouttier et al. (2012) find that using the Stochastically Perturbed Parameterization Tendencies (SPPT) scheme in the underdispersive AROME ensemble is an effective technique for enhancing spread. Keil et al. (2019) study the relative contributions of soil moisture heterogeneities, a stochastic boundary-layer perturbation scheme and varied aerosol concentrations representing microphysical uncertainties on the diurnal cycle of convective precipitation and its spatial variability. They observe that in the COSMO model the stochastic boundary-layer perturbations leads to the largest spatial variability impacting precipitation from initial time onwards with an amplitude comparable to the operational ensemble spread. Similarly, the results of Jankov et al. (2017) indicate that a WRF ensemble combining three stochastic methods consistently produces the best spread–skill ratio and generally outperforms the multiphysics ensemble (see also Jankov et al., 2019), suggesting that using a single-physics ensemble together with stochastic methods should be considered in the design of future high-resolution regional and global ensembles.

In recent years, efforts have been made to develop and test stochastic boundary layer turbulence schemes that reintroduce the missing small-scale variability. Kober and Craig (2016) developed a physically-based stochastic perturbation (PSP) scheme that uses turbulent kinetic energy and flux information from the model's turbulence parameterization to compute the corresponding variances in temperature, moisture and vertical velocity. Spatially and temporally correlated stochastic increments are then added to the model fields to introduce the resolved portion of this turbulent variability. Using the scheme in the COSMO model, they find that stochastic perturbations lead to triggering of additional convective cells and improve precipitation amounts in simulations of two days with weak synoptic forcing of convection. In a case with strong forcing, the boundary layer perturbations have little impact, as expected, since the amount of precipitation is controlled by the mesoscale and synoptic environment. The PSP scheme has been revised and improved by Hirt et al. (2019), whose version is used in this work (for details see Section 2.2).

Clark et al. (2021) implemented a similar, physically consistent stochastic boundary layer scheme in the Met Office's Unified Model, that introduces temporally correlated multiplicative Poisson noise with a scale-dependent distribution. They evaluate the scheme using small ensemble forecasts of two case studies of severe convective storms over the UK. They find that with horizontal grid lengths around 1 km temporal correlation is far more important than spatial. They also show that the scheme produces sufficient differences between ensemble members at the scale of convective cells. Fleury et al. (2022) test two process-oriented perturbation schemes in a single-column version of the convection-permitting AROME model. They study three idealized boundary layer cases using a Planetary Boundary Layer (PBL) turbulence scheme and a shallow convection scheme. They find that these schemes do not produce enough spread to represent the small-scale variability in temperature and humidity seen in large eddy simulations for the same cases. For wind, the variability compares favorably due to perturbations generated by the stochastic turbulence scheme.

In this chapter, we use the physically-based stochastic perturbation scheme PSP to represent small-scale model error in the boundary layer in the operational ICON-D2-EPS for a three-month period in summer 2021 over Germany. The large number of forecasts in the parallel trial allows for a systematic analysis of the impact of the scheme and thereby infer properties of forecast uncertainty in different weather regimes. A better understanding of these properties and their dependence on the weather regime will allow setting up a more optimal prediction system to represent the future state of the atmosphere and the uncertainty associated with it.

The chapter is structured as follows. Section 4.2 introduces the experimental setup, the simulation period and the used perturbation scheme. In section 4.3, the effect of the scheme on the diurnal cycle of precipitation in different forcing regimes is presented, including its beneficial impact on the spread in weak forcing conditions. Then the effect on spatial uncertainty of precipitation is shown, as measured by the Fractions Skill Score (FSS), as well as the probabilistic verification of other near-surface variables, which indicates a general improvement in the spread to skill relationship. Section 4.4 summarizes the conclusions of this work, discusses its limitations and offers a basis for future investigations.

4.2 Results and discussion

4.2.1 Synoptic forcing regime classification

From the perspective of forecast uncertainty at the convective scale, the type of convective forcing is important. Hence, studying separately the strong and weak forcing regimes allows to infer properties of forecast error and uncertainty evolution conditional to the weather regime. To make the distinction between strong and weak synoptic forcing, we applied the convective adjustment timescale τ_c (equation 4.1). It is an estimate of the time-scale for the removal of conditional instability, measured by Convective Available Potential Energy (CAPE), by convective heating (Done et al., 2006; Keil et al., 2014). The convective adjustment timescale is defined as

$$\tau_c = \frac{CAPE}{\mathrm{d}CAPE/\mathrm{d}t} = \frac{1}{2}\left(\frac{\rho_0 c_p T_0}{L_v g}\right)\frac{CAPE}{P} \tag{4.1}$$

In accordance with Done et al. (2006), the second definition hinges on the expression of the rate of change of CAPE with the vertically integrated latent heat release. This latent heat release can be directly inferred from the precipitation rate P (kg s^{-1} m^{-2}), with quantities within the brackets being constants, where ρ_0 and T_0 represent reference values of density and temperature, c_p the specific heat of air at constant pressure, L_v the latent heat of vaporization, and g represents the acceleration due to gravity (e.g. Zimmer et al., 2011). A small value of τ_c compared to the timescale of the synoptic flow (about 12 hours) means that CAPE is removed by the convection as soon as it is created and the large-scale flow controls the amount of convection. If the value of τ_c is large, the removal of the CAPE by convection is too slow, so small scale factors drive the convection. The computation of τ_c is only done on days when at least once the threshold of 1mm/h was exceeded in more than 100 grid points over Germany (as in Kühnlein et al., 2014).

In summer 2021 the weather was characterized by abundant precipitation, with the largest accumulations in the last 10 years on average over Germany (DWD, 2022). Several high impact weather events occurred, including the floods in western Germany (13-14 July) and the hailstorms in southern Germany in the last third of June, including a squall line with widespread severe winds on 29 June. Daily average values of the convective adjustment timescale, averaged over Germany, vary between less than an hour to more than 5 hours (dots in Fig. 4.1). Most of the strongly forced days have large amounts of domain averaged accumulated precipitation. In contrast, weakly forced days typically feature smaller domain averaged precipitation sums, while its spatial distribution is highly variable (see Fig. 5.4). To partition out the days with the strongest and the weakest synoptic control we take the lowest 20% of average τ_c values and classify these as strong forcing, while the highest 20% of daily values are considered as weak forcing. An advantage of this approach over setting certain fixed thresholds is the creation of equally populated samples containing 16 days for each regime.

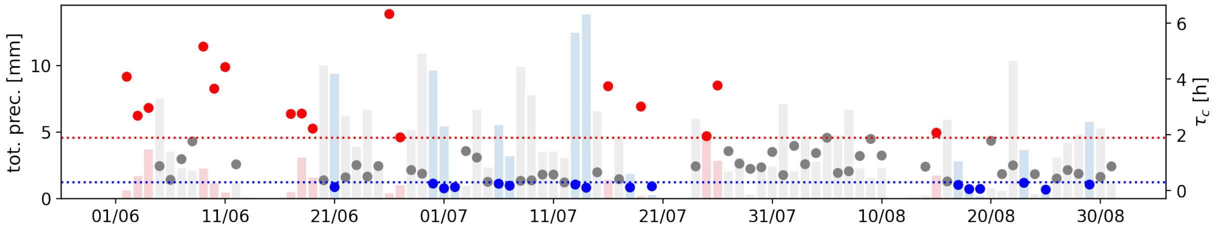

Figure 4.1: Time series of daily total precipitation (bars) and daily averaged convective adjustment timescale (dots), both for the reference experiment, averaged over Germany, for summer 2021. Red indicates weakly forced days, blue strongly forced days. The horizontal dotted lines show the threshold value of the convective adjustment timescale for weak forcing (red) and strong forcing (blue).

4.2.2 Diurnal cycle of precipitation

One of the key challenges in convective scale weather prediction is an accurate forecast of precipitation amount, timing, and uncertainty. This holds especially true in the absence of larger-scale forcing, when local processes in the boundary layer drive the convection. Figure 4.2 shows composite time series of domain-averaged precipitation amounts and its variability for both regimes based on 16 strongly and 16 weakly forced days. The PSP scheme has an overall higher impact during weak forcing, both in terms of average precipitation amounts and ensemble spread of precipitation. During weak forcing days the diurnal cycle is clearly evident and peaks in the afternoon (around 15 UTC). The PSP scheme shifts the maximum about an hour earlier due to more efficient triggering of convection, caused by buoyant air bubbles in the boundary layer, created by the PSP scheme. The onset of perturbations is directly connected to the subgrid standard deviation of selected variables, which increases as the solar radiation heats the surface. Hence, convection is formed earlier than in the absence of the PSP scheme, which is one of the goals of the scheme. For a more detailed discussion about the role of PSP in triggering mechanisms, the reader is referred to Hirt et al. (2019).

The earlier shift of the diurnal cycle of precipitation is beneficial since precipitation in convective-scale models usually lags the observed precipitation maximum in these flow conditions (e.g. Keil et al., 2019). Moreover, the magnitude of the peak is higher, especially in spread. Thus, physically-based perturbations in the boundary layer lead to a reduction of the underdispersion of precipitation (not shown), which is a general issue of convection-permitting ensemble prediction systems. This includes the ICON-D2 EPS as well, according to spread-skill ratios mostly below 0.5 in the DWD operational verification based on rain-gauge data.

On strong forcing days, there is no clear diurnal cycle in precipitation amount and the PSP perturbations have little effect. While the average amount of precipitation is higher in strong forcing, the magnitude of the spread is higher in weak forcing. These results are consistent with those of earlier case studies using the PSP scheme in the COSMO model (Kober and Craig, 2016; Keil et al., 2019; Hirt et al., 2019).

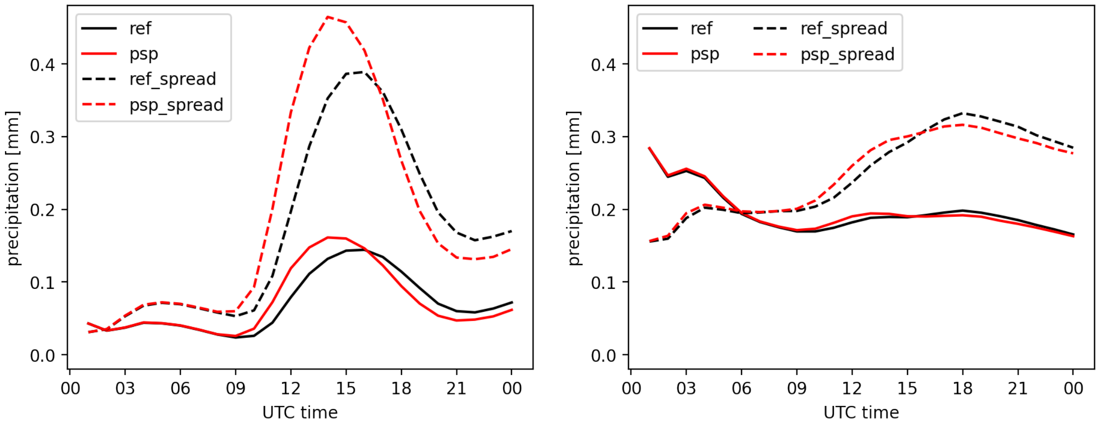

Figure 4.2: Composite time series of hourly precipitation amount (continuous lines) and spread (dashed lines) for weak forcing (left) and strong forcing days (right), for the reference experiment (black) and the PSP experiment (red), averaged over Germany.

4.2.3 Spatial uncertainty of precipitation

To assess the predictive skill of the precipitation forecasts we apply a spatial verification method to account for the spatiotemporal highly variable nature of precipitation. The widely-used Fractions Skill Score (FSS) directly compares the fractional coverage of the events in windows surrounding the observations and forecasts (Roberts and Lean, 2008).

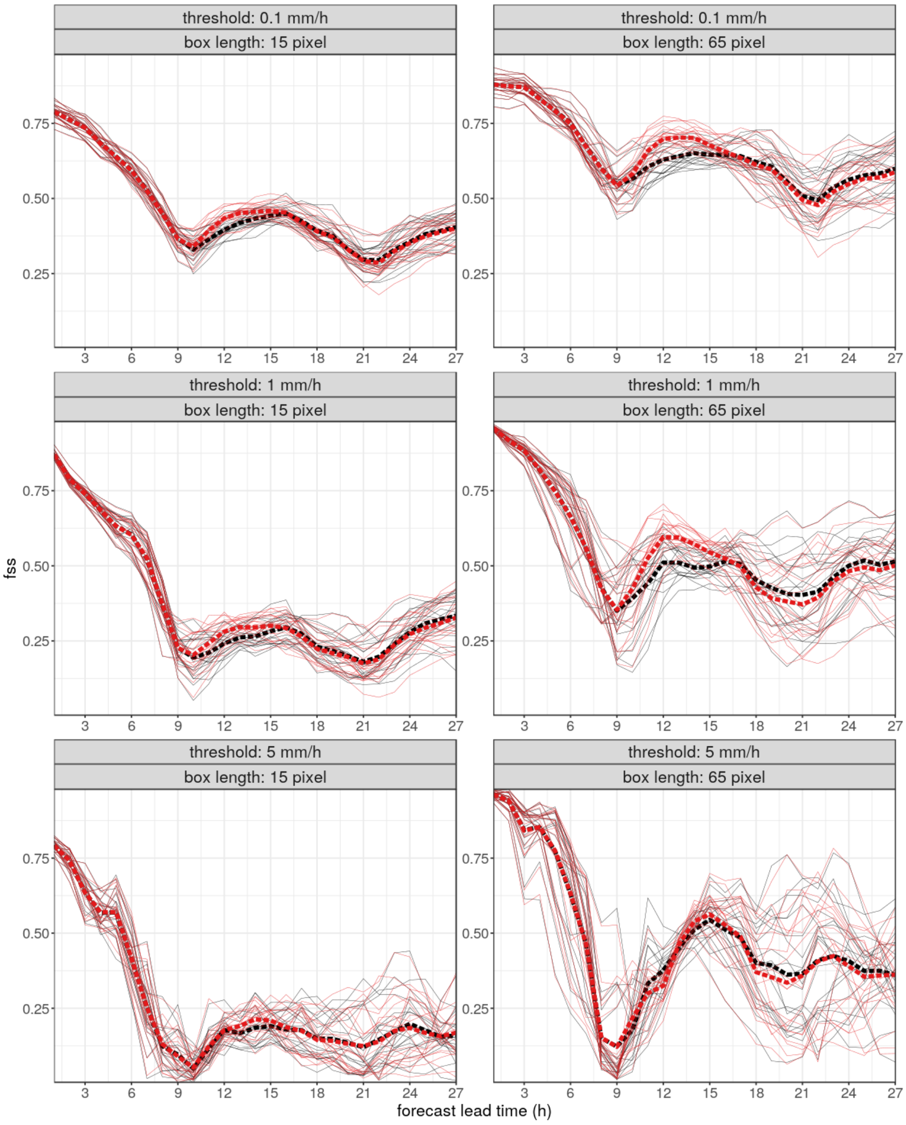

Figure 4.3: Spatial forecast skill as measured by the FSS for each ensemble member (thin lines) of the two ensembles (PSP in red and reference in black) as a function of the lead time averaged over the 16 days with weakly forced convection. The mean over the members is plotted as thick dashed line. The scores are shown for the exceedance of 0.1 mm/h (top), 1 mm/h (middle) and 5 mm/h (bottom) and for two aggregation window sizes with dimension of 15 pixels (about 30km, left) and 65 pixels (about 140km, right)

Observations are provided by the quality-controlled precipitation field estimated by the German radar network on a 1km grid every 5 minutes. The observed data is upscaled to the ICON-D2 grid and accumulated to hourly values for comparison with the model output.

In Figure 4.3, the FSS is shown for each ensemble member (thin lines) of both ensembles

(PSP experiment in red and reference in black) as a function of the lead time on days classified as weakly-forced according to the convective adjustment time scale (see Fig. 4.1). The scores are shown for the exceedance of three selected hourly precipitation thresholds (0.1 mm/h, 1 mm/h and 5 mm/h, respectively upper, middle and lower row) and for two aggregation window sizes, with dimension of 15 pixels (left column), corresponding to about 30 km, and 65 pixels (right column), corresponding to about 140 km. For easiness of reading, the average value of the FSS of the members is also plotted, as thick dashed line, for both the PSP experiment and the reference one.

Generally, the FSS is higher at short forecast lead times and decreases over time. However, after convective initiation and the generation of precipitation from 9 UTC onwards, the FSS increases and attains higher FSS values in the central part of the day, between 9 and 18 UTC, when the maximum of convection occurs. A relative minimum is observed around 21 UTC. At the smaller aggregation scale (left column), the lines related to individual members of both experiments tend to stay close together, showing similar performance of the two experiments. Between 12 and 15 UTC, the period of most active convection, the mean score shows a slightly better performance for the PSP experiment. Both ensembles become more disperse at longer forecast lead times with a higher precipitation threshold, as shown by the larger difference in FSS between the members.

When the larger aggregation window is considered (right column), the difference in performance of the two ensembles becomes more marked. The individual ensembles are more disperse, and the PSP experiment outperforms the reference one, as shown by the larger mean FSS value, in particular for the 0.1 and 1 mm/h thresholds. The difference is clearly evident during strong convection between 12 and 15 UTC. Interestingly, the FSS of the 5 mm/h threshold is slightly higher than of the 1 mm/h threshold at peak precipitation at 15 UTC. This is presumably caused by the sample size and averaging effects. After 18 UTC the FSS of the PSP experiment is slightly lower than that of the reference run in the last part of the day, in agreement with the behavior shown when representing the diurnal cycle of the spread (see Fig. 4.2). During strong forcing, the time series of the FSS do not show a significant impact of the PSP scheme and predominantly show a steady decrease with lead time (not shown).

4.2.4 Probabilistic verification of near-surface variables

An objective verification of the performance of both experiments has been carried out for a wide range of meteorological variables using a standard set of indices for the whole trial period. The results are shown in Figure 4.4 for a selection of variables: wind speed at 10m above the ground (first column), cloud cover for low clouds (second column), temperature and dew-point temperature at 2m above the ground (third and fourth column, respectively). The scores are computed against observations at the SYNOP stations over Germany. Due to the representativity error of these observations we exclude the verification of precipitation in this section. In the first row, the Continuous Ranked Probability Score (CRPS) is shown as a measure of quality of the ensemble forecasts (negatively oriented), in the second row the Root Mean Square Error (RMSE) of the ensemble mean, in the

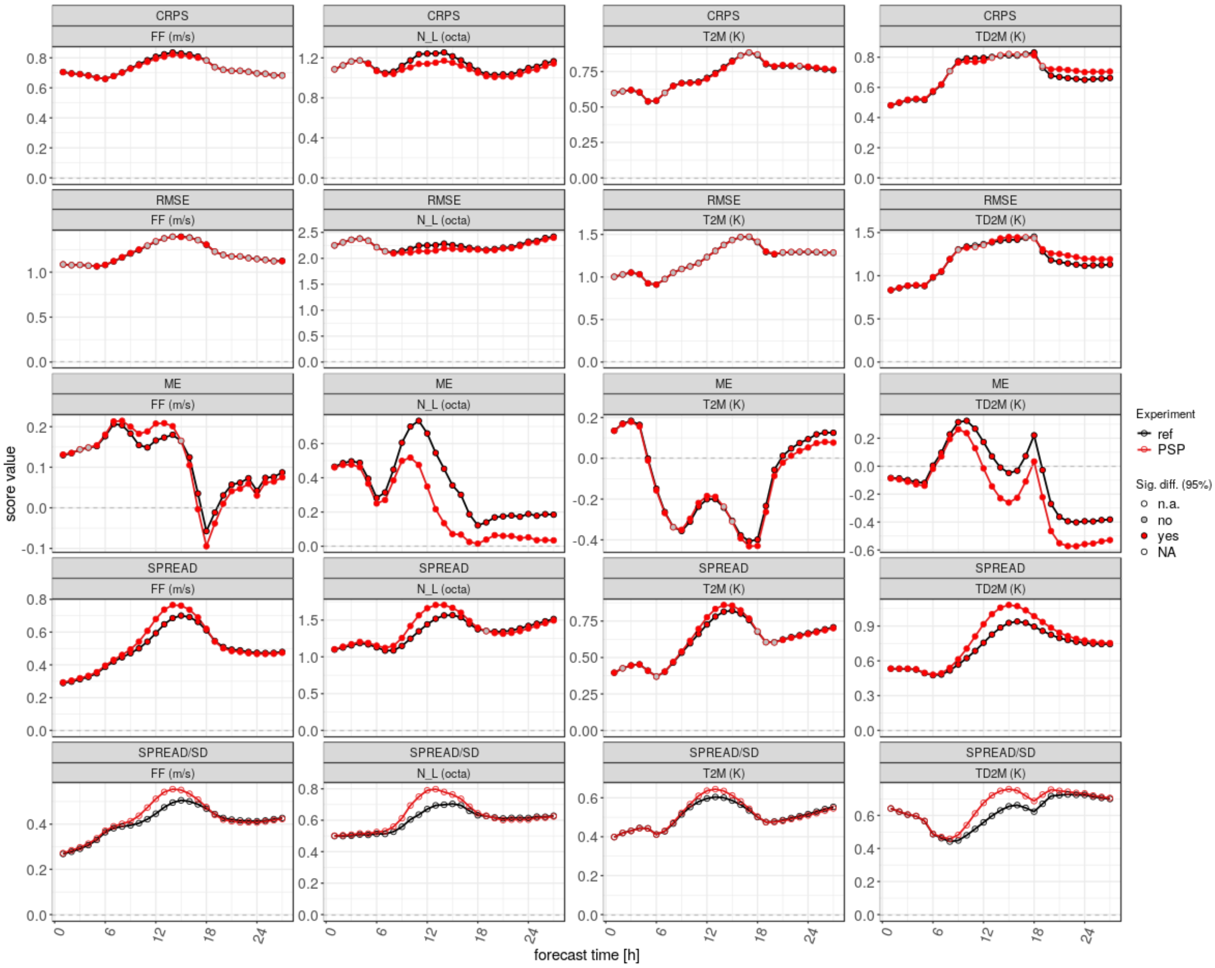

Figure 4.4: Diurnal cycle of domain averaged CRPS, RMSE, mean error, ensemble spread and the spread to skill relationship (see text for details) for 10m wind (FF), low cloud cover (NL), 2m temperature (T2M) and 2m dewpoint temperature (TD2M) of the reference (black) and PSP experiments (red), verified against SYNOP observations, for the period between 26 May and 31 August 2021.

third row the mean error of the ensemble mean (ME), and in the fourth row the ensemble standard deviation (SPREAD). Finally, in the fifth row the spread to skill relationship is shown, expressed as the ratio between the spread and the standard deviation (SD) of the error of the ensemble mean. This measure has been chosen because the ensemble spread should match the random component of the RMSE of the ensemble mean, having subtracted the bias, so that values less than one indicate underdispersion. For a description of the indices, the reader is referred to Wilks (2019). In each plot the red color is for the PSP experiment, while the black color is for the reference experiment. The dots are filled with color when the difference between the scores of the two experiments is significant, as computed following a bootstrap method. No significance estimation has been performed for the spread to skill relationship.

Probabilistic verification scores over the whole period are improved for a wide range of variables when using the PSP scheme, especially the spread and spread to skill relation. The CRPS is also slightly improved for the PSP experiment, compared to the reference, with the exception of the 2m dew-point temperature between 18 and 24 UTC. The 2m dew-point temperature scores are also slightly deteriorated in terms of RMSE in this part of the day, while for other variables the RMSE is either smaller (low cloud cover) or not significantly different for the PSP experiment, compared with the reference experiment. The combination of larger spread and equal to smaller RMSE leads to an increased, beneficial spread/skill relation. This is a particularly positive result, given the general issue of models being underdispersive for near-surface variables.

However, a more detailed look at the verification scores shows some issues, specifically for 2m dew-point temperature and low cloud cover. The mean error of the 2m dew-point temperature points to a marked drying effect in the considered period. Moreover, the (at first sight) beneficial behavior of the PSP scheme causing a reduction of the mean error of low clouds turns out to be mainly caused by increased patchiness of the low cloud field on non-rainy days, resulting in a reduced mean error. Visual inspection indicates that the increased patchiness is unrealistic (not shown). Preliminary inspection of a few case studies with low cloud cover, but no rain, indicates that the current vertical profile of the perturbations in PSP entrains too much dry air from aloft into the boundary layer causing a dry bias in the boundary layer (see mean error of the 2m dew-point temperature, too). A modification of the vertical profile at the upper boundary of the PBL determining the perturbation strength of PSP leads to improved results for a non-rainy and a rainy case study and will be pursued in future PSP applications (a systematic investigation is beyond the scope of the present trial). For medium and high clouds, the impact of PSP is almost neutral (not shown).

The PSP scheme also shows an increase of the mean error of 10m wind speed, which was detected also in the verification of wind gusts (not shown). This is likely caused by a double counting of turbulence effects that are taken into account when diagnosing near-surface wind speed. We should point out that the verification in Fig. 4.4 was performed against SYNOP observations, a traditionally used observation type for near-surface variables that has limitations when estimating forecast errors of cloud cover and wind gusts. Therefore, a different kind of observations should be used in future to improve the diagnostics, including e.g. satellite observations to directly compare visible reflectances (using a satellite forward operator, e.g., Scheck et al., 2020).

4.3 Conclusions

Increasing resolution of NWP models makes traditional parameterizations of boundary layer turbulence inadequate, because the assumption that the gridbox size is much larger than the size of eddies does not hold anymore in kilometer-scale models. In this chapter, we use the recently implemented physically-based stochastic perturbation scheme PSP in

ICON-D2-EPS as a representation of model error originating from the subgrid scale in the boundary layer, but affecting the smallest resolved scales. The experimental period spans a whole summer season, which allows for a systematic analysis of the impact of the scheme in different synoptic forcing conditions. The main conclusions of this work are the following.

1. The PSP scheme provides a good representation of the effect of subgrid scale turbulence in ICON-D2 and has realistic, beneficial effects in ensemble forecasts, especially on the ensemble spread. It helps triggering convection, while preserving the intensity of single convective cells and does not produce spurious convection. PSP reduces the underdispersion of precipitation.

2. Small-scale perturbations, introduced by PSP, have a larger impact on convective precipitation in weak than in strong synoptic forcing, especially on its spread. This is in line with the hypothesis of local processes in the boundary layer driving the convection on weakly forced days, whereas on strong forcing days, the synoptic pattern controls the convection.

3. The PSP scheme slightly improves the spatial distribution of precipitation (FSS) around the peak of its diurnal cycle in weak synoptic forcing, compared to radar observations. Its impact is neutral during strong forcing.

4. The probabilistic verification of near-surface variables predominantly shows a neutral to slightly beneficial forecast performance. The systematic assessment indicates few issues that deserve further research (namely the 2m dewpoint temperature and wind gusts at the surface) on the way towards operational implementation of PSP in ICON-D2-EPS.

A general issue with physically-based schemes are interactions between different schemes and the double-counting of physical processes. Further work will therefore examine the effects of combining PSP with the stochastic shallow convection scheme developed by Sakradzija and Klocke (2018) in ICON. The two schemes should act independently by design, although both would likely affect the layers around the top of the PBL, where we found a detrimental impact of PSP in its current implementation.

The results of this work are encouraging: PSP improves the spread to skill ratio of the ensemble for several variables, especially those near the surface, for which the forecast is often underdispersive. This is a promising step on the way to operational use of the scheme and in general for the development of physically-based stochastic schemes. Limitations of certain observation types and the deterioration of the forecasts for few variables provide a basis for further research.

Chapter 5

Flow dependence of forecast uncertainty in a large ensemble

5.1 Background

The atmosphere is by nature a chaotic system and there will always be a certain degree of uncertainty in the prediction of its future state (Lorenz, 1969; Selz et al., 2022). Moreover, a variety of errors occur in the weather forecasting process, from inaccurate observations to imperfect models, which further limit the predictability (Buizza, 2021a). To account for all these sources of uncertainty in the forecast, ensemble prediction systems (EPS) are used. They offer a useful probabilistic forecasting method that allows for a probability to be attached to the meteorological prediction (Leutbecher and Palmer, 2008; Buizza, 2021b). Typically, the different sources of uncertainty in a limited-area ensemble include perturbed initial and boundary conditions, as well as some representation of model error, due to incomplete description of physical processes and an insufficient representation of the subgrid-scale variability in numerical weather prediction (NWP) models (Clark et al., 2016). The errors arising from these sources quickly grow and propagate, due to highly nonlinear processes at these scales (Hohenegger and Schär, 2007). Therefore, there have been several endeavors to incorporate representations of these processes, such as boundary layer turbulence and convective initiation, into kilometer-scale models (e.g. Leoncini et al., 2010; Kober and Craig, 2016; Hirt et al., 2019; Clark et al., 2021).

