

PREDICTABILITY OF PRECIPITATION-RELATED
HAZARDS: INSIGHTS FROM ENSEMBLE
ARCHIVES WITH ADVANCED STATISTICAL
METHODS



JOHANNES GUTENBERG
UNIVERSITÄT MAINZ

Yuliya Kazachkova

Born: 21/01/1997 in Murmansk, Russia

Dissertation for the award of the title
Doctor of Natural Science
submitted to the Institute of Physics in the
Faculty of Physics, Mathematics and Computer Sciences (FB 08)
of the Johannes Gutenberg-Universität Mainz
on 12/06/2026

DECLARATION

I hereby declare that I have written this thesis independently and without the use of sources other than those indicated. All passages and ideas taken from other sources, either literally or in substance, are appropriately cited. Additionally, I declare that I have used generative AI tools as an aid in the preparation of this thesis. These tools were employed to support tasks such as language refinement, text clarity, grammar suggestions, and formatting advice. All scientific content, interpretations, and conclusions are my own, and I take full responsibility for the originality and accuracy of the work presented.

Mainz, 12/06/2026

Yuliya Kazachkova

Yuliya Kazachkova

Predictability of Precipitation-Related Hazards: Insights from Ensemble Archives with Advanced Statistical Methods

Oral examination: 22/05/2026

First advisor: Jun.-Prof. Annette Miltenberger

Second advisor: Prof. Dr. Peter Spichtinger

Institut für Physik der Atmosphäre

Johann-Joachim-Becher-Weg 21

Johannes Gutenberg Universität

55128 Mainz

ABSTRACT

Accurate precipitation forecasts are essential for hydrology and weather-related risk management, yet they remain challenging even in convection-permitting numerical weather prediction. Limiting factors are the incomplete knowledge of initial conditions as well as simplifications in the representation of physical processes. While strategies to deal with initial condition uncertainty are well-established, isolating and representing model uncertainty remains challenging. This thesis examines practical predictability of hazardous precipitation in convection-permitting forecasting systems by extracting flow-conditioned systematic model behaviour and hazard-relevant information from long operational ensemble records. It combines two complementary perspectives, regime-dependent precipitation error structures over complex terrain and probabilistic hail guidance inferred from precipitation patterns when hail is not forecast explicitly.

In the first study, winter orographic precipitation errors over the Harz exhibit a robust low-dimensional structure that recurs under similar large-scale conditions. A principal component analysis (PCA) of precipitation differences between radar data (RADKLIM) and high-resolution numerical weather prediction data (COSMO-DE-EPS) identifies three dominant spatial modes that together explain roughly two thirds of the winter error variance. The modes represent ridge-centred amplification, synoptic placement mismatches with different spatial extent, as well as a mismatch between upstream and downstream precipitation amounts. ERA5 composites and the moist Froude number link these modes to distinct flow and stability regimes, while ICON case studies and semi-idealised experiments relate them to physically distinct precipitation responses under different terrain-flow regimes and show that microphysical perturbations mainly affect intensity rather than the dominant spatial error structure. The results indicate that isolating the underlying issues with the model representation remains challenging even when controlling for the large-scale flow configuration and when excluding cases where initial condition uncertainty (as represented in the ensemble system) controls forecast error.

In the second study, high-resolution precipitation patterns contain a reproducible signal that supports skilful probabilistic discrimination between hail-producing and non-hail convective situations, with best test accuracies of 85.91% (CNN) and 85.71% (RF) using RADKLIM precipitation field snapshots alone. Transferring this precipitation-based signal to COSMO-DE-EPS requires bias correction, and quantile mapping enables a physically plausible application to model precipitation fields. Verification against hail tracks shows strong sensitivity to displacement-tolerant neighbourhood settings in space and time. Under such settings, the COSMO-RF product reproduces the main hail-active regions and captures interannual variability. Analysis of dynamic, thermo-

dynamic and aerosol (dust) conditions during mis-forecasts provides complementary context, but does not yield a single clear discriminator between hits, misses, and false alarms in the present setup.

Overall, the thesis shows that useful predictability can be extracted in two complementary ways. Over complex terrain, predictability emerges through the potential for flow-conditioned and physically interpretable structure in systematic precipitation error that may be applied to correct model precipitation fields. For hail, predictability emerges through a transferable hazard signal in precipitation patterns that can be converted into calibrated probabilities once systematic intensity biases are corrected, even though event-scale localisation remains limited.

ACKNOWLEDGMENTS

The years I spent in Mainz gave me far more than just this thesis. They were filled with experiences, challenges, growth, and unforgettable moments that shaped me both personally and professionally. Looking back, I realise how many wonderful people accompanied me along the way, and how much this journey depended on their support, kindness, and presence.

I would therefore like to take a moment to thank the people who made these years so special.

First of all, I would like to thank my Ph.D. supervisor, Annette Miltenberger, for her support, guidance, and valuable scientific input throughout this thesis. I am deeply grateful for her competence, her readiness to help, and the time and care she always gave not only to me, but to all of her students. I am also grateful for the opportunities she gave me, for her trust in me and my work, and for making this journey possible.

I would also like to thank my colleagues, especially Cornelis and Ziyang, for always being willing to help and for creating such a friendly and fun atmosphere in the office. It made everyday work much easier and much more enjoyable.

My thanks also go to my flatmates, Marios and Fadi, for all the conversations, Saturday coffees, shared moments, and the memories we made during our trips together.

I am especially grateful to my best friend Kir, who was always there for me despite the distance. His support, encouragement, and friendship carried me through homesickness, self-doubt, and the emotional weight of building a life abroad.

I would like to thank the friends I met here for bringing joy, warmth, and meaning into my life in another country, and for helping turn an unfamiliar place into something that felt lived in, shared, and real.

Finally, I am deeply grateful to my parents for always believing in me. Even if the details of my work were often far from familiar to them, they always trusted me and supported my choices.

CONTENTS

I	Introduction	1
1	Introduction and predictability background	3
2	Predictability and probabilistic forecast quality	7
2.1	Predictability as a flow- and scale-dependent property	7
2.2	Sources of forecast uncertainty and the role of ensembles	8
2.3	Probabilistic forecast quality and verification concepts	10
2.4	Systematic model error, regime dependence and predictable structure	12
2.5	Data-driven methods and practical predictability	12
2.6	Summary	13
3	Datasets	15
3.1	Observations of surface precipitation: RADKLIM dataset	15
3.2	Ensemble forecasts of surface precipitation: COSMO-DE-EPS	16
3.3	Large-scale atmospheric conditions: ERA5	17
3.4	Radar-based hail tracks: TRACE ₃ D hail	17
3.5	CAMS dataset	17
II	Identification of Systematic Biases in Ensemble Orographic Precipitation Forecasts	19
4	Introduction	21
4.1	Study area: the Harz Mountains	24
4.2	ICON model setup	25
4.3	Principal Component Analysis (PCA)	26
5	Model performance	29
6	Dominant precipitation error patterns and associated large-scale conditions	35
7	Icon case studies	41
7.1	PC1	44
7.1.1	Case 13.12.2011 06:00 UTC	44
7.1.2	Case 10.12.2014 12:00 UTC	47
7.2	PC2	52
7.3	PC3	53
7.4	Summary of case analyses	54
8	Semi-idealised Simulations	57
8.1	Model setup	57
8.1.1	Initial conditions	57
8.1.2	Topography adjustment	58
8.2	Results	58

8.2.1	Spatial distribution of precipitation	59
8.2.2	Cross-section comparison with COSMO and RADKLIM . . .	60
8.2.3	Moist Froude number	63
8.3	Conclusion	70
III	Precipitation-Based Probabilistic Hail Prediction	73
9	Introduction	75
10	Data Preprocessing and Methods	81
10.1	5 min RADKLIM data	81
10.1.1	Convective/Stratiform Classification	81
10.1.2	Snapshot extraction and labeling	83
10.2	Hourly RADKLIM data	84
10.2.1	Temporal and spatial harmonisation of RADKLIM data . .	84
10.3	COSMO-DE data processing	88
10.3.1	Hourly COSMO-DE precipitation fields	88
10.3.2	COSMO-DE snapshots	89
10.4	Random forest (RF) and convolutional neural network (CNN) . . .	90
10.4.1	Random forest (RF) setup	90
10.4.2	CNN setup	91
11	Nowcasting	93
12	Implementation in COSMO-DE model	99
12.1	Bias correction	99
12.2	Random forest benchmark on 1 h RADKLIM snapshots	102
12.3	Random forest results on COSMO snapshots	103
13	Verification and predictability	105
13.1	Sensitivity to spatial radius ($\Delta t = 30$ min)	106
13.2	Sensitivity to temporal window Δt	110
14	Large-scale conditions associated with detected and missed hail tracks .	115
14.1	Aerosol load in predicted and observed hail cases	118
15	Summary and conclusions	121
IV	Conclusions & Perspectives	125
16	Conclusions	127
17	Discussion and perspectives	131
V	Appendix	133
A	ICON sensitivity experiments	135
	Bibliography	141

Part I

INTRODUCTION

INTRODUCTION AND PREDICTABILITY BACKGROUND

Modern numerical weather prediction forms the basis for a wide range of decisions in hydrology, agriculture, transport and energy. Depending on the lead time, forecasts support different types of decisions. On very short time scales of minutes to a few hours, nowcasting of convective cells and intense showers is needed to anticipate flash floods, large hail and local wind gusts. On time scales of a few hours up to several days, short-range forecasts guide the operation of flood protection systems, the scheduling of irrigation and spraying in agriculture, the planning of aircraft routing and the management of wind and solar power in the electrical grid. Further ahead, medium-range and subseasonal information contributes to reservoir management, crop planning and assessments of energy demand. Across all of these applications, the value of a forecast depends not only on its availability but also on how reliable it is for the specific weather situation and for the variables that matter most to the user, such as near-surface wind, temperature or precipitation.

From a physical point of view, this reliability is constrained by the predictability of the atmosphere. The atmosphere is a chaotic system in which small uncertainties in the initial conditions, the lateral boundary conditions or the model formulation can grow with time. Error growth and forecast skill depend on the flow situation, so that some regimes are more predictable than others.

Over recent decades, numerical weather prediction (NWP) has undergone a steady increase in skill that is often described as a quiet revolution (Bauer, Thorpe, and Brunet, 2015). This progress reflects advances in model dynamics and physics, data assimilation, ensemble forecasting and high-performance computing. Deterministic forecasts now provide useful guidance up to about a week ahead for many variables (Bauer, Thorpe, and Brunet, 2015). At the same time, it is widely recognised that forecasts should be probabilistic, because forecast uncertainty is flow dependent and changes from day to day. Ensemble prediction systems were developed to represent this uncertainty by sampling a range of plausible atmospheric evolutions and providing probabilistic guidance (Leutbecher and Palmer, 2008; Buizza, 2018).

Convection-permitting models (CPMs) with grid spacings of a few kilometres represent an important step in this development. They resolve many aspects of mesoscale circulations and convection that are not captured by coarser models. This has led to improvements in the representation of intense rainfall, diurnal cycles and storm structures in many regions (Clark et al., 2016). At national weather services, CPMs such as COSMO-DE-EPS and ICON are now used operationally in both deterministic and ensemble configurations and provide key guidance for short-range precipitation fore-

casting over central Europe (Baldauf et al., 2011; Zängl, Reinert, Rípodas, et al., 2015). Despite these advances, CPMs still show systematic biases in the location and amount of precipitation, and they do not usually predict surface hail directly (Clark et al., 2016; Baldauf et al., 2011). Understanding how these model limitations depend on the flow regime, and how model output can be translated into useful hazard information, is therefore closely linked to the broader predictability problem.

Accurate precipitation forecasts are particularly important because they link the atmospheric circulation to the land surface and to many aspects of human activity. Precipitation influences runoff and river discharge, shapes soil moisture and vegetation, and is responsible for a wide spectrum of hydrological hazards that range from prolonged river floods to short-lived flash floods. Over continental Europe, the spatial distribution of precipitation is strongly modulated by orography and by convective storms, so that accumulations and extremes can vary sharply over distances of only a few to a few tens of kilometres. This large regional variability is central for both water management and weather-related risk, yet it also makes precipitation one of the most difficult variables to predict in NWP models.

In numerical models, precipitation is governed by processes that span a wide range of scales and are only partially resolved. Large-scale ascent and frontal systems interact with mesoscale circulations that are often shaped by topography, as well as with cloud dynamics and microphysical processes such as condensation, riming and melting. Many of these processes occur on scales smaller than the model grid spacing or on time scales shorter than the model time step, so their effects must be represented by parameterisations. As a result, precipitation prediction is especially sensitive to small-scale processes and model errors, in particular to uncertainties in cloud microphysics, convection and the representation of orographic forcing (Stephens et al., 2010). Even in convection-permitting models, which resolve much of the convective circulation, the timing, intensity and location of rainfall remain challenging, especially for short-lived convective events and extremes (Clark et al., 2016; Prein et al., 2015). These difficulties are particularly pronounced in situations where small-scale forcing and complex terrain play a central role, such as orographic precipitation and deep moist convection that can produce large hail. This combination of strong societal relevance and persistent modelling difficulties makes precipitation a natural focus for studies of predictability in high-resolution NWP systems.

Within this broader picture, a large part of the small-scale variability in precipitation is associated with orographic precipitation. Conceptually, when moist air is forced to flow over hills and mountains, uplift and cooling enhance cloud formation and precipitation on the windward slopes, while rain shadows and local minima can occur in the lee (Smith, 1979; Roe, 2005). This topographic imprint appears in many forms. In winter, persistent upslope flow over low mountain ranges can generate extended periods of stratiform precipitation and heavy snowfall. In the warm season, the same terrain can trigger or organise showers and thunderstorms and can influence where

convective systems intensify or decay. As a result, orographic precipitation plays a key role both in shaping long-term precipitation climatologies and in controlling where heavy precipitation events are most likely (Rauthe et al., 2013; Schneider and Barthlott, 2018).

In contrast to the more persistent orographic precipitation, hail occurs in brief, localised events that are tied to deep moist convection with strong updrafts and active mixed-phase microphysics. Large hailstones are produced only in a subset of storms and are locally rare, yet they can cause severe damage to buildings, crops and infrastructure. For Europe, insurance statistics and event-based studies show that hail contributes a substantial share of the insured losses associated with convective storms, and single events can reach losses of hundreds of millions of euros (Punge and Kunz, 2016). Reliable information on where and when hail is likely is therefore important for short-range warning and for longer-term risk assessment, even though the events themselves occupy only a small fraction of space and time.

This thesis examines practical predictability of hazardous precipitation in convection-permitting forecasting systems using two contrasting test cases, wintertime orographic precipitation over complex terrain and warm-season hail over Germany. The overarching goal is to explore what information on systematic model errors and initial-condition uncertainty can be extracted from long time series of operational high-resolution ensemble forecasts, and how this information can be used to improve the interpretation and evaluation of precipitation-related hazards.

These two test cases were chosen because they represent different manifestations of precipitation predictability. In the orographic case, forecast errors can contain a systematic, flow-dependent component that may recur under similar large-scale conditions. In the hail case, the hazard is a rare and highly local outcome of deep convection that is not predicted explicitly by operational models and must therefore be inferred indirectly from model output. Together, they provide a useful pair of examples for examining how predictable structure can emerge either in systematic model error or in hazard-relevant precipitation signatures.

Against this background, four research questions guide the thesis:

- **RQ1:** To what extent do systematic precipitation errors over complex terrain recur under similar large-scale flow conditions, and can they be described by a small number of robust spatial patterns that are linked to specific regimes?
- **RQ2:** Which physical mechanisms are associated with these recurring systematic model errors in the orographic precipitation case, and is there evidence for a unique dominant mechanism behind the model failures?
- **RQ3:** How much information about hail occurrence is contained in precipitation fields alone, and how far can this signal be used to construct probabilistic hail products from convection-permitting model precipitation?

- **RQ4:** Do hail forecast successes and failures cluster under specific large-scale meteorological conditions, and are forecast outcomes linked to different aerosol loadings?

The thesis addresses these questions from two complementary angles that span different aspects and scales of predictability while drawing on a common framework of high-resolution ensemble forecasting, observational reference data, and probabilistic evaluation.

In the first part, the Harz Mountains in central Germany are used as a regional testbed for wintertime orographic precipitation in COSMO-DE-EPS and ICON. The analysis focuses on systematic errors in accumulated precipitation relative to the RADKLIM radar-based precipitation product and on how these errors depend on large-scale flow and static stability. Principal component analysis (PCA) is used to characterise the dominant spatial structures of the model error, and the resulting modes are related to environmental conditions using ERA5 composites. Selected regimes are then examined in more detail using ICON case studies and semi-idealised experiments in order to connect the statistical patterns to physical mechanisms of terrain-flow interaction and model error formation.

In the second part, the focus shifts to warm-season hail over Germany. Because operational forecast systems such as ICON-D2 and ICON-EU do not provide an explicit surface hail forecast, hail information must be obtained indirectly from proxy quantities or post-processing that relates model output to observed hail. Here, the analysis examines how far high-resolution radar precipitation fields alone can separate hail-producing from non-hail convective situations, and to what extent this hail-related signal can be transferred to COSMO-DE-EPS ensemble precipitation. TRACE3D hail tracks are linked to RADKLIM and COSMO-DE-EPS precipitation snapshots, machine learning methods are used to derive probabilistic hail information, and the resulting predictions are evaluated against the hail-track dataset and against the observed hail corridors over Germany. The analysis also examines the large-scale meteorological and aerosol context of forecast hits, misses and false alarms.

Overall, the thesis connects regime-based error diagnosis and hazard-oriented post-processing within a single predictability framework. The central aim is to identify robust, flow-conditioned structures in precipitation errors and hazard signals that can be anticipated from the large-scale environment and used to support improved evaluation and prediction of precipitation-related hazards.

This chapter provides the predictability framework that underpins the analyses in the following parts. The focus is on practical predictability in operational forecasting, meaning the extent to which a forecast system can provide reliable probabilistic guidance for a given variable, spatial scale and lead time. I first summarise how predictability is constrained by atmospheric dynamics and why it depends on the flow situation. I then discuss the main sources of forecast uncertainty in modern prediction systems and the role of ensembles. Finally, I introduce key concepts for probabilistic forecast evaluation and outline why regime dependence and systematic model errors are central when predicting precipitation-related hazards.

2.1 Predictability as a flow- and scale-dependent property

Atmospheric predictability is constrained by the nonlinear nature of the governing dynamics. Small uncertainties in the atmospheric state tend to grow with lead time, which limits the horizon of deterministic forecasts and motivates a probabilistic description of future weather states (Lorenz, 1963). In this context, predictability is closely tied to the rate of error growth. If errors grow rapidly, they reach amplitudes comparable to the natural variability of the atmosphere more quickly, so that the forecast becomes weakly constrained at an earlier lead time. If errors grow more slowly, useful deterministic skill can persist for longer (Lorenz, 1969; Palmer, 2000).

Error growth is not uniform across situations. It depends on the large-scale circulation and on the physical processes that dominate the evolution. As a result, some flow regimes exhibit comparatively slow error growth, while others amplify uncertainties rapidly (Palmer, 2000; Buizza, 2018). In operational terms, this implies that forecast reliability is state dependent. The same prediction system can be highly reliable in one regime and much less reliable in another, even at the same lead time.

Predictability also depends strongly on spatial scale. Lorenz's multi-scale perspective emphasises that smaller scales tend to have shorter intrinsic predictability, meaning a shorter time over which their future evolution is constrained even under a near-perfect model, and that errors can spread upscale and affect larger-scale features (Lorenz, 1969). This scale dependence is particularly pronounced in the presence of moist processes. Idealised and real-case experiments show that small-scale differences associated with moist convection can grow quickly and then project onto larger scales as the forecast evolves (Zhang, Snyder, and Rotunno, 2003). For precipitation, these ideas have direct implications. Regional accumulations can be constrained by synoptic forcing and

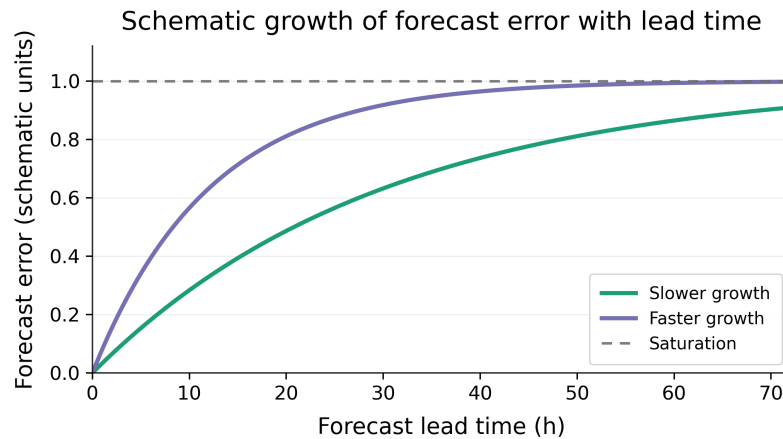


Figure 2.1: Schematic growth of forecast error with lead time. Error growth rates depend on the flow situation, so that some regimes show slower growth and remain predictable for longer than others. Errors eventually approach a saturation level comparable to climatological variability, which limits deterministic predictability.

large-scale moisture transport, whereas local extremes and the evolution of individual convective cells depend on processes near the model resolution limit and can lose predictability rapidly (Prein et al., 2015).

Figure 2.1 summarises this conceptual picture. Although it is schematic, it highlights two aspects. First, predictability should be discussed relative to a reference variability level, because forecast errors eventually become comparable to the natural variability of the system (Lorenz, 1969). Second, the rate at which predictability is lost can depend on the large-scale flow and on the active physical processes, which motivates regime-aware evaluation and post-processing (Palmer, 2000; Buizza, 2018).

2.2 Sources of forecast uncertainty and the role of ensembles

In practice, forecast uncertainty arises from several sources. A first contribution comes from uncertainty in the initial state. Even with dense observing networks and sophisticated data assimilation, the atmospheric state is not known perfectly at forecast initialisation (Kalnay, 2003; Leutbecher and Palmer, 2008). A second contribution comes from uncertainty at the lateral boundaries in limited-area models, because these models inherit uncertainty from the driving global model and from the coupling strategy (Leutbecher and Palmer, 2008). A third contribution comes from limitations of the numerical model itself. Unresolved or poorly resolved processes must be represented by parameterisations, which introduce structural assumptions and uncertain parameters. Even when mean biases are small, deficiencies in such representations can lead to

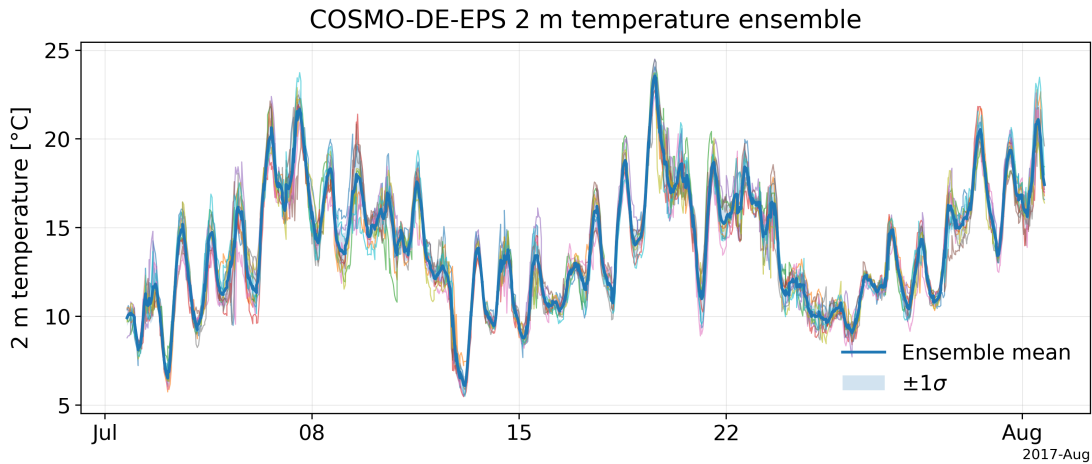


Figure 2.2: Illustration of ensemble forecasting using COSMO-DE-EPS 2 m temperature at a single grid point near the Harz region (51.80°N, 10.60°E) for July 2017. Thin lines show individual ensemble members (m01–m20), the thick line indicates the ensemble mean, and shading denotes $\pm 1\sigma$ across members.

systematic errors in the location, timing or intensity of predicted phenomena (Stephens et al., 2010; Prein et al., 2015; Clark et al., 2016).

Ensemble prediction systems provide a practical framework to represent forecast uncertainty. Instead of issuing a single forecast, ensembles generate a set of forecasts that sample a range of plausible atmospheric evolutions (Leutbecher and Palmer, 2008; Buizza, 2018). Operational ensembles typically perturb the initial state and, for limited-area systems, the lateral boundaries. Some systems also account for model uncertainty through stochastic physics or parameter perturbations (Buizza, 2018). The ensemble distribution is then interpreted probabilistically, which supports both uncertainty-aware decision making and systematic verification of reliability.

A key point is that uncertainty sources interact. Model error can amplify or mask the growth of initial-condition errors, and the separation of these contributions is not straightforward in real forecast systems (Slingo and Palmer, 2011). For precipitation this interaction is particularly important. Uncertainty in convective initiation, microphysical pathways or orographic forcing can strongly influence precipitation outcomes, and these sensitivities can feed back on the resolved dynamics. This motivates methods that diagnose and correct systematic components of error in addition to sampling initial-condition uncertainty. Although the examples in this thesis focus on precipitation-related hazards, the ensemble concept is easiest to illustrate with a smoothly varying variable such as 2 m temperature. Temperature highlights flow-dependent changes in ensemble spread without the additional intermittency and threshold behaviour that characterise precipitation. Figure 2.2 illustrates the ensemble concept in a compact way.

2.3 Probabilistic forecast quality and verification concepts

A probabilistic forecast is useful only if it is both reliable and informative. Reliability, also referred to as calibration, means that forecast probabilities are statistically consistent with observed frequencies. For example, events predicted with probability 0.3 should occur in about 30% of those cases. Reliability is commonly assessed with reliability diagrams, which compare forecast probability bins to the observed event frequency in those bins, and with proper scoring rules such as the Brier score (Wilks, 2011). Sharpness describes how strongly forecast probabilities deviate from climatology. A forecast can be reliable but still provide limited guidance if it issues probabilities close to the climatological event frequency in most cases. Resolution describes the ability to distinguish between situations with different outcome frequencies. In practice, useful probabilistic forecasts combine reliability with sufficient sharpness and resolution (Wilks, 2011).

These concepts are particularly relevant for rare or localised hazards, where unconditional event frequencies are low and the main value of the forecast lies in identifying high-risk situations. In such settings, reliability needs to be interpreted together with how often high probabilities are issued. Figure 2.3 illustrates this connection. In the reliability diagram (top), perfect calibration lies on the diagonal, while deviations indicate systematic miscalibration, for example overconfidence when high forecast probabilities verify too rarely. The accompanying histogram (bottom) shows the distribution of issued probabilities and therefore provides a compact view of sharpness. A forecast that is both reliable and useful should lie close to the diagonal while also allocating a non-negligible fraction of cases to elevated probability levels rather than clustering near the climatological frequency. Because rare-event bins can be noisy, it is also important to interpret these diagnostics together with sample size information and uncertainty estimates when possible.

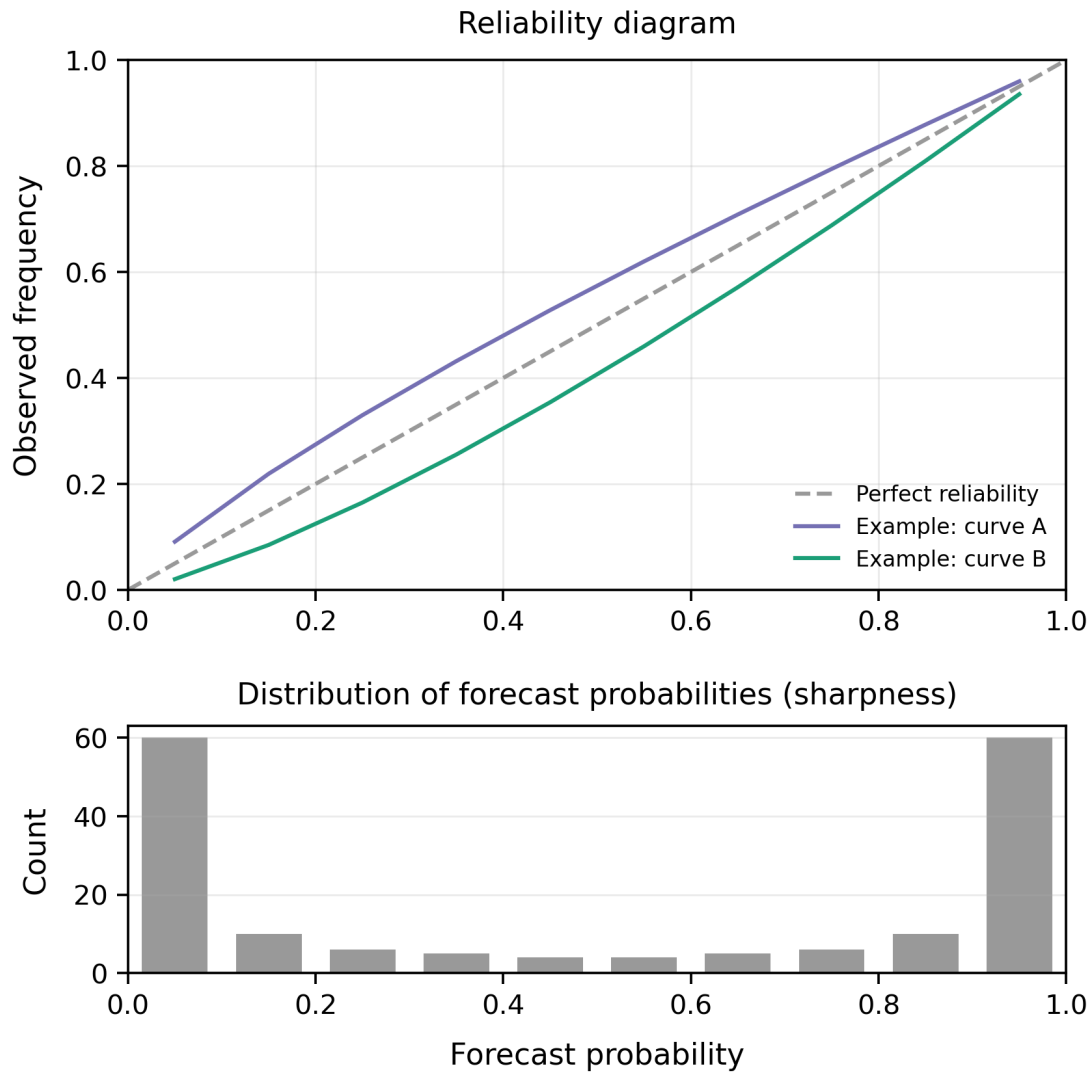


Figure 2.3: Schematic reliability diagram (top) and distribution of forecast probabilities (bottom). Perfect reliability lies on the diagonal. The accompanying histogram visualises sharpness by showing how often different probability levels are issued (Wilks, 2011).

A further aspect of verification is the dependence on spatial and temporal scale. Small displacement errors can lead to poor pointwise scores even when the forecast captures the overall structure of an event. For high-resolution precipitation forecasts, neighbourhood-based verification methods can provide a more physically meaningful assessment because they account for spatial uncertainty and represent the scale at which a forecast can be expected to be reliable (Prein et al., 2015). This is relevant for the second part of the thesis, where event-scale matching is sensitive to spatial and temporal tolerances.

2.4 Systematic model error, regime dependence and predictable structure

Forecast errors can be separated conceptually into case-to-case variability and systematic components. Case-to-case variability reflects random fluctuations, sampling uncertainty and intrinsically sensitive processes. Systematic components appear as biases in mean behaviour or as recurring spatial structures that persist across many cases. Such systematic errors often arise from structural deficiencies in parameterisations, unresolved dynamics, or the representation of surface and topographic forcing (Stephens et al., 2010; Clark et al., 2016).

An important question for practical predictability is whether systematic error depends on the large-scale flow. If similar flow situations repeatedly lead to similar error structures, then part of the model error is conditioned on the environment and is therefore potentially predictable in a statistical sense. This perspective motivates regime-aware diagnostics and post-processing approaches that explicitly condition on circulation types, stability measures or blocking metrics (Photiadou et al., 2016). In the context of precipitation, regime dependence is particularly plausible because the balance between large-scale forcing, orographic uplift and convection changes markedly between circulation patterns and stability states.

A practical implication is that systematic errors may vary not only in magnitude but also in spatial structure. When recurring patterns exist, they can be summarised through dominant modes or composites and related to environmental conditions, providing an interpretable framework for diagnosing regime dependence and assessing predictable components of model error (Wilks, 2011).

2.5 Data-driven methods and practical predictability

In parallel to advances in dynamical modelling, data-driven methods have become increasingly influential in weather prediction. Machine learning models can emulate parts of the atmospheric evolution at very low computational cost, which enables rapid generation of large forecast ensembles and supports exploration of uncertainty in new

ways (Pathak et al., 2022; Bi et al., 2023). Beyond direct forecasting, machine learning is widely used for post-processing. In this setting, statistical models learn relationships between raw model output and observations, correct systematic biases, and improve probabilistic calibration for specific variables and locations (Schulz and Lerch, 2022).

These developments do not remove the intrinsic predictability constraints imposed by atmospheric chaos. They do, however, reshape the practical question of forecast reliability by expanding the range of tools available to represent uncertainty and to extract predictable structure from large forecast archives. In particular, data-driven post-processing can complement ensemble-based uncertainty sampling by targeting flow-dependent and regime-dependent components of model error, and by producing calibrated probability estimates for hazards that are not explicitly predicted by the underlying NWP model (Schulz and Lerch, 2022).

2.6 Summary

Predictability is limited by the growth of small uncertainties and varies with flow regime and spatial scale. Ensemble prediction systems provide a practical way to represent this situation dependence and to issue probabilistic forecasts. Assessing such forecasts requires verification concepts that distinguish reliability from sharpness and consider the effects of spatial and temporal uncertainty, especially for localised hazards. Finally, systematic model error and regime dependence motivate diagnostics that look beyond case-by-case variability and seek recurring, physically interpretable structure in forecast deficiencies. With this framework in place, the following chapter describes the datasets that are used throughout the thesis.

This chapter introduces the observational, model, and reanalysis datasets used in the thesis. Some datasets appear in more than one part and with different temporal subsets. RADKLIM is used in both the winter-focused analysis of the Harz region (Part II) and the hail-focused analysis (Part III). COSMO-DE-EPS provides the ensemble forecasts used in both parts for 2011–2018. ERA5 supplies large-scale context for the Harz analysis. TRACE3D hail tracks are used in Part III with their full April–September record from 2005 to 2023. CAMS mineral dust reanalysis is used in the aerosols analysis and aligned with the COSMO-DE period.

3.1 Observations of surface precipitation: RADKLIM dataset

The RADKLIM dataset of the German Weather Service (DWD) provides observational estimates of surface precipitation at kilometre-scale resolution over Germany. The product combines radar-derived precipitation with surface rain-gauge data, and the postprocessing is tailored for climatological applications, which ensures consistent correction and interpretation of the raw radar fields. A detailed description of the dataset and the applied postprocessing is given in Winterrath et al. (2018) and Lengfeld et al. (2019). I use RADKLIM version 2017.002, which is available at 5 min and hourly temporal resolution. For comparisons with the COSMO-DE-EPS forecasts (Section 3.2), RADKLIM fields are regridded from the native 1 km grid to the COSMO grid of about 2.8 km using a first-order conservative remapping.

In this thesis the dataset is used in two contexts with different temporal subsets and resolutions:

- Part II: accumulated hourly RADKLIM for the winter months (December–February) of 2011–2018, aligned with the COSMO-DE ensemble availability, for direct comparison in the Harz region.
- Part III: April–September, when hail occurrence in Germany is most frequent. Both hourly fields (for COSMO comparison) and native 5 min RADKLIM fields are used to investigate whether radar-derived precipitation structures provide insight into hail occurrence. To match the period of the TRACE3D hail-track data (Section 3.4), the extended RADKLIM record 2005–2023 is also used where required.

3.2 Ensemble forecasts of surface precipitation: COSMO-DE-EPS

Ensemble forecasts at kilometre-scale resolution are an important source of precipitation guidance (Leutbecher and Palmer, 2008). Ensemble prediction systems aim to represent uncertainty in the initial atmospheric state and, to some degree, uncertainties arising from model formulation through parameterisations. Archives of operational high-resolution ensemble forecasts therefore provide valuable information about the predictability of surface precipitation and convective processes.

The COSMO-DE-EPS (COSMO-DE Ensemble Prediction System) was developed and operated by the Deutscher Wetterdienst (DWD) to provide short-range probabilistic forecasts over Germany and surrounding regions (Gebhardt et al., 2011). The ensemble is based on the convection-permitting COSMO-DE configuration (Baldauf et al., 2011). The model domain covers Germany and neighbouring areas with a horizontal grid spacing of about 2.8 km, comprising 421×461 grid points on a rotated latitude–longitude grid (Deutscher Wetterdienst, 2026b). The operational COSMO-DE configuration uses 50 vertical levels (Deutscher Wetterdienst, 2026b).

The model physics follow the COSMO framework (Baldauf et al., 2011) and include a one-moment bulk microphysics scheme, a Tiedtke-type mass-flux convection scheme for shallow convection (Tiedtke, 1989), the Rasch–Kristjánsson radiation scheme (Rasch and Kristjánsson, 1998), and the TERRA land-surface model (Doms et al., 2011). Turbulent mixing is represented with a level-2.5 closure following Mellor–Yamada (Mellor and Yamada, 1982).

The ensemble has 20 members. Member diversity is introduced through perturbations of initial and lateral boundary conditions and through variations in model physics (Gebhardt et al., 2011; Theis, Gebhardt, and Ben Bouallègue, 2015).

COSMO-DE-EPS is initialized every 3 h (8 cycles per day) and provides forecasts up to a maximum lead time of +21 h in the archive used here (Theis, Gebhardt, and Ben Bouallègue, 2013). In this study, the archived forecasts for 01.01.2011–15.05.2018 are used, and precipitation fields are analysed for lead times up to +21 h. The same COSMO archive underpins the winter analysis in the Harz region (Part II) and the hail-season comparison for April–September (Part III), with the COSMO grid serving as the common target for collocation with RADKLIM and hail-related diagnostics. COSMO-based convection-permitting forecasts were discontinued and replaced operationally by ICON-D2-EPS on 10 February 2021 (Deutscher Wetterdienst, 2021). A comparable long-term archive of ICON-D2-EPS forecasts was not available for this project, therefore the analysis is limited to the COSMO period to maintain consistency.

3.3 Large-scale atmospheric conditions: ERA5

ERA5 is the fifth-generation global reanalysis produced by the European Centre for Medium-Range Weather Forecasts (Hersbach et al., 2020). It combines short-range forecasts with a wide range of in-situ and satellite observations using a four-dimensional variational data assimilation system in IFS. ERA5 provides hourly fields that are distributed on a regular latitude–longitude grid at 0.25° resolution (regridged from the native reduced-Gaussian grid) and a vertical discretisation with 137 hybrid sigma–pressure levels from the surface to 0.01 hPa.

In this thesis ERA5 provides the large-scale thermodynamic and kinematic context for the Harz analysis in 2011–2018 (Part II). Three-dimensional temperature, specific humidity, and horizontal wind components (u , v) are used to characterise the background flow and moisture distribution. ERA5 is distributed via the Copernicus Climate Data Store (Copernicus Climate Change Service, 2023), but the model-level fields used here were retrieved directly from the ECMWF MARS archive (Copernicus Climate Change Service, 2023).

3.4 Radar-based hail tracks: TRACE₃D hail

The hail-track dataset developed at the Karlsruhe Institute of Technology (KIT) (Mohr, Kunz, and Schmidberger, 2024) contains continuous tracks of radar-detected convective cells that are likely to produce hail. The dataset is based on volumetric DWD radar data and covers April–September from 2005 to 2023. Tracks are identified with the TRACE₃D cell detection and tracking algorithm (Handwerker, 2002), which links cells across successive 15-minute radar volumes by exploiting three-dimensional reflectivity. The current version includes improved filters that reduce false detections and address split and merge artefacts. For each track the dataset provides start and end time, position, track length, lifetime, mean propagation direction, and translation speed. The spatial coverage is the full DWD composite at 1 km resolution.

In this thesis the TRACE₃D record is used for the hail-focused analysis in Part III. The full April–September period 2005–2023 provides climatological context for hail occurrence. For verification against COSMO-DE-EPS output and collocation with RADKLIM, the TRACE₃D tracks are restricted to the common period 2011–2018.

3.5 CAMS dataset

The Copernicus Atmosphere Monitoring Service global reanalysis (EAC4) provides observation-constrained fields of atmospheric composition produced with 4D-Var data assimilation (Inness et al., 2019; Copernicus Atmosphere Data Store, 2020). It delivers global fields on a $0.75^\circ \times 0.75^\circ$ grid at 3-hourly resolution from 2003 onward. For this

study a Germany subdomain is extracted to match the hail study region, and the period is limited to 2011–2018 to align with the COSMO-DE evaluation. COSMO-DE timestamps are paired to the nearest available CAMS valid time. The analysis focuses on the vertically integrated mass of mineral dust in three particle-size modes, which provides a concise view of dust loads accompanying the COSMO-DE precipitation snapshots. CAMS is used in the aerosols analysis in Section 14.1 and aligned with both the hail region and the COSMO-DE period.