The relative importance of different sources of uncertainty depends on the weather regime. Initial and boundary conditions are the dominant source of uncertainty when the synoptic forcing of convection is strong. On the other hand, local uncertainty sources have a significant impact on the amount and spread of precipitation in the ICON-D2 EPS ensemble when synoptic forcing of convection is weak and triggering of storms is driven by small-scale turbulence (Keil et al., 2014; Puh et al., 2023). Furthermore, the spatial predictability of precipitation strongly depends on the prevailing convective forcing regime. During weak forcing, spatial error and spread largely depend on the diurnal cycle of precipitation and are more affected by introducing stochastic perturbations in the boundary layer compared with the strong forcing regime, when these perturbations have little effect (Matsunobu et al., 2024).

Another important factor that determines how well the ensemble captures the sources of uncertainty is the size of the ensemble, that is the number of ensemble members that construct the probabilistic forecast and define a probability distribution of a variable. Although most operational data assimilation systems assume and produce a Gaussian error distribution, its shape is actually multi-modal or shows other kinds of non-Gaussianity. These complex shapes arise due to the nonlinear evolution of the atmosphere and large ensembles are needed to capture all their complexity (Leutbecher, 2019). Operational ensemble prediction systems typically have a few tens of members for global models (e.g. 51 members in the ECMWF medium-range EPS), while the ensemble size for limited area models is even smaller (e.g. 20 members in the ICON-D2 EPS). This is not enough to accurately capture infrequent, extreme events (e.g. Leutbecher, 2019). Kondo and Miyoshi (2019) showed that up to 1000 ensemble members are needed to represent characteristics of non-Gaussian distributions. Craig et al. (2022) explored the errors in numerical weather forecasts resulting from limited ensemble size using 1,000-member forecasts of convective weather over Germany at 3-km resolution. They examined sampling error and found that an asymptotic convergence behavior was observed for most distribution properties. Although this convergence was reached with as few as 10 ensemble members for the mean of forecast variables, sizes of up to 100 were required for the convergence law to apply for the standard deviation and there was no clear sign of convergence for the 95^{th} percentile even with 1,000 members.

However, the computational cost of large NWP ensembles is very high. Tempest et al. (2023) studied how ensemble size affects the uncertainty in ensemble forecasts using an idealized, computationally efficient model, which replicates the properties of cumulus convection, allowing ensemble sizes of up 100,000 members. It was found that for all computed distribution properties, including mean, variance, skewness, kurtosis, and several quantiles, the sampling uncertainty scaled as $\mathrm{n}^{-1/2}$ for sufficiently large ensemble size n. The Central Limit Theorem predicts that the uncertainty in a determined statistic is influenced by the distribution shape and is greater for those relying on rare events. This expected behavior was confirmed, along with finding that larger ensemble sizes are needed for such statistics to enter the asymptotic regime. Through evaluating asymptotic behavior in small ensembles, it was shown that the asymptotic theory can be applicable to certain forecast quantities even for the currently used operational ensemble sizes.

The idealized model was then expanded by Tempest et al. (2024) to include weak and strong forcing convective weather regimes in order to examine differences in sampling uncertainty convergence for each regime. Differences in distribution shape between the weak and strong forcing regimes affected the Convergence Measure, leading to significant disparities. Notably, substantial spread differences between weak and strong forcing runs over 24 hours resulted in considerable variations in the sampling uncertainty of mean and standard deviation. In extreme statistics like the 95^{th} percentile and cases with precipitation, moisture variables in weak forcing showed the highest sampling uncertainty, necessitating a greater number of members for convergence. This was due to the low density in the tails of weak forcing moisture variables. They concluded that different ensemble sizes are required depending on the convective weather regime.

This chapter builds on the previous work by Craig et al. (2022), Tempest et al. (2023) and Tempest et al. (2024), but with some important differences. Firstly, the fully-fledged ICON-D2 EPS model is used, while Tempest et al. (2024) used an idealized model, so their conclusions are not necessarily valid for a NWP model. Secondly, the flow dependence of forecast distributions is analyzed, which is new compared with Craig et al. (2022), who used a comparably complex model, but did not investigate the flow-dependence of their results. Moreover, the analysis of uncertainty convergence is expanded to surface variables, which are compared to mid-troposphere variables in a flow-dependent framework. Our hypothesis is that the evolution of surface variables is more influenced by convection than that of the mid-tropospheric variables in the weak forcing regime. In the strong forcing regime, we expect this influence to be comparable between the surface and the middle troposphere. Finally, novel maps of uncertainty allow for a more detailed analysis of its spatial pattern and an easier interpretation of the sources and evolution of forecast uncertainty in different convective forcing conditions and for different variables.

This chapter is structured as follows. Section 2 introduces the used NWP model, the experimental setup and summarizes some of the used methods introduced by Craig et al. (2022). Section 3 presents the results of the analyzes including the spatial distribution of uncertainty, focusing on the differences between a weak and a strong synoptic forcing case, the convergence of sampling uncertainty and its time evolution as represented by the evolving probability distributions of surface and mid-troposphere variables. Conclusions are drawn in section 4.

5.2 Results and discussion

5.2.1 Convergence of uncertainty with ensemble size

If a statistic follows the Central Limit Theorem, its sampling uncertainty scales as $\mathrm{n}^{-1/2}$ for sufficiently large n, where n is the size of the sample. This is expected in a large ensemble from which smaller samples are selected, as documented by Craig et al. (2022) and Tempest et al. (2023). In this section, we apply this method on an extended dataset, which includes more variables and is produced by an operational NWP model.

To analyze how sampling uncertainty decreases with ensemble size, we focus on a location near Reutlingen. Figures 5.1 (weak forcing regime) and 5.2 (strong forcing regime) show the convergence of uncertainty for temperature at 2 m and at 500 hPa for one gridpoint and a 10-kilometer neighborhood, for the mean, the standard deviation and the 95^{th} percentile. Figures for all analyzed variables can be found in Appendix A.

For most quantities, the width of the confidence interval follows the power law at large ensemble sizes, but the number of ensemble members needed to reach this asymptotic regime is the lower for the mean (less than 5 - see Figures 5.1a,d and 5.2a,d) than for the standard deviation (around 20 - see Figures 5.1b,e and 5.2b,e) and even higher for the

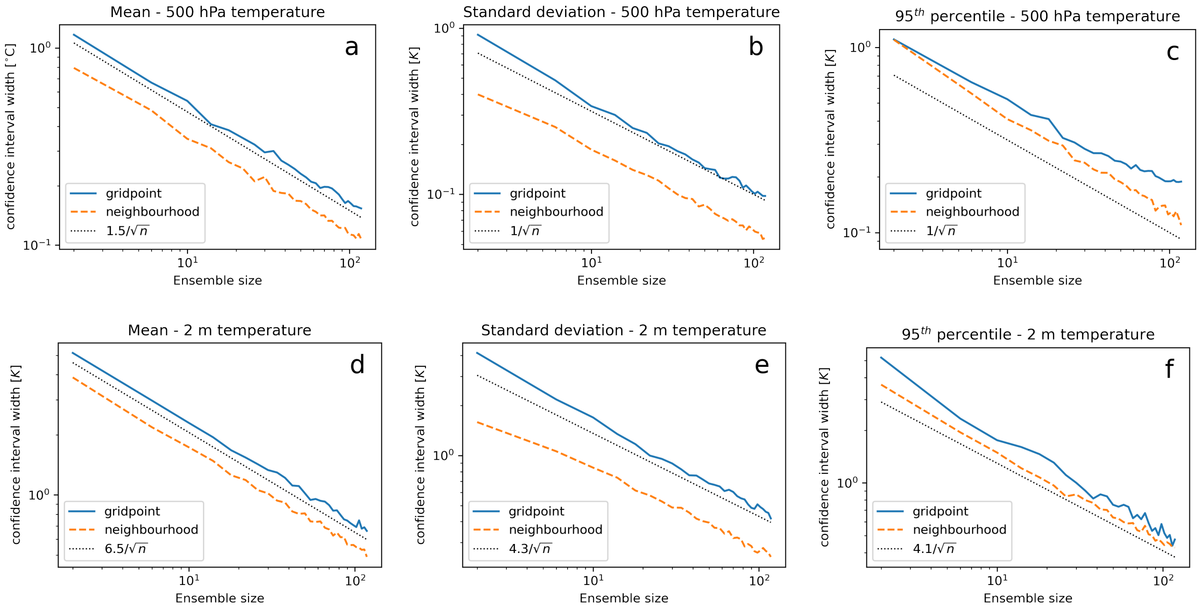

Figure 5.1: Width of the 95% confidence intervals for the mean, standard deviation and 95^{th} percentile (columns), for temperature at 500 hPa (a, b, c) and at 2 m (d, e, f). Forecast quantities are computed for 10 June (weak forcing) at 14 UTC for a gridpoint near Reutlingen. The dashed line shows the reference curve with slope $N^{-1/2}$, fitted by eye. Gridpoint: single gridpoint forecast, neighborhood: 10-km radius neighborhood forecast.

95^{th} percentile (more than 120 - see Figures 5.1c,f and 5.2c,f). This is in agreement with previous studies (e.g. Craig et al., 2022) and confirms the universality of this behavior. The width of the confidence interval is generally smaller for the neighborhood, since its effective ensemble size is three orders of magnitude larger, which is the purpose of the neighborhood method. Furthermore, convergence is reached with a smaller ensemble size compared to the same variable at a single gridpoint (as in Tempest et al., 2024).

The width of the confidence interval in weak forcing is always smaller for the neighborhood distribution, compared to the gridpoint distribution. In strong forcing, this difference is smaller, especially for surface variables. This is likely a consequence of the larger influence of synoptic scale systems in strong forcing, namely the cold front, which increases variability over a larger area, compared to the local effects of single convective cells, which are increasingly negligible as the size of the neighborhood increases. Craig et al. (2022) looked at this effect more in detail and drew a similar conclusion.

While the convergence law holds for most variables, there are a few exceptions. For instance, the convergence for the 95^{th} percentile for a single gridpoint is either not clear or the uncertainty is evidently not converging, especially in strong forcing, e.g. the 95^{th} percentile of temperature at 500 hPa, for which the slope of apparent convergence is flatter (Figure 5.2c). The reason for this exception is the long tail in the distribution (see Figure 5.7, top right panel), which affects the 95^{th} percentile more than the mean or the standard

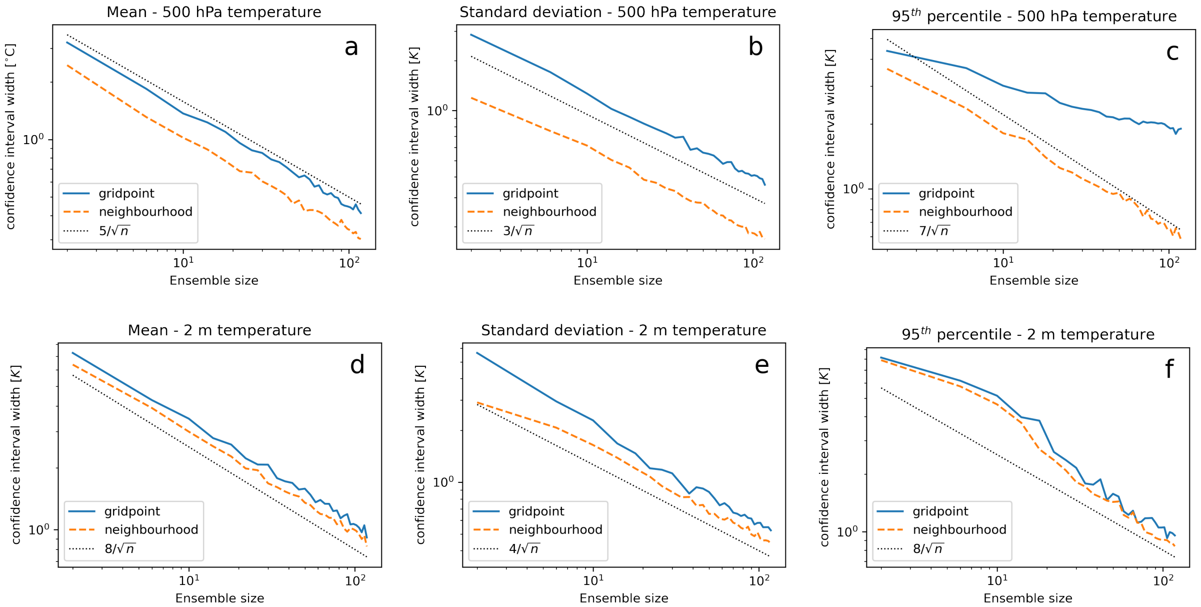

Figure 5.2: As in fig. 5.1, but for 29 June (strong forcing).

deviation. The reason for the skewed distribution could be latent heat release, which is produced in ensemble members with exceptionally strong convection, thereby increasing the temperature at 500 hPa, or advection of very warm air in the warm sector before the cold front. A similar behavior is observed for the zonal wind. In the case of the zonal wind at 500 hPa, convergence is not reached not only for the 95^{th} percentile, but also for the mean in both forcing regimes. In this case the bulk of the distribution is the cause, rather than the tails (Figure 5.7, bottom row).

An unusual behavior is also found in uncertainty convergence for the 95^{th} percentile of relative humidity in the strong forcing case (Figure 5.3 top left panel). There is a large uncertainty for small ensemble sizes that suddenly decreases and flattens again, converging as $\text{n}^{-1/2}$. This happens more abruptly and at a larger ensemble size for one gridpoint than for the neighborhood in ICON data. The cause for this "two-phase" behavior is the bimodal shape of the distribution (Figure 5.3 bottom left panel) with one large peak and one smaller peak, which strongly affects the value of the 95^{th} percentile. With an insufficient number of samples, this peak might be completely missed, leading to a different value of the 95^{th} percentile. However, once both peaks are sufficiently represented in the majority of the samples, the uncertainty of the 95^{th} percentile sharply decreases.

A similar performance is observed for the height distribution in the idealized model used by Tempest et al. (2024)(Figure 5.3 top right panel). Although the distributions are not of the same shape, since one has the larger peak at small values and one at large values, we can conclude that such a bimodal distribution shape requires more ensemble members to reach convergence of uncertainty, which agrees with previous work (Craig et al., 2022; Tempest et al., 2023). The significant difference between the curve for a single gridpoint in idealized model panel and the ICON panel (blue lines in Figure 5.3) is hard to interpret, but it should be pointed out that the two models have a substantially different structure, one being a 1-dimensional idealized model and the other being a fully fledged NWP model. Although this does not allow for a direct comparison, the qualitative behavior of the idealized model was found to be sufficiently realistic by Tempest et al. (2024).

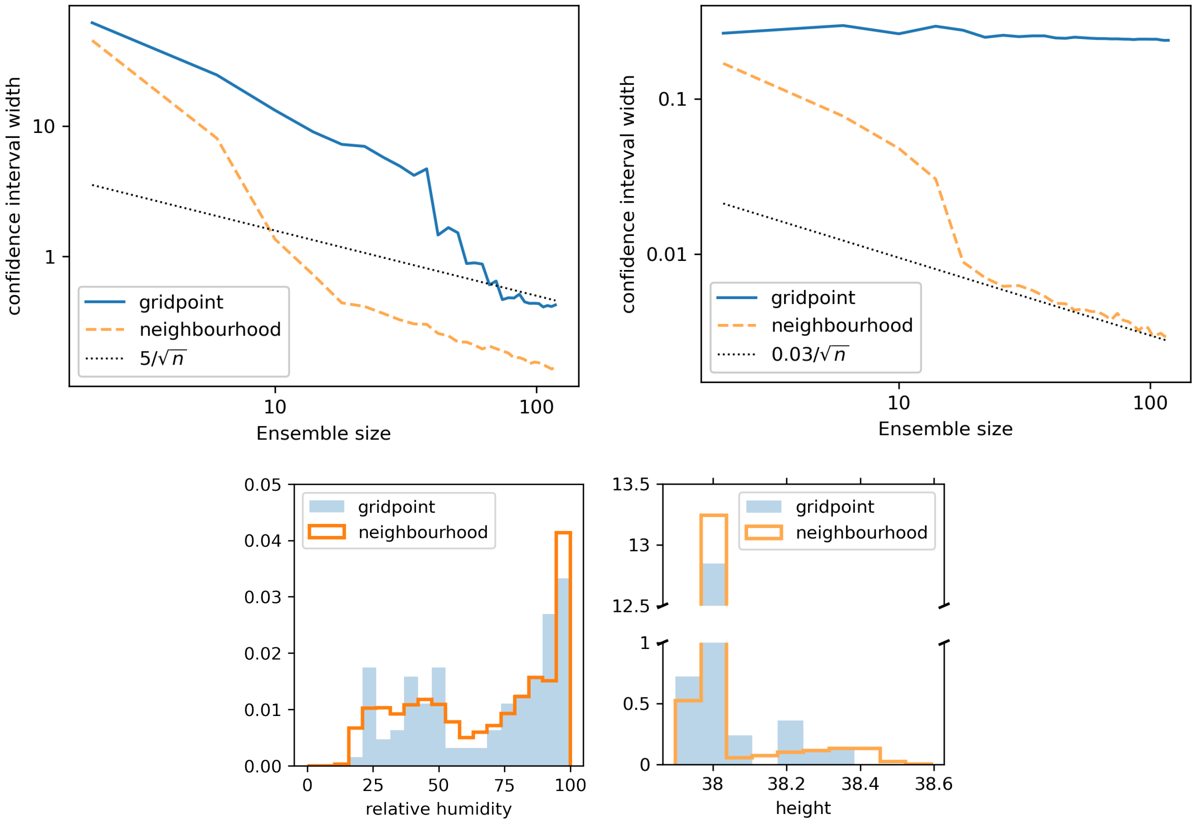

Figure 5.3: (Top row) Width of the 95% confidence intervals for the 95^{th} percentile of relative humidity of the ICON ensemble (left) and height in the 1-D idealized model by Tempest et al. (2024) (right). (Bottom row) histogram of relative humidity at 500 hPa (left, ICON ensemble) for the gridpoint and the 10-km neighborhood and height (idealized model, right) used to calculate the confidence interval plotted in the top row.

5.2.2 Maps of uncertainty

Convection increases the uncertainty of most forecast variables, since it is one of the most uncertain and unpredictable phenomena at small scales. To visualize the impact of convection on the spatial distribution of forecast uncertainty, we use the bootstrapping approach to create maps of uncertainty for several variables and statistics: the mean, standard deviation and 95^{th} percentile. Due to computational resources, uncertainty is calculated for a 10 km neighborhood around every second gridpoint for a 14h lead time in the two case studies, using 120-member samples.

We first consider the spatial pattern of uncertainty for the mean of 2m temperature in Figure 5.4 as an example before going on to examine other variables. Histograms of 2m temperature at indicated locations are included as a reference for the interpretation of the uncertainty map. The correlation between precipitation (i.e. convection) and 2m temperature uncertainty is clearly visible in the top panels. This is expected, since precipitation lowers the surface temperature in summer and therefore a fraction of ensemble members with precipitation will make the mean of surface temperature more uncertain.

In the weak forcing case, the pattern of uncertainty is patchy and the impact of individual convective cells is visible. At location number 2, most ensemble members predict convection, but slightly displaced, which leads to a skewed distribution with a longer tail towards lower temperatures for the strongest occurrences of convective storms. On the other hand, the uncertainty is smaller at location 3, because most members predict a clear sky and the distribution is narrower and unimodal. One exception to this framework is location 1, where uncertainty is relatively large, despite the absence of convection and precipitation. Further examination shows that this is again a consequence of an underlying bimodal distribution, in this case in cloud cover: most members predict clear skies, but a considerable number of them has low clouds over the North Sea (not shown), which is represented by the peak at lower temperatures. This makes the sampling uncertainty of the mean of 2 m temperature larger, as Tempest et al. (2023) showed more generally for bimodal distributions.

In the strong forcing case, areas of higher uncertainty are concentrated around MCSs, like the squall line in SW Germany (location 5) and other areas of widespread convection (Austria, NW Italy). This is mainly a consequence of the uncertain location of the individual MCS, since the majority of members predict it, but slightly displaced, which leads to a high uncertainty. In the case of the squall line, its uncertain timing and positioning at location 5 is clearly visible in the histogram, with a peak at relatively low temperatures and a smaller secondary peak at higher temperatures. The source of bimodality in this case is not the separation between cloudy and clear sky gridpoints, since the vast majority of data points are associated with cloudy skies. On the other hand, before the arrival of the front at location 7, the distribution is very narrow, indicating that the vast majority members agree that the squall line has not yet arrived. This is also the case where precipitation is not directly caused by convection, like at location number 4, while at location 6 the uncertainty is larger due to a more convective nature of precipitation, although the ensemble mean precipitation is about the same (between 2 and 5 mm).

Uncertainty maps for the mean, standard deviation and 95^{th} percentile of dew-point temperature at 2m and relative humidity at 500 hPa are shown in Figure 5.5. These are compared for the two forcing regimes and provide an example of surface and mid-tropospheric variable. Figures with uncertainty maps of other analyzed variables can be found in the Appendix, since they do not provide significant additional insight.

Firstly, the largest uncertainty of the mean is not necessarily at the same location as the largest uncertainty of other statistics. For example, one of the regions with larger

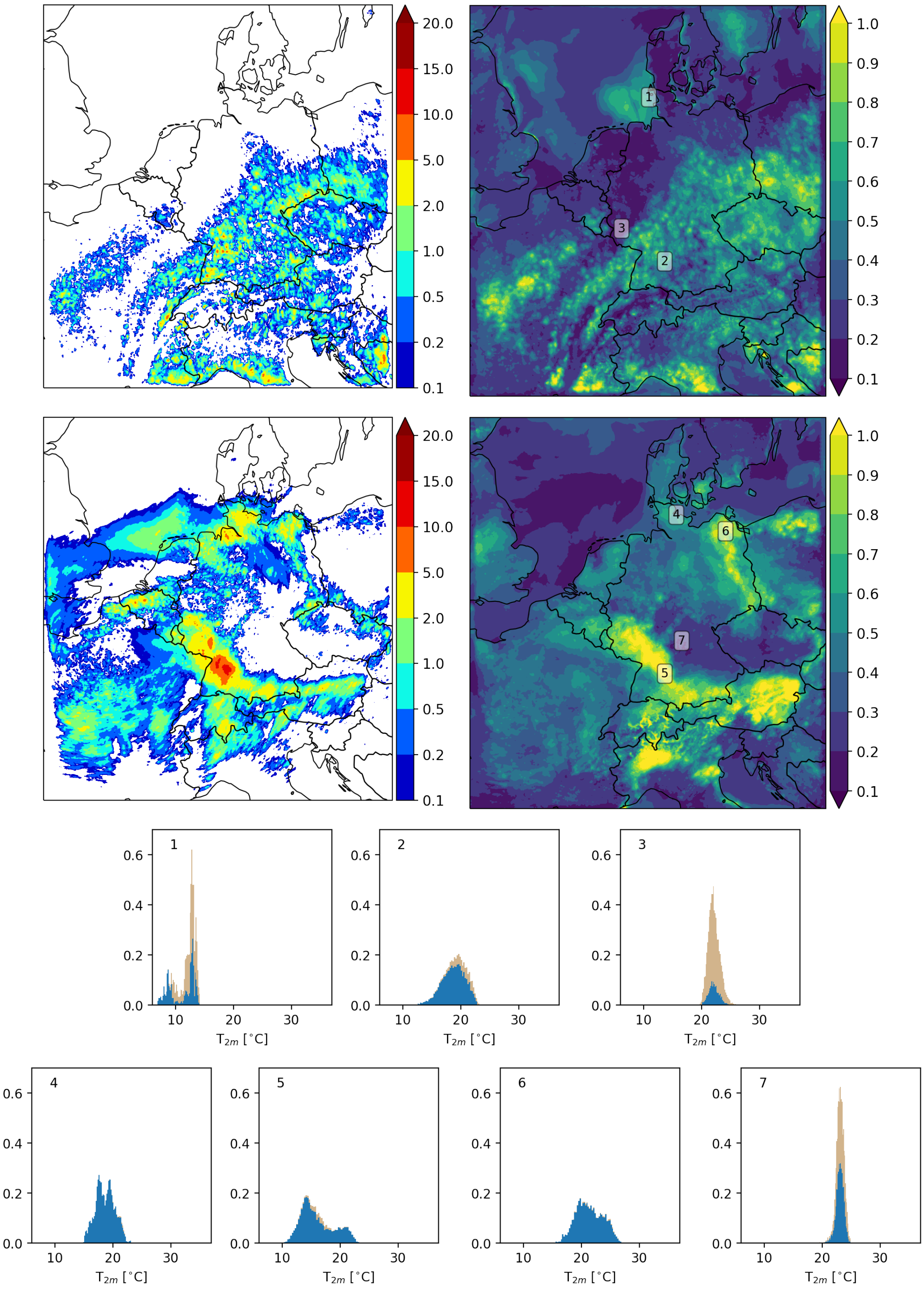

Figure 5.4: (Top left): ensemble mean of hourly precipitation in millimeters for a weak forcing (10 June 2021 at 14 UTC, top row) and a strong forcing case (29 June 2021 at 14 UTC, bottom row). (Top right): width of the 95% confidence interval for the mean of 2m temperature in Kelvin for the respective cases. (Bottom): histograms of 2m temperature for a 10km neighborhood around selected gridpoints, numbered on the uncertainty maps. Blue indicates cloudy gridpoints, brown cloud-free gridpoints, separated by the threshold of 95% cloud cover.

uncertainty in the mean of relative humidity at 500 hPa (Figure 5.5b) is around Salzburg, while in the uncertainty map of the standard deviation (Figure 5.5f) the magnitude is below average compared with other regions, like eastern Germany. This can again be attributed to the shape of the underlying distribution (not shown), which is often bimodal for relative humidity, like in the case of location 1 in Figure 5.4, as discussed above. The uncertainty of the standard deviation is not heavily affected because it does not depend that much on the relative importance of one or the other peak of the bimodal distribution, but rather on its total width.

The uncertainty pattern of the 95^{th} percentile is less coherent and shows a much larger sampling uncertainty, since 120 members are not enough to accurately estimate extreme values, as seen in the previous section. Nevertheless, the impact of convection is identifiable, both in weak and strong forcing and for all variables. It is interesting that convection not only impacts the uncertainty of surface variables, but also of mid tropospheric variables, even in weak forcing conditions. An example can be seen in Figure 5.5j, where the convective cells in southern Germany are recognizable in the uncertainty map of the 95^{th} percentile of relative humidity at 500 hPa.

A closer look at specific variables reveals a few interesting aspects. The strong gradients observed in the uncertainty pattern of relative humidity, especially for the 95^{th} percentile (Figure 5.5j and l), are caused by the typical source of bimodality for this variable: the cloudy versus clear sky separation. This has already been shown to play an important role in the context of forecast uncertainty by Craig et al. (2022). If ensemble members largely agree on one or the other side of the distribution of cloudiness, the uncertainty of relative humidity is low, because the underlying distribution is unimodal and narrow. As soon as a couple of members disagree, the uncertainty quickly grows, which is manifested in strong gradients in the uncertainty map.

The uncertainty of the statistics of dew-point temperature at 2m in the weak forcing regime (Figure 5.5a, e, i) is highest at the boundaries between regions with and without convection and is generally larger in regions without convection over land. A hypothesis is that soil moisture variability strongly influences the uncertainty of surface humidity, while at the boundary between wet and dry weather, the uncertainty is the highest due to the uncertain position of this boundary. Soil-atmosphere interactions are, however, beyond the scope of this work.

Additionally, to better quantify the influence of convection on the uncertainty of other variables, the correlation between uncertainty maps of different variables was computed for the weak and strong forcing regime (not shown). In the case of weak forcing, the correlation between uncertainty fields of different variables is strongest between precipitation and 2m temperature, as well as between precipitation and 10m wind, which confirms our hypothesis of a stronger influence of convection on surface variables in the weak forcing regime. For most other variable combinations it is stronger in the strong forcing case. Correlation is generally strongest for the uncertainty of the mean and weakest for the uncertainty of the 95^{th} percentile. This is a consequence of the patchier structure of the latter, due to its sensitivity to random sampling.