Part II

IDENTIFICATION OF SYSTEMATIC BIASES IN ENSEMBLE OROGRAPHIC PRECIPITATION FORECASTS

INTRODUCTION

Orographic precipitation is a major reason why rainfall and snowfall vary strongly over short distances in midlatitude regions. When numerical weather prediction and climate models misrepresent orographic processes, the result can be large local errors in accumulated precipitation, runoff, and related hazards, even if the large-scale circulation is captured reasonably well. This is a particular concern in low mountain ranges, where modest differences in flow orientation, stability, or cloud microphysics can shift the precipitation maximum by only a few tens of kilometres, yet substantially change local accumulations (Smith, 1979; Rotunno and Houze, 2007). High-resolution precipitation climatologies over Central Europe confirm that such terrain can imprint sharp spatial gradients in both long-term totals and extremes, with local maxima often tied to preferred flow directions and terrain-modified storm structures (Rauthe et al., 2013; Schneider and Barthlott, 2018). Understanding how models behave over low mountain ranges is therefore important for both day-to-day forecasting and longer-term impact studies.

The physical mechanisms governing orographic precipitation are well established in principle, but they remain challenging to represent accurately in numerical models. In the simplest picture, moist air forced to rise over topography cools and condenses, producing enhanced precipitation on the windward slope and reduced precipitation in the lee (Smith, 1979). In practice, however, the location and intensity of the precipitation maximum depend on the interplay between synoptic-scale forcing and the terrain-scale flow response. A useful synthesis is given by Roe (2005), who summarises key pathways in a schematic (Fig. 4.1). These range from statically stable upslope ascent and seeder–feeder enhancement on the windward slope (Fig. 4.1a–c) to thermally driven convection and flow convergence over elevated terrain and in the lee (Fig. 4.1d–f). Panel 4.1g illustrates a mixed situation in which large-scale ascent coexists with embedded convection. In all cases, forced ascent initiates condensation and cloud formation, followed by hydrometeor growth and fallout, but the relative importance of these processes and the position of the precipitation maximum depend sensitively on stability, wind speed, moisture supply, flow orientation, and the degree of low-level blocking or deflection by the terrain (Roe, 2005; Smith, 1979).

From a dynamical perspective, much of the spatial structure of orographic precipitation reflects whether the low-level flow can cross the barrier or is partially blocked and diverted around it. This regime behaviour is controlled by the upstream wind speed, the effective stratification, the mountain height and width, and moisture-related stability effects (Smith, 1979). A compact way to summarise the likelihood of blocking

is through nondimensional flow parameters that compare the kinetic energy of the incoming flow to the effective stability and the terrain height. The most widely used example is the moist Froude number, which indicates whether the flow tends to be blocked ($F_w < 1$) or can cross the ridge more freely ($F_w \gtrsim 1$). In more strongly blocked regimes, ascent and precipitation often peak upstream of the crest, whereas less blocked regimes favour maxima closer to the ridge or extending into the lee (Galewsky and Sobel, 2005). In this thesis, F_w is used as a regime descriptor, and its definition and computation are given in Section 8.2.3.

Other nondimensional measures are used in the literature depending on the application, including dry-stratification analogues of the Froude number, nondimensional mountain height measures, and metrics linked to mountain-wave response and flow deflection (Smith, 1979; Roe, 2005). Together, these parameters provide a physically motivated link between the large-scale environment and the spatial structure of orographic precipitation, and they motivate the regime-based perspective adopted in this thesis.

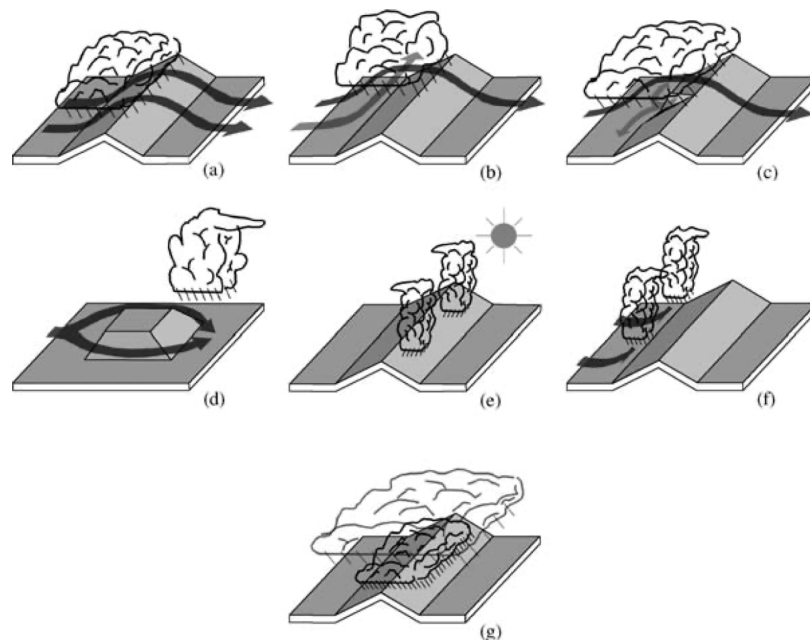


Figure 4.1: Schematic illustration of several orographic precipitation mechanisms, adapted from Roe (2005). Panels (a)–(c) represent mechanically forced uplift and windward enhancement, (d)–(f) illustrate thermally and convergence driven convection over orography, and (g) shows a composite situation with both large scale ascent and embedded convection.

Microphysical processes add another layer of complexity. Orographic clouds can span a wide range of precipitation pathways, from warm-phase systems in which rain forms efficiently through collision–coalescence of cloud droplets, to mixed-phase

and ice-dominated clouds where deposition growth, riming, aggregation and melting contribute substantially to surface precipitation (Morrison et al., 2020). Which pathway dominates depends on the thermodynamic environment (especially temperature and humidity), the supply of cloud condensation and ice-nucleating particles, and the time hydrometeors spend within the uplift region, which is controlled by the flow and the geometry of the terrain (Morrison et al., 2020).

Idealised and case-study simulations show that modifying microphysical properties and assumptions in models can change both the intensity and the location of orographic precipitation, even under similar synoptic forcing (Morales, Morrison, and Posselt, 2018; Thomas, Barrett, and Hoose, 2023). Examples include differences in droplet and ice size distributions, fall speeds, and conversion rates between hydrometeor categories, which together alter growth efficiencies and precipitation fallout. In this sense, systematic model errors over complex terrain can arise through multiple pathways: (i) biases in the dynamical response to the mountain, such as excessive blocking, misplaced ascent, or overly ridge-locked lifting, which shift where condensate is produced, and (ii) biases in the microphysical representation, which affect how efficiently condensate is converted into precipitation and where it falls out. Because dynamics and microphysics are coupled through latent heating and hydrometeor loading, errors in either component can reinforce each other and lead to recurring, flow-dependent error patterns.

Low mountain ranges in Central Europe have been examined in a number of field campaigns and modelling studies that provide context for the Harz. Studies focusing on the Black Forest and the Vosges, as well as on other midlatitude mountain ranges, show that terrain-induced lifting interacts with boundary-layer structure, land-surface heterogeneity, and mesoscale circulations, which together can modify where ascent is concentrated and where precipitation forms (Wulfmeyer et al., 2008; Schneider and Barthlott, 2018). Importantly, the relevant mechanisms depend on the weather situation. Under statically stable conditions, precipitation patterns are often shaped by stratified flow responses such as blocking, upstream lifting, and mountain-wave dynamics, for which nondimensional flow measures like the moist Froude number provide a useful regime description. In warmer and less stable conditions, the same terrain can additionally act as a trigger or organising influence for deep convection through differential heating and low-level convergence, so that the precipitation signal reflects orographic convection rather than purely forced upslope ascent (Schneider and Barthlott, 2018). For the Harz in particular, recent process-oriented simulations with ICON indicate that the representation of terrain and surface properties can influence near-surface winds and convergence, with implications for the location and intensity of precipitation in convective situations (Singh and Kalthoff, 2022).

Against this background, the present part of the thesis examines how a convection-permitting model represents orographic precipitation over the Harz Mountains and how its precipitation errors depend on the large-scale flow and microphysical conditions. In this setting, convection-permitting resolution is valuable not primarily because deep

convection is explicitly resolved, but because it improves the representation of terrain-induced mesoscale circulations and mountain-wave structure that shape uplift and precipitation patterns.

A central question is whether part of the precipitation error over low mountains is systematic in the sense that similar large-scale flow conditions repeatedly produce similar spatial error structures. If so, these errors are not only a verification problem but also a source of physically interpretable information about model limitations. This motivates the following research questions for this part of the thesis:

- Do precipitation errors over the Harz exhibit a small number of recurring spatial patterns, rather than only case-specific variability?
- To what extent are these recurring error patterns linked to distinct large-scale flow regimes and stability conditions?
- What physical processes, including flow blocking, terrain-induced uplift, and microphysical assumptions, help explain the dominant error patterns?

To address these questions, I analyse differences between COSMO precipitation and the RADKLIM radar-based precipitation product and apply principal component analysis (PCA) to extract recurring spatial error structures. By conditioning these patterns on the environmental state, the analysis targets the systematic component of model error, meaning error structures that recur under similar flow regimes rather than case-specific noise. The resulting modes are then interpreted using environmental composites from ERA5, detailed ICON simulations for representative cases, and semi-idealised experiments that probe the sensitivity of the orographic response to changes in flow and microphysics. The following sections describe the study area and datasets, introduce the PCA framework, and then present case studies and experiments that connect the statistical patterns to specific orographic processes over the Harz.

4.1 Study area: the Harz Mountains

The Harz is a low mountain range in central–northern Germany, spanning the federal states of Lower Saxony and Saxony-Anhalt, with a small portion extending into Thuringia. Elevations rise sharply from surrounding foothills (about 230–270 m a.s.l.) to the Brocken summit (1,141 m a.s.l.), producing pronounced orographic contrasts over short horizontal distances.

Climatically, the Harz lies within the belt of prevailing westerlies and frequently experiences windward precipitation enhancement and a lee-side rain-shadow pattern, with substantially reduced precipitation in the lowlands to the east under typical west to northwesterly flow (Smith, 1979; Rauthe et al., 2013). This spatial asymmetry is a

recurring feature of precipitation in German low mountain ranges and provides a clear setting for evaluating terrain-related forecast errors.

At the Brocken summit (DWD station, 1,142 m a.s.l.), 1961–1990 climate normals indicate a mean annual precipitation of 1,766.5 mm, with precipitation (≥ 1 mm) on about 205 days per year, and a mean annual temperature of 2.9°C (Deutscher Wetterdienst, 2025). In contrast, lee-side lowlands to the east are markedly drier. Regional assessments report that annual totals can drop well below 500 mm in dry years, highlighting the strength of the local orographic gradient (Fabig, 2007).

The Harz lies downstream of the North Atlantic storm track and is regularly influenced by westerly to southwesterly flow associated with extratropical cyclones. These synoptic systems transport moist maritime air toward central Germany, promoting windward ascent and widespread stratiform precipitation over the range (Ulbrich, Leckebusch, and Pinto, 2009; Raible et al., 2020). From autumn through winter, increased baroclinic activity leads to frequent frontal passages and enhances the likelihood of strong low-level winds impinging on the terrain. In spring and summer, baroclinicity weakens and precipitation is more often shaped by mesoscale and convective processes whose organisation and preferred tracks depend on the prevailing synoptic flow regime (Ulbrich, Leckebusch, and Pinto, 2009; Raible et al., 2020). Atmospheric blocking over the North Atlantic or Scandinavia can further modify this circulation by deflecting the westerlies and reducing upstream moisture transport into central Europe, which tends to suppress windward precipitation and can strengthen local rain-shadow contrasts downstream of the terrain (Trigo et al., 2004).

4.2 ICON model setup

Simulations of the Harz Mountains and surroundings (Section 4.1) are performed with the Icosahedral Nonhydrostatic model (ICON, v2.6.2.2) for both the case studies (Section 7) and the semi-idealised experiments (Section 8). ICON is used rather than COSMO because the German Weather Service replaced COSMO-DE and COSMO-DE-EPS with ICON-D2 and ICON-D2-EPS on 10 February 2021, and ICON provides improved scalability and efficient nesting on an icosahedral grid while retaining broadly comparable physics (Deutscher Wetterdienst, 2021; Zängl, Reinert, Rípodas, et al., 2015).

The limited-area configuration consists of two nested regional domains centred on the Harz. The outer domain (DOMo1) uses an R19Bo8 grid and spans 8.67–12.72°E and 48.08–52.12°N with a mean edge length of about 1.6 km. The inner nest (DOMo2) uses an R19Bo9 grid (9.68–11.70°E, 49.09–51.11°N) with a mean edge length of about 0.8 km. Two-way nesting is enabled, so DOMo2 receives boundary information from DOMo1 and feeds back to the parent domain during the integration (Pham et al., 2021; Zängl, Reinert, and Prill, 2022).

Initial conditions and time-dependent external lateral boundary conditions are generated with the standard ICON preprocessing and nesting workflow (Zängl, Reinert,

and Prill, 2022). In brief, ECMWF analysis fields are used to drive an ICON integration with an embedded European nest, and output from the European nest is subsequently remapped to the DOMo1/DOMo2 grids to provide the initial and lateral boundary files for the limited-area simulations. External lateral boundary conditions for the outermost domain (DOMo1) are updated hourly, while the boundary exchange between DOMo1 and DOMo2 is performed internally by the nesting algorithm at the model time step. The dynamical time step of DOMo1 is 20 s.

The physics configuration follows a convection-permitting setup, including mixed-phase two-moment cloud microphysics, a TKE-based turbulence scheme, the TERRA land-surface model, and RRTM-based shortwave and longwave radiation. Shallow convection is parameterised, while deep convection is explicitly simulated (no deep-convection parameterisation) at these grid spacings. External surface parameters (e.g. topography, land–sea mask, soil and vegetation) are provided on the native mesh for each domain.

Model output is evaluated primarily against the RADKLIM radar-based precipitation data (Section 3.1), while large-scale environmental conditions are characterised using ERA5 reanalysis (Section 3.3). Further observational and auxiliary datasets used in later chapters are summarised in Chapter 3. The next section introduces the principal component analysis (PCA) framework used to extract dominant spatial structures in the model error.

4.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA), also known as Empirical Orthogonal Functions, is a standard method in the geosciences for extracting dominant patterns from high dimensional spatiotemporal data (Wilks, 2011; Von Storch and Zwiers, 2002). In this study PCA is used to identify systematic differences between COSMO precipitation and RADKLIM observations over the Harz region in terms of a small number of spatial patterns and associated time series.

The starting point is a data matrix with time along one axis and space along the other. At each grid point the temporal mean is removed in order to focus on anomalies rather than absolute values. PCA then identifies orthogonal spatial structures that maximise the variance of these anomalies. In practice this is achieved by computing a singular value decomposition

$$\mathbf{X}' = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^{\top}, \quad (4.1)$$

where $\mathbf{X}'$ denotes the anomaly matrix, columns of $\mathbf{V}$ represent spatial patterns, and the corresponding time series are given by the columns of $\mathbf{U}\mathbf{\Sigma}$. The squared singular values in $\mathbf{\Sigma}$ determine how much of the total variance each mode explains. This SVD based formulation is mathematically equivalent to an eigenanalysis of the covariance matrix and is widely used in EOF applications because it operates directly on the data

matrix and is numerically stable for large problems (Bretherton, Smith, and Wallace, 1992; Hannachi, Jolliffe, and Stephenson, 2007).

In the present work the input field is the precipitation difference between RADKLIM and COSMO on a common COSMO grid that covers the Harz and the adjacent areas. RADKLIM, which is available on its native 1 km grid, is first remapped to the COSMO grid with a horizontal resolution of 2.8 km using first-order conservative regridding, as implemented in the `cf-python` library (Hassell et al., 2017). The PC analysis is then performed on the remapped RADKLIM minus COSMO difference field, so that positive values correspond to situations in which the model underestimates precipitation relative to the radar product. Using the difference rather than the raw fields directs the analysis toward systematic model–observation departures. Temporal means are removed within the Eof solver, so that the leading modes describe spatially coherent deviations from the time averaged error pattern rather than the mean bias itself. The decomposition is carried out with the `Eof` class from the `eofs` standard Python library, which implements the SVD based EOF methodology for climate data and returns both the spatial patterns and the associated principal component time series (Dawson, 2016).

PCA has a long history in climate and weather research because it compresses complex fields into a small number of leading structures that can then be related to underlying processes (Wilks, 2011; Von Storch and Zwiers, 2002). Classic applications include the identification of large scale teleconnection patterns such as the North Atlantic Oscillation and related North Atlantic–European circulation modes, which are routinely obtained as leading EOFs of sea level pressure or geopotential height fields (Wallace and Gutzler, 1981; Hurrell, 1995). PCA and EOF methods have also been used to analyse sea surface temperature variability in the Pacific and to define indices for El Niño–Southern Oscillation and other coupled ocean–atmosphere modes (Deser et al., 2010; Hannachi, Jolliffe, and Stephenson, 2007). On regional scales, EOF based analyses are widely applied to precipitation, temperature, and circulation fields in order to extract dominant patterns of variability and to link these patterns to changes in extremes, blocking frequency, or storm track behaviour (Ulbrich, Leckebusch, and Pinto, 2009).

In this thesis, the method is applied in a less conventional forecast-verification setting and is used to analyse systematic model error. PCA provides a compact description of RADKLIM–COSMO precipitation differences over the Harz. The resulting modes isolate a small number of recurring spatial error patterns that can be examined in the context of orographic flow, stability, and microphysics. These modes then form the basis for the environmental composites, case studies, and semi-idealised simulations that are discussed in the following chapters.

MODEL PERFORMANCE

This chapter provides a baseline evaluation of COSMO-DE-EPS winter precipitation over the Harz, using RADKLIM as reference. The aim is to characterise how the ensemble behaves in terms of mean precipitation, bias growth with lead time, and basic probabilistic performance before turning to more detailed analyses. All analyses are restricted to the extended winter season and to a nominal forecast lead time of 21 h unless stated otherwise. This choice corresponds to the lead time used later in the PCA and case-study analysis and provides enough spin-up for the COSMO-DE-EPS microphysics while still lying within the short-range forecast window that is most relevant for high-impact precipitation events.

Before computing any statistics, periods with missing or broken data are removed. Occasional gaps occur in both COSMO-DE-EPS and RADKLIM. For the winter analysis at the nominal 21 h lead time, 66 of 5528 COSMO-DE-EPS forecast times (1.19%) were classified as broken across all ensemble members and excluded. In the corresponding hourly RADKLIM data set, covering the same winter months over 2011–2017, 11 of 15 168 timestamps (0.07%) were missing because of unavailable files and 10 (0.07%) were discarded because the radar field was entirely invalid. These dates were removed consistently from model and observations, so that only times with a complete set of ensemble members and valid radar data were retained for verification.

The dependence of the winter-mean precipitation bias on lead time is shown in Fig. 5.1. For each winter and each ensemble member, I compute a spatial mean over the Harz domain and then form the difference between COSMO-DE-EPS and RADKLIM for every 3 h lead time. The boxplots summarise the distribution of these winter-mean biases across all winters in the 2011–2018 period and 20 ensemble members. At very short lead times the median bias is close to zero or slightly negative, which suggests that the earliest forecasts are reasonably well calibrated or even a bit too dry on average. With increasing lead time, the distribution gradually shifts towards positive values and the spread increases. From about 10–12 h onward the median bias becomes clearly positive and continues to grow up to 21 h. This behaviour is consistent with a slowly amplifying wet bias as the forecast integrates forward, a feature that has also been reported for other convection permitting models over complex terrain (Prein et al., 2015; Morrison et al., 2020). Such growth of a wet bias can arise from accumulated effects of parameterised cloud and precipitation processes and their coupling to terrain-forced ascent. Even at convection-permitting grid spacings, uncertainties in microphysics remain a key error source, while at coarser resolutions additional errors stem from smoothed orography and stronger reliance on convection parameterisation.