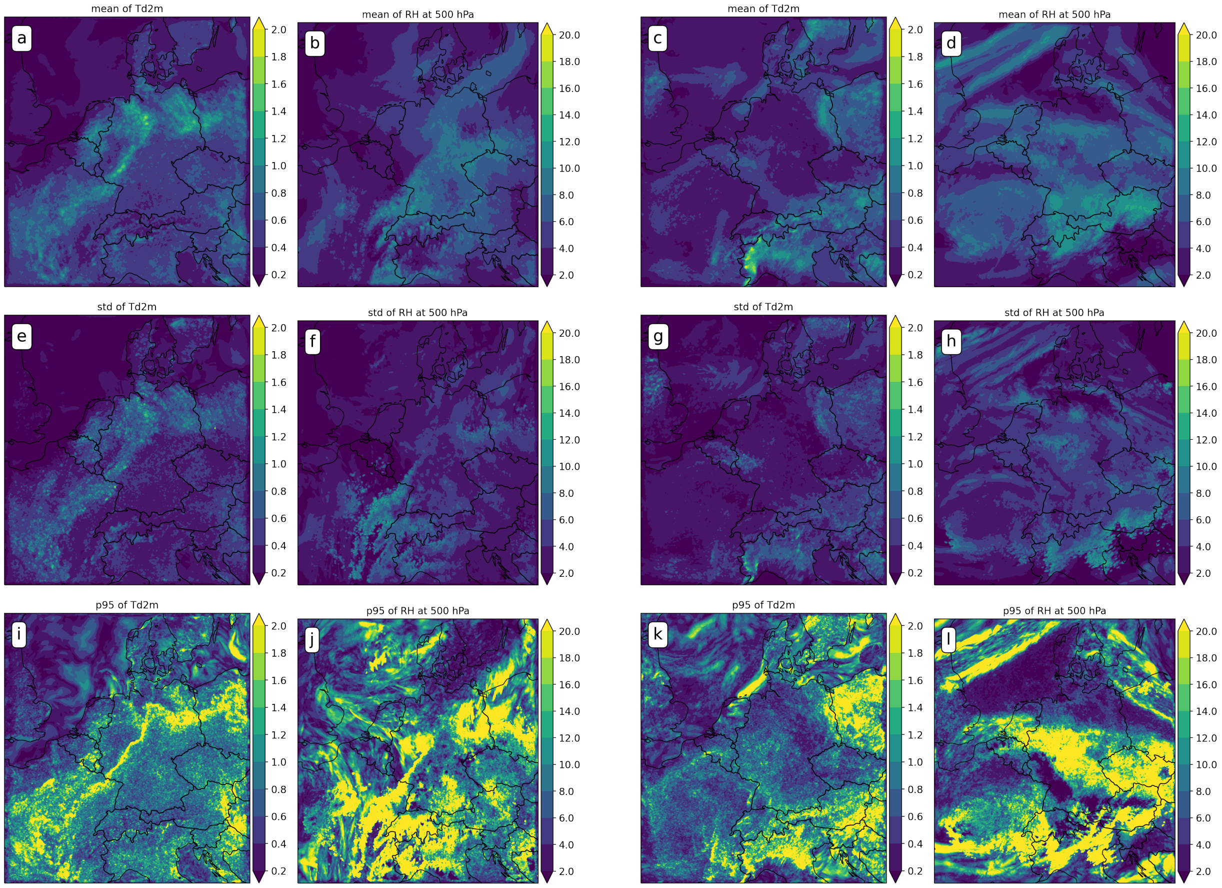

Figure 5.5: Width of the 95% confidence interval for the mean (top row), standard deviation (middle row) and 95th percentile (bottom row) of dew-point temperature at 2m in Kelvin (a, e, i and c, g, k) and relative humidity at 500 hPa (b, f, j and d, h, l) on 10 June 2021 at 14 UTC (left half of the panels) and 29 June 2021 at 14 UTC (right half of the panels).

5.2.3 Flow-dependent evolution of forecast uncertainty

Finally, the evolution of forecast distributions is investigated with particular attention to the dependence on the forcing regime of convection. This is expected to emerge from the direct impact of the diurnal cycle of convection and the frontal passage on the distributions of forecast variables.

Two-dimensional histograms of several variables are drawn for the location near Reutlingen for the two case studies, shown in Figures 5.6 and 5.7. They show similar properties as in Craig et al. (2022), which confirms the classification of variables in three categories: quasi-normal, highly skewed and multimodal. Compared to the dataset used in Craig et al. (2022), the ICON experiments allow for more custom output, which made the analysis of additional variables possible.

Surface variables can be classified in two categories, depending on the convective regime and the time. Apart from the precipitation distribution, which is always highly skewed, the distributions of other variables are quasi-normal or multimodal. In weak forcing, the dominant influence on the evolution of the distributions is the daily cycle of convection. In the first 10 hours of the forecast, when the atmosphere is mostly stable, the distributions are quasi-normal. With the onset of convection, the spread of temperature and dew-point temperature at 2m increases and the shape of the distributions is more complex, at certain times even bimodal, e.g. temperature at 12 UTC. As convection decays in the evening and the atmosphere stabilizes again, the distributions become narrower and the shape is again quasi-normal. This evolution is in line with that of the idealized model by Tempest et al. (2024).

The bimodality of the temperature distribution at midday cannot be explained by the cloudy and clear sky gridpoint discrimination, as initially hypothesized. An additional inspection of the neighborhood in question reveals that the bimodality is caused by orography inside the 10-kilometer neighborhood (not shown). Before the onset of larger convective cells, most gridpoints in the neighborhood are still cloud-free and the solar radiation heats the valley more than the surrounding hills, resulting in a temperature difference of a few degrees. This example shows that a neighborhood of 10 km is too large for 2 meter temperature in complex terrain.

Mid-troposphere temperature and wind are less influenced by the daily cycle in weak forcing. On the other hand, relative humidity closely follows the daily cycle of convection, although with a delay of around two hours, which corresponds to the time between the onset of convection and the time when significant moisture is transported by convection from the boundary layer to the middle troposphere. During this time, the distribution shape changes from quasi-normal to bimodal and even highly skewed when the peak is close to saturation. In the evening, the peak moves back to drier values, but a long tail of moister values remains.

In strong forcing, the main source of uncertainty is the timing of the cold front passage around 14 UTC with the associated squall line. This coincidentally matches with the time of the greatest extent of convection in the weak forcing case, at 14 UTC. In the histogram of precipitation, the higher values associated with the squall line have a significant impact on the distribution shape with a much denser tail for values larger than 10 mm. The peak in the first few hours of the forecast is still in the spin-up phase, so it is not of interest for our study. The temperature at 2 m and the wind at 10 meters have a clear step change in the distribution at the passage of the front, due to the abrupt change in temperature and wind speed and direction near the surface. The dew-point temperature, however, experiences a more gradual decrease, but the spread increases when the front passes.

The distributions of mid-tropospheric temperature and wind evolve in a smoother manner than surface variables. Compared to weak forcing, the spread in the strong forcing case is larger throughout the forecast, which is a consequence of the uncertain location of synoptic systems that directly affect the middle troposphere. The temperature gradually decreases with lead time, while the spread increases, with a temporary tail of higher temperatures around 14 UTC, probably caused by advection just before the front or latent heat release, as mentioned in section 5.2.1. The distribution of zonal wind does not evolve as monotonically as for the temperature, but the spread also increases with lead time.

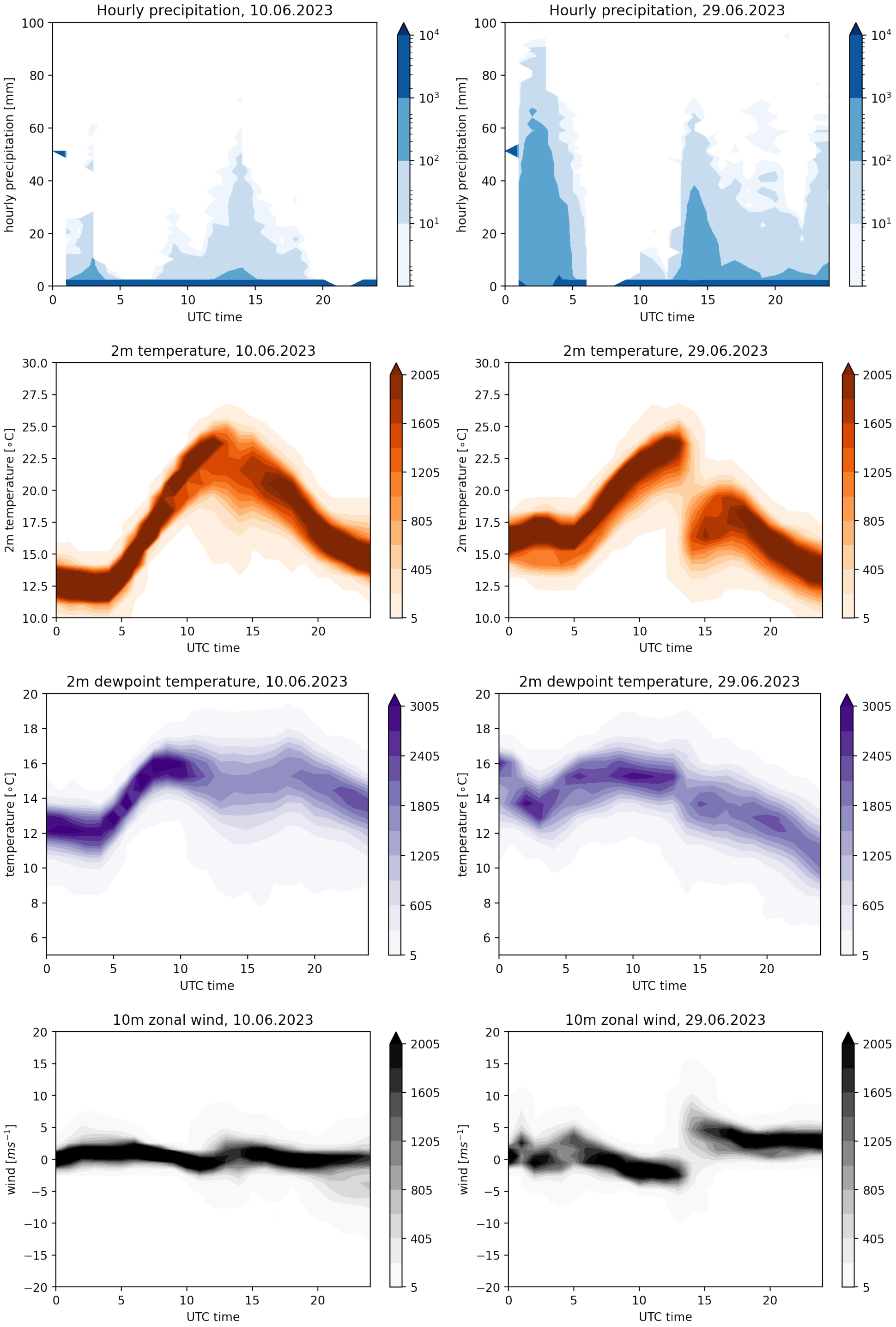

Figure 5.6: Two-dimensional histograms of surface variables for a 10km neighborhood around Reutlingen on 10 June 2021 (left column) and 29 June 2021 (right column). The horizontal axis shows forecast lead time and the vertical axis shows the range of values. Shading represents the frequency of occurrence. From top to bottom: hourly precipitation, temperature at 2 meters, dew-point temperature at 2 meters and zonal wind speed at 10 meters.

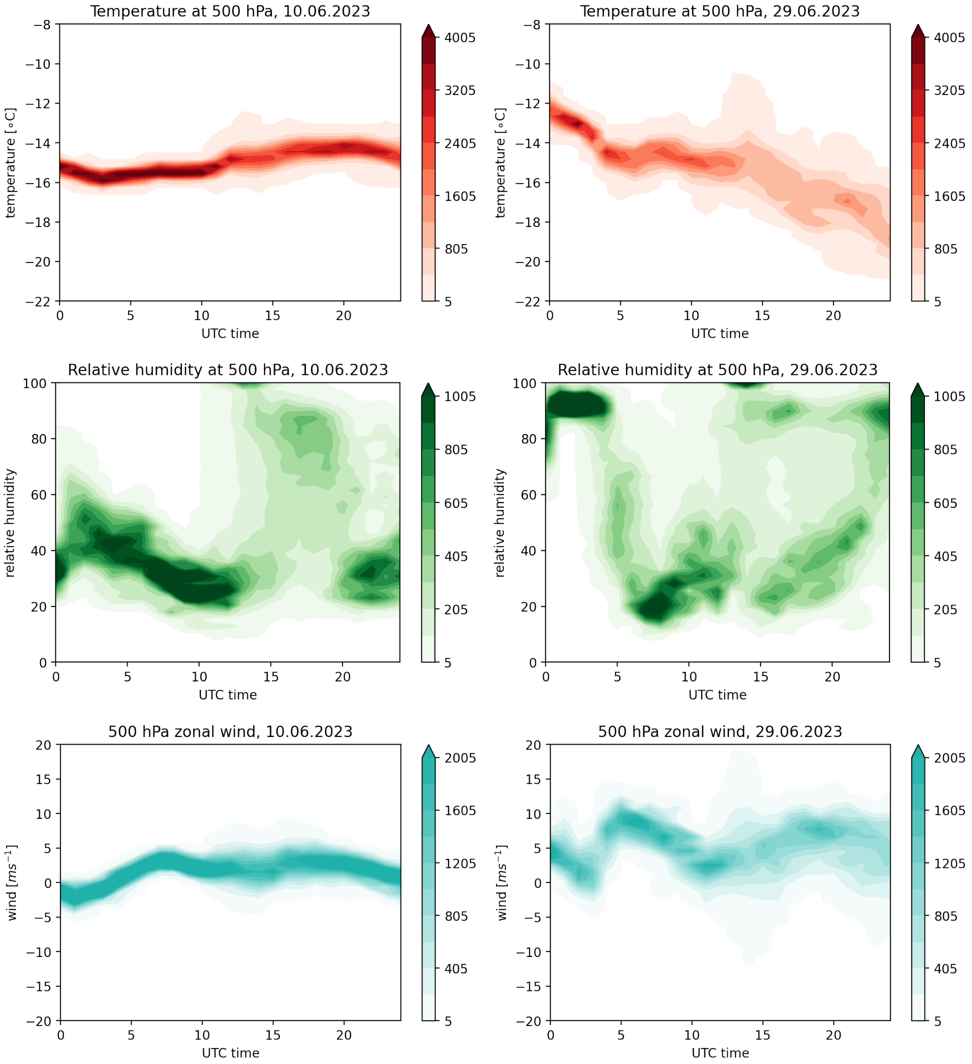

Figure 5.7: As in Figure 5.6, but for mid-tropospheric variables: temperature, relative humidity and zonal wind speed, all at 500 hPa.

Similarly to temperature, two long tails develop at 14 UTC, indicating the large variation of the wind speed and direction (the negative values) between some ensemble members. The evolution of the relative humidity distribution is the most dramatic. From a skewed distribution with a peak at around 20% in the morning it quickly changes to a bimodal distribution with the highest peak at saturation around 14 UTC. From then on, the lower peak of the distribution gradually moves towards higher values to eventually merge with the peak near saturation in the late evening. As anticipated in previous sections and studies, humidity-related variables have the most complex distribution properties and require a larger ensemble to be sufficiently resolved and represented. This confirms what Tempest et al. (2024) found using their idealized model, who pointed out that this is due to the tails in moisture variables' distributions, which are longer and less dense in weak forcing, compared to strong forcing. This translates into a larger sampling uncertainty, which demands more ensemble members for a desired level of uncertainty and to reach convergence of sampling error at all. It is worth noting, however, that the idealized model did not simulate a frontal passage or squall line, which determines the strong forcing case in this chapter. Therefore, the impact of this phenomenon on the distributions of variables and the convergence of sampling error are a novel contribution to the understanding of flow-dependent forecast uncertainty, together with its spatial analysis.

5.3 Conclusions

The question explored in this chapter is how does the weather regime, namely the strength of synoptic forcing of convection, impact the evolution of forecast uncertainty and how is this represented with sampling uncertainty in a big ensemble. A 120-member ICON-D2 ensemble is employed, which is to our knowledge the largest ICON ensemble ever run in a limited-area setting. The bootstrapping method to evaluate sampling error is used in a full operational NWP model for the first time, after Craig et al. (2022) and Tempest et al. (2023) showed its advantages with simpler models. This allows to create maps of uncertainty, which facilitates the identification of the key factors determining how large the uncertainty was for different variables. Two representative cases of weak and strong convective forcing are chosen based on the systematic classification performed by Puh et al. (2023). Moreover, convergence of sampling error is investigated for a chosen location and it is compared to the time evolution of distributions of forecast variables for two synoptic forcing regimes. The main conclusions of this work are summarized here.

1. The sampling error convergence law introduced by Craig et al. (2022) holds for ICON output variables, for both surface and mid-tropospheric variables. However, convergence is not observed for the 95^{th} percentile, which shows an unexpected behavior with a "step" between two distinct quasi-converging sections in the strong forcing regime. This behavior is also found in an idealized model by Tempest et al. (2024).

2. Convection increases the uncertainly of model variables and is the key factor in determining the spatial pattern of uncertainty, which is heavily influenced by synoptic forcing. In weak forcing, the uncertainty pattern is patchy, with single convective cells emerging, while in strong forcing the structure is more coherent due to the larger scale of mesoscale convective systems.

3. The shape of the underlying distribution is heavily influenced by convection and its evolution in both forcing regimes, which dictates the properties of uncertainty. The flow-dependence of the distributions is reflected in the ubiquitous daily cycle of convection in weak forcing, while the passage of a squall line associated with a cold front is prevalent in strong forcing.

When interpreting the findings of this chapter, it is essential to consider several limitations. Although the two analyzed cases are representative, the results obtained may not hold for every strong or weak synoptic forcing case. Secondly, this chapter focuses on summertime convection over Germany, so the findings may not necessarily hold true for other regions, seasons and weather conditions. Additionally, the ensemble comprises only 120 ensemble members, which has been determined to sufficiently represent certain forecast distribution properties such as the ensemble mean or standard deviation. However, it is not sufficient for capturing more sensitive aspects, especially those related to extreme events. The chosen number of 1000 bootstrapping samples may also not be sufficient to determine whether convergence was reached, but it is approximately one order of magnitude larger than the size of the sampled ensemble, which Craig et al. (2022) found to be sufficient. Another limitation is that the sampling error convergence analysis is performed only for one specific location. Nevertheless, our results are expected to qualitatively hold for most of the model domain, since maps of uncertainty for smaller samples show a very similar pattern of uncertainty, which was tested by correlating the fields for smaller samples to the largest, 120-member sample. The correlation coefficient is mostly larger than 0.9, indicating that the uncertainty fields are qualitatively similar. Lastly, only a limited range of forecast variables, statistics and atmospheric layers is investigated for practical reasons. The vertical distribution of uncertainty is beyond the scope of this work. In conclusion, although this chapter has several limitations, the relationships identified here between the meteorological flow, the forecast distribution, and sampling uncertainty should be typical of those found in many contexts.

The main result of this chapter is the strong link between uncertainty of forecast variables and convection, which increases uncertainty by modifying the shape of the distributions. Moreover, we find that the synoptic forcing of convection plays a key role in determining the spatial distribution and the evolution of uncertainty through the interaction or lack thereof between the synoptic scale and the mesoscale at forecast lead times up to 24 hours. Investigating the effect of flow-dependent convection on the uncertainty of variables at longer lead times would be an intriguing path for future work. Further research on the sources and evolution of forecast uncertainty will eventually lead to a more accurate and informative probabilistic weather forecasting.

Chapter 6

Forecast uncertainty beyond 24 hours

6.1 Background

The atmosphere is often considered a deterministic system in the context of classical physics, as it follows well-defined physical laws governing the behavior of gases and fluids. At the same time, a number of reasons make it difficult to accurately predict its future state. Firstly, the governing laws of natural systems are not perfectly known. Additionally, observational uncertainties prevent us from knowing the precise state of the natural system. Consequently, our ability to forecast is limited to imperfect numerical models based on imperfect initial conditions. Errors originating from these initial conditions tend to amplify over time due to atmospheric system instabilities and become intertwined with errors stemming from the use of imperfect models. Since the true states of natural systems remain elusive, the exact error patterns in analysis fields are also unknown. Nonetheless, it is possible to investigate the evolution of potential forecast errors by examining the changes in various perturbations applied to the system's state.

An important consideration is that the magnitude of instabilities, and consequently the velocity or rate at which errors grow, are influenced by the scale of the movements (Lorenz, 1969). Since spatial and temporal scales of atmospheric phenomena are connected, small-scale systems evolve faster than large-scale systems. For small-scale systems, this results in much faster nonlinear perturbation growth that, simply due to their size, saturates at a low energy level. The instabilities driving the formation of characteristics within the flow are also accountable for the generation of errors, which either alter features, intensify them, or lead to their absence. Consequently, accelerated perturbation growth aligns with faster error growth, leading to a quicker decline in predictability for features at smaller scales (Hohenegger and Schär, 2007). For example, perturbations originating from the boundary layer, although small, may have a significant impact on the larger scale flow in an environment with high convective available potential energy (CAPE) and convective inhibition (Selz and Craig, 2015). In this context, perturbations may have the capacity to initiate the formation of extra convective cells rather than simply shifting those already present. Moreover, there is evidence suggesting that the broader influence of convection on the geostrophically balanced flow could be heavily influenced by the prevailing convective regime (Done et al., 2006). Hence, the existence of moist convection by itself doesn't always indicate low predictability, as it heavily relies on the prevailing weather conditions. This holds true especially for lead times longer than about 12 hours, when smaller-scale errors have saturated and predictability decays to a background distribution determined by the synoptic-scale flow. If the evolution of convection is determined by the synoptic-scale pattern, like in the case of a cold front, its predictability will be higher than on average (Keil et al., 2014). Nevertheless, moist convection strongly promotes rapid error growth, with typical timescales of the order of an hour (Leoncini et al., 2010).

The influence of the larger-scale flow on convection itself is of key importance in the context of predictability. The impact of dynamical forcing on convection in mid-latitudes is often characterized as either strong or weak. Typically, this classification relies on the presence or absence of synoptic or mesoscale dynamical features capable of inducing upward motion and the formation of CAPE. When inhibition of convection is weak and triggering disturbances are abundant, convection occurs whenever instability exists, leading to the rapid consumption of CAPE. Conversely, when inhibition is strong and triggering disturbances are scarce, CAPE can accumulate, indicating a non-equilibrium state. Equilibrium conditions often coincide with strong forcing because dynamical ascent weakens inversions, while widespread convection supplies numerous triggering disturbances. However, exceptionally strong capping inversions or other inhibitory factors can prevent convection from reaching equilibrium, causing rapid CAPE increases despite strong forcing.

However, there are exceptions to the lower predictability of weakly forced convection. Craig et al. (2012) found that the impact time of radar rainfall assimilation on the forecast is related to the large-scale control of convection, with shorter impact time-scales when convection is in equilibrium with the large-scale forcing, and longer impacts when instability is present but convection is inhibited, which corresponds to the non-equilibrium regime. In the equilibrium limit, the influence of the data assimilation increments on total precipitation is forgotten over a couple of hours, while in the case of non-equilibrium, the insertion of radar data is sufficient to initiate convective systems that are long-lived, prolonging their predictability. Therefore, a forecast initialization during active convection can significantly modify the forecast evolution and reduce its uncertainty later on.

Soil moisture (SM) was identified as another source of predictability and uncertainty for convection initiation and intensity. It influences the partitioning of net radiative flux into latent and sensible surface heat fluxes, eventually moistening and heating the boundary layer and playing an important role in cloud development as well as in the initiation and formation of convective precipitation (Wallace and Hobbs, 2006). In the majority of cases, there is a positive feedback between SM and the initiation of convective activity (Liu et al., 2022). Baur et al. (2022) showed that initial soil moisture influences the triggering of convection, but with dry conditions accelerating and moist conditions decelerating its onset within the first hours. Their results thus show a negative soil-moisture–precipitation coupling during the initiation phase, especially near soil moisture gradients, and a positive coupling afterwards. Schneider et al. (2019) also found that in the majority of their analyzed cases, the model produces a positive SM–precipitation feedback when averaged over the entire model domain. Furthermore, SM, along with aerosols, is responsible for the maximum precipitation response, while the sensitivity to terrain forcing always shows the smallest spread. Their results show that the impact of these perturbations, including those to SM, on precipitation is on average higher for weak than for strong synoptic forcing, as for other local factors in the literature (e.g. Keil et al., 2019).

Apart from soil moisture, atmospheric moisture and stability also serve as a memory mechanism for the evolution of convection. Petch (2004) uses a 2-dimensional cloud-resolving model to show that the differences in the atmospheric conditions at sunrise on the second day of the simulation lead to very large differences in the timing and intensity of convective rainfall in the ensemble. The amount of total daily rainfall is shown to have a major impact on the subsequent day's convection, due to significantly different amounts of moisture in the atmosphere at sunrise. Hence, if different ensemble members produce a wide variety of total rainfall during the day, the very different moisture and temperature profiles at sunrise on the second day increase the variability in the diurnal cycle of convection on the next day.

The question that we pose in this chapter is how much does the longer-lasting impact given by memory effects discussed above influence forecast uncertainty at the convective scale beyond the first 24 hours of the simulation, especially the precipitation. These include soil moisture and atmospheric stability, as well as forecast initialization during active convection. The role of small-scale stochastic perturbations in these mechanisms is studied in the flow-dependent framework of weak and strong forcing. The hypothesis is that the impact of memory effects is stronger in weak forcing.

In previous chapters, the evolution of forecast uncertainty was studied in the first 24 hours of the forecast, which for a case of weak forcing means investigating one diurnal cycle of convection. With a maximum forecast lead time of 48 h (36 h), the second diurnal cycle can be predicted and the influence of the different evolution on the first day, based on added perturbations or the different initialization time, can be inferred.

6.2 Results

6.2.1 Time evolution of precipitation

We first focus on the time evolution of precipitation in the weak and strong forcing cases. The description of the general weather pattern of the case studies is in section 3.3. The evolution of other variables does not provide additional insight, since precipitation is the least robust variable of the chosen set and the one that is most directly influenced by convection. The interested reader can see figures for other variables in the Appendix.

Figure 6.1a illustrates the time series of precipitation under weak forcing conditions, characterized by the presence of two distinct diurnal cycles. The initial cycle conforms to the typical pattern associated with weak forcing convection, initiating around 10 UTC with the onset of triggered convection. Conversely, the second cycle exhibits greater uncertainty within the forecasts, commencing as early as nighttime. When comparing the

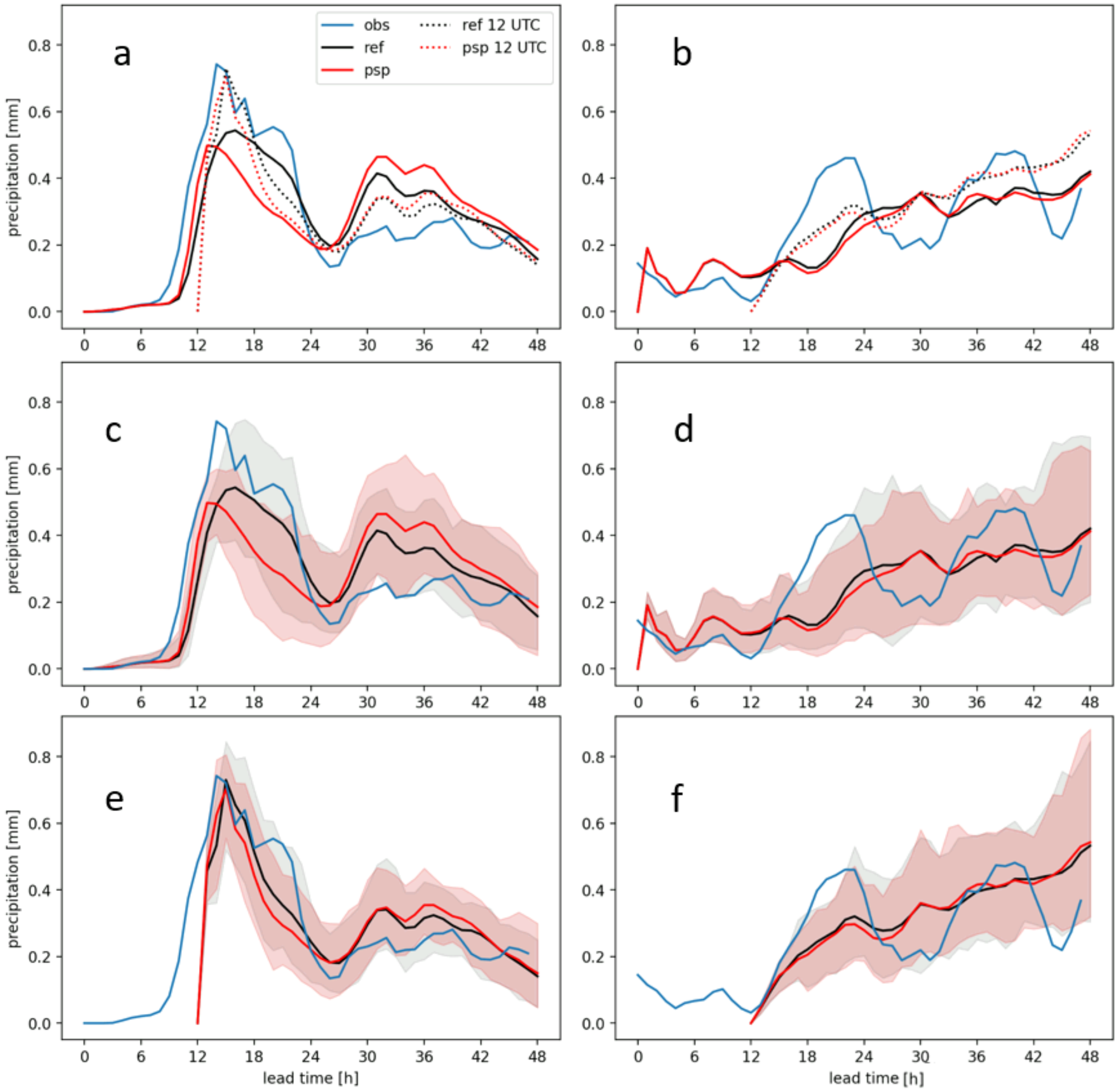

Figure 6.1: Time series of hourly precipitation (in mm) for the 4 simulations and the estimated amount from radar observations on weak forcing (a, c, e - 26 and 27 August) and strong forcing (b, d, f - 18 and 19 August), averaged over Germany. The lines show the ensemble mean, while the shading in panels c, d, e and f represents the range between the first and 3rd quartile of the distribution given by the ensemble members (c and d initialized at 00 UTC, e and f initialized at 12 UTC). The black lines show the reference experiment (00 UTC initialization continuous, 12 UTC dotted), while the red lines show the PSP experiment (00 UTC initialization continuous, 12 UTC dotted). The estimated precipitation from radar observations is shown with the continuous blue line. The x axis shows the forecast lead time in hours.

observations with the simulations, it is important to keep in mind that the curves for the different experiments are showing the ensemble mean, while the observation curve is to be interpreted as one realization of the forecast distribution and is therefore expected to experience more variability. However, it is still useful to make the comparison between forecasts and observations to get an idea of the skill and realism of the simulations.