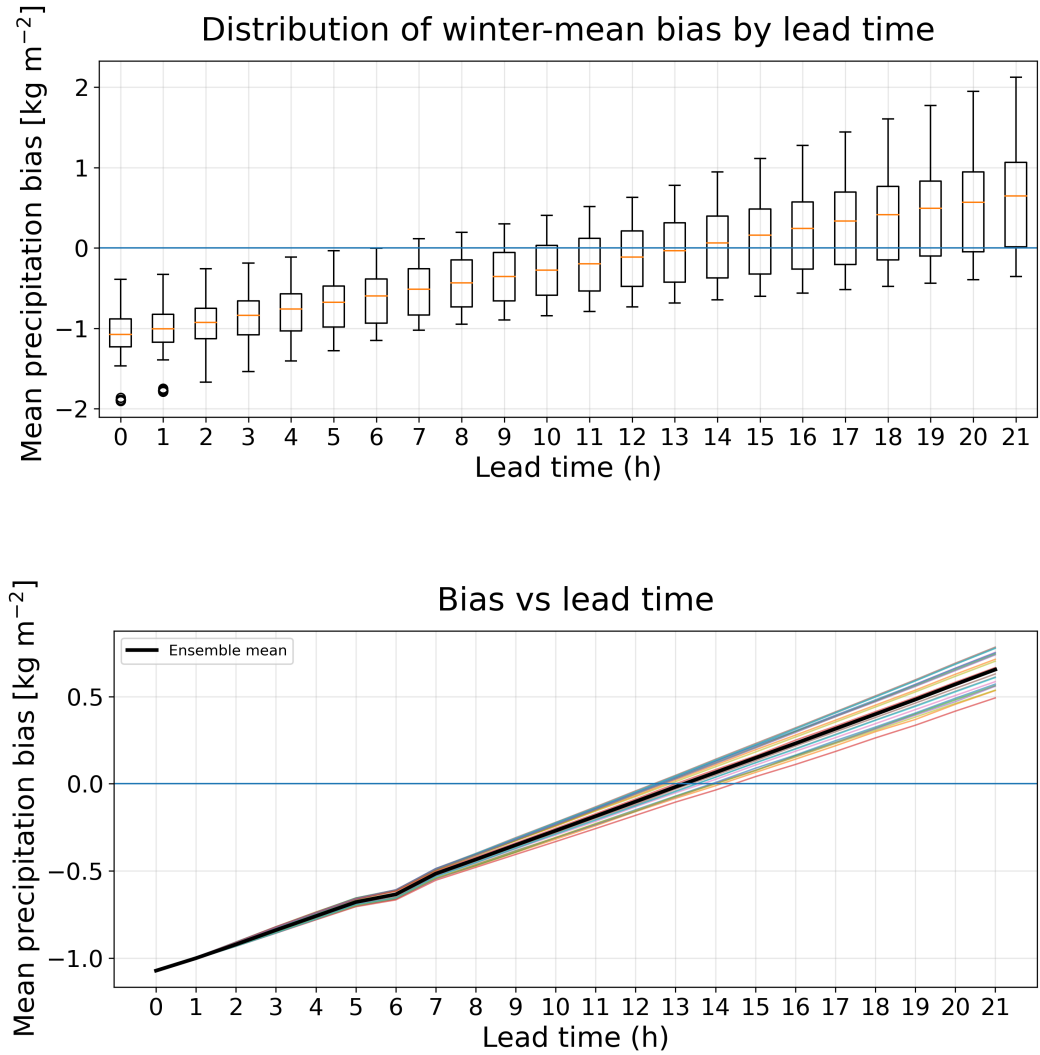


Figure 5.1: Winter-mean precipitation bias (COSMO-DE-EPS minus RADKLIM) as a function of lead time over the Harz region. The upper panel shows the distribution of bias across ensemble members and winters, while the lower panel illustrates the evolution of the ensemble-mean bias and individual member biases.

Reliability characteristics for a simple threshold event at 21 h lead time are illustrated in Fig. 5.2. The upper panel shows a reliability diagram for the probability that 3-hourly precipitation exceeds 1.0 kg m^{-2} per 3 h, with the one-to-one line indicating perfect reliability. COSMO-DE-EPS points lie systematically below this line across most probability bins. For example, events forecast with probabilities around 0.6–0.8 are

observed only about 25–35 % of the time. The lower panel shows the corresponding distribution of forecast probabilities and reveals a strong accumulation of cases near probability one. Together these diagnostics indicate that the ensemble is overconfident for this threshold, a behaviour that is consistent with underdispersive ensemble spread and/or systematic errors in the representation of precipitation processes (Wilks, 2011). Similar overconfidence in precipitation probabilities has been reported for other convection-permitting or high-resolution ensemble systems (Wulfmeyer et al., 2008). These reliability results are consistent with the lead-time dependence of the mean bias shown in Fig. 5.1. The boxplots indicate that the distribution shifts towards a positive bias at longer lead times, so that by 21 h COSMO-DE-EPS tends to be too wet on average. The reliability diagram shows that this tendency also appears in a probabilistic sense, with high forecast probabilities for exceeding 1.0 kg m^{-2} per 3 h occurring less frequently in RADKLIM than predicted. In other words, the bias analysis diagnoses the drift in the ensemble mean, while the reliability analysis highlights overconfidence for this threshold, consistent with limited ensemble spread and/or other systematic forecast errors.

A complementary view of the systematic error is obtained by comparing mean precipitation fields over the Harz (Fig. 5.3). Panel (a) shows the RADKLIM mean, which highlights a distinct orographic enhancement over the central ridge with much weaker values in the surrounding lowlands. The maximum is relatively compact and aligned with the highest terrain, consistent with persistent windward uplift during winter westerly flow. In COSMO-DE-EPS (panel b), the mean precipitation pattern is qualitatively similar, but the orographic maximum is stronger and spatially broader. The high-precipitation core extends further along and downstream of the ridge, which suggests that the model produces too much accumulation in terrain-influenced locations and spreads the enhancement over a wider area than observed. The difference field (panel c) summarises this behaviour. Positive values dominate over the mountains, with the largest excess centred over and just downstream of the highest elevations, indicating a systematic wet bias in COSMO relative to RADKLIM. Away from the ridge the bias is smaller and more mixed, implying that the largest contribution to the domain-mean error comes from the representation of terrain-forced precipitation rather than from a uniform offset. This spatial signature is consistent with the lead-time dependence of the bias and the reliability results shown earlier. If the model already overproduces precipitation in the orographic core, then increasing wet bias with lead time and overconfident exceedance probabilities at 21 h are a plausible consequence, because modest spatial over-amplification over the ridge translates directly into higher event frequencies when precipitation is aggregated over the Harz region.

In summary, these conventional verification diagnostics show that COSMO-DE-EPS reproduces the broad winter precipitation climatology of the Harz, yet develops an increasing wet bias with lead time and tends to place too much precipitation over the main ridge, particularly for high forecast probabilities. The PCA-based analysis in the

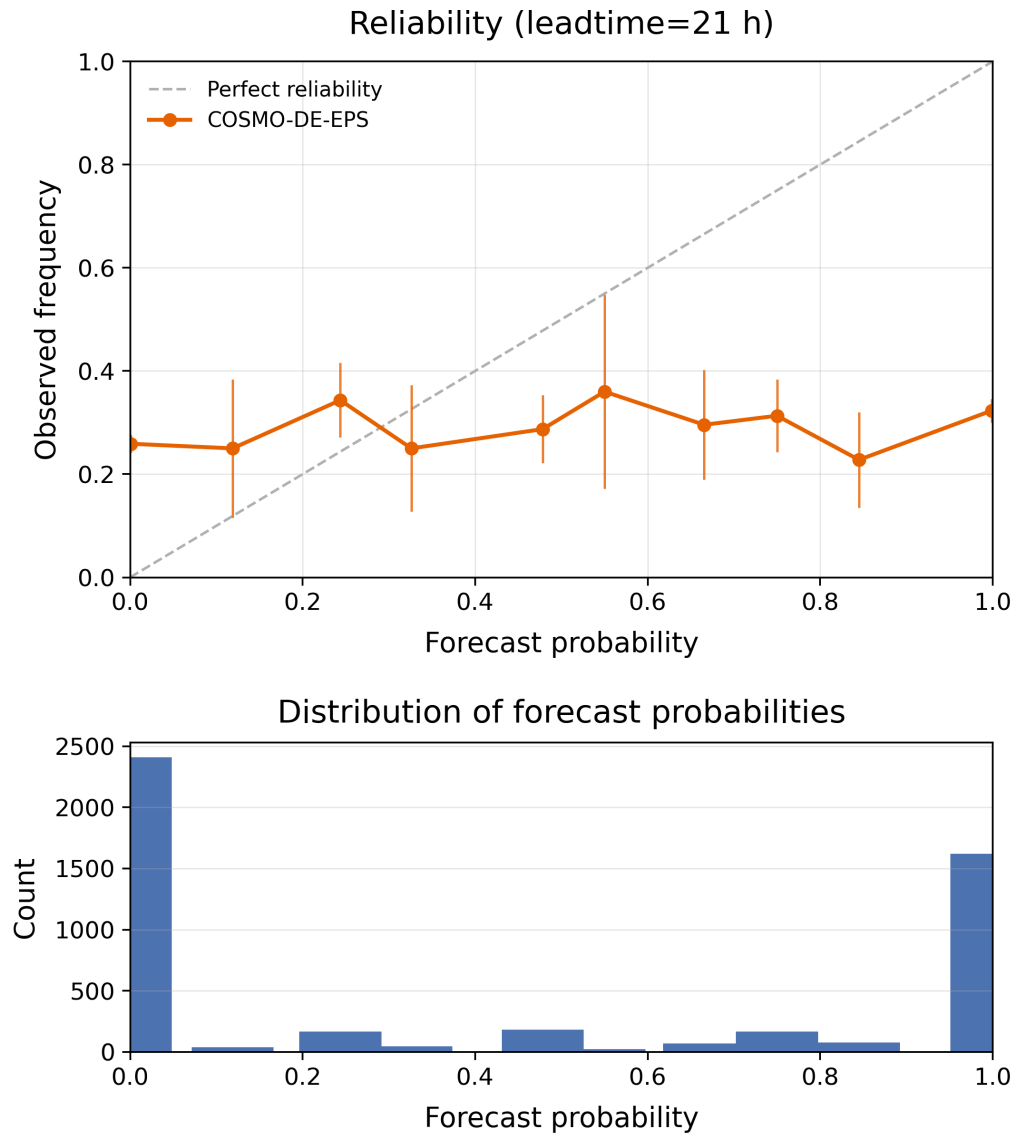


Figure 5.2: Reliability characteristics of COSMO-DE-EPS precipitation forecasts over the Harz at 21 h lead time for the event that 3 h precipitation exceeds 1.0 kg m^{-2} per 3 h. The upper panel shows the reliability diagram, with the diagonal indicating perfect reliability. The lower panel shows the corresponding distribution of forecast probabilities.

following chapters builds on this baseline picture. It asks how this bias decomposes into a small number of recurrent spatial patterns and how those patterns relate to the large-scale environment, flow regime, and orographic processes that govern precipitation over the Harz.

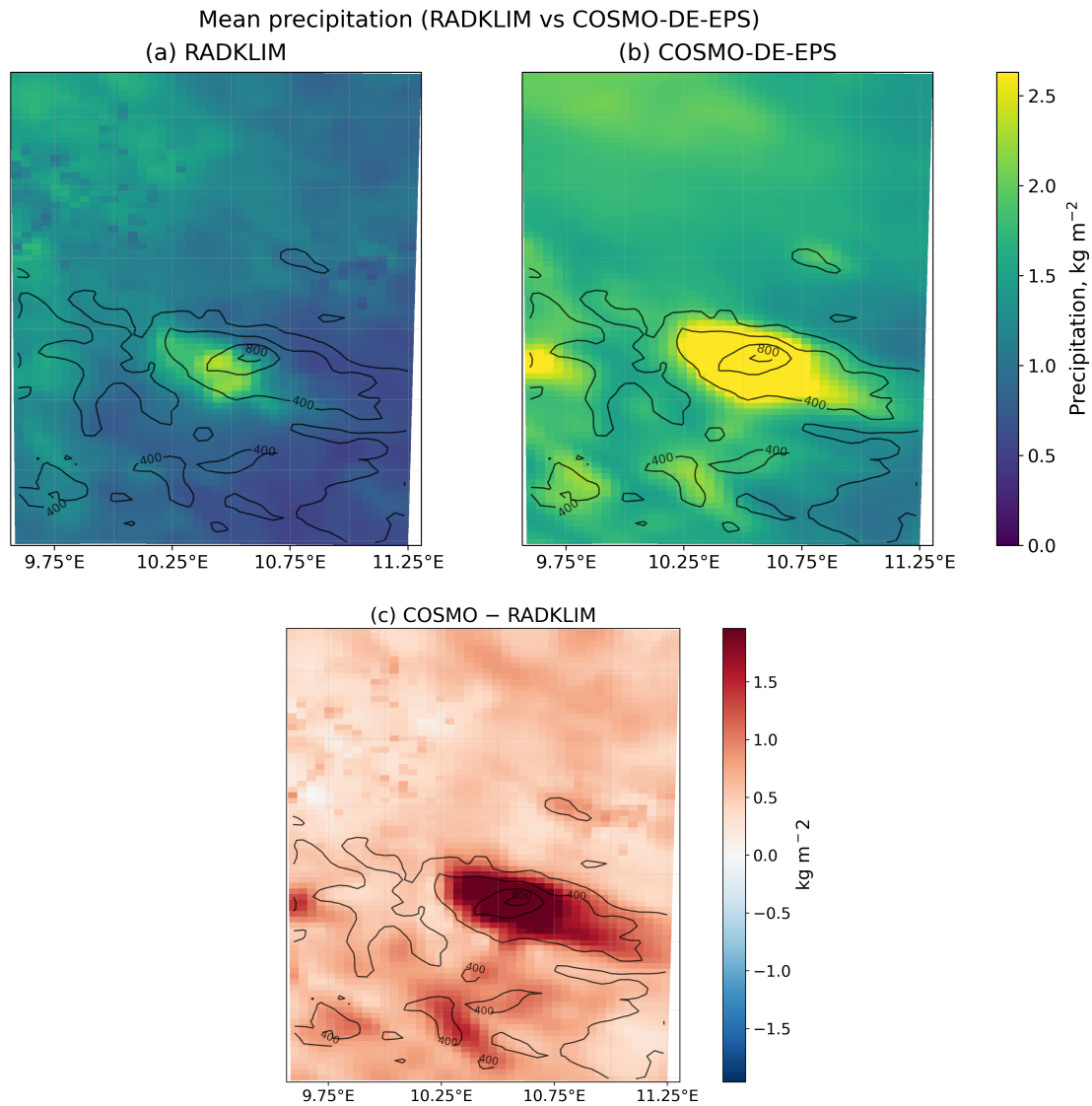


Figure 5.3: Mean precipitation over the Harz. (a) RADKLIM, (b) COSMO-DE-EPS, and (c) COSMO minus RADKLIM.

DOMINANT PRECIPITATION ERROR PATTERNS AND ASSOCIATED LARGE-SCALE CONDITIONS

Principal Component Analysis (PCA) is applied to the precipitation error field defined as the difference between precipitation accumulations from RADKLIM data (Section 3.1) and COSMO-DE-EPS forecasts (Section 3.2). Throughout this chapter, the error field is defined as RADKLIM minus COSMO, so that positive values indicate model underestimation and negative values indicate model overestimation relative to RADKLIM. Focusing on the error field rather than the raw precipitation fields directs the analysis toward systematic model–observation discrepancies in orographic precipitation. The PCA is restricted to winter (DJF), when precipitation over the Harz is dominated by synoptic-scale systems that are typically more coherent and easier to relate to upstream flow conditions than warm-season convective events.

Figure 6.1 shows the three leading PCA modes of the winter precipitation error. The first mode (PC₁) explains about 43% of the variance, the second (PC₂) about 13%, and the third (PC₃) about 6%. Together, they account for roughly two thirds of the total variability. PC₁ exhibits a broad ridge-centred structure with maxima over the main crest and along the range, indicating that the largest and most coherent model–observation differences frequently occur directly over the mountains. PC₂ has a more elongated structure with a shifted maximum and weaker flanking amplitudes, which suggests a secondary regime that is still organised at Harz scale but is more sensitive to flow orientation and the position of larger-scale precipitation systems. PC₃ displays a dipole-like structure across the ridge, with opposite-signed lobes on the windward and leeward sides. This mode is naturally interpreted as a displacement pattern, in which the dominant precipitation anomaly shifts between upstream and downstream sectors of the mountain range.

Ensemble information provides an additional perspective on the robustness of these modes. For each analysis time, the dominant PC is identified separately for all 20 COSMO-DE-EPS ensemble members. In about 60.8% of winter dates, all members project most strongly onto the same mode. This suggests that the corresponding error pattern is often relatively robust to small perturbations in initial and boundary conditions. In such cases, the large-scale evolution remains sufficiently consistent across members that the COSMO–RADKLIM mismatch is more plausibly linked to systematic model behaviour, including parameterized physics and the representation of terrain-related processes. The remaining 39.2% of dates show mixed dominant PCs across members, which indicates situations in which small perturbations shift the forecast evolution enough to move the simulation between different error regimes. In the

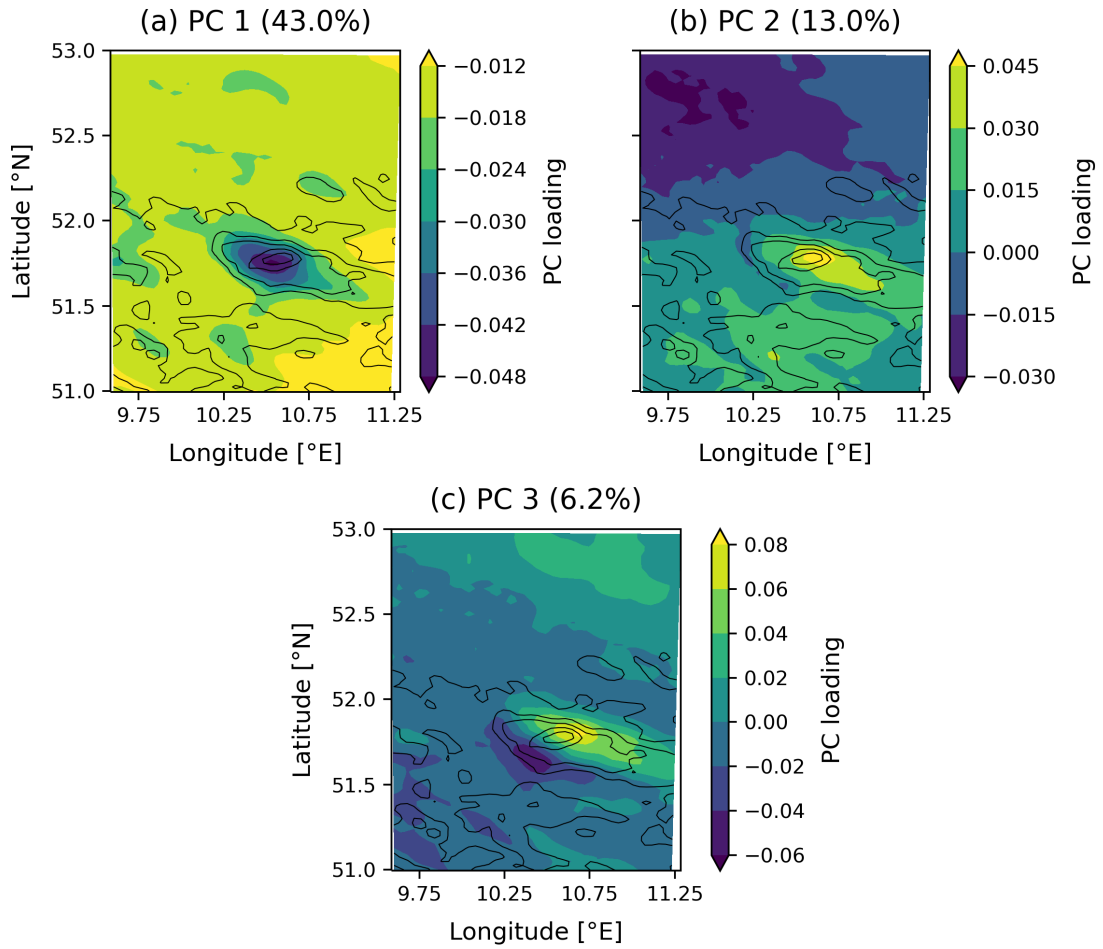


Figure 6.1: Leading PC patterns of winter RADKLIM–COSMO precipitation error over the Harz region. Panels (a)–(c) show the spatial patterns associated with PC 1, PC 2, and PC 3, which explain 43.0 %, 13.0 %, and 6.2 % of the total variance, respectively.

following, the analysis therefore focuses on dates for which all ensemble members share the same dominant PC. This selection reduces contamination by member-dependent flow evolution and highlights the systematic, regime-dependent component of the precipitation error.

To provide a complementary and physically intuitive view of the PCA modes, Fig. 6.2 shows the mean precipitation difference (RADKLIM–COSMO) computed separately for dates dominated by each PC. For PC1 and PC2, the regime-mean difference is negative over the ridge, which indicates that COSMO tends to produce larger accumulations than RADKLIM in the orographic core. The signal is most coherent over the central crest, consistent with the ridge-centred structures in Fig. 6.1. For PC3, the mean difference

exhibits a dipole-like pattern with opposite-signed anomalies on the windward and leeward sides, consistent with a displacement of precipitation relative to the ridge. While the PCA loadings in Fig. 6.1 describe the dominant directions of variability in the error field, the regime-mean composites in Fig. 6.2 illustrate how these modes translate into systematic biases when the atmosphere resides in each regime.