The comparison between the reference simulations and the observations is discussed first. Overall, the simulations initialized at 12 UTC demonstrate the closest alignment with radar observations. This is not surprising, since they contain information about ongoing convection up to 12UTC on the first day, which constrains the whole forecast evolution and lowers the uncertainty[1] especially on the second day of weak forcing (compare panel 6.1c and e). This is the first "memory mechanism" that we discuss in the introduction (Craig et al., 2012). As expected, this effect is weaker in the case of strong forcing, with uncertainty growing at a similar pace in the 00 UTC and the 12 UTC forecast, although reaching a slightly lower absolute value by the end of the forecast due to the shorter time for the growth of errors (compare panel 6.1d and f). The uncertainty band of the reference simulation encompasses the radar estimation, except for the peak in the evening of day 1. This peak is around midnight, which cannot be considered to be caused by the diurnal cycle of convection, although some influence is not excluded in this case. Moreover, the model domain does not include the main feature of the larger-scale forcing, which is a cyclone in the Mediterranean region (see Figure 3.4). As a consequence, clouds and convective clusters are advected from the south-east (as seen in Figure 6.7) and are therefore mainly determined by the coarser lateral boundary conditions, hence a direct quantitative comparison of observed and simulated precipitation requires caution. Additionally, the ensemble size of 40 members is likely not large enough to capture events that are further away from the ensemble mean, as seen in Chapter 5.

We now assess the impact of PSP on the time evolution of precipitation. The 00 UTC PSP simulation performs the worst compared to the observations in terms of precipitation amount, although it slightly improves the timing of the onset of precipitation and its highest peak. The more effective triggering of convection, however, increases the cloud cover too quickly (not shown), limiting successive triggering by the turbulence in the PBL, caused by solar heating of the surface, and significantly reducing the average precipitation later in the afternoon. This is a worsening of a bias that is already present in the reference simulation. This negative bias on the first day is likely the cause for the positive bias on the second day, as there is more CAPE to be consumed by generating more precipitation, together with other environmental effects which will be discussed in the following section. The uncertainty in both simulations quickly grows as the convection develops, but it is larger for the reference experiment until the early night, when the PSP experiment has larger uncertainty, with a lower first quartile due to the bias in precipitation. Afterwards, the uncertainty in the PSP experiment outgrows the reference simulation for most of the second day, until they roughly overlap again in the evening. On the second day, the larger

[1] The uncertainty bands in Figure 6.1 are defined by the first and third quartile, which contain 50% of the ensemble members.

uncertainty of the PSP simulation is a consequence of the higher upper quartile, since a portion of members have a significantly larger amount of precipitation. However, on neither of the days the PSP scheme improves the forecast quantitatively, as the radar estimation falls out of the uncertainty bend of the PSP simulation on both peaks of the diurnal cycles, while it is mostly contained by the band of the reference simulation. It is likely that this pronounced undesired effect of the PSP scheme is a characteristic of the chosen case, and therefore it is not possible to draw general conclusions about the effect of the scheme on longer forecasts, for which more cases are needed.

In strong forcing (Figure 6.1b), the PSP and reference experiments are much more similar. There are only slight differences in the ensemble- and area-average precipitation, as well as in the uncertainty. However, the 12 UTC forecasts are distinctly different, which confirms our hypothesis and the previous findings of the larger-scale conditions being more important in strong forcing, while in weak forcing they are equally important to the small-scale features in terms of the impact on precipitation and uncertainty.

Figure 6.2 shows the time evolution of the histogram of precipitation for a neighbourhood around the indicated location in Eastern Germany in figures in section 6.2.2. The analysis underscores the shift in the diurnal cycle by PSP and the denser tail in the reference simulation on day 1. Conversely, the PSP simulation has a denser and longer tail from the early morning hours of day 2, resulting in a higher average precipitation amount and uncertainty. A similar effect, although of a smaller magnitude, can be observed on day 2 in the 12 UTC initialized forecasts, comparing the PSP and the reference simulations. The diurnal cycle on day 2 is more pronounced in both simulations in comparison with the 00 UTC forecasts, slightly enhanced by PSP. Additionally, the spinup effect is discernible in the 12 UTC runs, characterized by a shorter tail and lower average precipitation levels (too low compared to observations, not shown). The appendix shows the same figure for strong forcing, since its inclusion does not yield significant additional insights, given the insignificant impact of PSP.

6.2.2 Spatial distribution of precipitation

The impact of the PSP scheme on precipitation is now analyzed more in detail, including its spatial distribution. As expected, the impact is more significant in the case of weak forcing of convection.

Weak forcing

The comparison of the maps of precipitation estimates from radar observations and the ensemble mean for the reference simulation for weak forcing is depicted in Figure 6.3, as well as the anomalies caused by PSP, relative to the reference. These are to be interpreted in a broader sense, since the ensemble mean is not designed to represent reality, but rather to show agreement between ensemble members on the spatial distribution of higher or lower amounts of precipitation. On day 1 of weak forcing, there is widespread convective activity. In Eastern Germany there is a pronounced reduction observed with PSP, relative

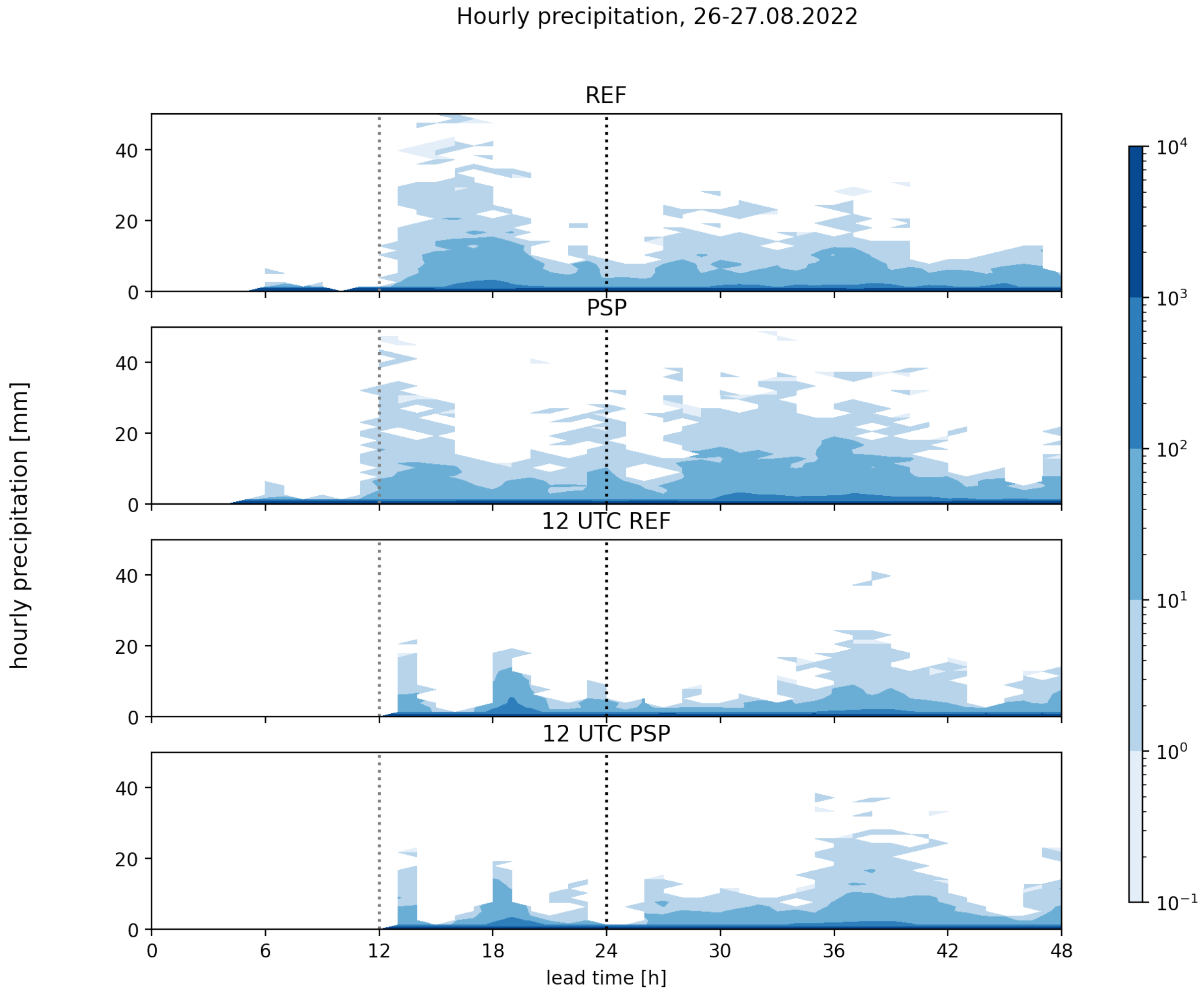

Figure 6.2: 2D histogram (the x axis represents lead time) of hourly precipitation for a 10-km neighbourhood around a location in Eastern Germany on 26 and 27 August (weak forcing), for the 4 simulations: reference, PSP and their 12 UTC equivalents. The shading indicates the frequency of occurrence.

to the reference simulation (panel 6.3c). This reduction can be attributed to the shading effect of the anvil clouds produced by preceding convection events, whose onset was earlier with PSP, as discussed and seen in the time series in the previous section. Moving to day 2, the convective patterns are more concentrated, primarily affecting southern and eastern regions of Germany. Although both simulations exhibit a mismatch in the exact location of the heaviest precipitation accumulation, PSP demonstrates a noteworthy improvement by shifting the precipitation southwards (panel 6.3g), aligning more closely with observed data. The predominant factor contributing to the substantial PSP impact on precipitation is identified as nighttime and morning convection, elucidated below with Figure 6.4. Interestingly, the forecast initialized at 12 UTC demonstrates a comparable effect of PSP, albeit to a lesser extent (panel 6.3h). The impact of the later initialization time is comparable in magnitude to that of PSP on the second day (panel 6.3d), but it is more widespread.

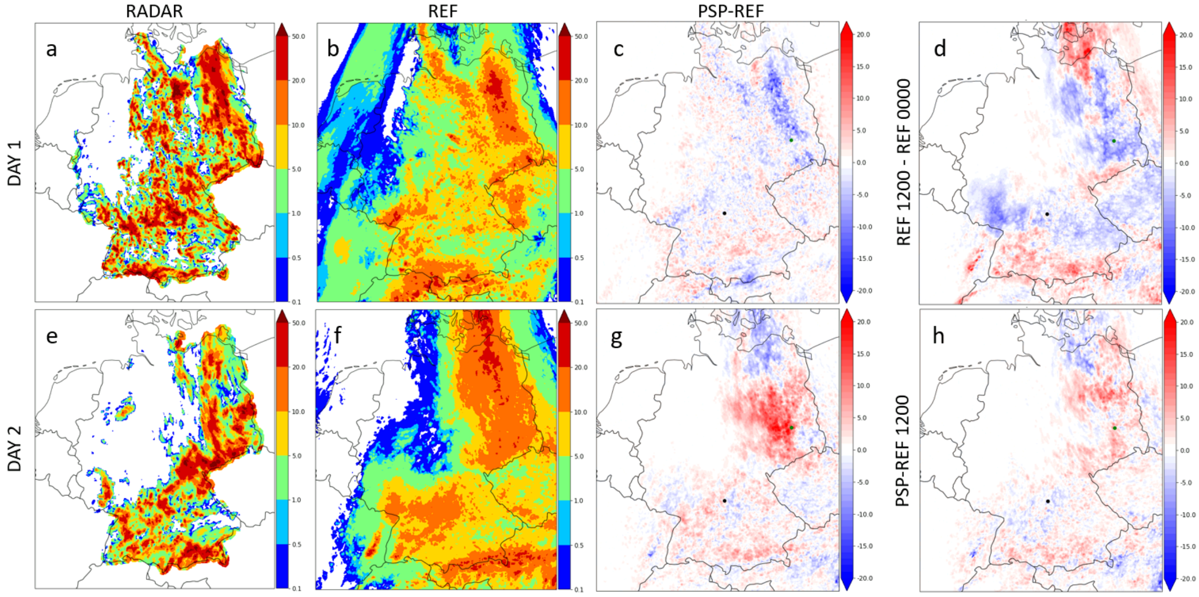

Figure 6.3: Total accumulated precipitation on day 1 (26 August) and day 2 (27 August) as estimated by the radar observations (RADAR, a and e), in the ensemble mean of the reference simulation (REF, b and f). Difference in total accumulated precipitation between the PSP and the reference simulation initialized at 00 UTC (PSP-REF, c and g), at 12 UTC (PSP-REF 1200, day 2, h) and between the 12 UTC and 00 UTC reference simulations (REF 1200 - REF 0000, day 2, d). The units for all panes are millimeters. The green dot indicates the location for the neighbourhood analysis in the previous section.

To better understand the mechanisms behind the impact of PSP on precipitation, we now focus on the early morning hours. During the night and early morning, all forecasts overlook a convective cluster positioned in central Germany, which moves eastwards during this time frame. However, PSP stands out by effectively shifting the focal point of convective activity towards the south, while intensifying its overall impact (top right panel in Figure 6.4). Notably, the 12 UTC forecast demonstrates a more accurate prediction of the spatial distribution of convection, although with a drawback of weaker intensity in the ensemble mean (lower middle panel in Figure 6.4).

Delving into the underlying causes of the substantial impact observed in PSP's nighttime convection in Figure 6.4, several factors come into play. We now investigate potential mechanisms that led to the observed effects, i.e. soil moisture memory and atmospheric stability. Firstly, the soil moisture anomaly with respect to the reference is analyzed. Although there is a clear pattern in the spatial distribution of anomalies (Figure 6.5, lower left panel), these closely follow the pattern already seen in the precipitation (see Figure 6.3c) after 24 h, with a dry signal where the precipitation was scarcer and vice versa. No clear signal is observed in the region with the highest precipitation increase on day 2. This

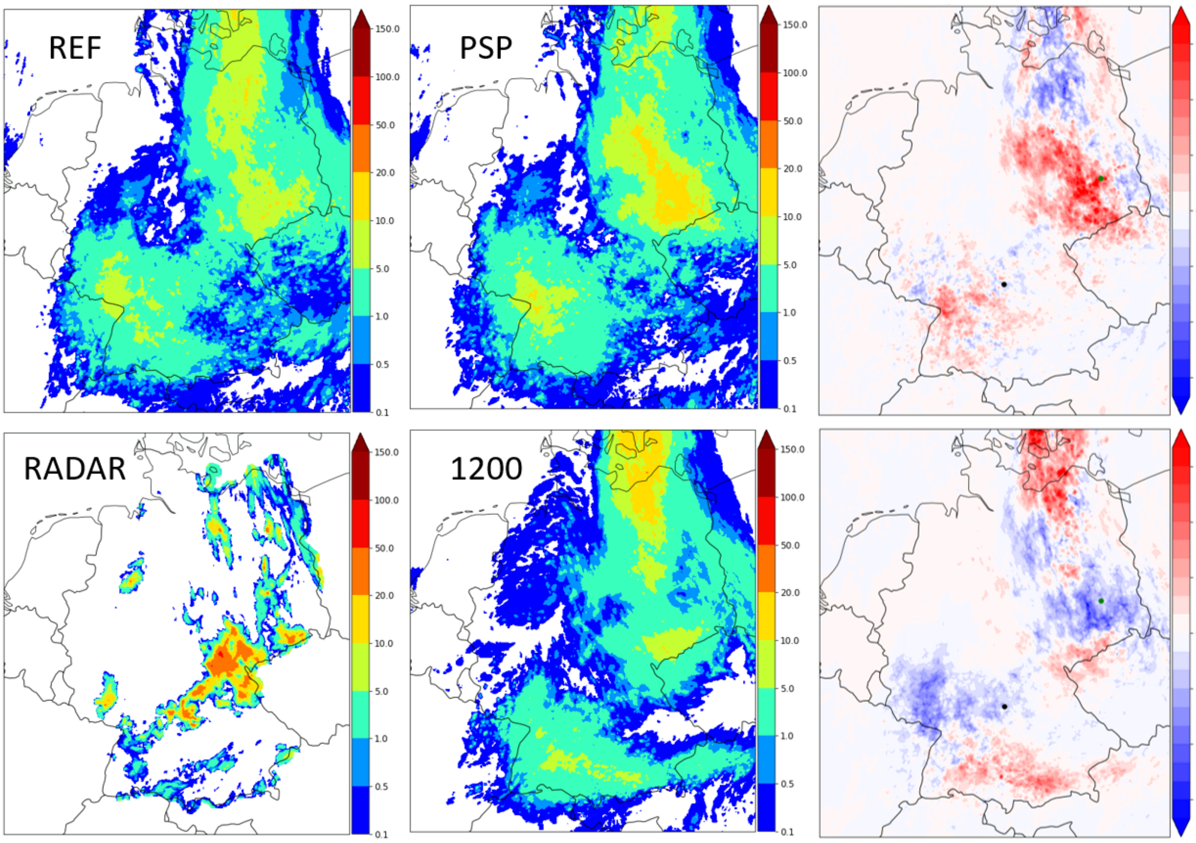

Figure 6.4: Total accumulated precipitation between 03UTC and 09UTC of day 2 (27 August) as estimated by the radar observations (RADAR), in the ensemble mean of the reference simulation (REF), the PSP simulation (PSP) and the simulation initialized at 12 UTC (1200). Difference in accumulated 6h precipitation between the PSP and the reference simulation initialized at 00 UTC (top right) and the 12 UTC and 00 UTC reference simulations (bottom right). The units for all panes are millimeters. The green dot indicates the location for the neighbourhood analysis in the previous section.

proves once again that it is hard to clearly assess the impact of soil moisture variability on convection (as in e.g. Hauck et al., 2011). Nevertheless, the apparent anti-correlation of the soil moisture anomaly with the humidity anomaly at 950 hPa at midnight would confirm the findings of Baur et al. (2022), if the 950 hPa level were above the boundary layer (not assessed).

However, a larger CAPE, attributable to a moister boundary layer, alongside increased Convective Inhibition (CIN) in Northeast Germany, as shown in Figure 6.5 (upper panels and lower right panel), clearly point towards the second possible mechanism. Concurrently, ongoing convection benefits from more favorable environmental conditions conducive to its sustenance and organization, characterized by enhanced CAPE, vertical wind shear, and low-level convergence (not shown). It is hard to attribute more precisely which factors con-

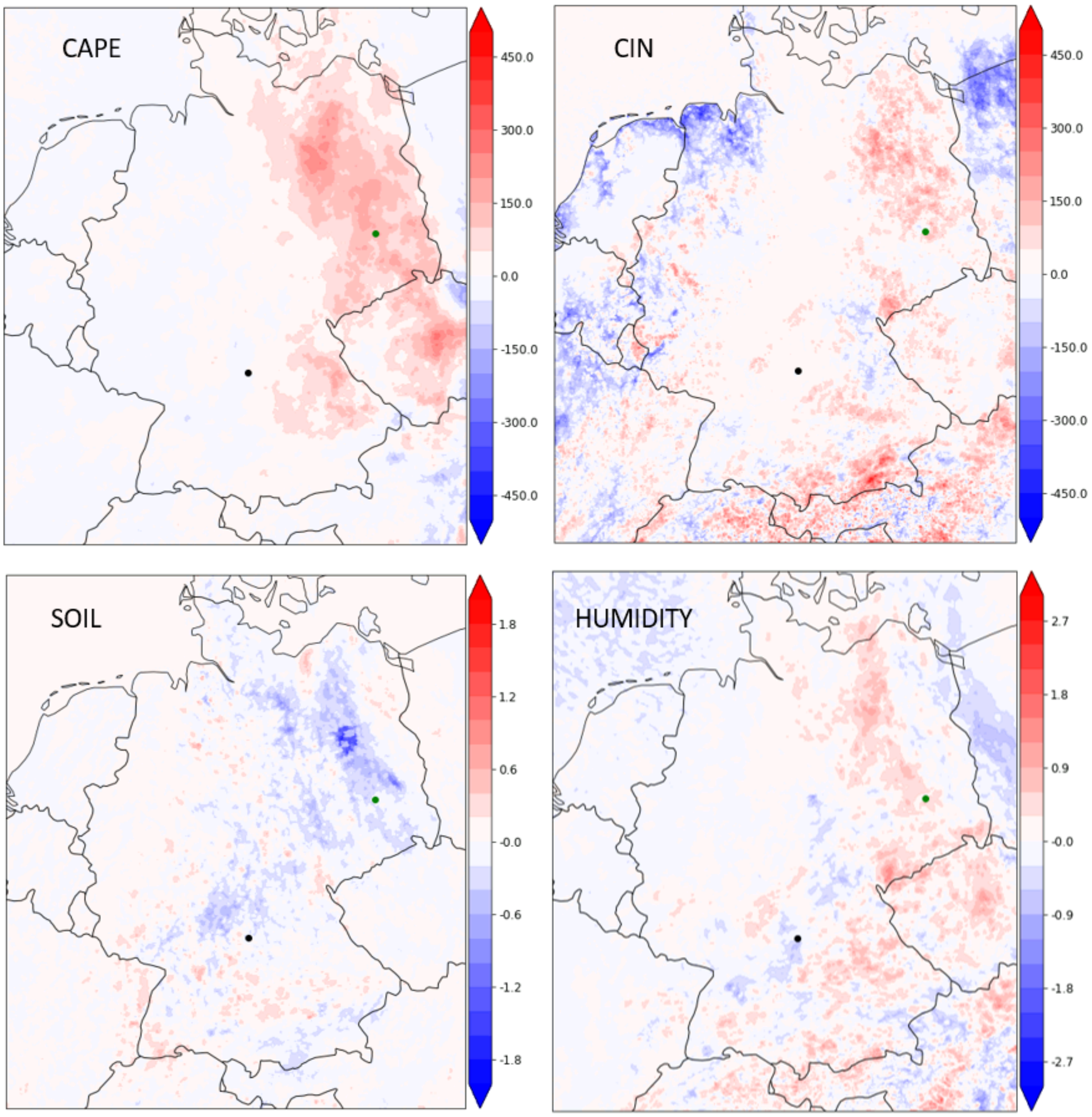

Figure 6.5: Maps of the difference in CAPE, CIN (in J/kg), soil moisture at 5 mm depth (in kg/m^2) and specific humidity at 950 hPa (in g/kg) at 00 UTC of 27 August, between the ensemble mean of the PSP simulation and the reference simulation. The green dot indicates the location for the neighbourhood analysis in the previous section.

tributed the most to the anomalies of the PSP simulation, but we can conclude that the perturbations introduced by the scheme could modify the convective environment. This "memory effect" represents an opportunity to increase the predictability horizon of convection in such weakly forced conditions if the location of convective activity is precisely predicted on the first day, as well as its intensity, which strongly influences the environmental conditions for the development of convection on the next day. Indeed, the time series of the difference in domain-average precipitable water (Figure B.1 in the Appendix) between the PSP and the reference simulations shows an opposite evolution, with the final amounts after 48 hours being almost equal, which confirms the findings of Petch (2004).

The enhanced convective dynamics facilitated by PSP also cause a discernible alteration in the synoptic-scale flow, as evidenced by the analysis presented in Figure 6.6. This figure delineates the pressure difference that emerged 24 and 48 hours into the simulations. A distinctive dipole-like impact observed in the pressure field signifies a notable shift in the synoptic pattern, primarily influenced by the intensified convective processes and latent heat release, particularly prevalent in Eastern Germany. This enhanced convective activity corresponds to a more robust outflow within the upper troposphere, further impacting the broader circulation pattern. A similar impact was also observed by Done et al. (2006) in a weak forcing case 12 hours after the initial triggering of convection, when the background 250 mb jet speed was increased by up to 20 ms^{-1}, compared to a simulation without convection. In this chapter, these mechanisms have not been fully investigated, therefore we cannot make solid conclusions and further work is needed to confidently link the interplay of these factors.

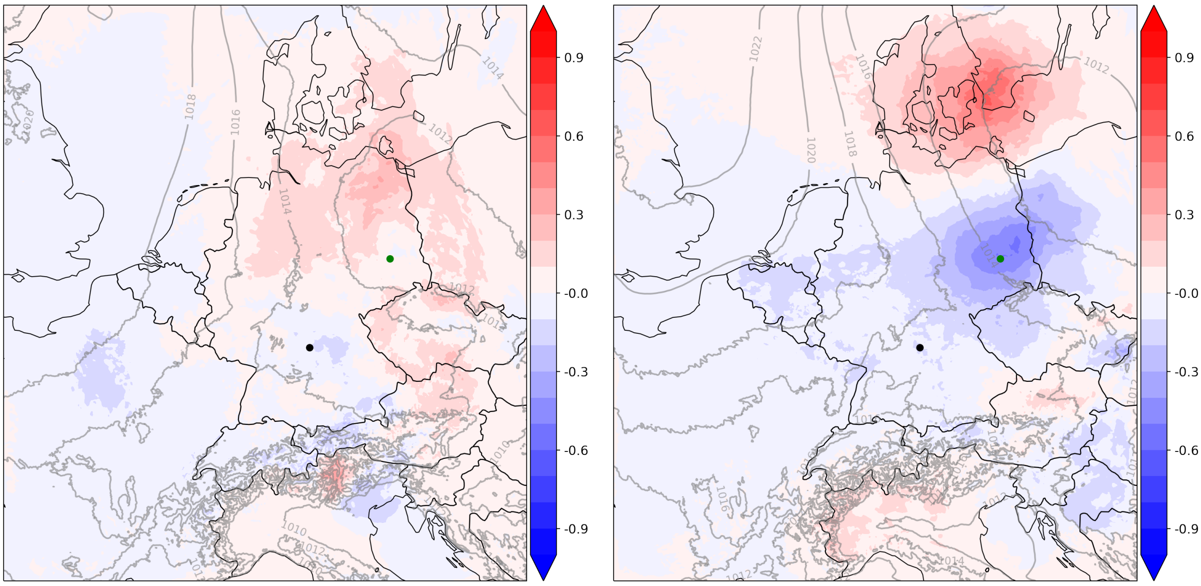

Figure 6.6: Ensemble mean of sea level pressure for the reference simulation (contours, in hPa) and difference between the PSP simulation and the reference simulation (shaded, in hPa) at 00 UTC on 27 August (24 h lead time, left) and at 00 UTC on 28 August (48 h lead time, right). The green dot indicates the location for the neighbourhood analysis in the previous section.

Strong forcing

For the strong forcing case, fewer results are presented, since the impact of PSP on precipitation is not significant. As shown in Figure 6.7, the predominant accumulations of precipitation are notably concentrated in the southern regions of Germany over the two-day span. However, on the initial day, forecasts initialized at 00 UTC failed to capture the heightened precipitation intensity observed in Central Germany, near the border with Czechia (panels 6.7a, b and c). Moreover, the nighttime maximum of precipitation in southern Germany, caused by quasi-stationary convection, is misplaced and positioned on the northern edge of the Alps instead in the reference simulation (panels 6.7e and f), while the forecast initialized at 12 UTC improved significantly the predicted amount and spatial distribution of precipitation in this region (panel 6.7d). The implementation of PSP does not yield a substantial impact (panel 6.7g, h), as its effect appears comparatively minor when compared with the outcomes of the later initialization time (panel 6.7d). Specifically, the impact of PSP on the disparity in accumulated precipitation is quantified to be approximately two to three times smaller than its impact on scenarios characterized by weaker forcing dynamics. This observation underscores once again the flow-dependence of the upscale growth of small-scale perturbations and of forecast uncertainty.

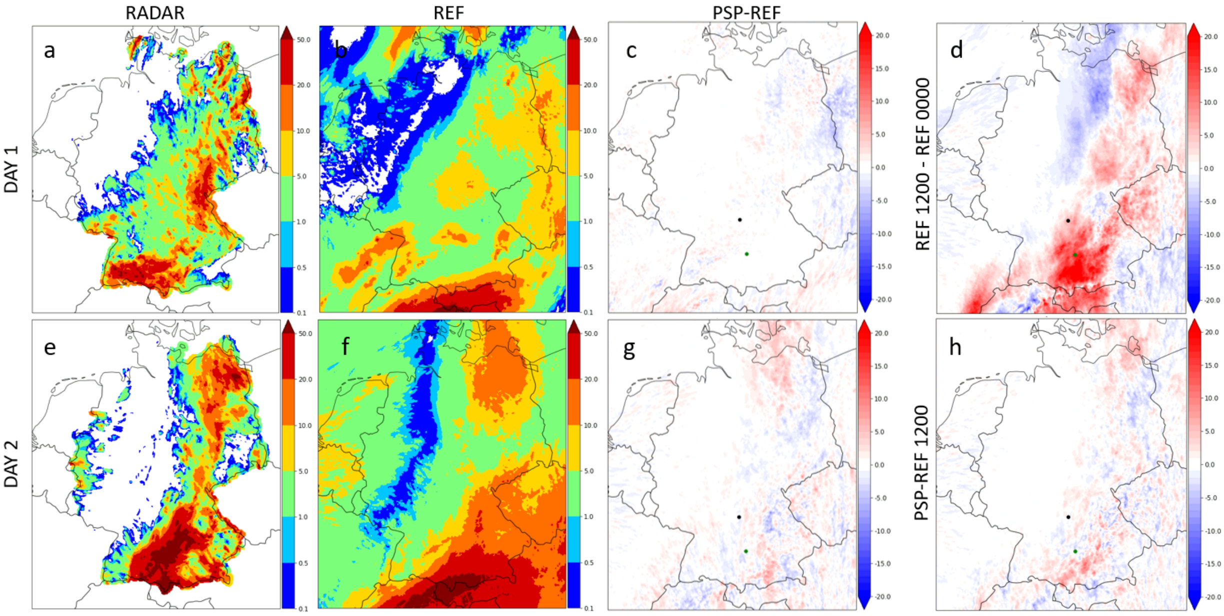

Figure 6.7: Total accumulated precipitation on day 1 (18 August) and day 2 (19 August) as estimated by the radar observations (RADAR, a and e), in the ensemble mean of the reference simulation (REF, b and f). Difference in total accumulated precipitation between the PSP and the reference simulation initialized at 00 UTC (PSP-REF, c and g), at 12 UTC (PSP-REF 1200, day 2, h) and between the 12 UTC and 00 UTC reference simulations (REF 1200 - REF 0000, day 2, d). The units for all panes are millimeters. The green dot indicates the location for the neighbourhood analysis (in Appendix B for strong forcing).

6.3 Summary and conclusions

This chapter delves into the influence of small-scale stochastic perturbations on forecast uncertainty within a 48-hour time frame, examining two cases of strong and weak forcing of convection. It contrasts the impact of PSP with that of altering the initialization time, where existing convection is assimilated into the forecast, initiating convective systems that can be long-lived. This represents one of the memory effects that are investigated, along with soil moisture and atmospheric stability. While prior chapters primarily focused on analyzing forecast uncertainty within the initial 24-hour period, extending the maximum forecast lead time to 48 hours enables the prediction of a second diurnal cycle. This expanded scope allows for a deeper understanding of how added perturbations or changes in initialization time influence the evolution of forecast uncertainty during the second day of the forecast, with a focus on the responsible memory mechanisms that cause an observed change in the precipitation with the addition of PSP. The main conclusions are summarized here.

In weak forcing conditions, the PSP scheme can systematically influence mesoscale conditions and flow, thereby exerting a discernible impact on subsequent convective activity, particularly noticeable on the second day of the simulations. This influence can be delineated through the following mechanism. Firstly, triggered convection depletes CAPE while inhibiting the development of additional nearby convection due to the shading by the cloud cover it created. Secondly, any remaining CAPE is advected and transferred to the following day, shaping the subsequent environmental conditions for the development of convection. Thirdly, alterations in low-level convergence and upper-level divergence impact wind fields and vertical wind shear, further modulating convective dynamics. Additionally, if the ensuing environmental conditions are favorable to enhanced convection, it results in increased precipitation and the formation of a positive anomaly, and vice versa. Soil moisture anomalies between the two simulations were shown to not significantly influence subsequent convective activity. Nevertheless, we conclude that the complex interplay between thermodynamic stratification, state of the soil and available net radiation determines the importance and behaviour of different sources of uncertainty in the forecast (as in Keil et al., 2019).

The impact of initialization time, alongside PSP, is also larger in weak forcing, when the magnitude of their influence on the precipitation is comparable, which confirms our hypothesis. In strong forcing, the impact of a different initialization time is significantly stronger than that of PSP, although slightly weaker than its impact in weak forcing.

When assessing the implications of this research, it is crucial to acknowledge inherent limitations. Firstly, while the two cases studied serve as representative examples, it's important to recognize that the findings may not universally apply to all instances of strong or weak synoptic forcing scenarios. Secondly, given the focus on summertime convection specifically within Germany, the generalizability of the results to other regions, seasons, and weather conditions may be limited. Moreover, the ensemble utilized in this study consists of only 40 members, which may not be sufficiently robust for capturing extreme events, but was shown to be adequate for estimating the uncertainty of the mean. Finally, the memory mechanisms discussed in this chapter are not fully investigated and therefore need to be elucidated in more detail in future work. In summary, despite these acknowledged limitations, the relationships elucidated here between weather regime, forecast distributions, and their evolution are expected to be broadly applicable across various contexts.

The most important conclusion of this chapter is that although the uncertainty on the convective scale quickly grows and the predictability left comes mostly from the larger-scale flow, there are mechanisms on the smaller scale that can still influence the forecast beyond the usual influence time. Moreover, if the factors that contribute to this extended influence are assimilated and well represented in the initial conditions, the resulting forecast is more accurate. Therefore, several weather prediction centers developed forecasting systems that are initialized with an hourly frequency, like SINFONY (Seamless INtegrated FOrecastiNg sYstem) at DWD using a rapid update cycle (RUC) in the ICON-LAM-EPS (Blahak, 2023) or the High-Resolution Rapid Refresh (HRRR) model (Dowell et al., 2022) at NOAA (National Oceanic and Atmospheric Administration). As shown in this chapter, a more recent initialization time, including information on local environmental conditions, reduces uncertainty in weak forcing situations and prolongs the predictability of convection beyond the typical timescale when local predictability is lost, which confirms the findings of Craig et al. (2012). This is particularly beneficial for early warnings of high impact weather caused by convection, which can cause damage to a wide spectrum of society.

Chapter 7

Conclusions

The accurate prediction of atmospheric phenomena is hindered by the inherent chaos of the Earth's atmosphere, coupled with a multitude of uncertainties stemming from observational limitations and model imperfections. These challenges necessitate sophisticated forecasting methodologies capable of quantifying and addressing uncertainties to provide reliable meteorological predictions. Ensemble Prediction Systems (EPS) have emerged as indispensable tools in this regard, offering a probabilistic framework that accommodates the inherent variability of atmospheric processes.

However, there are several limitations to ensembles too. Firstly, the size of ensembles is constrained by cost, often limiting them to smaller sizes. Assessing errors stemming from these limited sizes proves challenging because the exact distribution of a forecast variable, as observed by a larger ensemble, is unknown, making it uncertain how many ensemble members are necessary for accurate sampling. Moreover, an ongoing challenge is to develop ensemble methodologies that effectively capture the multitude of uncertainties in natural systems and combine them to create physically consistent and effective variability in the ensemble forecast.

The aim of this thesis is to address these challenges by providing a more complete understanding of the nature and the evolution of forecast uncertainty in a convection-permitting ensemble. There are three central research questions to this thesis. Firstly, does the PSP scheme systematically improve the probabilistic skill of convection-permitting ensemble forecasts over Germany? Secondly, is a 120-member ensemble sufficiently large to observe convergence of sampling error with a fully-fledged NWP ensemble? Lastly, how does the convective weather regime affect the evolution of uncertainty of forecast variables and how does it influence its spatial distribution?

To answer these questions, three different convection-permitting ensemble experiments are performed using the ICON Limited Area Model (LAM). To begin with, the impact of the PSP scheme is assessed over a 3-month period in a 20-member ensemble, focusing on its flow-dependent behaviour in strong and weak forcing regimes. Two cases from this period are then chosen for a large, 120-member ensemble experiment, which allowed for a more detailed analysis of the evolution of forecast uncertainty and the convergence of sampling error, as well as the spatial distribution of uncertainty and the role of convection therein. Lastly, this analysis is extended to longer lead times with 40-member ensemble experiments spanning up to 48 hours, allowing the forecast of a second diurnal cycle of convection in weak forcing conditions.

Conclusions from each part of the thesis are now summarized. A closing remark will highlight and discuss limitations and implications for future applications and research.

Physically-based stochastic perturbation scheme The PSP scheme is used in ICON-D2-EPS as a representation of model error originating from the subgrid scale in the boundary layer, but affecting the smallest resolved scales. The experimental period spans a whole summer season, which allows for a systematic analysis of the impact of the scheme in different synoptic forcing conditions.

The scheme provides a good representation of the effect of subgrid scale turbulence in ICON-D2 and has realistic, beneficial effects in ensemble forecasts, especially by increasing the ensemble spread and therefore reducing the underdispersion of surface variables. It helps triggering convection, while preserving the intensity of single convective cells and does not produce spurious convection.

The small-scale perturbations introduced by PSP have a larger impact on convective precipitation in weak than in strong synoptic forcing, especially on its spread. This is in line with the hypothesis of local processes in the boundary layer driving the convection on weakly forced days, whereas on strong forcing days, the synoptic pattern controls the convection.

The main result of this chapter is the improvement of the spread to skill ratio of the ensemble with PSP for several variables, while preserving the same level of skill. Combined with its physical foundation, the scheme is a good representation of model error stemming from unresolved eddies in the PBL and can therefore be used in large ensemble experiments.

Flow dependence of forecast uncertainty A 120 member ICON-D2 ensemble with PSP is used to answer the question of how the weather regime, namely the strength of synoptic forcing of convection, impacts the evolution of forecast uncertainty and how this is represented by sampling uncertainty in a big ensemble. To our knowledge this is the largest ICON ensemble ever run in a limited-area setting. The bootstrapping method to evaluate sampling error is used in a full operational NWP model for the first time, after Craig et al. (2022) and Tempest et al. (2023) showed its advantages with simpler models. This allows the creation of maps of uncertainty, which facilitates the identification of the key factors determining how large the uncertainty was for different variables. Two representative cases of weak and strong convective forcing are chosen based on the systematic classification performed in the first part of the thesis. Moreover, convergence of sampling error is investigated for a chosen location and it is compared to the time evolution of distributions of forecast variables for two synoptic forcing regimes.

It was found that the sampling error convergence law introduced by Craig et al. (2022) holds for ICON output variables, for both surface and mid-tropospheric variables. However, convergence is not observed for the 95^{th} percentile, which shows an unexpected behaviour with a "step" between two distinct quasi-converging sections in the strong forcing regime. This behaviour is also found in an idealized model by Tempest et al. (2024).

Secondly, convection increases the uncertainly of model variables and is the key factor in determining the spatial pattern of uncertainty, which is heavily influenced by synoptic forcing. In weak forcing, the uncertainty pattern is patchy, with single convective cells emerging, while in strong forcing the structure is more coherent due to the larger scale of mesoscale convective systems.

The shape of the underlying distribution of forecast variables is heavily influenced by convection and its evolution in both forcing regimes, which dictates the properties of uncertainty. The flow-dependence of distributions is reflected in the ubiquitous daily cycle of convection in weak forcing, while the passage of a squall line associated with a cold front is prevalent in strong forcing.

The main result of this chapter is the strong link between uncertainty of forecast variables and convection, which increases uncertainty by modifying the shape of the distributions. Moreover, synoptic forcing of convection plays a key role in determining the spatial distribution and the evolution of uncertainty through the interaction or lack thereof between the synoptic scale and the mesoscale at forecast lead times up to 24 hours.

Forecast uncertainty beyond 24 hours In the last part of the thesis, the analysis of the flow dependence of forecast uncertainty is extended to 48 hours with another pair of cases with weak and strong forcing of convection. We study the longer-lasting impact given by the memory effects of soil moisture and atmospheric stability on forecast uncertainty at the convective scale beyond the first 24 hours of the simulation. While prior chapters primarily focused on analyzing forecast uncertainty within the initial 24-hour period, extending the maximum forecast lead time to 48 hours enables the prediction of a second diurnal cycle. This expanded scope allows for a deeper understanding of how added perturbations or changes in initialization time influence the evolution of forecast uncertainty during the second day of the forecast.

Investigations showed that the PSP scheme can systematically influence mesoscale conditions and flow in weak forcing conditions, thereby exerting a discernible impact on subsequent convective activity, particularly noticeable on the second day of the simulations. This influence can be delineated through the following mechanism. Firstly, triggered convection depletes CAPE while inhibiting the development of additional nearby convection due to the shading by the cloud cover it created. Secondly, any remaining CAPE is advected and transferred to the following day, shaping the subsequent environmental conditions for the development of convection. Thirdly, alterations in low-level convergence and upper-level divergence impact wind fields on a larger scale, further modulating convective dynamics. Additionally, if the ensuing environmental conditions are favorable to enhanced convection, it results in increased precipitation and the formation of a positive anomaly, and vice versa. Soil moisture anomalies between the two simulations were shown to not significantly influence subsequent convective activity. Nevertheless, we conclude that the complex interplay between thermodynamic stratification, state of the soil and available net radiation determines the importance and behaviour of different sources of uncertainty in the forecast.

The impact of initialization time, alongside PSP, is larger in weak forcing, when the magnitude of their influence on the precipitation is comparable. In strong forcing, the impact of a different initialization time is significantly stronger than that of PSP, although slightly weaker than its impact in weak forcing.

The most important conclusion of this chapter is that although the uncertainty on the convective scale quickly grows and the predictability left comes mostly from the larger-scale flow, there are mechanisms on the smaller scale that can still influence the forecast beyond the usual influence time. Moreover, if the factors that contribute to this extended influence are assimilated and well represented in the initial conditions, forecast uncertainty is decreased and the resulting forecast is more accurate. Understanding the processes that modulate forecast uncertainty beyond 24 hours is important in the context of ensemble design, since these can significantly impact our interpretation of the resulting ensemble spread, which is still the most common measure of uncertainty.

There are a number of limitations with the results presented in this thesis. Firstly, although the analyzed cases are representative, the results obtained may not hold for every strong or weak synoptic forcing case. Secondly, this work focuses on summertime convection over Germany, so the findings may not necessarily hold true for other regions, seasons and weather conditions. Additionally, the ensembles comprise only 20, 40 or 120 ensemble members, depending on the chapter, which has been determined to sufficiently represent certain forecast distribution properties such as the ensemble mean or standard deviation for the latter. However, it is not sufficient for capturing more sensitive aspects, especially those related to extreme events. The chosen number of 1000 bootstrapping samples may also not be sufficient to determine whether convergence was reached, but it is approximately one order of magnitude larger than the size of the sampled ensemble, which Craig et al. (2022) found to be sufficient. Another limitation is that the sampling error convergence analysis is performed only for one specific location. Nevertheless, our results are expected to qualitatively hold for most of the model domain, as explained in chapter 5. Finally, the memory mechanisms discussed in the last part of the thesis are not fully investigated and therefore need to be studied in more detail in future work. In summary, despite these acknowledged limitations, the relationships elucidated here between weather regime, forecast distributions, forecast uncertainty and its evolution are expected to be broadly applicable across various contexts.

This thesis provides significant progress towards understanding the evolution of forecast uncertainty, however there is scope for further investigation. To begin with, improvements in the design of the PSP scheme should be tested, aimed at containing unwanted effects caused by the scheme, like excessive drying of the boundary layer. This could be done, for example, by adapting the vertical tapering profile of the perturbations to prevent excessive mixing around the top of the boundary layer. Some preliminary tests show promising results, but a more systematic assessment is needed, which the DWD is currently working on. Moreover, the effect of the scheme in combination with other model perturbations, especially if stochastic, should be studied before operational usage. Since the scheme is designed to represent only a specific source of model uncertainty, it not expected to substitute established approaches like parameter perturbations, but rather to be combined with other physically-based schemes, e.g. the stochastic shallow convection scheme developed by Sakradzija and Klocke (2018) in ICON. Matsunobu et al. (2024) recently found that complementing PSP by perturbed parameters in the microphysics scheme shows an additive effect on spatial error and spread for a characteristic case study. Therefore, an ensemble design with such complementing uncertainty sources seems to be the way forward in the field of model uncertainty representation.

This thesis, alongside Craig et al. (2022), Tempest et al. (2023) and Tempest et al. (2024), showed that the required ensemble size to accurately estimate the properties of forecast distributions, and hence uncertainty, strongly depends on the property of interest. Extreme, rare events will become more frequent in a warming climate, which calls for a larger ensemble size in probabilistic forecasting. Furthermore, an efficient transition to renewable energy production demands accurate probabilistic forecasts of different variables, like cloud cover and wind speed. Such multi-variate distributions were not studied in this thesis, but a larger ensemble size is expected to be needed for the sampling uncertainty to decrease to a useful level. Another example of such a multi-variate problem is the prediction of storm surge, which is becoming a more dangerous threat as the sea level rises, affecting millions of people around the globe.

The recent fast development of data-driven models, which show a forecast skill comparable, or even superior to conventional forecast models (e.g. Bi et al., 2023; Lam et al., 2022), combined with the huge advantage that, once trained, Artificial Intelligence (AI) models require much less computational effort to compute a weather forecast, seems to open new opportunities for ensemble forecasting. Ensemble sizes of thousands of members do not sound as impossible as before, which could be a significant breakthrough in the history of weather prediction. This thesis shows that the sampling error can be minimized in a very large ensemble, even for extreme, rare events. However, the question arises whether these events can be represented, since state-of-the-art data-based models usually learn from the past evolution of the weather, and future extreme events may be different, especially in a changing climate. Nevertheless, from early warnings of high-impact weather events to extremely accurate short-term forecasts with many practical applications, AI could help mitigate the impacts of climate change on humanity and make it more resilient.

Appendix A

Additional figures for chapter 5

The interested reader can see Figures 5.1, 5.2 and 5.5 with the full set of variables in this appendix. Figures A.1 and A.2 show the convergence of the confidence interval for the weak and strong forging regimes. Figures A.3 and A.4 show the maps of uncertainty for surface variables, while Figures A.5 and A.6 show the maps of uncertainty for mid-tropospheric variables, for the two forcing regimes.

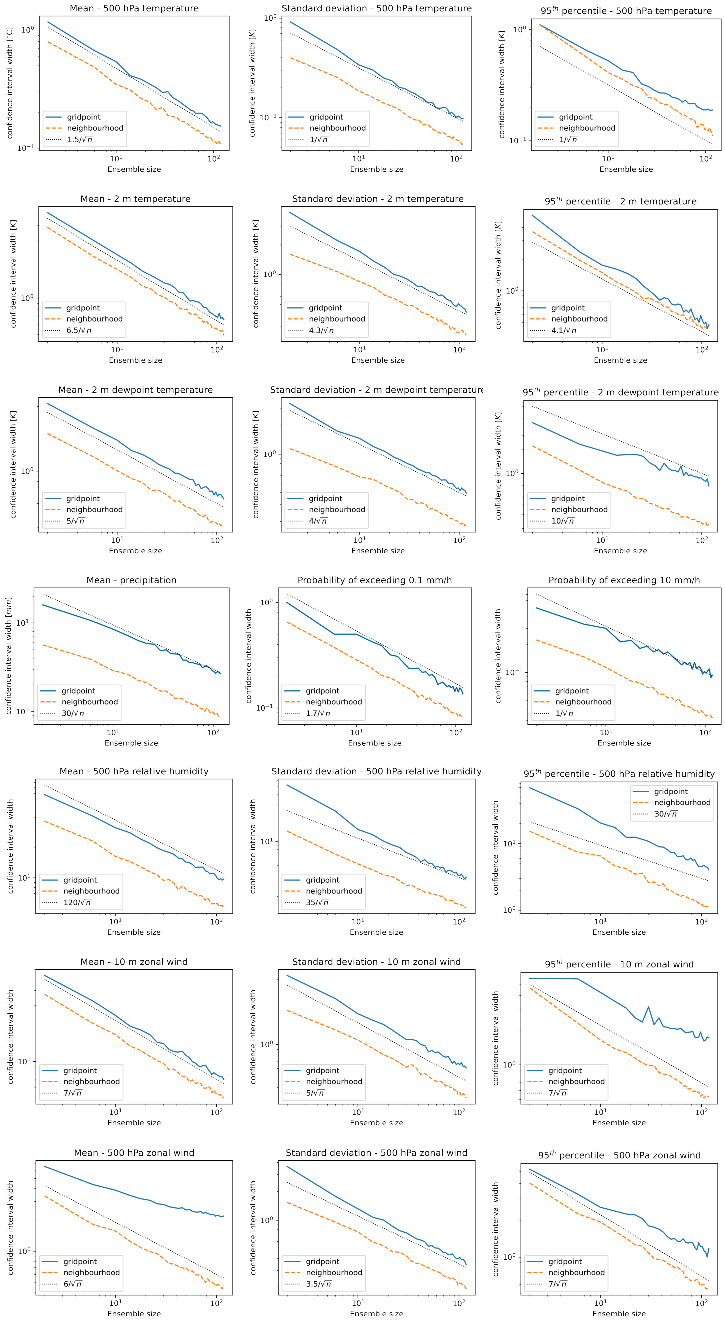

Figure A.1: Width of the 95% confidence intervals for the mean, standard deviation and 95^{th} percentile (columns), for a range of variables (rows, see panel titles). For precipitation, probability of exceeding 0.1 mm and 10 mm is shown instead of the standard deviation and the 95^{th} percentile. Forecast quantities are computed for 10 June at 14 UTC for a gridpoint near Reutlingen. The dashed line shows the reference curve with slope $N^{-1/2}$, fitted by eye. Gridpoint: single gridpoint forecast, neighbourhood: 10-km radius neighbourhood forecast.

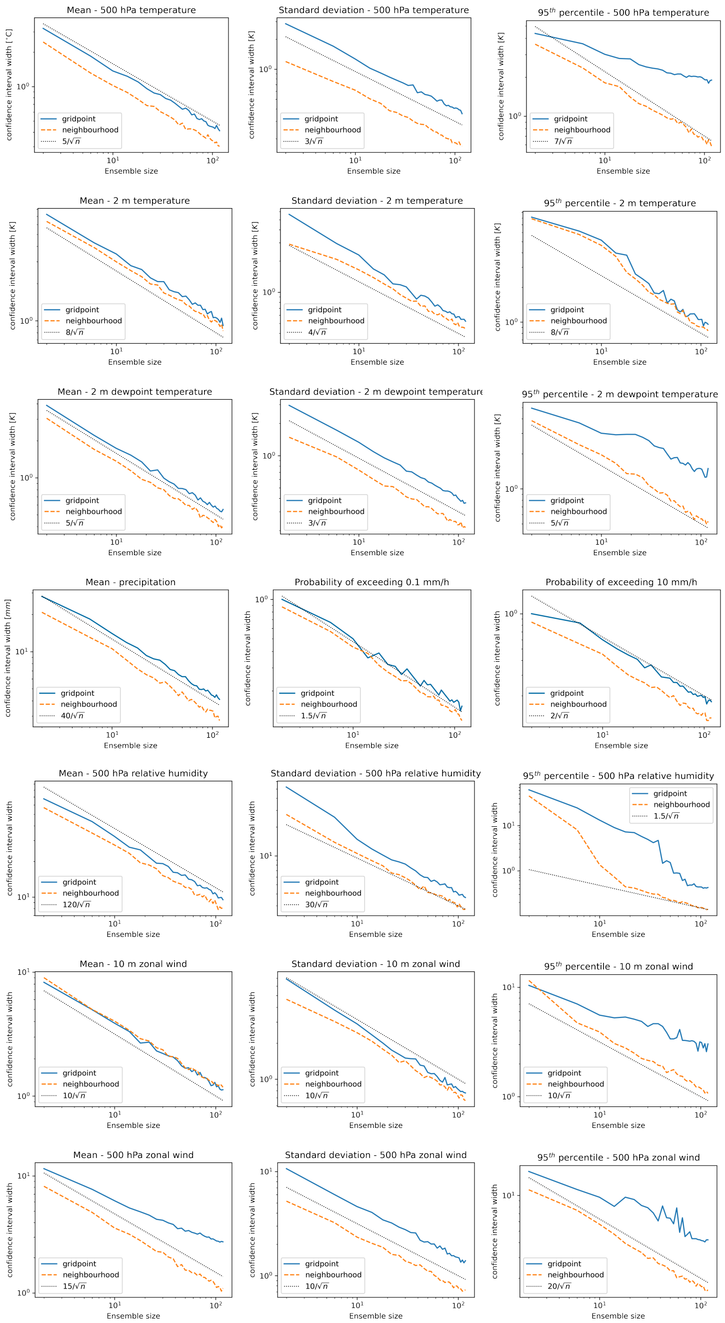

Figure A.2: As in fig. A.1, but for 29 June 2021.

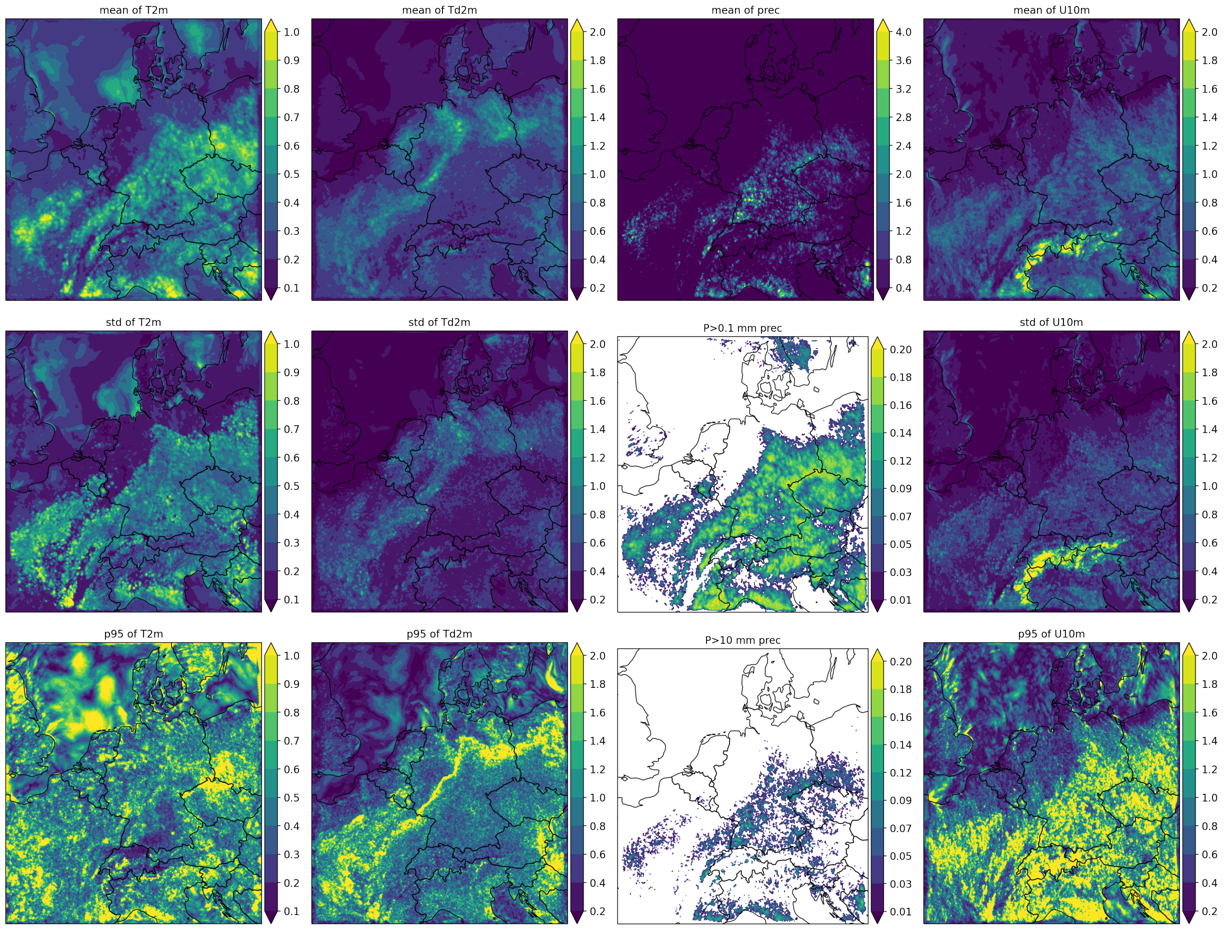

Figure A.3: Width of the 95% confidence interval for the mean (top row), standard deviation (middle row) and 95^{th} percentile (bottom row) of (from left to right) temperature and dew-point temperature at 2m in Kelvin, precipitation in millimeters per hour and zonal wind at 10 meters in meters per second on 10 June 2021 at 14 UTC. For precipitation, probability of exceeding 0.1 mm and 10 mm is shown instead of the standard deviation and the 95^{th} percentile. In white regions, none of the ensemble members exceeded the threshold. In white regions, none of the ensemble members exceeded the threshold.