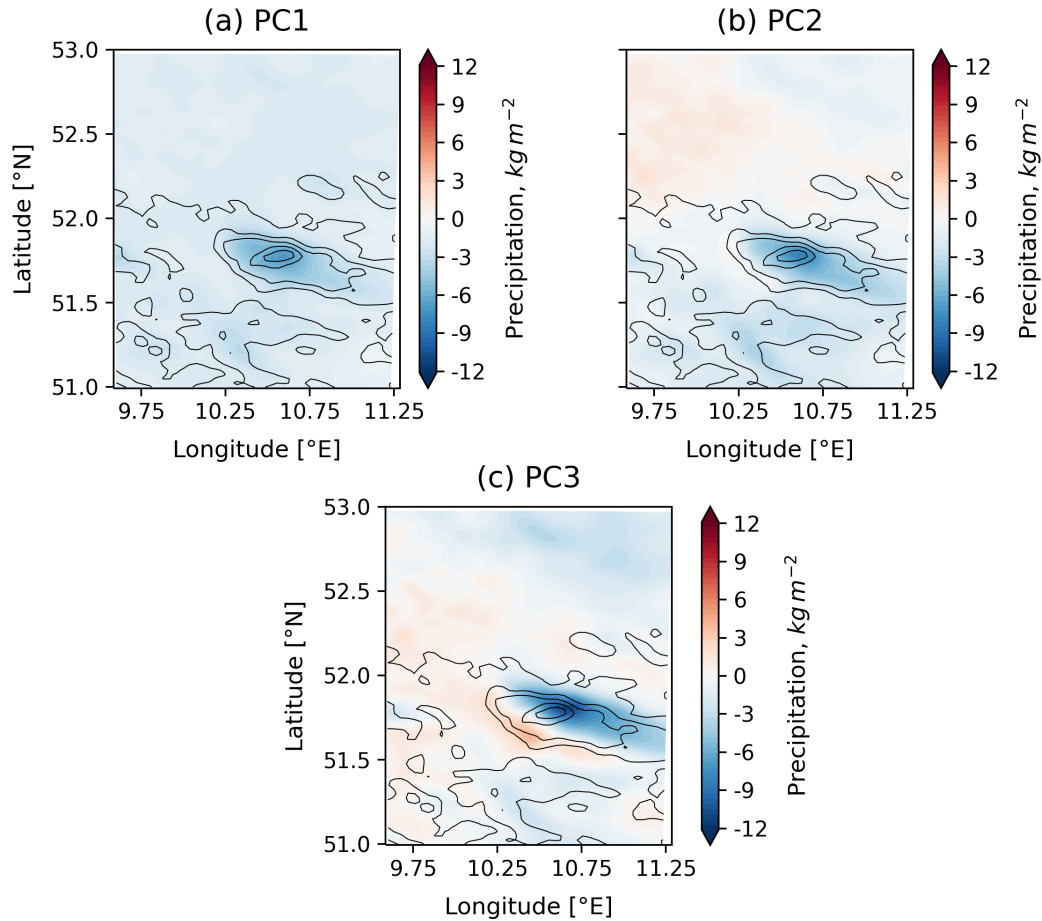


Figure 6.2: Mean RADKLIM–COSMO precipitation difference for winter dates dominated by each principal component. Panels (a)–(c) show the regime-mean difference for PC1, PC2, and PC3, respectively. Negative values indicate larger precipitation amounts in COSMO than in RADKLIM. Black contours indicate topography.

To relate the PCA modes to the large-scale environment, ERA5 reanalysis fields are used to construct composites of vertical profiles of temperature, specific humidity, wind direction, and wind speed (Fig. 6.3). For each mode, ERA5 data are averaged over the ERA5 grid box covering the Harz region (about 8×8 grid points) and over all dates on which that mode dominates. Profiles are shown up to about 4 km. Figure 6.3 displays

the median profiles for PC₁, PC₂, and PC₃, with shading indicating the interquartile range (25th–75th percentile).

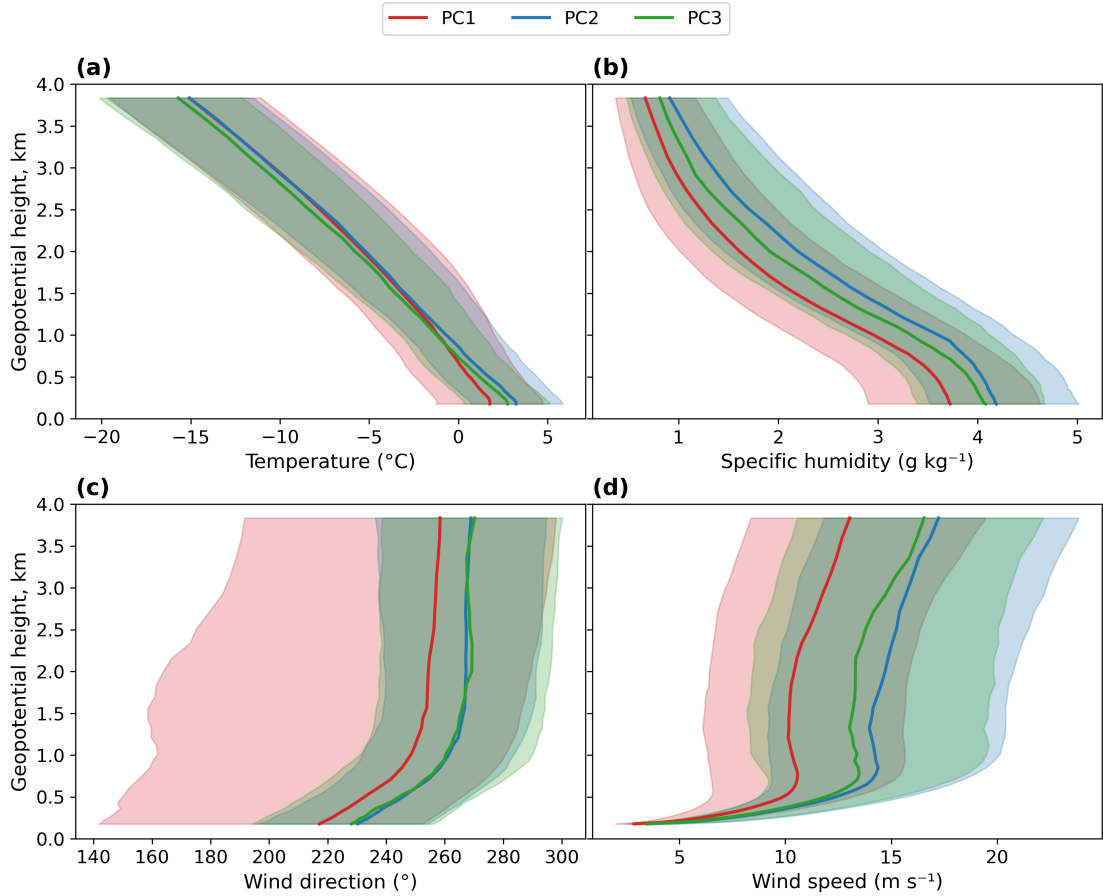


Figure 6.3: Environmental profiles associated with the three leading principal components (PCs) of precipitation error. Panels (a)–(d) show vertical profiles of temperature, specific humidity, wind direction, and wind speed, respectively, for PC₁ (red), PC₂ (blue), and PC₃ (green). Shading shows the spread between 25th and 75th percentiles for each PC.

The three modes are associated with clearly distinct environmental scenarios. PC₁ corresponds to the coldest and driest profiles, with comparatively low temperatures and low specific humidity throughout the column (Fig. 6.3a,b). Wind speeds are also generally weakest (Fig. 6.3d), while the wind direction profile is broadly similar to PC₂, with predominantly southwesterly flow and only modest directional shear (Fig. 6.3c). Taken together, these characteristics are consistent with a relatively weakly forced cold-air regime in which moisture supply is limited. In such situations, orographic uplift can still organise precipitation over the Harz, but the associated error pattern remains broad and ridge-centred rather than strongly displaced upstream or downstream.

PC2 is associated with the warmest and most humid environments of the three modes. Temperatures are higher throughout the column and specific humidity is largest, especially below about 2 km (Fig. 6.3a,b). Wind speeds are also strongest, reaching and exceeding 20 m s^{-1} in the lower and middle troposphere (Fig. 6.3d). The wind direction profile is fairly uniform with height and clustered around westerly to southwesterly flow (Fig. 6.3c), which is typical of well-organised synoptic systems such as fronts or deep low-pressure systems crossing central Europe. In this regime, strong moisture supply and strong dynamical forcing act together, so orographic uplift over the Harz is superimposed on already intense large-scale precipitation. PC2 can therefore be interpreted as a strongly forced moist regime in which strong synoptic systems interacting with the mountains produce large and spatially coherent COSMO–RADKLIM differences.

PC3 occupies an intermediate position between these two modes. Temperature and specific humidity lie between the cold, dry PC1 regime and the warm, moist PC2 regime (Fig. 6.3a,b). Wind speeds are also intermediate, stronger than in PC1 but weaker than in PC2 (Fig. 6.3d). The wind direction profile remains broadly similar to PC2, with predominantly southwesterly flow and no pronounced vertical rotation (Fig. 6.3c). This mode therefore represents moderately forced and moderately moist situations in which the large-scale environment is less extreme than in PC2 but still distinct from the cold, weakly forced PC1 regime. Under such conditions, comparatively small changes in flow and stability can shift where ascent is focused relative to the ridge, which is consistent with the dipole structure of PC3 and its interpretation as a windward–lee displacement mode.

In summary, the PCA indicates that a relatively small set of large-scale environments gives rise to distinct and recurring patterns of RADKLIM–COSMO precipitation error over the Harz. Cold and weakly forced situations favour broad ridge-centred overestimation (PC1), strongly forced moist regimes are associated with large and spatially organised errors (PC2), and intermediate regimes produce displacement-type errors between windward and leeward sectors (PC3). These three modes provide a physically interpretable framework for the more detailed case analyses that follow.

ICON CASE STUDIES

In this chapter, the PCA-based framework is used to examine individual events in more detail. For each of the three leading principal components, one or two representative dates are selected whose ERA5 upstream profiles closely match the corresponding composite profiles shown in Fig. 7.1 - 7.3. These cases are then simulated with ICON in a convection-permitting setup in order to diagnose how the model's microphysics and dynamics respond under PC1–PC3 type environments and to assess to what extent the error patterns identified in COSMO are reproduced or modified.

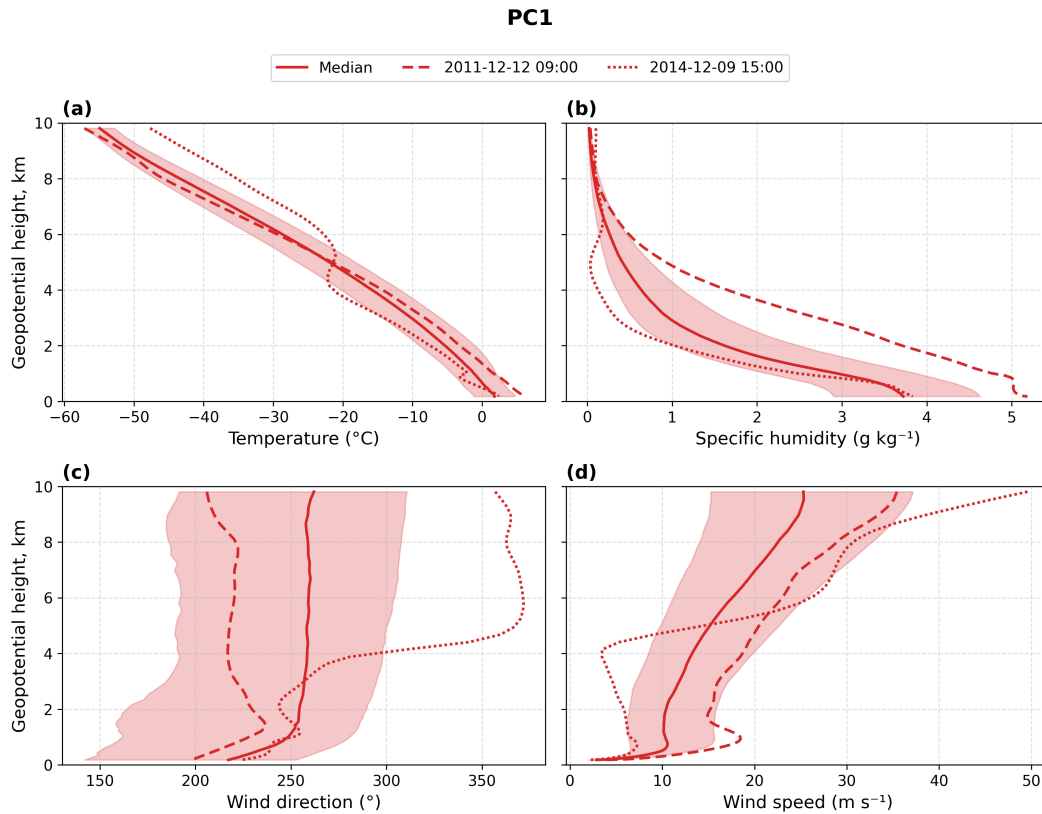


Figure 7.1: ERA5 vertical profiles for the representative ICON case study compared with the corresponding PC1 composite. The solid line shows the median profile for all dates dominated by PC1, shading indicates the 25th–75th percentile range, and dashed/dotted lines show the profiles for the specific case dates listed in Table 7.1.

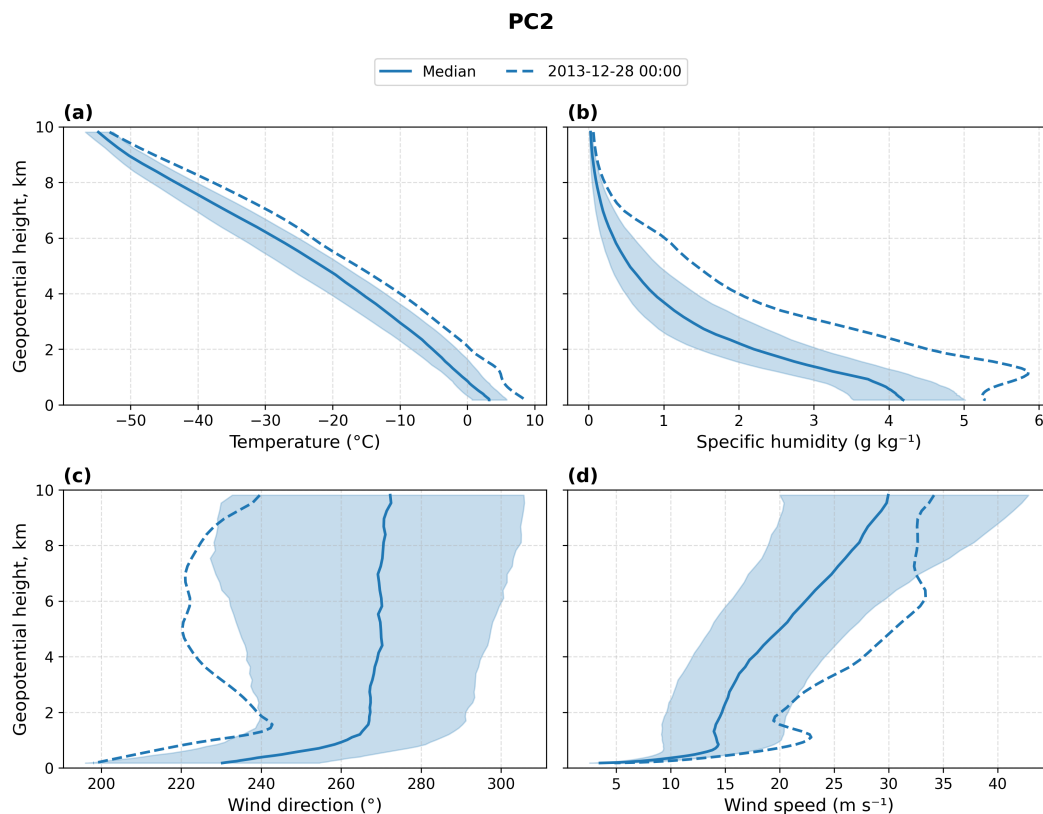


Figure 7.2: Same as Fig. 7.1, but for PC2.

For each representative case identified from the PCA analysis, I ran several ICON configurations to test the sensitivity of simulated precipitation to microphysical parameterizations (Table 7.1). The experiments varied three elements: (i) the number of prognostic moments in the microphysics scheme (single and double moment), (ii) the prescribed cloud condensation nuclei concentration (CCN_type=6 and CCN_type=8), and (iii) the inclusion or exclusion of ice phase processes.

In the single moment scheme, the model predicts only the mass mixing ratio for each hydrometeor class and assumes fixed number concentrations and size distribution parameters. The double moment scheme also prognoses number concentration, allowing droplet and particle sizes, as well as autoconversion and collection efficiencies, to respond to environmental conditions.

The CCN_type setting controls the background aerosol available for droplet activation. CCN_type=6 represents a lower-CCN environment (cleaner, more maritime-like), whereas CCN_type=8 represents a higher-CCN environment (more continental or polluted). Higher CCN concentrations generally imply more numerous but smaller droplets and typically delay warm-rain formation.

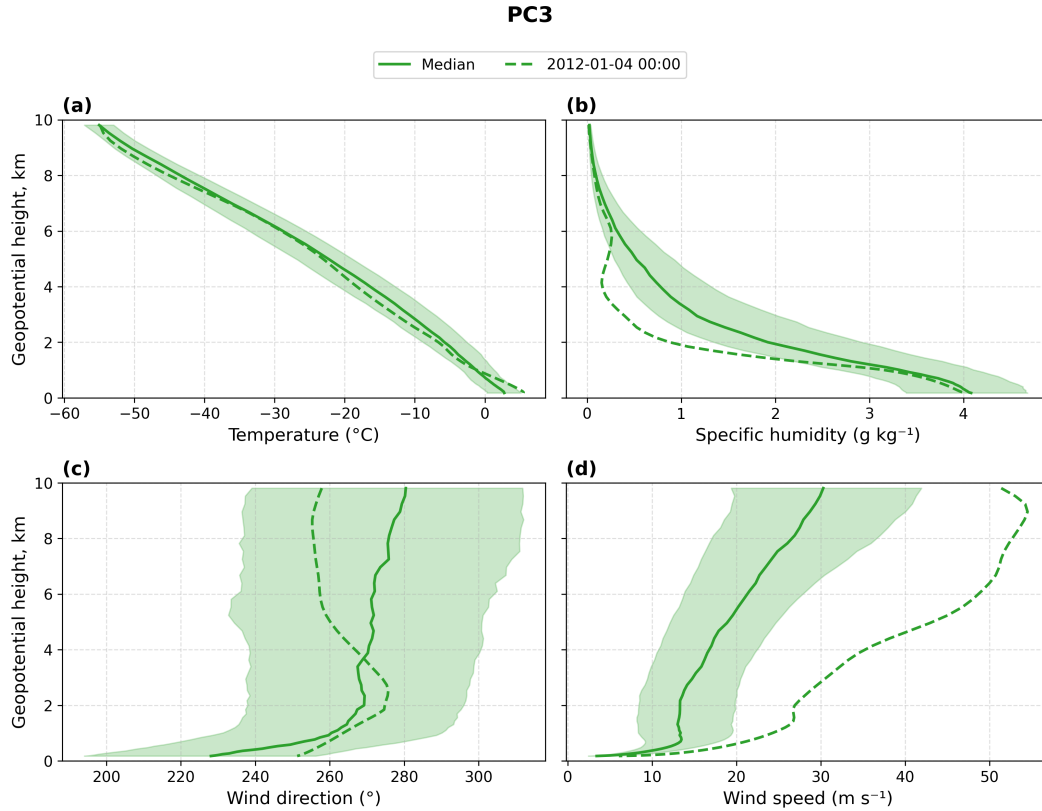


Figure 7.3: Same as Fig. 7.1, but for PC3.

The no-ice experiments disable ice phase processes (nucleation, deposition, riming, aggregation, melting) to isolate warm rain formation. Precipitation then arises from condensation, collision and coalescence, and sedimentation of liquid hydrometeors. Ice hydrometeors were also removed from the initial and boundary conditions, ensuring that no ice phase enters the domain during the integration.

An additional set of simulations at 800 m horizontal resolution was performed for the PC1 cases to test whether finer grid spacing improves the representation of terrain-induced gravity waves and corrects the precipitation bias. Differences relative to the 1.6 km baseline were minor, which indicates that the systematic bias is not primarily controlled by horizontal resolution in this setup. All subsequent sensitivity analyses therefore use the 1.6 km configuration, which provides sufficient spatial detail for the Harz while remaining computationally affordable.

This experiment design separates the effects of droplet activation and ice microphysics from those associated with the model response to orographic forcing. All sensitivity experiments performed with the 1.6 km configuration are documented in the Appendix

(Fig. A.1 – Fig. A.4). In the following chapter, only the physically consistent and diagnostically relevant cases are discussed in detail.

Table 7.1: Overview of ICON simulation configurations for the representative cases of PC₁–PC₃.

Principal Component	Date / Time (UTC)	Scheme	CCN type	Ice-phase switch
PC ₁	13.12.2011 06:00	1-moment, 2-moment	6, 8	on / off
	10.12.2014 12:00	1-moment, 2-moment	6, 8	on / off
PC ₂	28.12.2013 21:00	1-moment, 2-moment	6, 8	on / off
PC ₃	04.01.2012 21:00	1-moment, 2-moment	6, 8	on / off

7.1 PC₁

For PC₁, two representative cases were selected for detailed analysis: 13 December 2011 at 06:00 UTC and 10 December 2014 at 12:00 UTC. These dates were chosen because their ERA5 vertical profiles closely match the median conditions of the PC₁ regime.