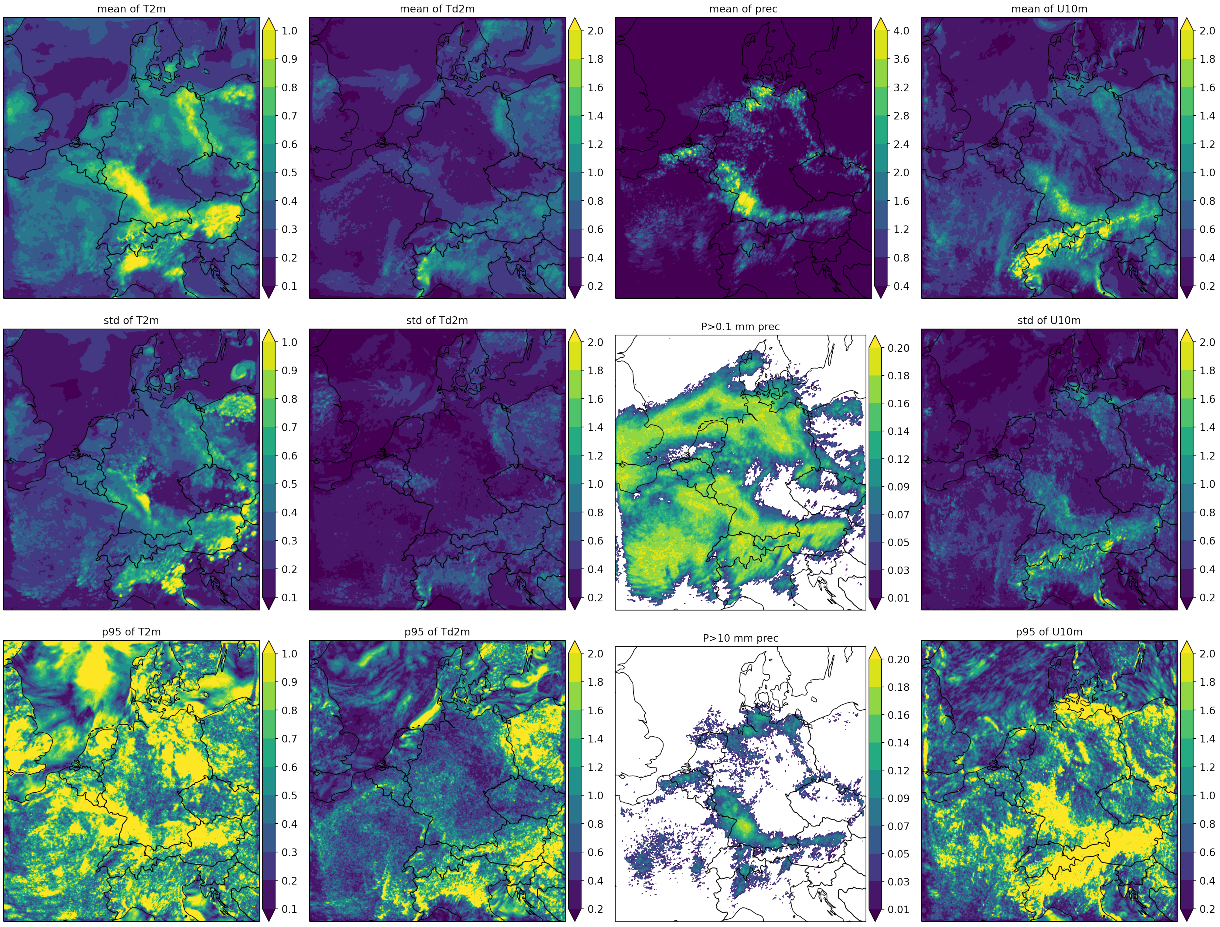

Figure A.4: As in figure A.3, but for 29 June 2021 at 14 UTC.

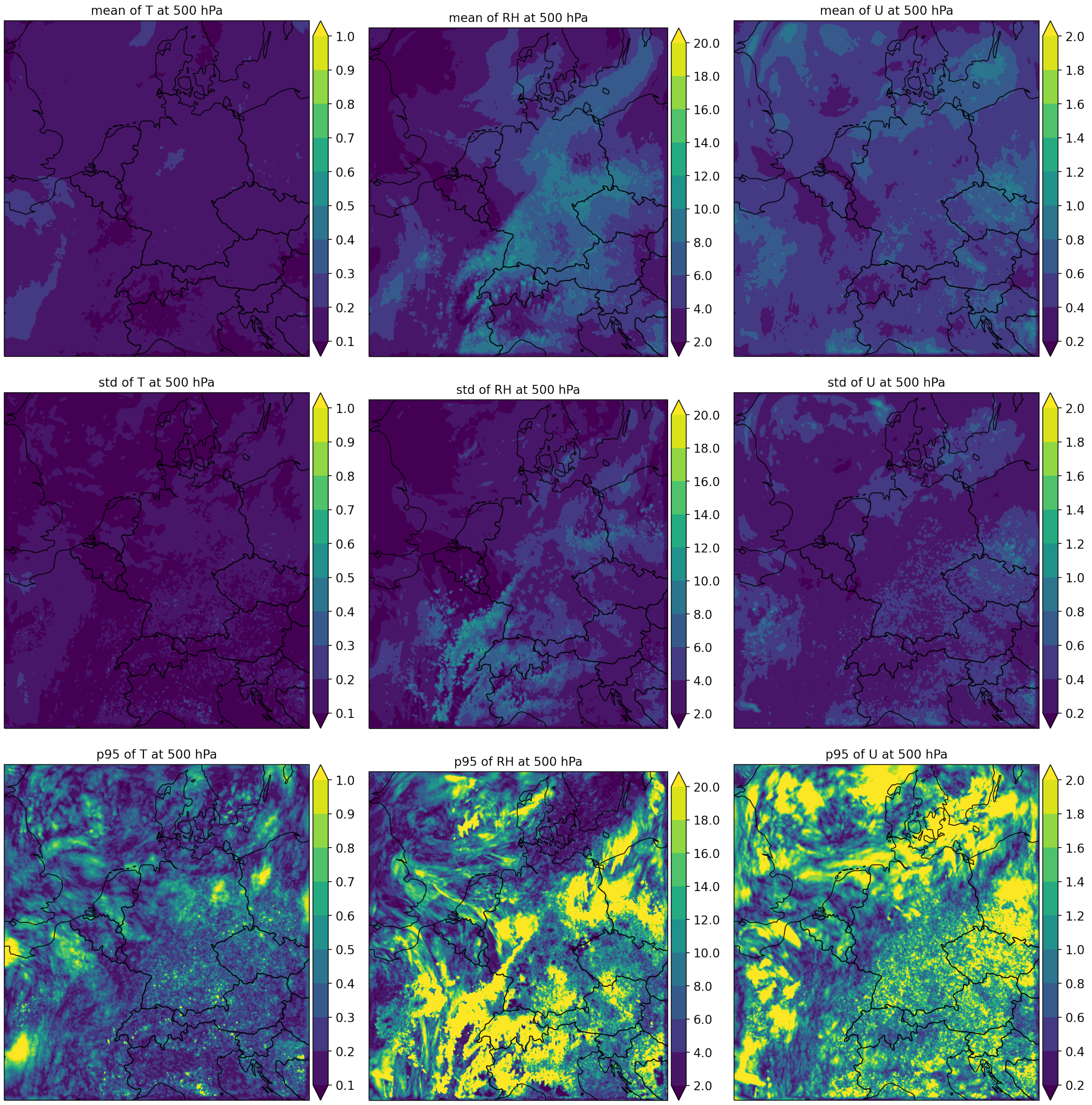

Figure A.5: Width of the 95% confidence interval for the mean (top row), standard deviation (middle row) and 95^{th} percentile (bottom row) of (from left to right) temperature in Kelvin, relative humidity and zonal wind in meters per second, all at 500 hPa, on 10 June 2021 at 14 UTC

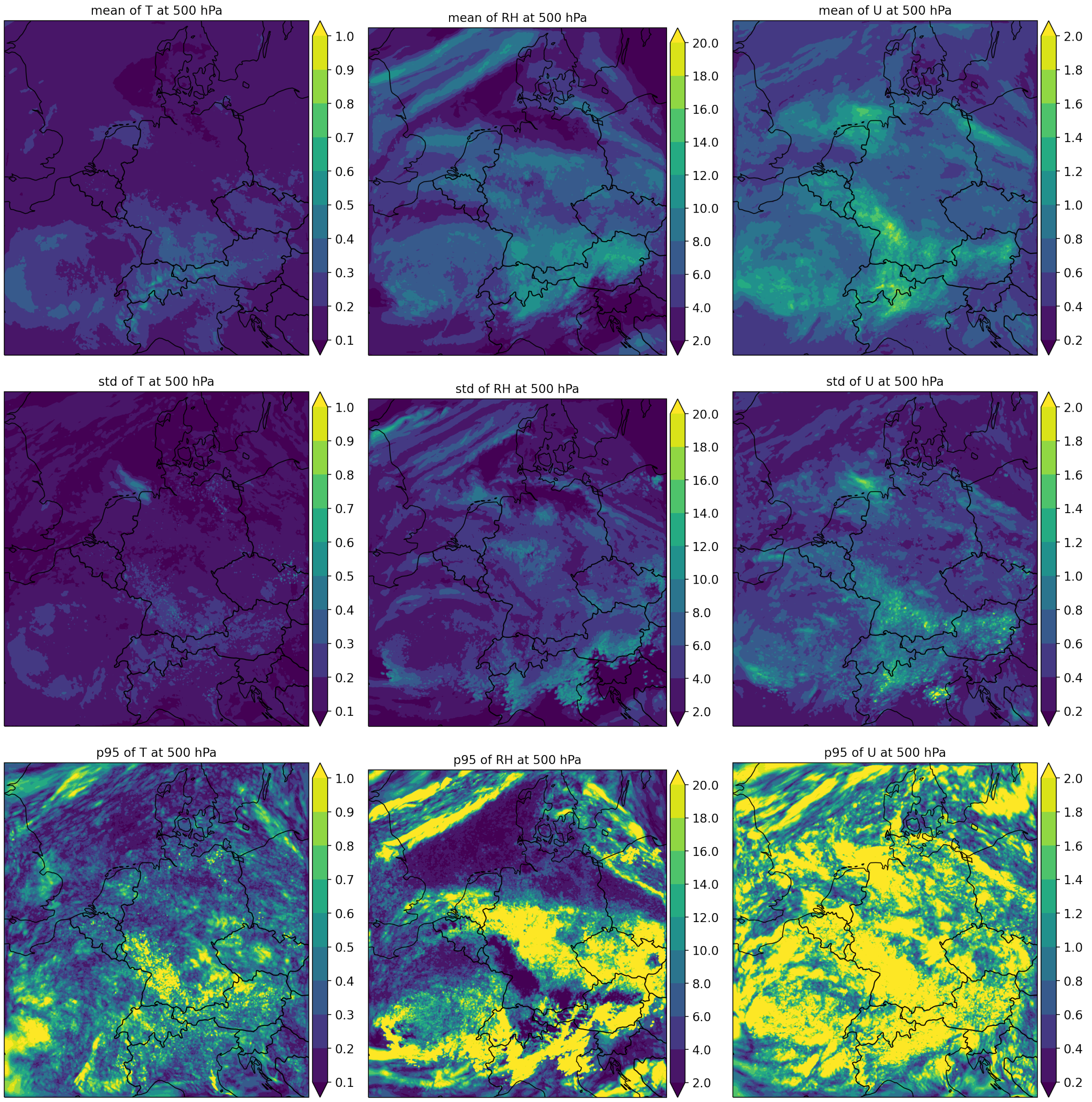

Figure A.6: As in Figure A.5, but for 29 June 2021 at 14 UTC.

Appendix B

Additional figures for chapter 6

The interested reader can see the time series of precipitable water difference between the PSP and the reference simulations in this appendix. Additionally, figures like Figure 6.2 are shown for the full set of variables and both cases .

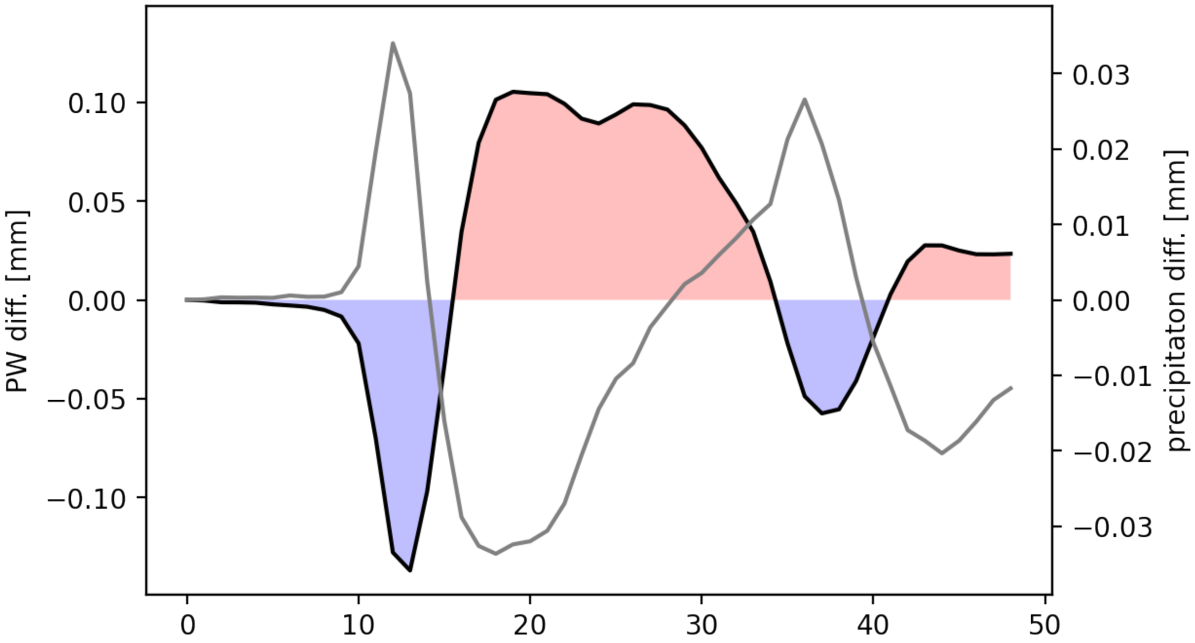

Figure B.1: Time series of the difference in precipitable water (black line with coloured area) and hourly precipitation (grey line) between the PSP and reference simulations.

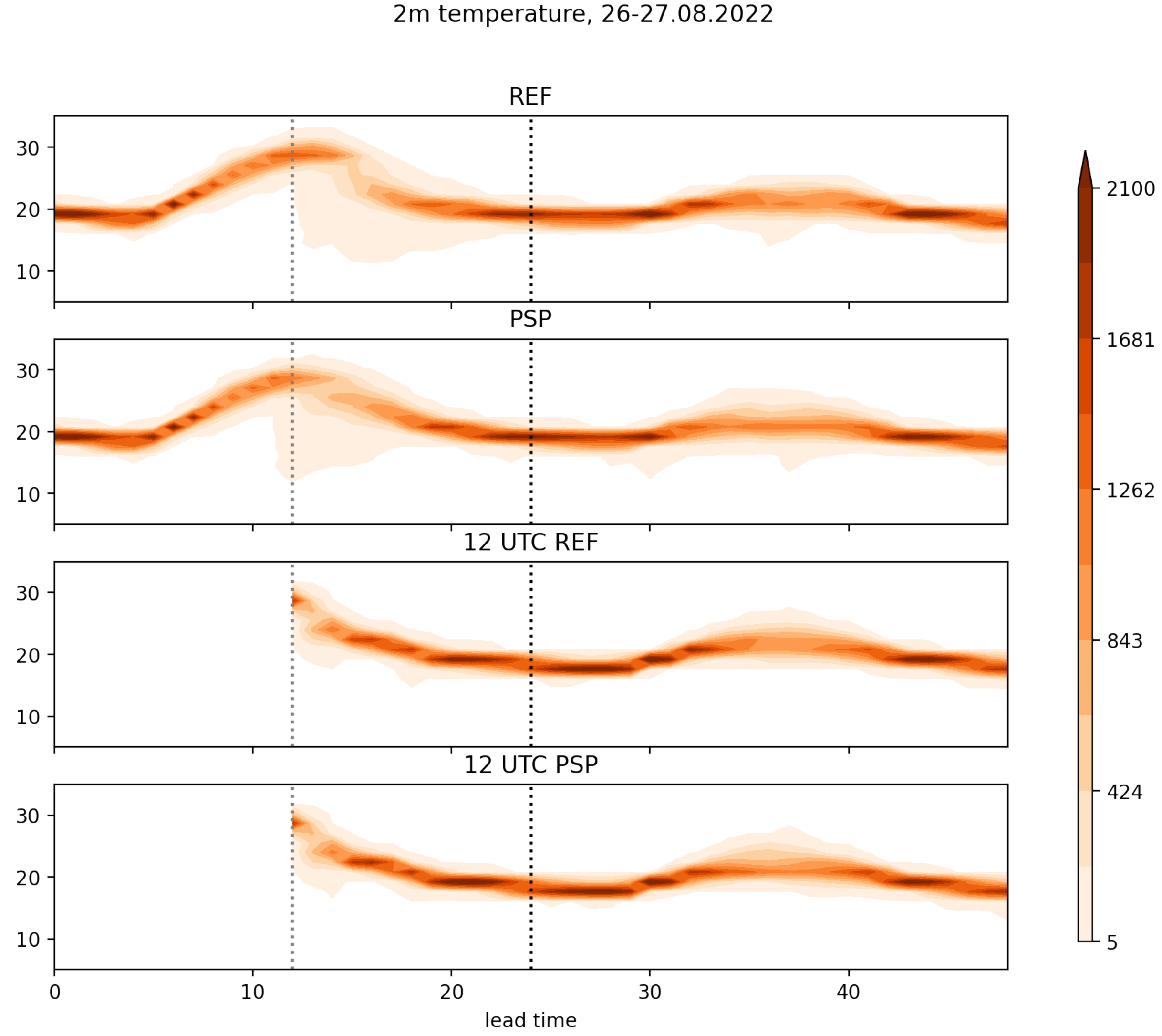

Figure B.2: 2D histogram of 2m temperature (in degrees Celsius) for a 10-km neighbourhood around a location in Southern Germany on 18 and 19 August (weak forcing), for the 4 simulations: reference, PSP and their 12 UTC equivalents. The shading indicates the frequency of occurrence.

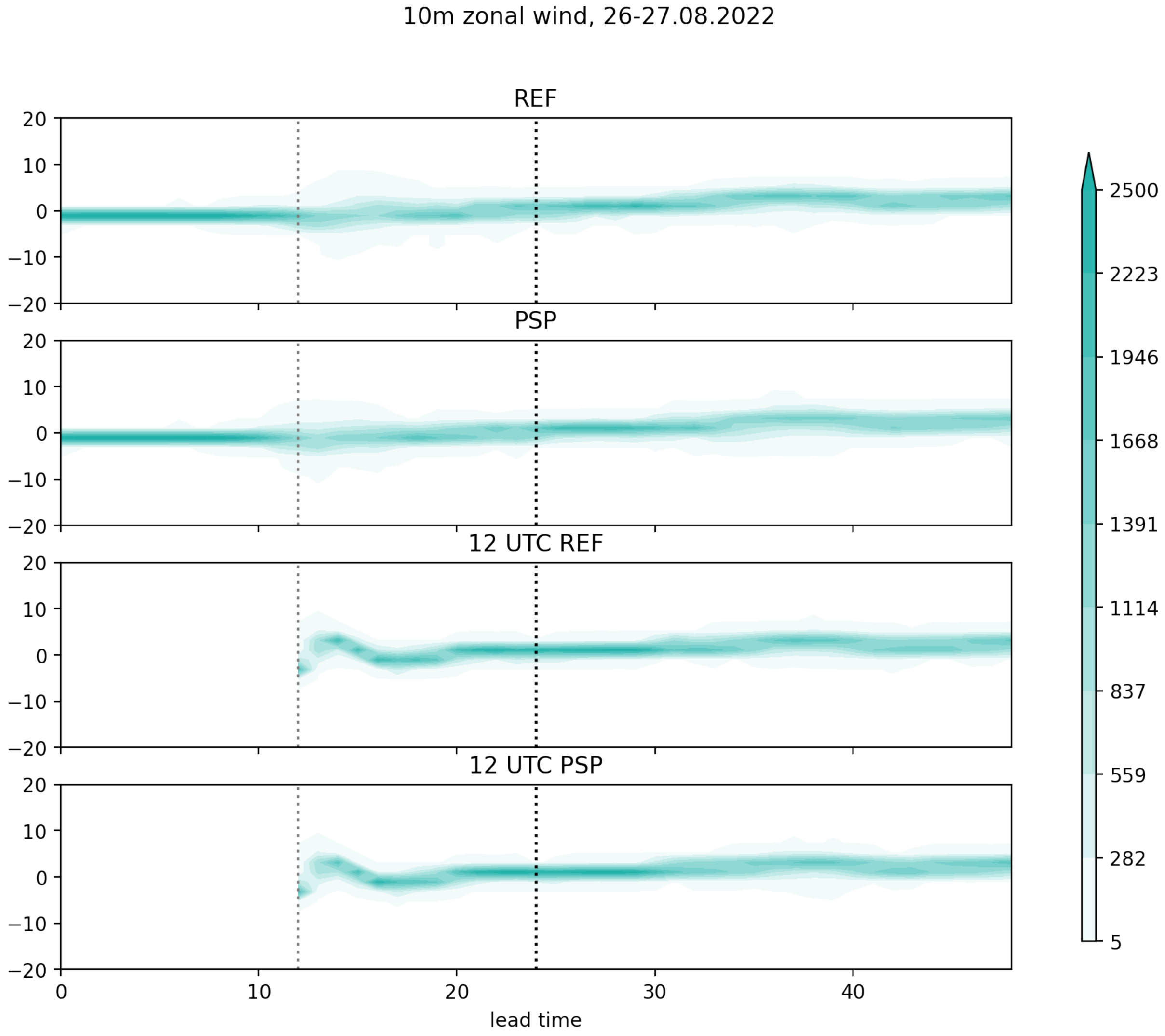

Figure B.3: 2D histogram of 10m wind (in m/s) for a 10-km neighbourhood around a location in Southern Germany on 18 and 19 August (weak forcing), for the 4 simulations: reference, PSP and their 12 UTC equivalents. The shading indicates the frequency of occurrence.

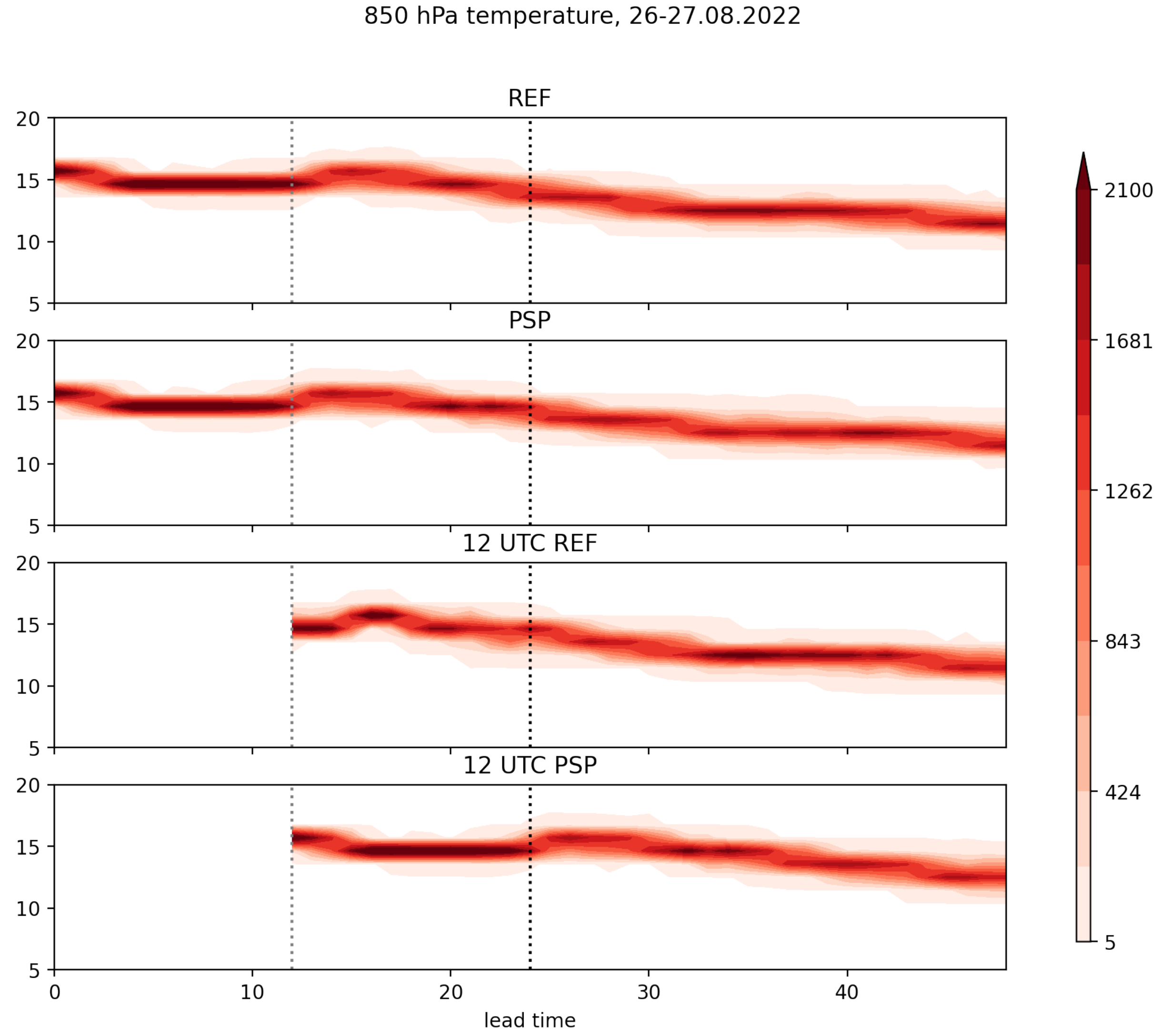

Figure B.4: 2D histogram of temperature at 850 hPa (in degrees Celsius) for a 10-km neighbourhood around a location in Southern Germany on 18 and 19 August (weak forcing), for the 4 simulations: reference, PSP and their 12 UTC equivalents. The shading indicates the frequency of occurrence.