7.1.1 Case 13.12.2011 06:00 UTC

Figure 7.4 shows the 21 h accumulated precipitation for 13 December 2011, 06:00 UTC, from RADKLIM observations, COSMO, and ICON. The ICON simulation reproduces the main spatial structure seen in COSMO, especially the localisation of precipitation near the mountainous region and the enhanced accumulation over the crest. This similarity supports the use of ICON sensitivity experiments to investigate the physical processes that may also contribute to the systematic COSMO bias.

A clear difference appears between the model fields and RADKLIM. COSMO overestimates precipitation over the central Harz and concentrates the maximum close to the main ridge, whereas RADKLIM shows a broader precipitation pattern with a wider extension over the windward side. ICON also shows a wet bias in this case, but the precipitation maximum is slightly less concentrated than in COSMO.

The wet bias in this PC₁-type situation may result from several factors. One possibility is that the model produces too much precipitation from stratiform cloud processes under cold-season conditions with weak to moderate flow. Another possibility is that terrain-forced ascent is too strongly focused near the ridge, which would enhance

condensation and hydrometeor production locally. To test whether the bias is primarily linked to the horizontal representation of terrain-induced flow, an additional simulation with 500 m grid spacing was performed. The resulting precipitation pattern changed only slightly relative to the 1.6 km baseline, indicating that the PC1 bias is not primarily controlled by horizontal resolution in this setup.

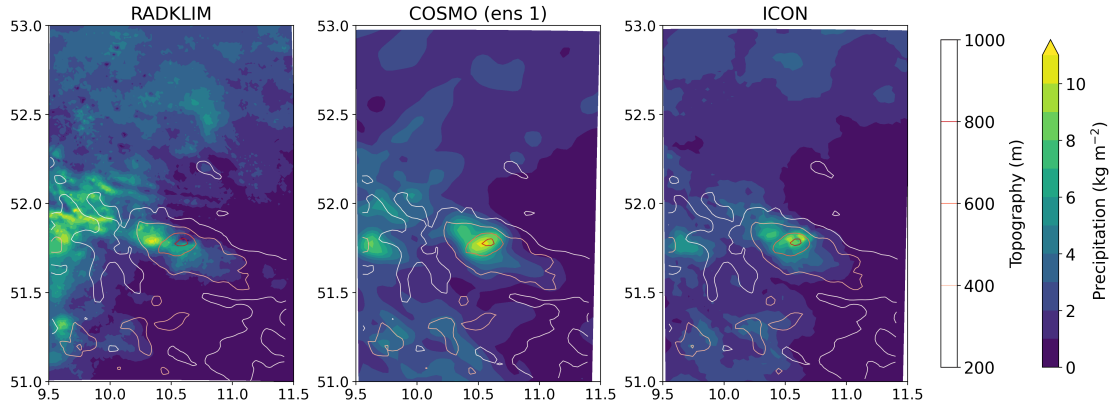


Figure 7.4: 21 h accumulated precipitation for the PC1 case (13 December 2011, 06:00 UTC) from RADKLIM, COSMO, and ICON (1-moment microphysics).

To further examine the physical structure of the precipitation-producing cloud, additional hydrometeor diagnostics were analysed, including total column integrated snow (TQS), rain (TQR), cloud water (TQC), and cloud ice (TQI). In this case, the diagnostics indicate a substantial contribution from ice-phase hydrometeors.

Figure 7.5a shows the spatial distribution of total column integrated cloud ice (TQI). Enhanced TQI occurs near and downstream of the higher terrain, which is consistent with persistent ascent and cloud-ice production in the orographically forced flow. Figure 7.5b shows a vertical cross-section of specific cloud ice content (qi) along 10.49°E . The largest qi values occur aloft, mainly in the mid- to upper troposphere, while near-surface values remain small. This vertical structure is consistent with cold-cloud and mixed-phase processes contributing to the simulated precipitation in the PC1 regime.

Together, these fields support the interpretation that ice-phase processes are active in the crest-centred precipitation maximum and may contribute to the positive precipitation bias in this case. They do not by themselves prove the dominant process pathway, but they motivate the sensitivity experiments in which ice processes are modified or removed.

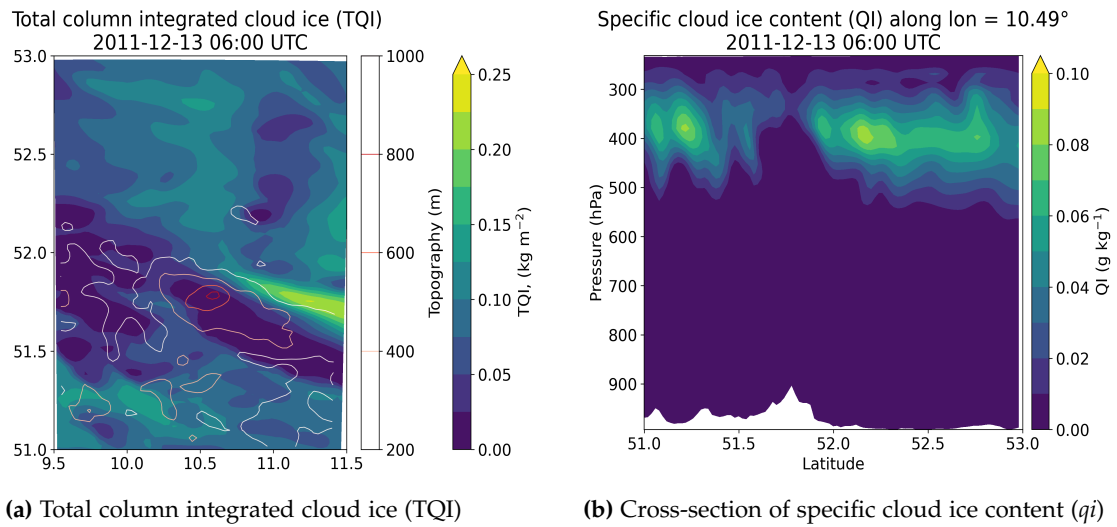


Figure 7.5: Cloud-ice diagnostics for the PC1 case (13 December 2011, 06:00 UTC) from the ICON 1-moment simulation. Panel (a) shows total column integrated cloud ice (TQI), and panel (b) shows a latitude–pressure cross-section of specific cloud ice content (q_i) along 10.49°E . Topography is shown as contours in panel (a) and as the lower boundary in panel (b).

To isolate the role of ice-phase microphysics, an additional “no-ice” sensitivity experiment was performed for the same PC1 case using `CCN_type=6` (Fig. 7.6). In this setup, ice nucleation and subsequent ice-phase growth processes are disabled, so precipitation can form only through warm-rain pathways. This experiment therefore provides a direct test of whether the crest-centred precipitation maximum in the PC1 regime depends primarily on ice processes or whether it is already imposed by the dynamical response to the terrain.

The no-ice simulation still produces a precipitation maximum over the central Harz crest, which indicates that the location of the main maximum is largely controlled by orographic lifting and the associated flow response. However, compared with the standard ICON simulation and with COSMO, the no-ice field is much weaker and less extensive, with strongly reduced precipitation outside the crest region. This shows that ice-phase microphysics substantially amplifies the precipitation amount and spatial extent in this case, even though it is not required to generate the crest-centred pattern itself.

Relative to RADKLIM, the no-ice experiment reduces the excessive precipitation intensity over the crest, but it also becomes too localized and does not reproduce the broader observed precipitation distribution on the windward side. Taken together, the sensitivity test suggests that the PC1 bias reflects a combination of dynamical focusing over the terrain and overly efficient precipitation production in the full mixed-phase simulations.

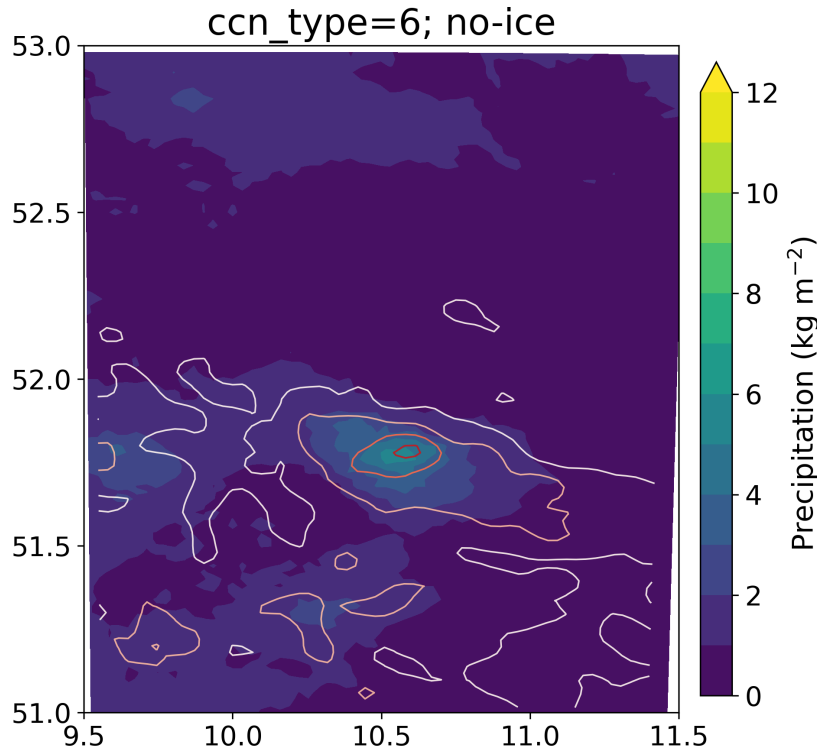


Figure 7.6: ICON precipitation field from the “no-ice” sensitivity experiment for the PC1 case (13 December 2011, 06:00 UTC) with CCN_type=6.

7.1.2 Case 10.12.2014 12:00 UTC

Figure 7.7 shows the 21 h accumulated precipitation for the second PC1 representative case (10 December 2014, 12:00 UTC) from RADKLIM, COSMO-DE-EPS, and the baseline ICON simulation. As in the first PC1 case, the precipitation maximum is located near the Harz crest. At the same time, the observed RADKLIM field is comparatively patchy and less spatially extended, whereas COSMO and ICON both produce a broader ridge-parallel precipitation band with larger areal coverage.

The common structure in COSMO and ICON suggests that the main error is not specific to one model implementation only, but reflects a more general difficulty in representing precipitation formation and distribution in this PC1-type environment. In this regime, the large-scale forcing is relatively weak and the moisture supply is limited, so moderate differences in cloud microphysical conversion rates and hydrometeor growth can strongly affect where precipitation is released and how concentrated it becomes over the terrain. The simulations reproduce the crest-related enhancement, but they maintain too much precipitation along and downstream of the ridge compared

with RADKLIM, which is consistent with the ridge-centred wet bias associated with PC1.

To examine the role of microphysics, several ICON sensitivity experiments were performed for this case (Table 7.1). The baseline simulation shown in Fig. 7.7 uses the one-moment microphysics configuration. Relative to this baseline, the two-moment scheme improves the spatial structure of the precipitation field (Fig. 7.8), mainly by reducing the overly smooth and elongated ridge-parallel pattern and by producing a more realistic localisation of precipitation near the observed maximum. This indicates that allowing hydrometeor number concentrations to evolve prognostically improves the response of the model to the weakly forced orographic precipitation regime represented by PC1.

Additional sensitivity experiments with CCN_type=6 and CCN_type=8 indicate that CCN assumptions further modulate the precipitation distribution. Higher CCN concentrations tend to reduce warm-rain efficiency and can weaken or slightly shift the accumulated precipitation maximum, which in some configurations moves the simulated pattern closer to RADKLIM. The two-moment formulation also improves the spatial structure in this case, mainly in terms of the placement and extent of the ridge-centred maximum. Because some sensitivity plots are shown with different colour scales, this comparison is interpreted primarily in terms of pattern shape rather than absolute intensity. In contrast, disabling the ice phase produces only small changes in the accumulated precipitation field for this event, suggesting that warm-phase processes exert the stronger control on the precipitation response here than in the first PC1 case. Taken together, these results support the interpretation that, for this event, the spatial bias is more sensitive to the representation of liquid-phase microphysics (especially the one- vs two-moment formulation) than to ice-phase processes alone.

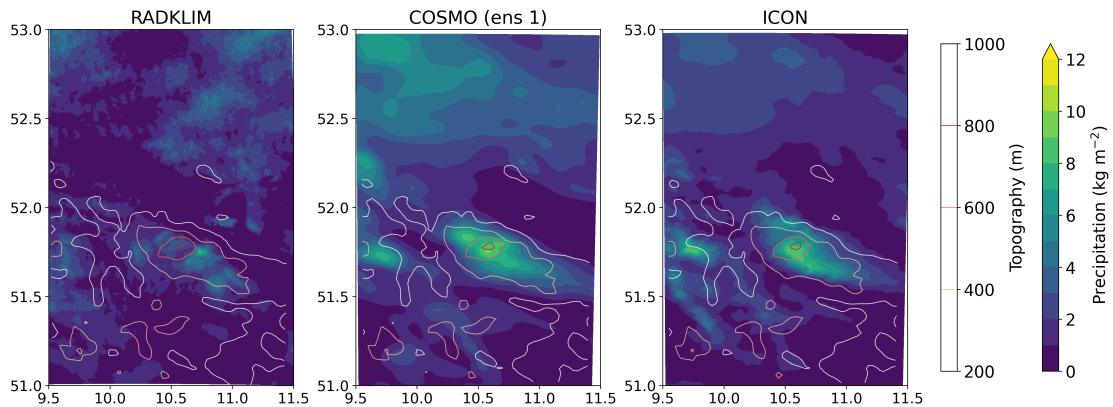


Figure 7.7: 21 h accumulated precipitation for the PC1 case (10 December 2014, 12:00 UTC) from RADKLIM (left), COSMO-DE-EPS (centre; one representative member), and ICON (right; baseline one-moment configuration).

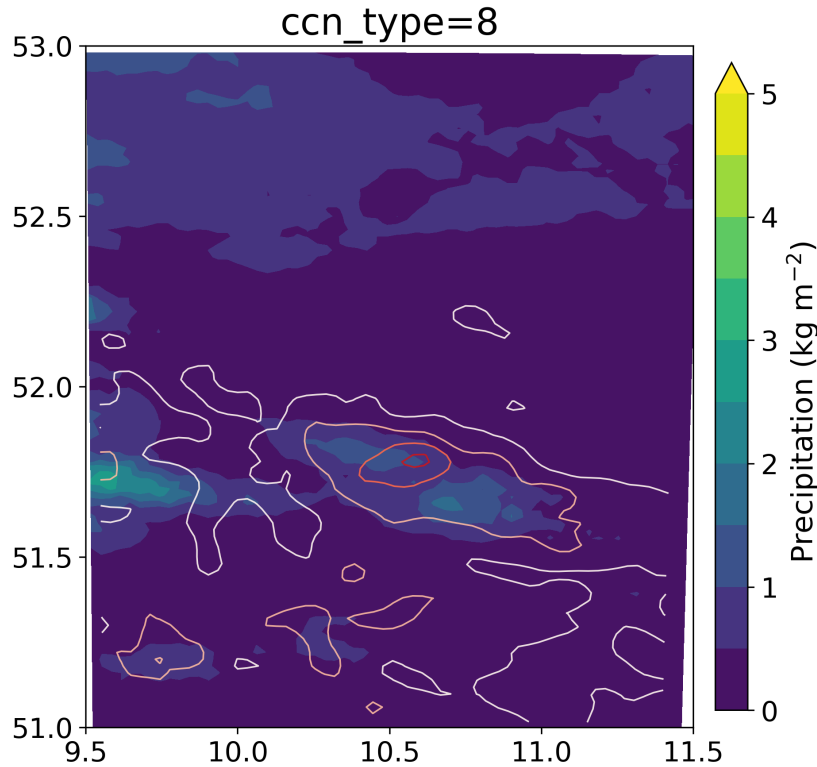


Figure 7.8: ICON precipitation field for the PC1 case (10 December 2014, 12:00 UTC) using the two-moment microphysics scheme with CCN_type=8. Compared with the baseline one-moment run, the precipitation pattern is less compact and better matches the observed spatial structure.

PC1 case summary

Both PC1 cases exhibit the same characteristic bias, namely an overestimation of precipitation that is too tightly confined to the orographic ridge (Figs. 7.4, 7.7). At the same time, they occur in contrasting thermodynamic and kinematic environments, as indicated by the domain-mean profiles (Fig. 7.9).

The 13 December 2011 case (blue) features a markedly colder free troposphere and generally moister conditions aloft. Above about 800 hPa, temperatures are clearly lower than in the 2014 case, while specific humidity remains comparatively elevated through much of the mid to upper troposphere. Winds also strengthen strongly with height and are overall stronger through a large part of the lower to mid-troposphere. These conditions are consistent with deeper cloud development and enhanced ice-phase growth aloft, in line with the elevated tq_i/q_i signals (Fig. 7.5) and an ice-dominated contribution to the precipitation excess.

The 10 December 2014 case (orange) is warmer and generally drier in the free troposphere, while being slightly cooler near the surface. Compared with 2011, it shows weaker winds through much of the lower and mid-levels, together with a more westerly flow direction in the mid to upper troposphere. This combination is consistent with shallower cloud structures and a relatively larger rain contribution. The more westerly flow also provides a plausible explanation for the observed downstream (north-easterly) displacement of the precipitation maximum relative to the ridge.

Taken together, the profiles support the physical interpretation developed above. Colder and moister upper-level conditions in 2011 favour deeper, more ice-rich clouds, whereas the warmer and drier free-tropospheric environment in the 2014 case favours lower cloud tops and a stronger warm-rain contribution. Despite these regime differences, the same ridge-centred overamplification persists in both cases. This robustness suggests that the bias is not controlled by thermodynamics alone. Instead, it suggests a structural issue in the coupling between terrain-induced ascent and microphysics, in which precipitation production may be too efficient for a given orographic forcing in both cold and warm regimes. In other words, the model response of the dynamics–microphysics system to uplift appears too strong, largely independent of the dominant hydrometeor phase.

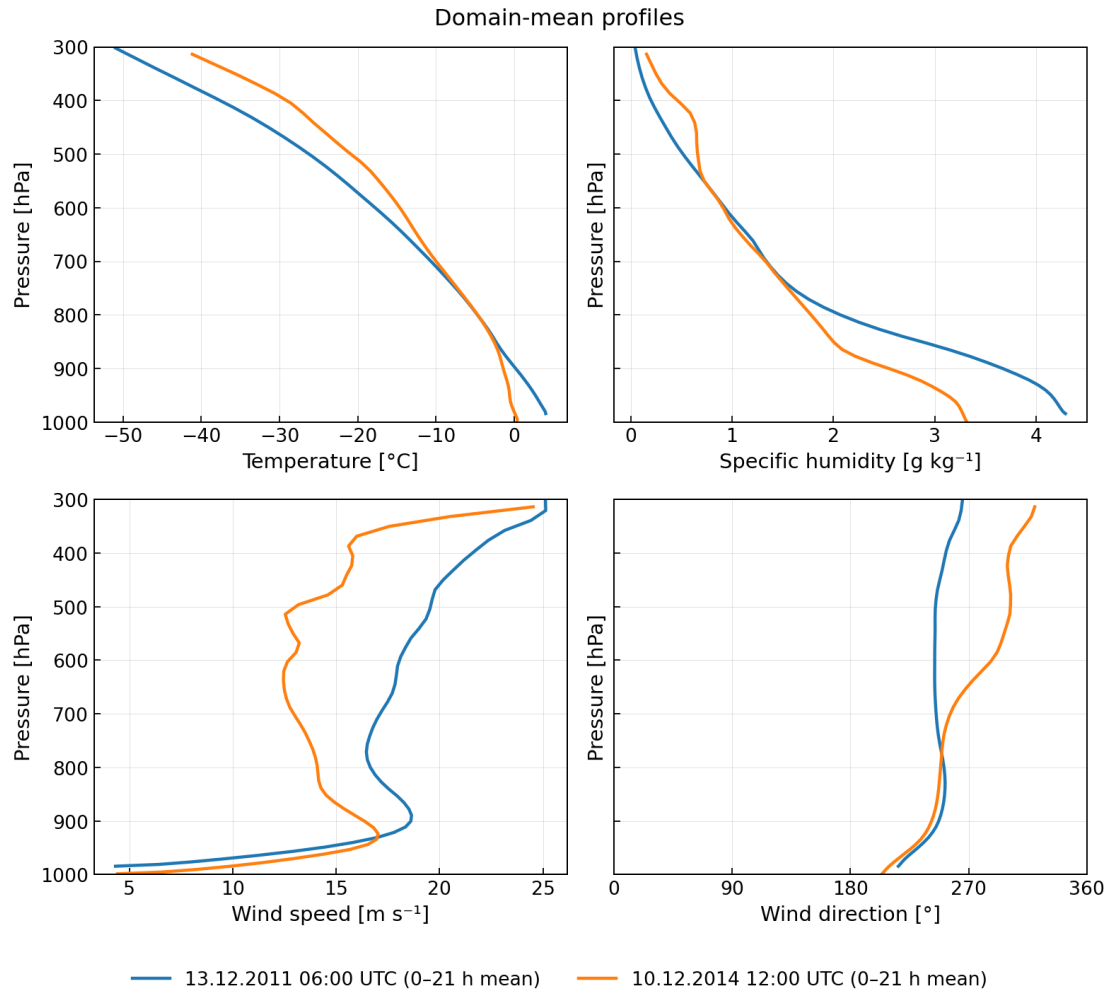


Figure 7.9: Domain-mean upstream profiles for the two PC1 representative cases (0–21 h mean): (a) temperature, (b) specific humidity, (c) wind speed, and (d) wind direction. Blue denotes 13 December 2011 06:00 UTC and orange denotes 10 December 2014 12:00 UTC.