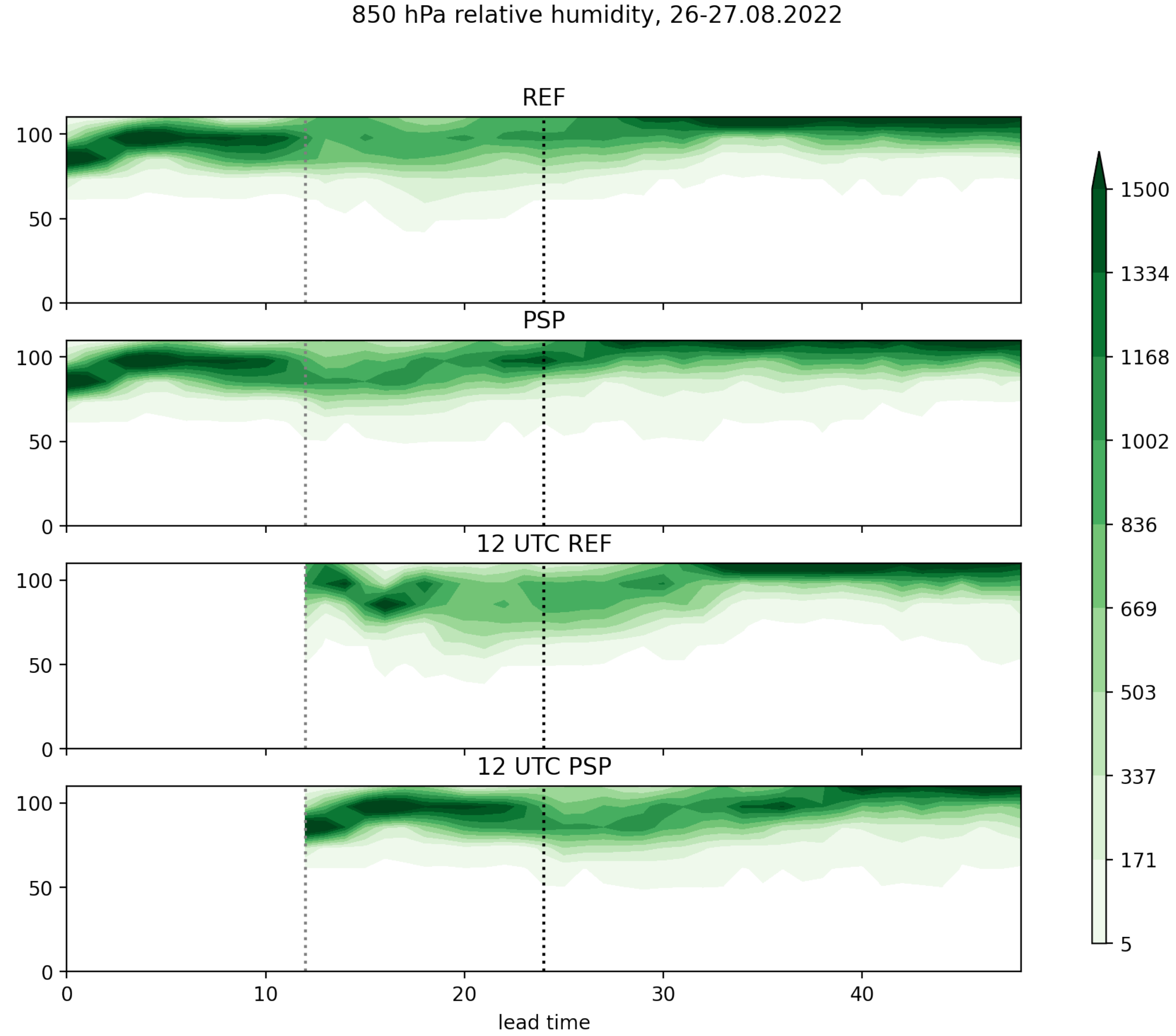

Figure B.5: 2D histogram of relative humidity at 850 hPa for a 10-km neighbourhood around a location in Southern Germany on 18 and 19 August (weak forcing), for the 4 simulations: reference, PSP and their 12 UTC equivalents. The shading indicates the frequency of occurrence.

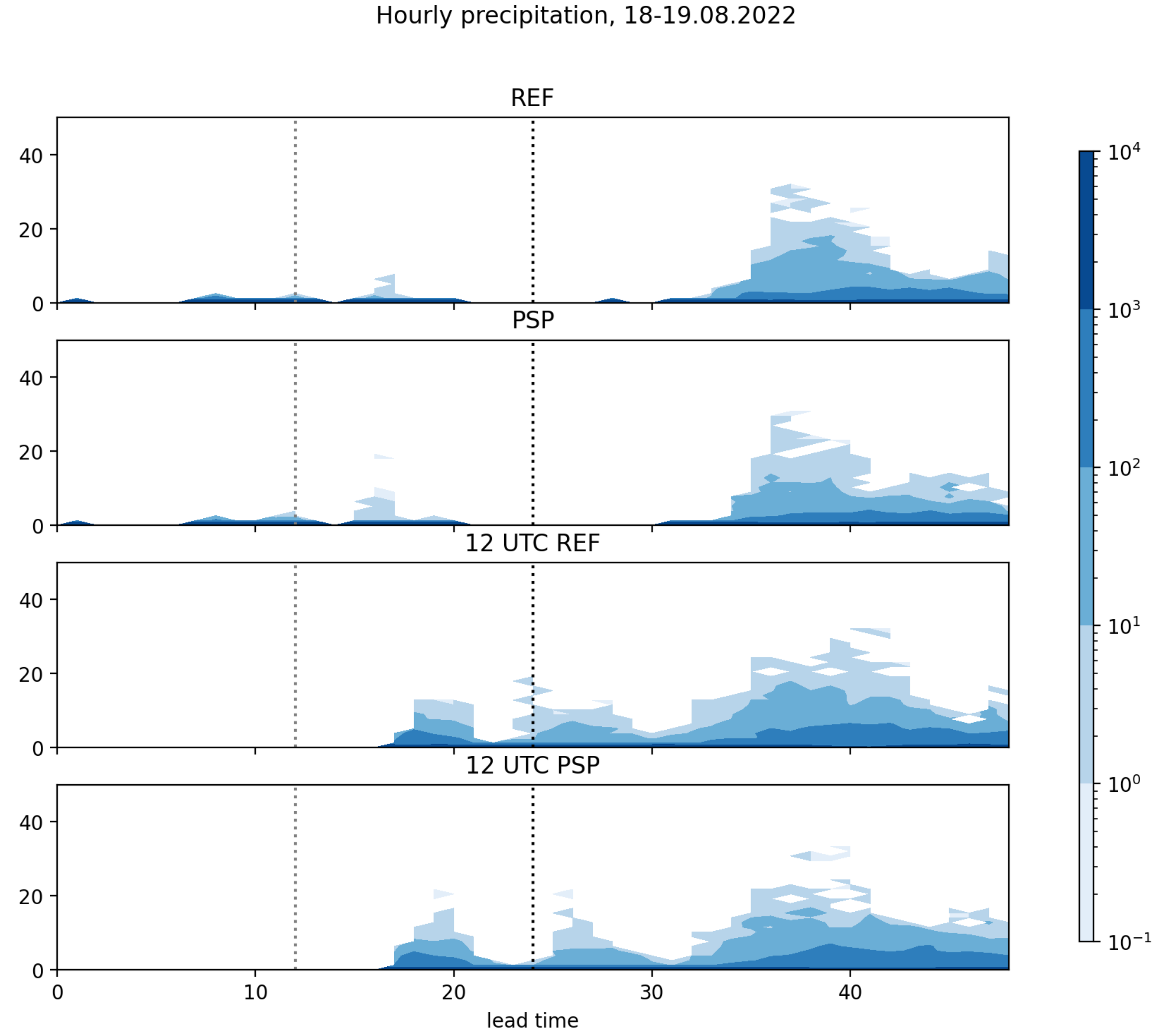

Figure B.6: 2D histogram (the x axis represents lead time in hours) of hourly precipitation (in mm) for a 10-km neighbourhood around a location in Southern Germany on 18 and 19 August (strong forcing), for the 4 simulations: reference, PSP and their 12 UTC equivalents. The shading indicates the frequency of occurrence.

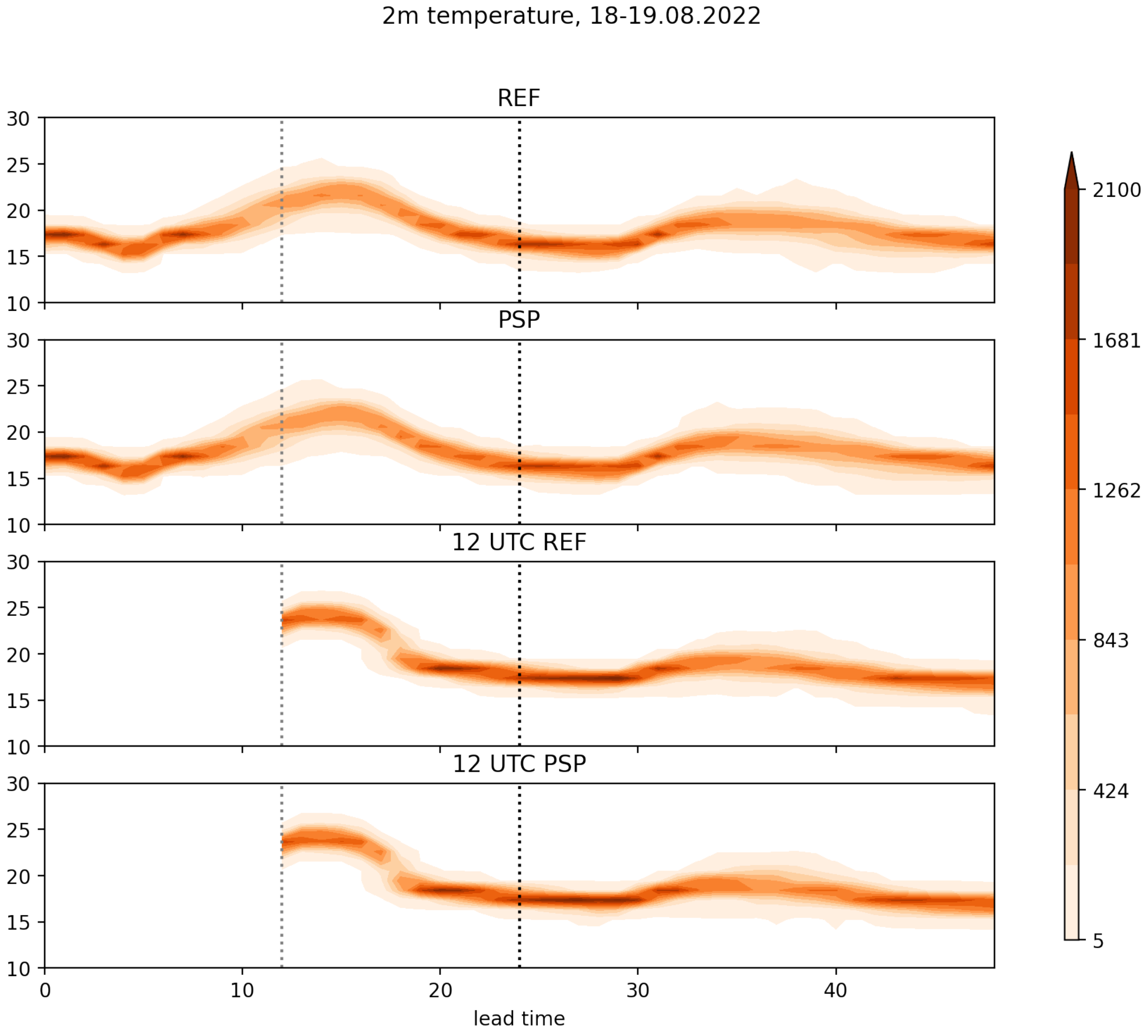

Figure B.7: 2D histogram of 2m temperature (in degrees Celsius) for a 10-km neighbourhood around a location in Southern Germany on 18 and 19 August (strong forcing), for the 4 simulations: reference, PSP and their 12 UTC equivalents. The shading indicates the frequency of occurrence.

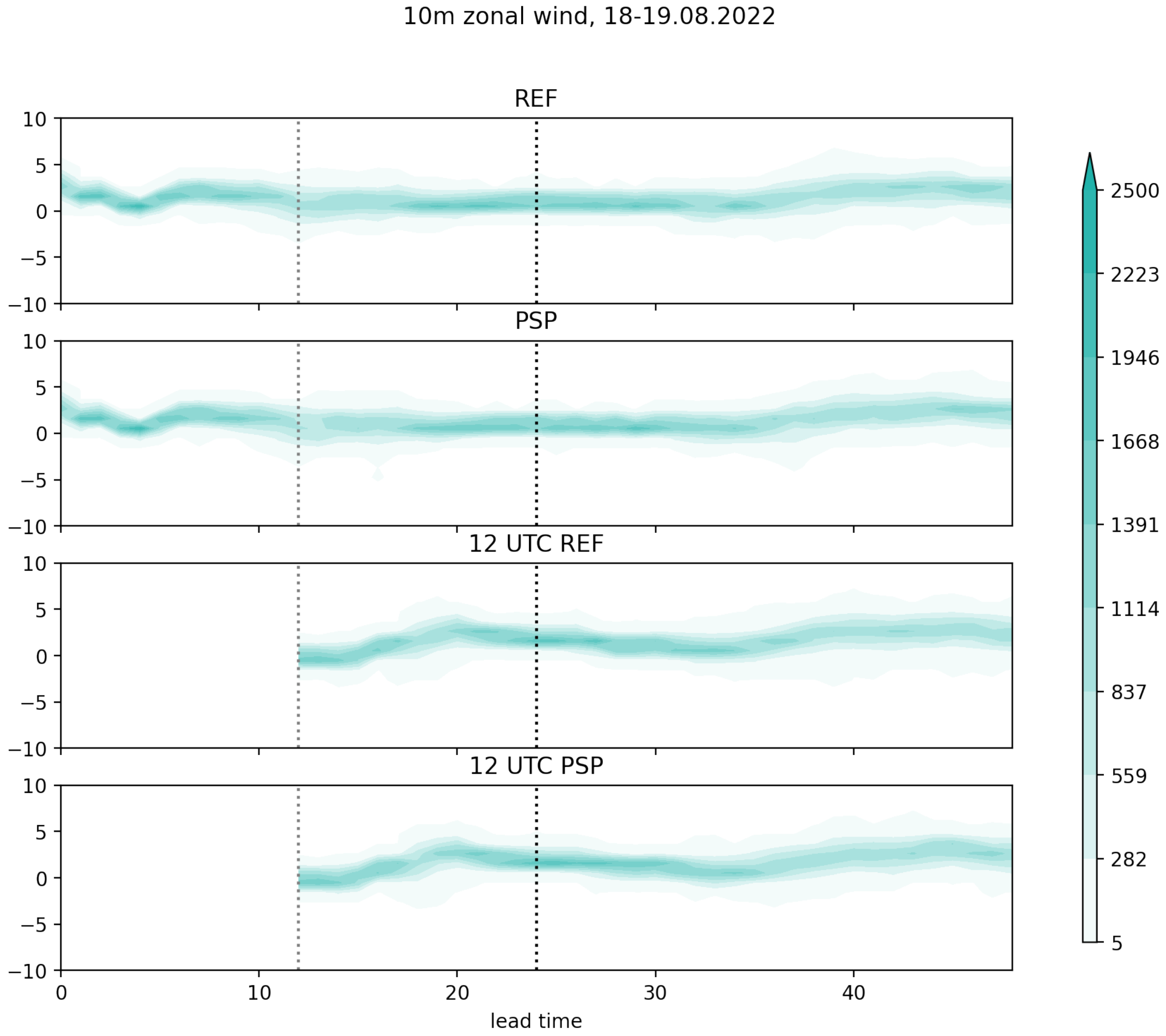

Figure B.8: 2D histogram of 10m wind (in m/s) for a 10-km neighbourhood around a location in Southern Germany on 18 and 19 August (strong forcing), for the 4 simulations: reference, PSP and their 12 UTC equivalents. The shading indicates the frequency of occurrence.

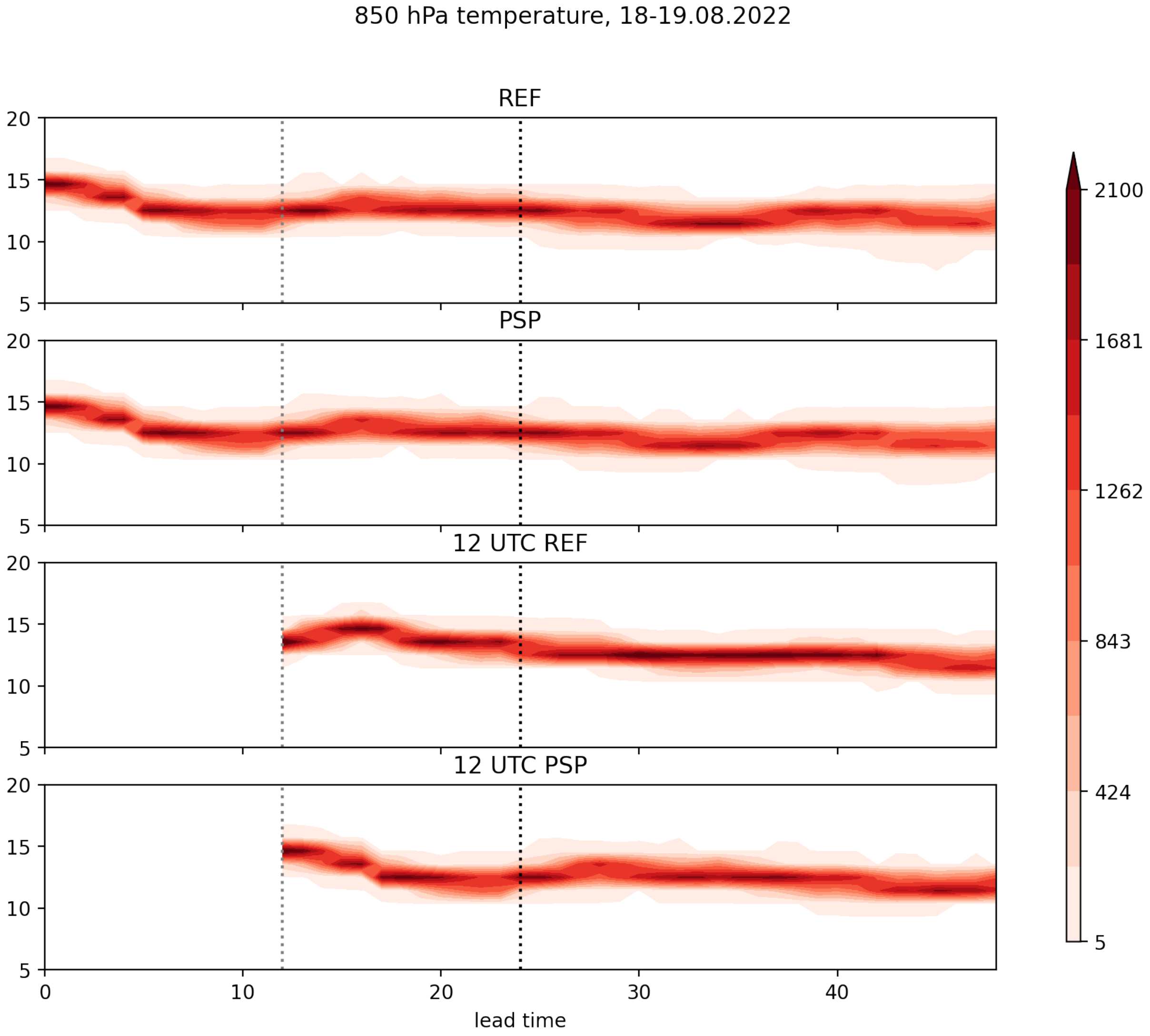

Figure B.9: 2D histogram of temperature at 850 hPa (in degrees Celsius) for a 10-km neighbourhood around a location in Southern Germany on 18 and 19 August (strong forcing), for the 4 simulations: reference, PSP and their 12 UTC equivalents. The shading indicates the frequency of occurrence.

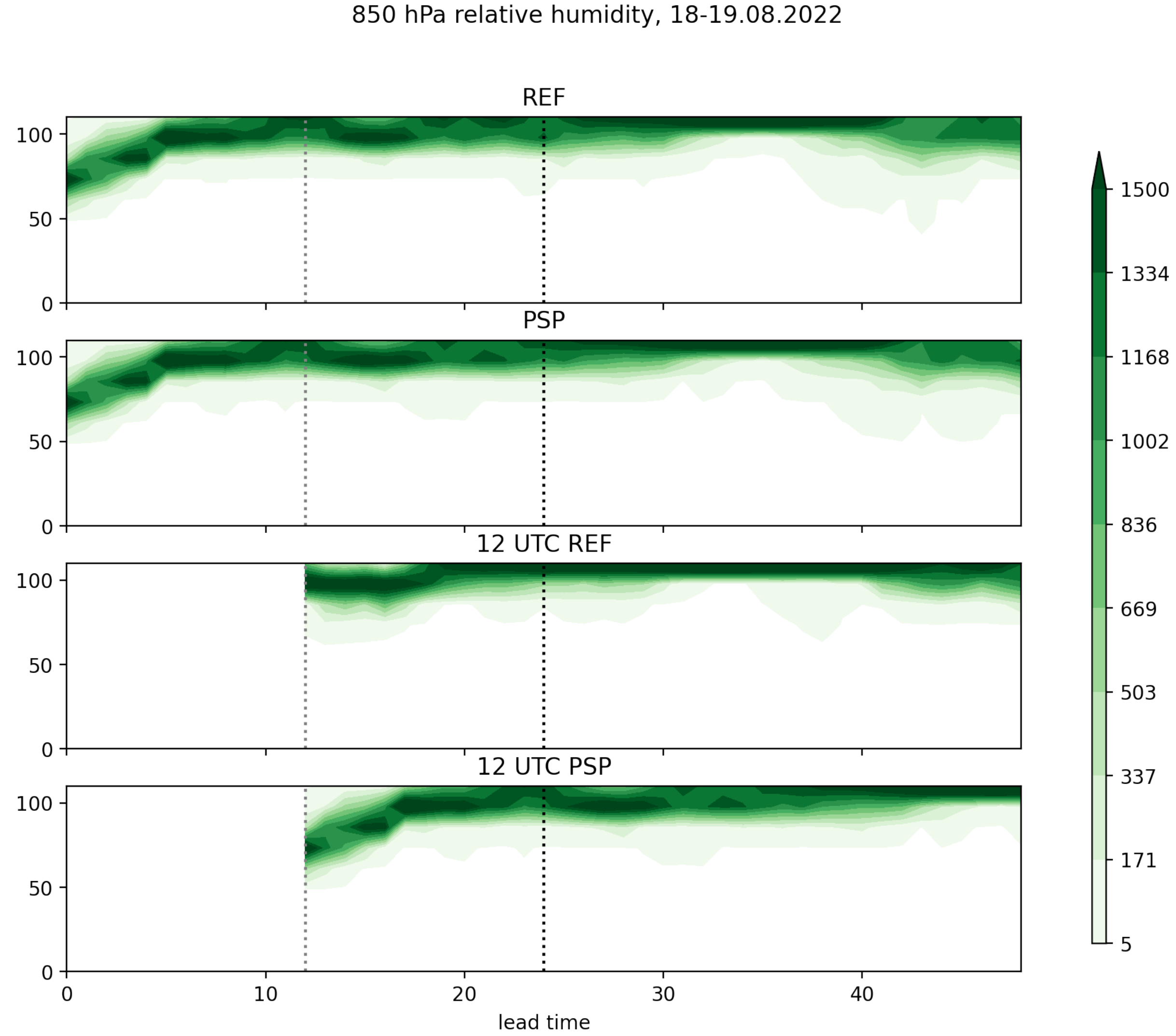

Figure B.10: 2D histogram of relative humidity at 850 hPa for a 10-km neighbourhood around a location in Southern Germany on 18 and 19 August (strong forcing), for the 4 simulations: reference, PSP and their 12 UTC equivalents. The shading indicates the frequency of occurrence.

List of Figures

1.1 Conceptual representation of an ensemble forecast 4
1.2 Conceptual model of the evolution of a distribution 7
1.3 Three distribution types 11
1.4 Time evolution of the distribution of saturation deficit 12
1.5 Sampling error convergence example 13

2.1 ICON-D2 domain 18
2.2 Schematic of PBL turbulence 19
2.3 Example of the perturbation field created by the PSP scheme 21

3.1 Overview of daily precipitation and convective forcing classification 26
3.2 Synoptic charts for the case studies of the big ensemble simulations 28
3.3 Flow chart of 48h simulations 29
3.4 Synoptic charts for the case studies with extended forecasts 29

4.1 Overview of daily total precipitation and convective adjustment timescale . 34
4.2 Composite time series of hourly precipitation amount and spread - weak and strong forcing 35
4.3 Time series of FSS of precipitation - weak forcing 36
4.4 Probabilistic verification scores of near-surface variables 38

5.1 Convergence of sampling error of temperature - weak forcing 44
5.2 Convergence of sampling error of temperature - strong forcing 45
5.3 Convergence comparison - ICON and idealized model 46
5.4 Uncertainty maps for the mean of T2m with the corresponding histograms at various locations 48
5.5 Uncertainty maps for humidity variables 50
5.6 2D histograms of surface variables 52
5.7 Two-dimensional histograms of mid-tropospheric variables 53

6.1 Time series of hourly precipitation 60
6.2 2D histogram of precipitation 63
6.3 Maps of total accumulated precipitation and difference PSP-REF - weak forcing 64

6.4 Maps of nighttime precipitation 65
6.5 Maps of nighttime CAPE, CIN and humidity difference 66
6.6 Sea level pressure difference after 24 and 48 hours 67
6.7 Maps of total accumulated precipitation and difference PSP-REF - strong forcing 68

A.1 Convergence plots of sampling error - weak forcing 78
A.2 Convergence plots of sampling error - strong forcing 79
A.3 Maps of uncertainty for surface variables - weak forcing 80
A.4 Maps of uncertainty for surface variables - strong forcing 81
A.5 Maps of uncertainty for mid-tropospheric variables - weak forcing 82
A.6 Maps of uncertainty for mid-tropospheric variables - strong forcing 83

B.1 Time series of precipitable water difference 85
B.2 2D histogram of T2m - weak forcing 86
B.3 2D histogram of U10m - weak forcing 87
B.4 2D histogram of T850 - weak forcing 88
B.5 2D histogram of rh850 - weak forcing 89
B.6 2D histogram of precipitation - strong forcing 90
B.7 2D histogram of T2m - strong forcing 91
B.8 2D histogram of U10m - strong forcing 92
B.9 2D histogram of T850 - strong forcing 93
B.10 2D histogram of rh850 - strong forcing 94

List of Tables

1.1 Classification of forecast variables 11

3.1 Description of the ensemble simulations 25

List of Abbreviations

AI Artificial Intelligence. 75

AROME Application of Research to Operations at Mesoscale. 32

CAPE Convective Available Potential Energy. 9, 33, 34, 57, 58, 61, 65, 69, 73

CDF Cumulative Distribution Function. 21

CI Confidence Interval. 21

CIN Convective Inhibition. 65

COSMO Consortium for Small-scale Modeling. 20, 21, 32, 35

CRPS Continuous Ranked Probability Score. 37, 39

D2 Domain 2. 15, 17, 18, 26, 27, 33, 35, 36, 40–43, 54, 72, 95

DWD Deutscher Wetterdienst. 15, 17, 18, 25, 26, 35, 70, 74

ECMWF European Center for Medium-range Weather Forecasting. 3, 42

EDA Ensembles of Data Assimilations. 3

EPS Ensemble Prediction System. v, vi, 3, 5, 8, 17, 18, 26, 27, 31–33, 35, 40–43, 70–72

FSS Fractions Skill Score. 33, 35–37, 40

IBC Initial and Boundary Conditions. 18, 25

IC Initial Conditions. 5, 18

ICON Icosahedral Nonhydrostatic. v, 15, 17, 18, 20, 25–28, 33, 35, 36, 40–43, 45, 46, 50, 54, 70–72, 75, 95

KENDA Kilometer Scale Ensemble Data Assimilation. 18, 26, 27

LAM Limited Area Model. v, 70, 71

LBC Lateral Boundary Conditions. 5, 18

MCS Mesoscale Convective System. 27, 47

MOGREPS-UK Met Office convective-scale ensemble. 6

NCEP National Centers for Environmental Prediction. 3

NWP Numerical Weather Prediction. 3, 5–7, 14, 17, 19, 31, 39, 41–43, 46, 54, 71, 72

PBL Planetary Boundary Layer. 17, 19, 32, 39, 40, 61, 72

PDF Probability Distribution Function. 3, 4

PSP Physically Based Stochastic Perturbation Scheme. v, 8, 14, 15, 17, 20, 21, 25–28, 32–35, 37–40, 61–64, 66–69, 71–75, 95

RMS Root Mean Square. 14

RMSE Root Mean Square Error. 37–39

SM Soil Moisture. 58, 59

SPPT Stochastically Perturbed Parameterization Tendencies. 32

SV Singular Vectors. 3

SYNOP Surface Synoptic Observations. 37, 39

UTC Coordinated Universal Time. 28

WRF Weather Research and Forecasting. 32

Bibliography

Abbe, C., The physical basis of long-range weather forecasts, *Monthly Weather Review*, 29 (12):551–561, 1901.

Bachmann, K., Keil, C., Craig, G. C., Weissmann, M., and Welzbacher, C. A., Predictability of deep convection in idealized and operational forecasts: Effects of radar data assimilation, orography, and synoptic weather regime, *Monthly Weather Review*, 148(1): 63–81, 2020. doi: 10.1175/MWR-D-19-0045.1.

Bannister, R. N., Migliorini, S., Rudd, A. C., and Baker, L. H., Methods of investigating forecast error sensitivity to ensemble size in a limited-area convection-permitting ensemble, *Geoscientific Model Development Discussions*, 2017:1–38, 2017. doi: 10.5194/gmd-2017-260.

Bauer, P., Thorpe, A., and Brunet, G., The quiet revolution of numerical weather prediction, *Nature*, 525(7567):47–55, 2015.

Baur, F., Keil, C., and Barthlott, C., Combined effects of soil moisture and microphysical perturbations on convective clouds and precipitation for a locally forced case over central europe, *Quarterly Journal of the Royal Meteorological Society*, 148(746):2132–2146, 2022. doi: https://doi.org/10.1002/qj.4295.