7.2 PC2

The representative case for PC2 (28.12.2013, 21:00 UTC) corresponds to a frontal precipitation event. Figure 7.10 compares the 21-hour accumulated precipitation from RADKLIM, COSMO-DE output, and ICON simulations. In contrast to PC1, this component is characterised primarily by a displacement and structural error of the precipitation band, rather than a ridge-centred amplitude overresponse.

In the observations, precipitation is relatively localized, with a compact maximum near the central Harz. By contrast, COSMO produces a broader and more intense precipitation area across the central domain, including substantial accumulation where RADKLIM shows much weaker totals. ICON reproduces the same general large-scale band orientation but with an even more spatially diffuse field and widespread precipitation across the southern and central parts of the domain. Thus, the dominant error in this case is not simply an intensity bias at the ridge, but a misrepresentation of frontal position and spatial extent.

The spatial pattern in this event, together with the ERA5 composite profiles (Fig. 6.3), is consistent with the PC2 loading structure (Fig. 6.1), which supports the interpretation of PC2 as a frontal placement and organisation bias. In other words, PC2 mainly captures errors in where and how the frontal precipitation is distributed over the domain, rather than the orographic overamplification that dominates PC1.

PC2 therefore represents situations in which the model misplaces or overly broadens the synoptic-scale precipitation band, which then projects onto the Harz region as errors in precipitation location, timing, and spatial structure. This interpretation is consistent with a dynamically controlled bias mode linked to the large-scale flow and frontal advection.

Across the ICON sensitivity experiments (Fig. A.3), both single-moment and double-moment configurations reproduce the general frontal pattern while differing mainly in precipitation intensity and smoothness. The single-moment setup tends to retain sharper local maxima, whereas the double-moment configuration produces a smoother field with reduced peak intensity. Increasing the CCN concentration (CCN_type=8) or disabling ice processes further weakens precipitation amounts, but does not substantially correct the misplaced band structure. This indicates that microphysical changes modulate amplitude more than position. The persistence of the spatial error across configurations suggests that the dominant source of the PC2 mismatch is dynamical rather than microphysical.

Overall, this case highlights a distinct error mechanism from PC1. Whereas PC1 is associated with excessive terrain-locked enhancement, PC2 is dominated by synoptic-scale precipitation placement and spreading errors. The comparatively weak sensitivity of precipitation location to microphysical configuration supports the interpretation of PC2 as a large-scale dynamical bias mode.

7.3 PC₃

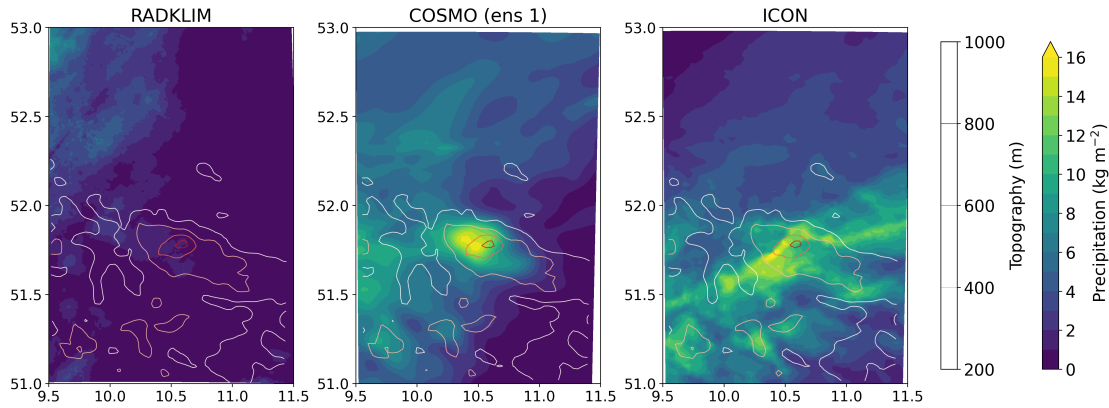


Figure 7.10: 21-hour accumulated precipitation for the PC₂ case (28.12.2013, 21:00 UTC) from RADKLIM, COSMO-DE, and ICON (shown here for the one-moment configuration). White contours denote topography.

7.3 PC₃

The representative case for PC₃ (04.01.2012, 21:00 UTC) corresponds to a localized precipitation event under comparatively weak synoptic forcing. Figure 7.11 compares the 21-hour accumulated precipitation from RADKLIM, COSMO-DE output, and ICON simulations. All three datasets show a compact precipitation maximum anchored near the Harz, which is consistent with the localized character of the PC₃ pattern. At the same time, notable differences remain in amplitude and spatial structure.

RADKLIM shows a relatively compact core near the ridge together with weaker surrounding precipitation and an additional band farther north. COSMO reproduces the central maximum but makes it too broad and too intense, with a pronounced elongation toward the east to southeast. ICON also captures the terrain-related maximum and its general placement, but with a smoother and slightly broader structure than observed. Thus, the dominant mismatch in this case is a combination of amplitude and local placement/shape errors rather than a large-scale frontal displacement.

The error structure is consistent with the PC₃ pattern (Fig. 6.1), which can be interpreted as a localized shift and redistribution of precipitation along the windward slope and ridge region. In particular, the models tend to place the strongest accumulation more directly over the ridge and downstream flank than in RADKLIM, while retaining the same terrain-locked character. This supports the interpretation of PC₃ as a localized displacement and organisation bias mode.

The ICON sensitivity experiments indicate that changes in microphysical configuration (single- vs. double-moment, different CCN_{type} settings, or disabling ice processes) primarily affect the peak intensity and smoothness of the precipitation core, but have only limited influence on its location. This suggests that microphysics modulates ampli-

tude, while the remaining positional mismatch is controlled mainly by dynamics and boundary-layer structure.

The persistence of the misplaced or overly broadened precipitation maximum across configurations suggests a dynamical or thermodynamic control, potentially related to the representation of low-level stratification, turbulent mixing, and the onset of terrain-induced ascent. In this regime, even small differences in near-surface stability or flow direction can shift the precipitation peak by a short distance and change how strongly the maximum is locked to the ridge. The precise contribution of these processes cannot be isolated from this case alone, so the interpretation remains a physically plausible hypothesis rather than a definitive attribution.

Overall, PC₃ represents a localized terrain-related bias mode that is distinct from both PC₁ and PC₂. Unlike the ridge-wide overamplification in PC₁ and the synoptic-scale placement error in PC₂, PC₃ is dominated by local errors in the position, shape, and intensity of a compact precipitation maximum near the mountain barrier.

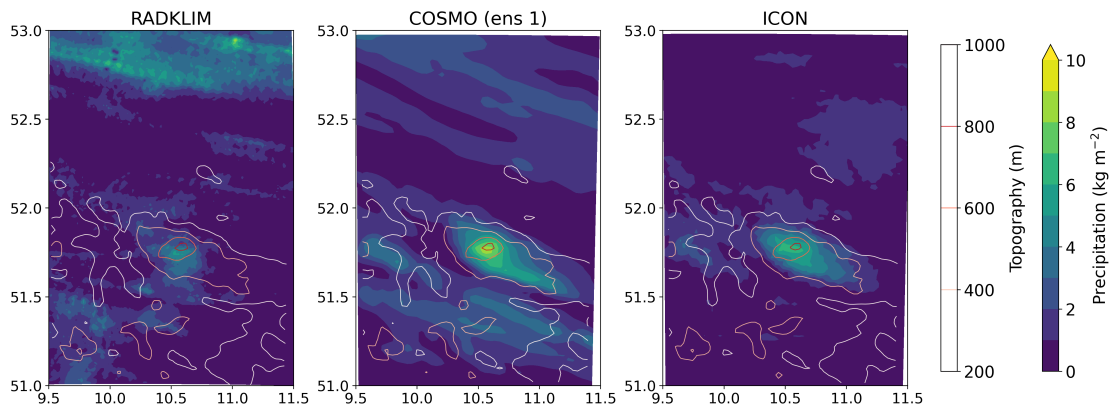


Figure 7.11: 21-hour accumulated precipitation for the PC₃ case (04.01.2012, 21:00 UTC) from RADKLIM, COSMO-DE, and ICON (shown here for the one-moment configuration). White contours denote topography.

7.4 Summary of case analyses

The three investigated cases represent distinct meteorological regimes that project onto different PCA-derived error patterns. PC₁ is associated with localized orographically enhanced precipitation and a systematic ridge-centred overestimation. PC₂ captures synoptic-scale frontal situations in which the main error is the misplaced and overly broad distribution of precipitation linked to frontal placement and organisation. PC₃ reflects a localized terrain-related error mode with small-scale shifts and redistribution of the precipitation maximum near the ridge. Together, these cases show that different

meteorological situations contribute to different systematic error structures in COSMO and ICON.

The case analyses further indicate that microphysical configuration alone cannot explain the identified biases. Changes in the number of moments, CCN concentration, and ice-phase treatment can modify precipitation intensity and smoothness, but they do not fundamentally alter the dominant spatial error patterns. This is particularly evident for PC₂, where the frontal placement error persists across configurations, and for PC₃, where the location of the compact terrain-locked maximum remains only weakly sensitive to microphysical changes. These results suggest that the leading error modes arise from coupled interactions between dynamics, thermodynamic structure, microphysics, and terrain forcing rather than from a single parameterization component.

At the same time, the present sensitivity experiments sample only a small subset of the model parameter space. Microphysical schemes contain a large number of tunable parameters, while only a limited set was varied systematically here. The findings should therefore be interpreted as a first step toward constraining the role of microphysics in systematic precipitation errors, not as a complete attribution. Overall, the results indicate that microphysics influences local precipitation amplitude, whereas the large-scale and terrain-locked error structures are more likely linked to coupled process interactions within the full model system.

To further isolate the governing mechanisms, semi-idealised simulations were initialized for representative cases from each principal component. These experiments are discussed in the following chapter.

SEMI-IDEALISED SIMULATIONS

To expand the analysis beyond individual case studies, I set up semi-idealised simulations with the ICON model. This approach makes it possible to study precipitation error patterns in a controlled setting and to assess whether certain model biases become more apparent under specific atmospheric conditions. These simulations help bridge the gap between the detailed insights from individual case studies and a broader understanding of how the model behaves across different principal components.

8.1 Model setup

I use the limited-area, non-hydrostatic ICON model (version 2.6.2.2) on a regular Cartesian grid with 500 m horizontal spacing. This resolution is fine enough to resolve the main ridge and its slope-scale flow features (lee waves, blocking, channeling) without the high cost of a large-eddy setup. The lateral boundaries are periodic. Periodic boundaries were chosen so that air leaving one side of the domain re-enters on the opposite side with the same properties. This avoids artificial inflow/outflow conditions and keeps the upstream state fixed by design, which is exactly what is wanted for semi-idealised tests focused on terrain-induced processes rather than on boundary forcing. This idealisation also removes realistic synoptic evolution and upstream moisture variability, but this is a deliberate trade-off in order to isolate the terrain-flow response under controlled background conditions. Apart from the terrain (Sec. 8.1.2), everything is horizontally uniform, there is no nudging and no time-varying lateral forcing.

8.1.1 Initial conditions

The initial atmosphere is intentionally simple: a horizontally homogeneous column taken from the median upstream composites for each principal component (PC)—i.e. the domain-mean profiles of temperature $T(z)$, specific humidity $q(z)$, and winds $(u(z), v(z))$ shown in Fig. 7.9 (and the analogous figures for other PCs). Using the PC medians keeps the cases representative while eliminating mesoscale variability, so any structure that develops can be attributed to the interaction between this “typical” upstream state and the ridge. Hydrostatic balance provides pressure and potential temperature, all hydrometeors start at zero. No synoptic tendencies or external forcing are applied.

8.1.2 Topography adjustment

Periodic lateral boundaries only work cleanly if the terrain also matches at the edges. In the original ICON external parameter file, the Harz ridge sits well inside the domain, but the foothills on the southern boundary are noticeably higher than those on the northern boundary. If one used this field directly with periodic boundaries, air leaving the domain in the north would re-enter over a different surface height in the south, producing artificial jumps in pressure and flow.

To avoid this, I slightly modified the topography before using it in the semi-idealised setup. First, the ICON topography was interpolated from its native unstructured triangular grid to a regular latitude–longitude grid covering the Harz region (9.5–11.7°E, 51–53°N). I then adjusted only the southern part of the domain: for each longitude column, the elevations in a narrow latitudinal band near the southern boundary were smoothly scaled so that the height at the very edge approaches the height at the northern edge of the domain. Above about 51.5°N the terrain is left unchanged, so the main ridge and its local structure are preserved. Finally, the adjusted field was interpolated onto the ICON torus grid and used in the simulations. Figure 8.1 illustrates the results. The adjustment mainly lowers and smooths the southern foothills, while the shape and amplitude of the Harz ridge remain almost identical to the original field.

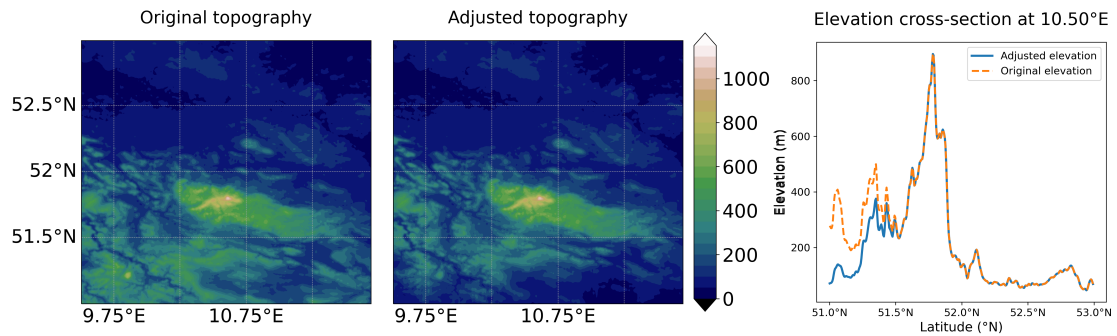


Figure 8.1: Comparison of the original and adjusted topography used for the semi-idealised simulations. The left and middle panels show elevation maps before and after the adjustment, while the right panel presents a north–south cross-section at 10.5°E.

8.2 Results

After preparing the adjusted topography and horizontally homogeneous initial states, the next step was to examine how ICON responds to the characteristic atmospheric conditions associated with each principal component. The semi-idealised simulations provide a controlled framework in which topography and model physics are held fixed

across runs, while only the initial thermodynamic and wind profiles vary according to the PC composites.

Each simulation was integrated for 21 hours. Precipitation developed during the first hours and reached a quasi-steady state after approximately 10–12 hours, with only minor additional accumulation thereafter. For consistency, all results shown below refer to 12-hour accumulated precipitation, which represents the mature phase of the simulated orographic response.

The following sections first describe the spatial and vertical precipitation structures for PC1–PC3 and compare them with COSMO and RADKLIM. The analysis then links these differences to flow regime using the moist Froude number, which provides a compact measure of the balance between flow strength, stability, and mountain forcing.

8.2.1 Spatial distribution of precipitation

Figure 8.2 shows the 12-hour accumulated precipitation for the semi-idealised ICON simulations initialised from the PC1–PC3 composite environments. Although all three simulations produce a terrain-related precipitation signal near the Harz ridge, the total amount and spatial extent differ between the setups.

The PC1 setup produces only weak precipitation, with maxima of about 0.7 kg m^{-2} . This is consistent with the colder and relatively drier composite environment of PC1 (Fig. 6.3), in which limited moisture and comparatively weak flow restrict the efficiency of sustained orographic precipitation formation. Precipitation remains confined to small and patchy areas near the ridge top, indicating a relatively weak mountain-wave and uplift response in this regime.

By contrast, the PC2 setup yields substantially larger accumulations, locally exceeding 3 kg m^{-2} , and a broader, more continuous precipitation pattern across the windward slope and ridge region. This behaviour is consistent with the warmer, moister, and faster-flowing PC2 environment (Fig. 6.3), which promotes stronger ascent and more efficient condensation over the terrain.

The PC3 setup also produces accumulations above 3 kg m^{-2} , but with a different spatial organisation. The main precipitation maximum is more tightly anchored to the windward slope and decreases more rapidly across the crest and into the lee than in PC2. This indicates a stronger confinement of precipitation to the immediate orographic forcing region, despite broadly comparable total intensity.

Overall, the semi-idealised simulations demonstrate that differences in upstream temperature, humidity, and flow profiles represented by the three leading PCs translate into distinct precipitation intensities and spatial structures, even in the absence of synoptic variability. This supports the interpretation of the PCs as physically meaningful modes of terrain–flow interaction under the simplified background conditions considered here.

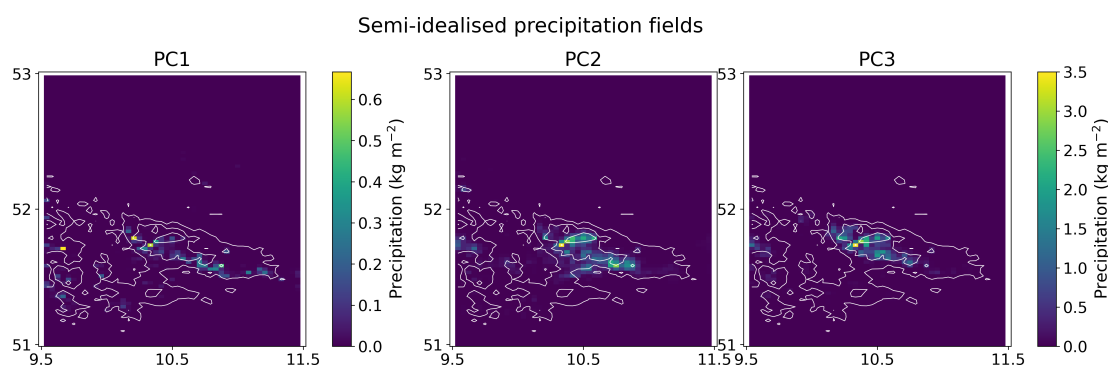


Figure 8.2: 12-hour accumulated precipitation generated by the ICON model in the semi-idealised simulation setup for each principal component.

8.2.2 Cross-section comparison with COSMO and RADKLIM

To place the semi-idealised results in context, Figure 8.3 compares cross-sections of precipitation rate along 10.5°E for COSMO, RADKLIM, the ICON case simulations, and the corresponding semi-idealised ICON runs. This comparison relates the simplified experiments to the full-physics model behaviour diagnosed in the real cases.

Across all PCs, COSMO tends to produce stronger ridge-centred precipitation than RADKLIM, particularly near the mountain peak. The semi-idealised simulations reproduce the tendency toward terrain-locked enhancement, but generally with lower intensities and simpler structures, especially downstream of the crest. This is expected because the semi-idealised setup removes synoptic variability, mesoscale heterogeneity, and time-varying boundary forcing.

For PC₁, the cross-sections are relatively weak overall, consistent with a weakly forced regime. For PC₂ and PC₃, COSMO and the case-based ICON simulations show sharper and more intense peaks than RADKLIM, while the semi-idealised runs reproduce the qualitative tendency for ridge-focused enhancement under stronger-flow conditions. In PC₃, the semi-idealised response also supports the interpretation that small shifts in the location of maximum ascent relative to the ridge can produce a displacement-type error signature.

Together, these comparisons confirm that the precipitation biases seen in the full COSMO simulations (overestimation of totals and excessive concentration near or just downstream of the ridge) can arise even in the absence of complex synoptic variability. This supports the point that part of the error is structural, linked to the representation of terrain-induced ascent and microphysical response, rather than purely case-specific or driven by boundary conditions.

Taken together, these comparisons indicate that key aspects of the ridge-centred precipitation bias can arise even in a highly simplified framework. This supports the

8.2 RESULTS

view that part of the bias is structural and linked to the representation of terrain-induced ascent and precipitation production, rather than being solely a consequence of case-specific synoptic evolution or boundary-condition errors.

The remaining differences in intensity and spatial structure across the three PCs point to differences in flow regime. To interpret these contrasts dynamically, the next section analyses the moist Froude number (F_w), which quantifies the balance between flow strength, stability, and mountain height.