Ben Bouallègue, Z., Theis, S. E., and Gebhardt, C., Enhancing COSMO-DE ensemble forecasts by inexpensive techniques, *Meteorologische Zeitschrift*, 22(1):49–59, 02 2013. doi: 10.1127/0941-2948/2013/0374.

Berner, J., Achatz, U., Batté, L., Bengtsson, L., de la Cámara, A., Christensen, H. M., Colangeli, M., Coleman, D. R. B., Crommelin, D., Dolaptchiev, S. I., Franzke, C. L. E., Friederichs, P., Imkeller, P., Järvinen, H., Juricke, S., Kitsios, V., Lott, F., Lucarini, V., Mahajan, S., Palmer, T. N., Penland, C., Sakradzija, M., von Storch, J.-S., Weisheimer, A., Weniger, M., Williams, P. D., and Yano, J.-I., Stochastic Parameterization: Toward a New View of Weather and Climate Models, *Bulletin of the American Meteorological Society*, 98(3):565 – 588, 2017. doi: 10.1175/BAMS-D-15-00268.1.

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., and Tian, Q., Accurate medium-range global weather forecasting with 3d neural networks, *Nature*, 619:533–538, 1 2023. doi: 10.1038/s41586-023-06185-3.

Bierdel, L., Friederichs, P., and Bentzien, S., Spatial kinetic energy spectra in the convection-permitting limited-area NWP model COSMO-DE, *Meteorologische Zeitschrift*, 21(3):245–258, jun 2012. ISSN 0941-2948. doi: 10.1127/0941-2948/2012/0319.

Bjerknes, V., Das problem der wettervorhersage, betrachtet vom standpunkte der mechanik und der physik, *Meteor. Z.*, 21:1–7, 1904.

Blahak, U. Current status of sinfony - the combination of nowcasting and numerical weather prediction on the convective scale at dwd. EMS2023, page 254, 2023. doi: https://doi.org/10.5194/ems2023-254.

Blake, B. T., Carley, J. R., Alcott, T. I., Jankov, I., Pyle, M. E., Perfater, S. E., and Albright, B., An Adaptive Approach for the Calculation of Ensemble Gridpoint Probabilities, *Weather and Forecasting*, 33(4):1063–1080, 2018. doi: 10.1175/WAF-D-18-0035.1.

Bolin, B., Numerical forecasting with the barotropic model 1, *Tellus*, 7(1):27–49, 1955.

Bouttier, F., Vié, B., Nuissier, O., and Raynaud, L., Impact of Stochastic Physics in a Convection-Permitting Ensemble, *Mon. Wea. Rev.*, 140(11):3706–3721, November 2012. ISSN 0027-0644, 1520-0493. doi: 10.1175/MWR-D-12-00031.1.

Buizza, R., Milleer, M., and Palmer, T. N., Stochastic representation of model uncertainties in the ecmwf ensemble prediction system, *Quarterly Journal of the Royal Meteorological Society*, 125(560):2887–2908, 1999. doi: https://doi.org/10.1002/qj.49712556006.

Buizza, R. Chapter 4 - predictability. In Ólafsson, H. and Bao, J.-W., editors, *Uncertainties in Numerical Weather Prediction*, page 119–146. Elsevier, 2021a. ISBN 978-0-12-815491-5. doi: 10.1016/B978-0-12-815491-5.00004-5.

Buizza, R. Chapter 3 - probabilistic view of numerical weather prediction and ensemble prediction. In Ólafsson, H. and Bao, J.-W., editors, *Uncertainties in Numerical Weather Prediction*, page 81–117. Elsevier, 2021b. ISBN 978-0-12-815491-5. doi: 10.1016/B978-0-12-815491-5.00003-3.

Buizza, R. and Palmer, T. N., The singular-vector structure of the atmospheric global circulation, *Journal of the Atmospheric Sciences*, 52(9):1434–1456, 1995.

Buizza, R., Leutbecher, M., and Isaksen, L., Potential use of an ensemble of analyses in the ecmwf ensemble prediction system, *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 134(637):2051–2066, 2008.

Charney, J. G., Fjörtoft, R., and Neumann, J. v., Numerical integration of the barotropic vorticity equation, *Tellus*, 2(4):237–254, 1950.

Clark, P., Roberts, N., Lean, H., Ballard, S. P., and Charlton-Perez, C., Convection-permitting models: A step-change in rainfall forecasting, *Meteor. Appl.*, 23:165–181, 2016.

Clark, P. A., Halliwell, C. E., and Flack, D. L. A., A Physically Based Stochastic Boundary Layer Perturbation Scheme. Part I: Formulation and Evaluation in a Convection-Permitting Model, *Journal of the Atmospheric Sciences*, 78(3):727 – 746, 2021. doi: 10.1175/JAS-D-19-0291.1.

Craig, G. C. and Cohen, B. G., Fluctuations in an Equilibrium Convective Ensemble. Part I: Theoretical Formulation, *Journal of the Atmospheric Sciences*, 63(8):1996–2004, aug 2006. ISSN 0022-4928. doi: 10.1175/JAS3709.1.

Craig, G. C., Keil, C., and Leuenberger, D., Constraints on the impact of radar rainfall data assimilation on forecasts of cumulus convection, *Quarterly Journal of the Royal Meteorological Society*, 138(663):340–352, 2012. doi: https://doi.org/10.1002/qj.929.

Craig, G. C., Puh, M., Keil, C., Tempest, K., Necker, T., Ruiz, J., Weissmann, M., and Miyoshi, T., Distributions and convergence of forecast variables in a 1,000-member convection-permitting ensemble, *Quarterly Journal of the Royal Meteorological Society*, 148(746):2325–2343, 2022. doi: https://doi.org/10.1002/qj.4305.

Davison, A. C. and Hinkley, D. V., *Bootstrap Methods and their Application*, Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

Done, J. M., Craig, G. C., Gray, S. L., Clark, P. A., and Gray, M. E. B., Mesoscale simulations of organized convection: Importance of convective equilibrium, *Quarterly Journal of the Royal Meteorological Society*, 132(616):737–756, 2006. doi: https://doi.org/10.1256/qj.04.84.

Dowell, D. C., Alexander, C. R., James, E. P., Weygandt, S. S., Benjamin, S. G., Manikin, G. S., Blake, B. T., Brown, J. M., Olson, J. B., Hu, M., Smirnova, T. G., Ladwig, T., Kenyon, J. S., Ahmadov, R., Turner, D. D., Duda, J. D., and Alcott, T. I., The High-Resolution Rapid Refresh (HRRR): An Hourly Updating Convection-Allowing Forecast Model. Part I: Motivation and System Description, *Weather and Forecasting*, 37(8): 1371–1395, 2022. doi: 10.1175/WAF-D-21-0151.1.

Du, J., Berner, J., Buizza, R., Charron, M., Houtekamer, P., Hou, D., Jankov, I., Mu, M., Wang, X., Wei, M., and Yuan, H. *Ensemble Methods for Meteorological Predictions*, page 99–149. Springer Berlin Heidelberg, Berlin, Heidelberg, 2019. ISBN 978-3-642-39925-1. doi: 10.1007/978-3-642-39925-1_13.

DWD. Deutschlandwetter im Sommer 2021. `https://www.dwd.de/DE/presse/pressemit--teilungen/DE/2021/20210830_deutschlandwetter_sommer2021_news.html`, 2022. Accessed: 18.05.2022.

Ebert, E. E., Neighborhood verification: A strategy for rewarding close forecasts, *Weather and Forecasting*, 24(6):1498–1510, December 2009. doi: 10.1175/2009waf2222251.1.

ESSL. The derecho and hailstorms of 18 August 2022. `https://www.essl.org/cms/the-derecho-and-hailstorms-of-18-august-2022/`, 2022. Accessed: 19.12.2023.

Felger, P. The impact of the psp2 scheme on humidity in the boundary layer during non-rainy conditions. Bachelor's thesis, LMU Munich, 2022.

Fleury, A., Bouttier, F., and Couvreux, F., Process-oriented stochastic perturbations applied to the parametrization of turbulence and shallow convection for ensemble prediction, *Quarterly Journal of the Royal Meteorological Society*, 148(743):981–1000, 2022. doi: https://doi.org/10.1002/qj.4242.

Frogner, I.-L., Singleton, A. T., Køltzow, M. Ø., and Andrae, U., Convection-permitting ensembles: Challenges related to their design and use, *Quarterly Journal of the Royal Meteorological Society*, 145(S1):90–106, April 2019. doi: 10.1002/qj.3525.

Gebhardt, C., Theis, S. E., Paulat, M., and Ben Bouallègue, Z., Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries, *Atmos. Res.*, 100(2):168–177, May 2011. ISSN 0169-8095. doi: 10.1016/j.atmosres.2010.12.008.

Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N., and Tennant, W., The met office convective-scale ensemble, mogreps-uk, *Quarterly Journal of the Royal Meteorological Society*, 143(708):2846–2861, 2017. doi: 10.1002/qj.3135.

Harnisch, F. and Keil, C., Initial Conditions for Convective-Scale Ensemble Forecasting Provided by Ensemble Data Assimilation, *Monthly Weather Review*, 143(5):1583–1600, 2015. ISSN 0027-0644. doi: 10.1175/MWR-D-14-00209.1.

Hauck, C., Barthlott, C., Krauss, L., and Kalthoff, N., Soil moisture variability and its influence on convective precipitation over complex terrain, *Quarterly Journal of the Royal Meteorological Society*, 137(S1):42–56, 2011. doi: https://doi.org/10.1002/qj.766.

Hirt, M. and Craig, G. C. PSP - Parameterizing boundary layer variability and subgrid scale orography. `https://download.dwd.de/pub/DWD/Forschung_und_Entwicklung/ICCARUS2018_pre--sentations_PDF/Tuesday/04_Hirt.pdf`, 2018. Accessed: 07.12.2023.

Hirt, M. and Craig, G. C., A cold pool perturbation scheme to improve convective initiation in convection-permitting models, *Quarterly Journal of the Royal Meteorological Society*, 147(737):2429–2447, 2021. doi: https://doi.org/10.1002/qj.4032.

Hirt, M., Rasp, S., Blahak, U., and Craig, G. C., Stochastic parameterization of processes leading to convective initiation in kilometer-scale models, *Monthly Weather Review*, 147 (11):3917 – 3934, 2019. doi: 10.1175/MWR-D-19-0060.1.

Hitchens, N. M., Brooks, H. E., and Kay, M. P., Objective Limits on Forecasting Skill of Rare Events, *Weather and Forecasting*, 28(2):525–534, 2013. doi: 10.1175/WAF-D-12-00113.1.

Hohenegger, C. and Schär, C., Predictability and Error Growth Dynamics in Cloud-Resolving Models, *Journal of the Atmospheric Sciences*, 64(12):4467–4478, 2007. doi: 10.1175/2007JAS2143.1.

Houtekamer, P. L., Lefaivre, L., Derome, J., Ritchie, H., and Mitchell, H. L., A system simulation approach to ensemble prediction, *Monthly Weather Review*, 124(6):1225–1242, 1996.

Jacques, D. and Zawadzki, I., The impacts of representing the correlation of errors in radar data assimilation. part ii: Model output as background estimates, *Monthly Weather Review*, 143(7):2637–2656, 2015. doi: 10.1175/MWR-D-14-00243.1.

Jankov, I., Berner, J., Beck, J., Jiang, H., Olson, J. B., Grell, G., Smirnova, T. G., Benjamin, S. G., and Brown, J. M., A Performance Comparison between Multiphysics and Stochastic Approaches within a North American RAP Ensemble, *Mon. Wea. Rev.*, 145 (4):1161–1179, April 2017. ISSN 0027-0644, 1520-0493. doi: 10.1175/MWR-D-16-0160.1.

Jankov, I., Beck, J., Wolff, J., Harrold, M., Olson, J. B., Smirnova, T., Alexander, C., and Berner, J., Stochastically Perturbed Parameterizations in an HRRR-Based Ensemble, *Monthly Weather Review*, 147(1):153 – 173, 2019. doi: 10.1175/MWR-D-18-0092.1.

Jolliffe, I. T., Uncertainty and inference for verification measures, *Weather and Forecasting*, 22(3):637–650, 06 2007. Copyright - Copyright American Meteorological Society Jun 2007; CODEN - WEFOE3.

Kalnay, E., *Atmospheric modeling, data assimilation and predictability*, Cambridge university press, 2003.

Kawabata, T. and Ueno, G., Non-gaussian probability densities of convection initiation and development investigated using a particle filter with a storm-scale numerical weather prediction model, *Monthly Weather Review*, 148(1):3–20, 2020. doi: 10.1175/MWR-D-18-0367.1.

Keil, C., Heinlein, F., and Craig, G. C., The convective adjustment time-scale as indicator of predictability of convective precipitation, *Quarterly Journal of the Royal Meteorological Society*, 140(679):480–490, 2014. doi: https://doi.org/10.1002/qj.2143.

Keil, C., Baur, F., Bachmann, K., Rasp, S., Schneider, L., and Barthlott, C., Relative contribution of soil moisture, boundary-layer and microphysical perturbations on convective predictability in different weather regimes, *Quarterly Journal of the Royal Meteorological Society*, 145(724):3102–3115, 2019. doi: 10.1002/qj.3607.

Keil, C., Chabert, L., Nuissier, O., and Raynaud, L., Dependence of predictability of precipitation in the northwestern mediterranean coastal region on the strength of synoptic control, *Atmospheric Chemistry and Physics*, 20(24):15851–15865, December 2020. doi: 10.5194/acp-20-15851-2020.

Kober, K. and Craig, G. C., Physically based stochastic perturbations (psp) in the boundary layer to represent uncertainty in convective initiation, *Journal of the Atmospheric Sciences*, 73(7):2893 – 2911, 2016. doi: 10.1175/JAS-D-15-0144.1.

Kondo, K. and Miyoshi, T., Non-gaussian statistics in global atmospheric dynamics: a study with a 10 240-member ensemble kalman filter using an intermediate atmospheric general circulation model, *Nonlinear Processes in Geophysics*, 26(3):211–225, 2019. doi: 10.5194/npg-26-211-2019.

Kühnlein, C., Keil, C., Craig, G. C., and Gebhardt, C., The impact of downscaled initial condition perturbations on convective-scale ensemble forecasts of precipitation, *Q.J.R. Meteorol. Soc.*, 140(682):1552–1562, July 2014. ISSN 1477-870X. doi: 10.1002/qj.2238.

Kunz, M., Abbas, S. S., Bauckholt, M., Böhmländer, A., Feuerle, T., Gasch, P., Glaser, C., Groß, J., Hajnsek, I., Handwerker, J., Hase, F., Khordakova, D., Knippertz, P., Kohler, M., Lange, D., Latt, M., Laube, J., Martin, L., Mauder, M., Möhler, O., Mohr, S., Reitter, R. W., Rettenmeier, A., Rolf, C., Saathoff, H., Schrön, M., Schütze, C., Spahr, S., Späth, F., Vogel, F., Völksch, I., Weber, U., Wieser, A., Wilhelm, J., Zhang, H., and Dietrich, P., Swabian moses 2021: An interdisciplinary field campaign for investigating convective storms and their event chains, *Frontiers in Earth Science*, 10, 2022. ISSN 2296-6463. doi: 10.3389/feart.2022.999593.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al., Graphcast: Learning skillful medium-range global weather forecasting, *arXiv preprint arXiv:2212.12794*, 2022.

Lazo, J. K., Morss, R. E., and Demuth, J. L., 300 billion served: Sources, perceptions, uses, and values of weather forecasts, *Bulletin of the American Meteorological Society*, 90(6):785–798, 2009.

Legrand, R., Michel, Y., and Montmerle, T., Diagnosing non-gaussianity of forecast and analysis errors in a convective-scale model, *Nonlinear Processes in Geophysics*, 23(1): 1–12, 2016. doi: 10.5194/npg-23-1-2016.

Leith, C. E., Theoretical Skill of Monte Carlo Forecasts, *Monthly Weather Review*, 102(6): 409–418, 1974. doi: 10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Leoncini, G., Plant, R. S., Gray, S. L., and Clark, P. A., Perturbation growth at the convective scale for csip iop18, *Quarterly Journal of the Royal Meteorological Society*, 136(648):653–670, 2010. doi: https://doi.org/10.1002/qj.587.

Leutbecher, M. and Palmer, T., Ensemble forecasting, *Journal of Computational Physics*, 227(7):3515–3539, 2008. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2007.02.014. Predicting weather, climate and extreme events.

Leutbecher, M., Ensemble size: How suboptimal is less than infinity?, *Quarterly Journal of the Royal Meteorological Society*, 145(S1):107–128, 2019. doi: 10.1002/qj.3387.

Liu, W., Zhang, Q., Li, C., Xu, L., and Xiao, W., The influence of soil moisture on convective activity: a review, *Theor Appl Climatol*, 149(149):221–232, 2022. doi: https://doi.org/10.1007/s00704-022-04046-z.

Lorenz, E. N., A study of the predictability of a 28-variable atmospheric model, *Tellus*, 17 (3):321–333, 1965. doi: https://doi.org/10.1111/j.2153-3490.1965.tb01424.x.

Lorenz, E. N., The predictability of a flow which possesses many scales of motion, *Tellus*, 21:289–307, 1969. doi: 10.1111/j.2153-3490.1969.tb00444.x.

Marsigli, C., Montani, A., and Paccagnella, T., Perturbation of initial and boundary conditions for a limited-area ensemble: multi-model versus single-model approach, *Quarterly Journal of the Royal Meteorological Society*, 140(678):197–208, 2014. doi: https://doi.org/10.1002/qj.2128.

Matsunobu, T., Zarboo, A., Barthlott, C., and Keil, C., Impact of combined microphysical uncertainties on convective clouds and precipitation in icon-d2-eps forecasts during different synoptic control, *Weather and Climate Dynamics Discussions*, 2022:1–25, 2022. doi: 10.5194/wcd-2022-17.

Matsunobu, T., Puh, M., and Keil, C., Flow- and scale-dependent spatial predictability of convective precipitation combining different model uncertainty representations, *Quarterly Journal of the Royal Meteorological Society*, 2024. doi: https://doi.org/10.1002/qj.4713.

Miyoshi, T., Kondo, K., and Imamura, T., The 10,240-member ensemble kalman filtering with an intermediate agcm, *Geophysical Research Letters*, 41(14):5264–5271, 2014. doi: 10.1002/2014GL060863.

Montani, A., Cesari, D., Marsigli, C., and Paccagnella, T., Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: main achievements and open challenges, *Tellus A: Dynamic Meteorology and Oceanography*, 63(3):605–624, January 2011. doi: 10.1111/j.1600-0870.2010.00499.x.

Palmer, T. N., Gelaro, R., Barkmeijer, J., and Buizza, R., Singular vectors, metrics, and adaptive observations, *Journal of the Atmospheric Sciences*, 55(4):633–653, February 1998. doi: 10.1175/1520-0469(1998)055<0633:svmaao>2.0.co;2.

Petch, J. C., The predictability of deep convection in cloud-resolving simulations over land, *Quarterly Journal of the Royal Meteorological Society*, 130(604):3173–3187, 2004. doi: https://doi.org/10.1256/qj.03.107.

Puh, M., Keil, C., Gebhardt, C., Marsigli, C., Hirt, M., Jakub, F., and C. Craig, G., Physically based stochastic perturbations improve high-resolution forecast of convection, *Quarterly Journal of the Royal Meteorological Society*, n/a(n/a), 2023. doi: https://doi.org/10.1002/qj.4574.

Raschendorfer, M. The new turbulence parameterization of LM. COSMO Newsletter No. 1, 89-97, 2001.

Raynaud, L. and Bouttier, F., The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts, *Quarterly Journal of the Royal Meteorological Society*, 143(709):3037–3047, 2017. doi: 10.1002/qj.3159.

Reinert, D., Prill, F., Denhard, H. F. M., Baldauf, M., C. Schraff, C. G., Marsigli, C., and Zängl, G., DWD Database Reference for the Global and Regional ICON and ICON-EPS Forecasting System, 2021. doi: 10.5676/DWD_pub/nwv/icon_2.1.7.

Richardson, L. F., *Weather prediction by numerical process*, University Press, 1922.

Roberts, B., Clark, A. J., Jirak, I. L., Gallo, B. T., Bain, C., Flack, D. L. A., Warner, J., Schwartz, C. S., and Reames, L. J., Model configuration versus driving model: Influences on next-day regional convection-allowing model forecasts during a real-time experiment, *Weather and Forecasting*, 38:99–123, 1 2023. ISSN 1520-0434. doi: 10.1175/WAF-D-21-0211.1.

Roberts, N. M. and Lean, H. W., Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events, *Monthly Weather Review*, 136(1): 78 – 97, 2008. doi: 10.1175/2007MWR2123.1.

Ruiz, J., Lien, G.-Y., Kondo, K., Otsuka, S., and Miyoshi, T., Reduced non-gaussianity by 30-second rapid update in convective-scale numerical weather prediction, *Nonlinear Processes in Geophysics Discussions*, 2021:1–13, 2021. doi: 10.5194/npg-2021-15.

Sakradzija, M. and Klocke, D., Physically constrained stochastic shallow convection in realistic kilometer-scale simulations, *Journal of Advances in Modeling Earth Systems*, 10 (11):2755–2776, 2018. doi: https://doi.org/10.1029/2018MS001358.

Scheck, L., Weissmann, M., and Bach, L., Assimilating visible satellite images for convective-scale numerical weather prediction: A case-study, *Q. J. Roy. Meteor. Soc.*, 146(732):3165–3186, oct 2020. ISSN 1477-870X. doi: 10.1002/QJ.3840.

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

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

Schwartz, C. S., Kain, J. S., Weiss, S. J., Xue, M., Bright, D. R., Kong, F., Thomas, K. W., Levit, J. J., Coniglio, M. C., and Wandishin, M. S., Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership, *Weather and Forecasting*, 25(1):263–280, February 2010. doi: 10.1175/2009waf2222267.1.

Schwartz, C. S., Romine, G. S., Fossell, K. R., Sobash, R. A., and Weisman, M. L., Toward 1-km ensemble forecasts over large domains, *Monthly Weather Review*, 145(8):2943–2969, August 2017. doi: 10.1175/mwr-d-16-0410.1.

Seifert, A. and Beheng, K. D., A two-moment cloud microphysics parameterization for mixed-phase clouds. part 1: Model description, *Meteorology and atmospheric physics*, 92 (1):45–66, 2006a.

Seifert, A. and Beheng, K. D., A two-moment cloud microphysics parameterization for mixed-phase clouds. part 1: Model description, *Meteorology and atmospheric physics*, 92 (1):45–66, 2006b.

Selz, T., Estimating the intrinsic limit of predictability using a stochastic convection scheme, *Journal of the Atmospheric Sciences*, 76(3):757–765, March 2019. doi: 10.1175/jas-d-17-0373.1.

Selz, T. and Craig, G. C., Upscale Error Growth in a High-Resolution Simulation of a Summertime Weather Event over Europe, *Monthly Weather Review*, 143(3):813–827, 2015. doi: 10.1175/MWR-D-14-00140.1.

Selz, T., Riemer, M., and Craig, G. C., The Transition from Practical to Intrinsic Predictability of Midlatitude Weather, *Journal of the Atmospheric Sciences*, 79(8):2013 – 2030, 2022. doi: https://doi.org/10.1175/JAS-D-21-0271.1.

Tempest, K. I., Craig, G. C., and Brehmer, J. R., Convergence of forecast distributions in a 100,000-member idealised convective-scale ensemble, *Quarterly Journal of the Royal Meteorological Society*, n/a(n/a), 2023. doi: https://doi.org/10.1002/qj.4410.

Tempest, K. I., Craig, G. C., Puh, M., and Keil, C., Convergence of ensemble forecast distributions in weak and strong forcing convective weather regimes, *Quarterly Journal of the Royal Meteorological Society*, 2024. doi: https://doi.org/10.1002/qj.4684.

Theis, S. E., Hense, A., and Damrath, U., Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach, *Meteorological Applications*, 12(3):257–268, 2005. doi: https://doi.org/10.1017/S1350482705001763.

Thompson, G., Berner, J., Frediani, M., Otkin, J. A., and Griffin, S. M., A Stochastic Parameter Perturbation Method to Represent Uncertainty in a Microphysics Scheme, *Monthly Weather Review*, 149(5):1481 – 1497, 2021. doi: 10.1175/MWR-D-20-0077.1.

Toth, Z. and Buizza, R. Chapter 2 - weather forecasting: What sets the forecast skill horizon? In Robertson, A. W. and Vitart, F., editors, *Sub-Seasonal to Seasonal Prediction*, pages 17–45. Elsevier, 2019a. ISBN 978-0-12-811714-9. doi: https://doi.org/10.1016/B978-0-12-811714-9.00002-4.

Toth, Z. and Buizza, R. Chapter 2 - Weather Forecasting: What Sets the Forecast Skill Horizon? In Robertson, A. W. and Vitart, F., editors, *Sub-Seasonal to Seasonal Prediction*, pages 17 – 45. Elsevier, 2019b. ISBN 978-0-12-811714-9. doi: https://doi.org/10.1016/B978-0-12-811714-9.00002-4.

Toth, Z. and Kalnay, E., Ensemble forecasting at NCEP and the breeding method, *Monthly Weather Review*, 125(12):3297–3319, December 1997. doi: 10.1175/1520-0493(1997) 125<3297:efanat>2.0.co;2.

Trentmann, J., Keil, C., Salzmann, M., Barthlott, C., Bauer, H.-S., Schwitalla, T., Lawrence, M. G., Leuenberger, D., Wulfmeyer, V., Corsmeier, U., Kottmeier, C., and Wernli, H., Multi-model simulations of a convective situation in low-mountain terrain in central europe, *Meteorology and atmospheric physics*, 103(1-4):95–103, 2009. ISSN 0066-6416, 0177-7971, 0259-8477, 1436-5065. doi: 10.1007/s00703-008-0323-6. 12.01.02; LK 01.

Vitart, F. and Robertson, A. W. Introduction: Why sub-seasonal to seasonal prediction (s2s)? In *Sub-seasonal to seasonal prediction*, pages 3–15. Elsevier, 2019.

Wallace, J. M. and Hobbs, P. V., *Atmospheric science: an introductory survey*, volume 92, Elsevier, 2006.

Wilks, D. S. Chapter 9 - forecast verification. In Wilks, D. S., editor, *Statistical Methods in the Atmospheric Sciences (Fourth Edition)*, pages 369–483. Elsevier, fourth edition edition, 2019. ISBN 978-0-12-815823-4. doi: https://doi.org/10.1016/B978-0-12-815823-4.00009-2.

Zimmer, M., Craig, G. C., Keil, C., and Wernli, H., Classification of precipitation events with a convective response timescale and their forecasting characteristics, *Geophysical Research Letters*, 38(5), 2011. doi: https://doi.org/10.1029/2010GL046199.

Zängl, G., Reinert, D., Rípodas, P., and Baldauf, M., The icon (icosahedral non-hydrostatic) modelling framework of dwd and mpi-m: Description of the non-hydrostatic dynamical core, *Quarterly Journal of the Royal Meteorological Society*, 141:563–579, 1 2015. ISSN 00359009. doi: 10.1002/qj.2378.

Acknowledgements

Firstly, I am deeply grateful to my supervisor Dr. George Craig for his unwavering support and guidance throughout this doctoral journey. His expertise and encouragement have been invaluable in shaping this thesis and pushing me to achieve my best.

Secondly, I would like to thank Dr. Christian Keil for all his prompt feedback, the discussions and the advice he gave me. I am convinced that by learning from his attention to detail and clarity, I have become a better scientist.

Thanks to all my colleagues, Kirsten Tempest, Jonas Späth, Tobias Selz to name a few, who have helped me in one way or another during my PhD journey. Special thanks to Takumi Matsunobu, this thesis would not have been finished without his help in setting up the experiments and when things were not going according to plan. I also want to give a shout-out to Oriol Tinto, Fabian Jakub and Robert Redl for their help with technical issues.

I am grateful for having been part of the Waves to Weather community, which made me feel like my work is part of a bigger effort for improvements in weather predictability. I will always keep fond memories of all the meetings, meaningful connections and friendships I made inside the W2W community. Special thanks to Audine Laurian for all the organisational work she did and her help with administrative issues. Her positive attitude made me feel like I was part of a big family.

I would additionally like to thank Christoph Gebhardt and Chiara Marsigli from DWD for the excellent collaboration which made the first part of this project possible. The ongoing work to include the PSP scheme in operational forecasting at DWD is a result of their effort and collaborative spirit.

Last but not least, a sincere "grazie"/"hvala" to my parents. Without their unconditional support I would not have been able to study in Munich and achieve what I did. I will forever be thankful for their sacrifices.

W2W This work was carried out in Project A6 of the Collaborative Research Centre Waves to Weather, funded by the DFG (Deutsche Forschungsgemeinschaft) grant SFB/TR165.