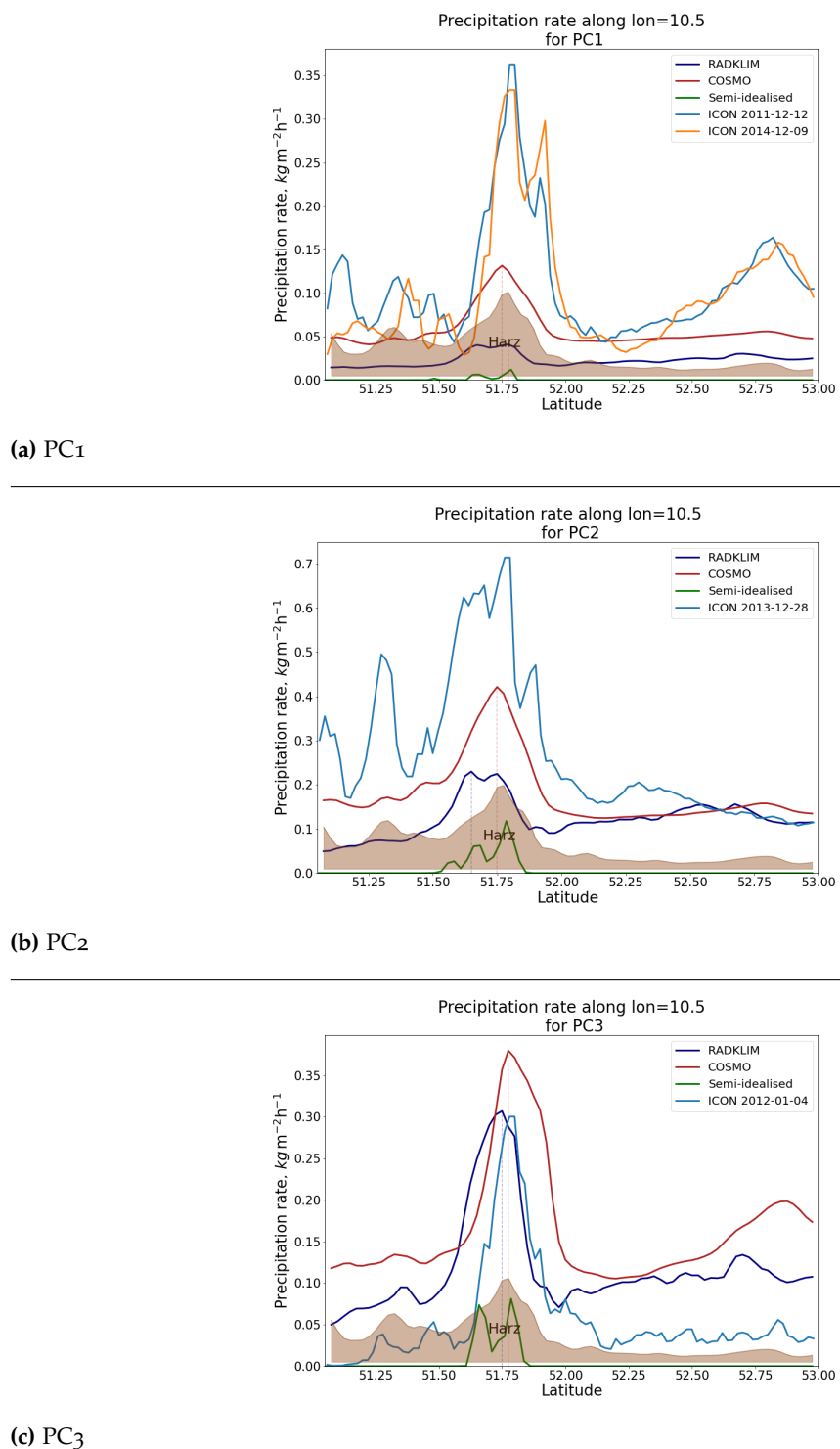


Figure 8.3: Cross-sections of precipitation rate along 10.5°E for COSMO, RADKLIM, and semi-idealised ICON simulations corresponding to PC₁–PC₃.

8.2.3 Moist Froude number

The moist Froude number (F_w) is an adaptation of the classical Froude number for moist atmospheric flows that interact with topography. It represents the balance between the flow strength and atmospheric stability and thus indicates how efficiently air can rise over the mountain barrier. It is defined as:

$$F_w = \frac{U}{N_m h}, \quad (8.1)$$

where:

- U is the mean wind speed approaching the terrain,
- N_m is the moist Brunt–Väisälä frequency, which accounts for atmospheric stability under saturated conditions,
- h is the height of the mountain.

The Brunt–Väisälä frequency (N_m) describes how strongly a displaced air parcel tends to return to its original level. For this analysis, I followed the formulation by Durran and Klemp (1982), which incorporates the effects of latent heating and provides a more realistic measure of stability in moist, orographically influenced flows. The quantities entering F_w are calculated from the composite upstream profiles used to initialise the semi-idealised simulations, so that the diagnosed flow regime is directly linked to the background conditions representative of each principal component:

$$N_m^2 = \frac{g}{T} \left(\frac{dT}{dz} + \Gamma_m \right) \left(1 + \frac{Lq_s}{RT} \right) - \frac{g}{1 + q_w} \frac{dq_w}{dz}, \quad (8.2)$$

where T is temperature, g is the gravitational acceleration, q_w is the total water mixing ratio, q_s is the saturation mixing ratio, L is the latent heat of vaporization, R is the gas constant for dry air, and Γ_m is the moist adiabatic lapse rate.

The value of F_w determines the flow regime. When $F_w < 1$, the flow is blocked, and the air is mostly deflected around the mountain, with precipitation forming mainly on the windward side. When F_w is close to 1, the flow is partially blocked, and part of the air can pass over the ridge, producing precipitation near the crest and extending into the lee. When $F_w > 1$, the flow becomes unblocked and can move freely across the mountain, shifting precipitation toward and beyond the ridge crest.

By analysing F_w for each PC, one can relate the simulated precipitation patterns to the flow dynamics and stability conditions that govern orographic precipitation formation.

Vertical and statistical distribution.

Figure 8.4 shows vertical profiles and histograms of the moist Froude number (F_w) for the three principal components.

For PC1 (Fig. 8.4a,b), F_w remains mostly below 1 throughout the column, with many values concentrated between about 0.4 and 0.8. This indicates predominantly blocked or partially blocked flow conditions. This behaviour is consistent with the colder, drier, and more weakly forced PC1 environment identified in the ERA5 composites (Fig. 6.3), and with the weak precipitation response in the semi-idealised PC1 simulation (Fig. 8.2).

For PC2 (Fig. 8.4c,d), F_w spans a broader range and frequently exceeds 1, indicating a transition toward partially blocked to unblocked flow. The histogram shows substantial probability across both sub-unity and super-unity values, which suggests that the PC2 regime contains multiple dynamical sub-states. This is consistent with the warmer, moister, and stronger-flow composite environment of PC2, which supports stronger orographic ascent and more widespread precipitation in the semi-idealised runs (Fig. 8.2).

For PC3 (Fig. 8.4e,f), the F_w distribution is also broad and appears partly bimodal, with one peak at lower values and another at higher values. This suggests that PC3 likewise contains more than one flow regime, despite its coherent PCA loading pattern. In physical terms, PC3 appears to include situations in which modest changes in wind speed or stability shift the flow between more blocked and less blocked states, which is consistent with the displacement-type precipitation behaviour associated with this mode.

8.2 RESULTS

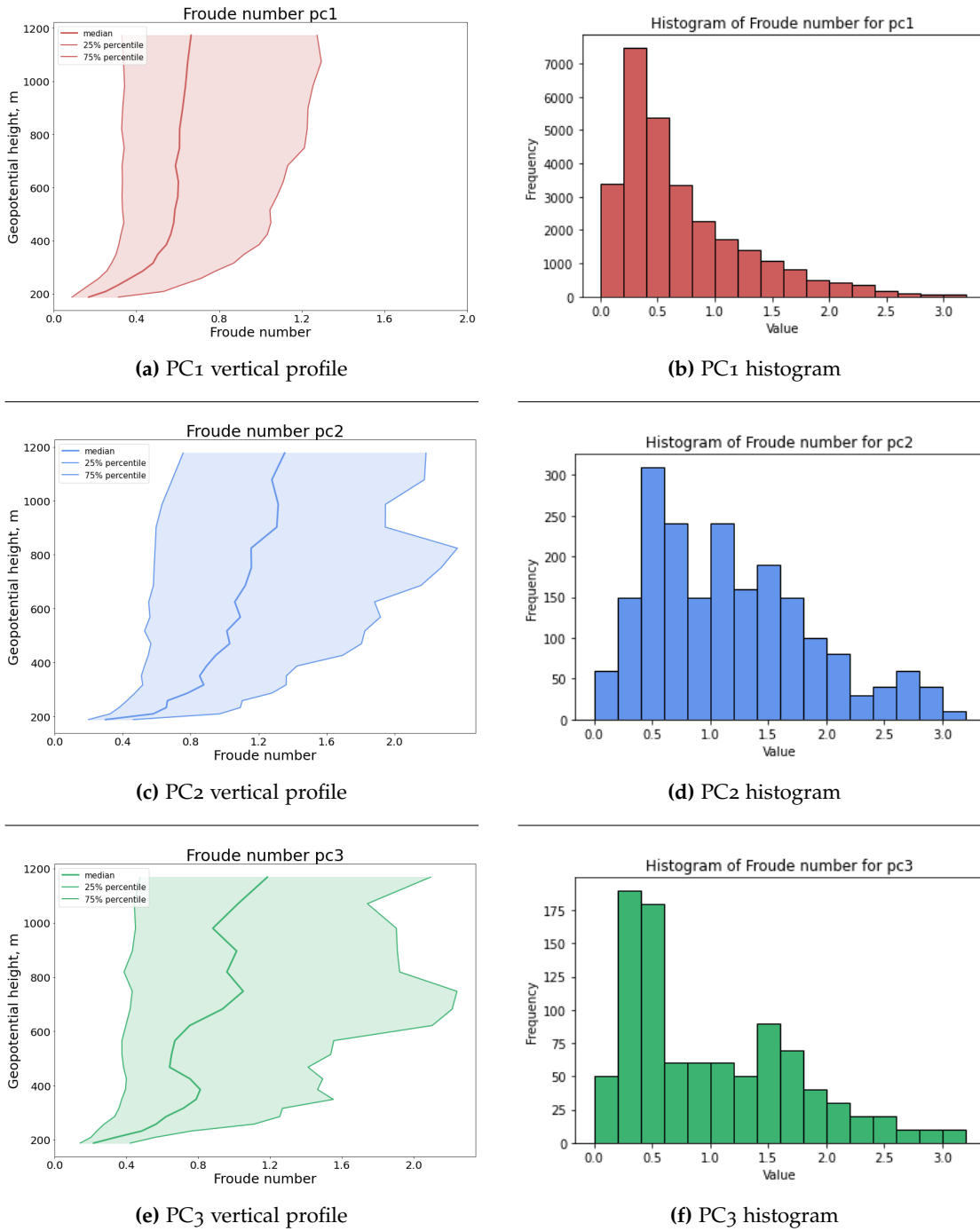


Figure 8.4: Vertical profiles (a,c,e) and histograms (b,d,f) of the moist Froude number F_w for PC1-PC3. Shading shows the 25th-75th percentile range.

The distributions in Fig. 8.4 therefore suggest that PC2 and PC3 each include dynamically distinct subsets. To examine whether this dynamical separation is reflected in precipitation bias patterns, the PC2 and PC3 samples are split into low- F_w ($F_w < 1$) and high- F_w ($F_w > 1$) subsets.

Link between F_w and precipitation bias.

To examine how F_w relates to precipitation bias in COSMO, the PC2 and PC3 cases were separated into low- F_w ($F_w < 1$) and high- F_w ($F_w > 1$) subsets. The resulting COSMO–RADKLIM precipitation differences are shown in Fig. 8.5.

In the high- F_w subsets (Fig. 8.5a,c), COSMO exhibits stronger positive precipitation differences relative to RADKLIM, with maxima concentrated near the mountain crest. These patterns are consistent with the stronger winds and generally moister conditions in the corresponding high- F_w profile subsets (Figs. 8.6 and 8.7), under which terrain-induced ascent is more vigorous and the model tends to over-amplify the precipitation response.

In the low- F_w subsets (Fig. 8.5b,d), precipitation differences are weaker and more spatially diffuse. The spatial structure is generally less sharply ridge-centred than in the high- F_w cases, which is consistent with more blocked or partially blocked flow. At the same time, COSMO still tends to produce positive precipitation differences in many areas. This indicates that the model’s wet bias is not confined to unblocked conditions, although it becomes more strongly concentrated over the crest at higher F_w .

8.2 RESULTS

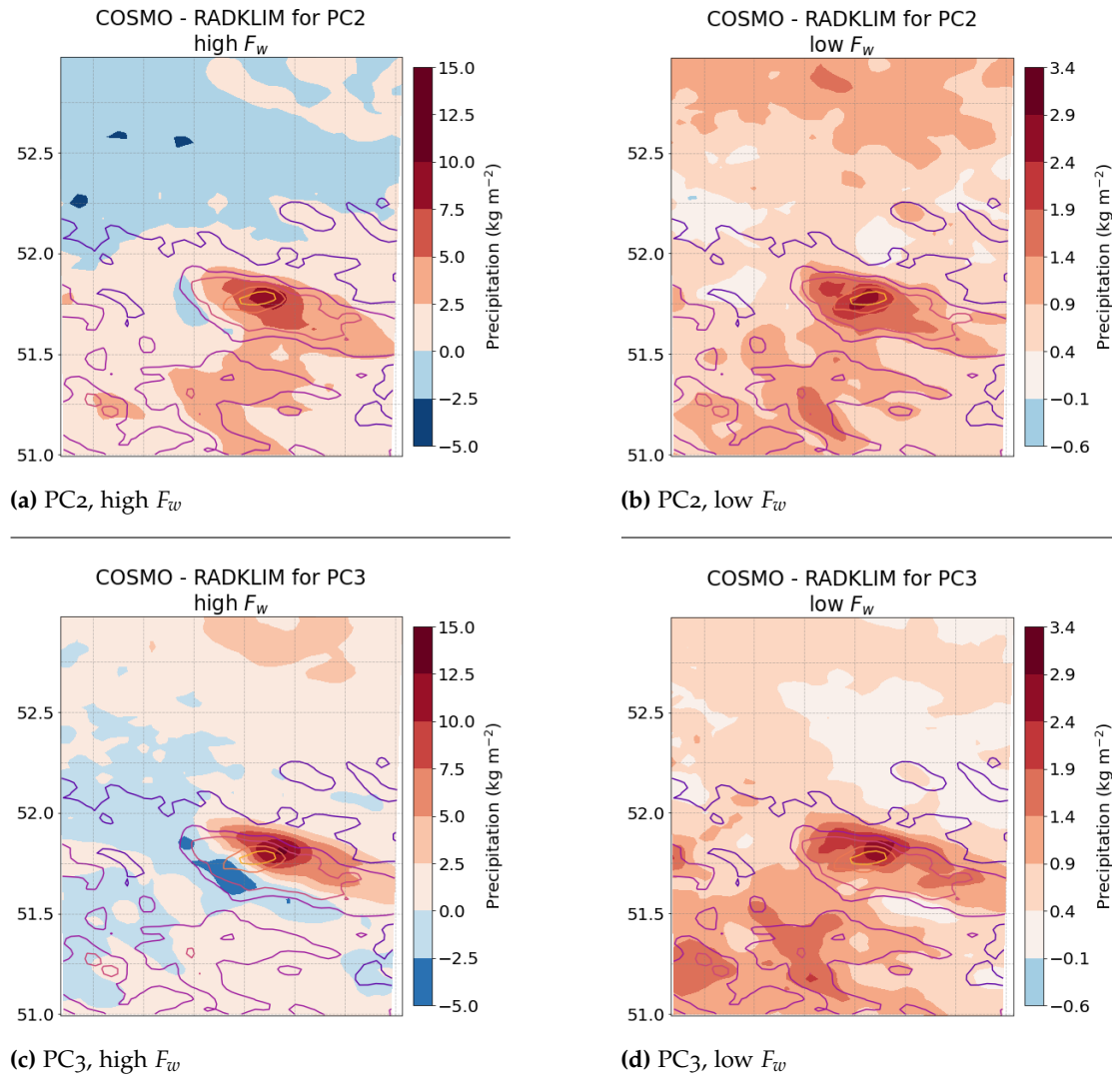


Figure 8.5: COSMO–RADKLIM precipitation difference for PC2 and PC3 under high and low Froude number conditions.

Thermodynamic and kinematic contrasts.

Figures 8.6 and 8.7 show vertical profiles of specific humidity, temperature, wind direction, and wind speed for the low- and high- F_w subsets in PC2 and PC3, respectively. In both PCs, the high- F_w subset is characterized by systematically stronger wind speeds through much of the lower and middle troposphere and, in many layers, higher specific humidity than the low- F_w subset. This is consistent with conditions that favour more efficient flow over the terrain and stronger orographic ascent.

The clearest separation occurs in wind speed, with high- F_w cases showing markedly larger values through much of the column in both PC2 and PC3 (Figs. 8.6d and 8.7d). Differences in specific humidity are also present, although the interquartile ranges overlap substantially (Figs. 8.6a and 8.7a). Temperature differences are generally smaller and remain partly overlapping (Figs. 8.6b and 8.7b).

PC3 shows broader spread and less vertically coherent separation, especially in wind direction (Fig. 8.7c), which is consistent with the broader and partly bimodal F_w distribution identified for this component (Fig. 8.4). By contrast, PC2 exhibits a clearer separation in flow strength and a more coherent vertical structure, particularly in wind speed (Fig. 8.6d). Overall, these profile contrasts support the interpretation that F_w separates dynamically distinct sub-regimes and that this separation is linked to differences in precipitation structure and bias amplitude.

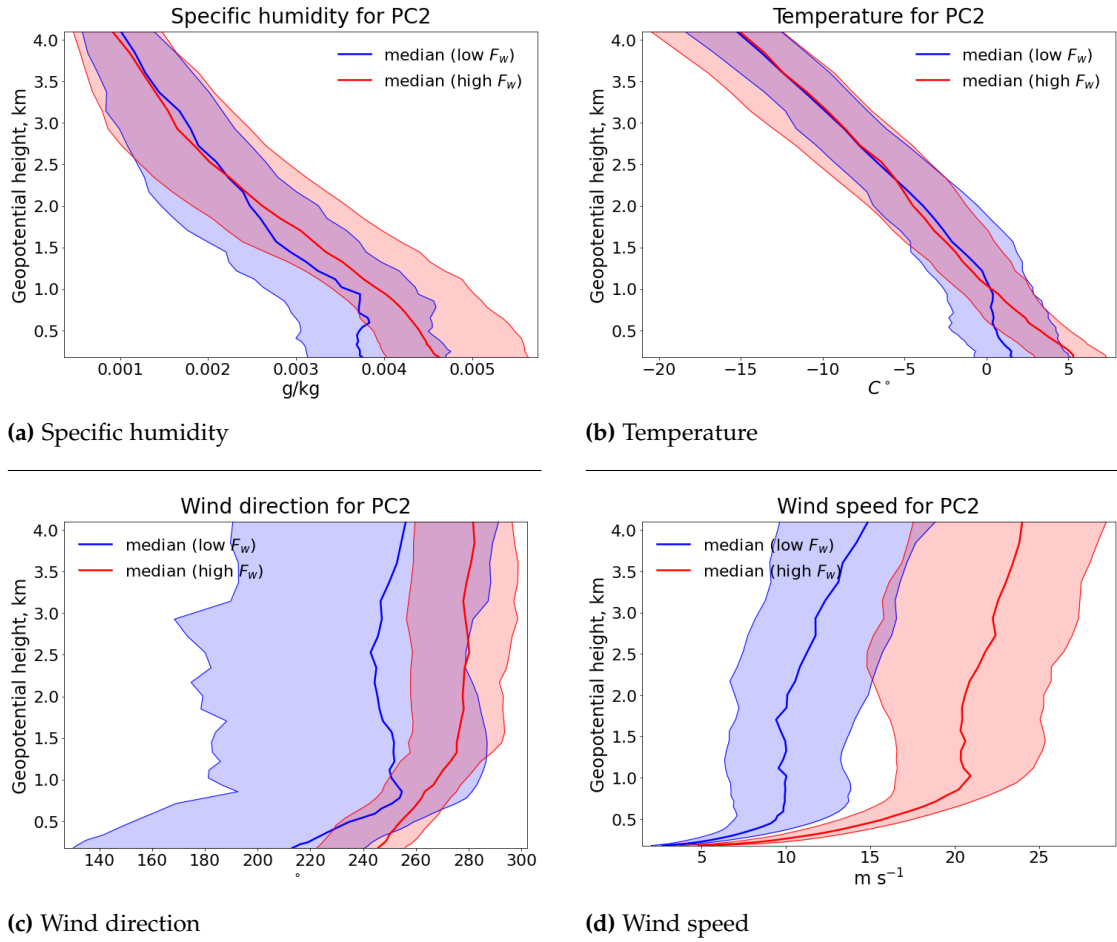


Figure 8.6: Vertical profiles for PC2 under high (red) and low (blue) Froude number regimes. Shading shows the 25th–75th percentile range.

8.2 RESULTS

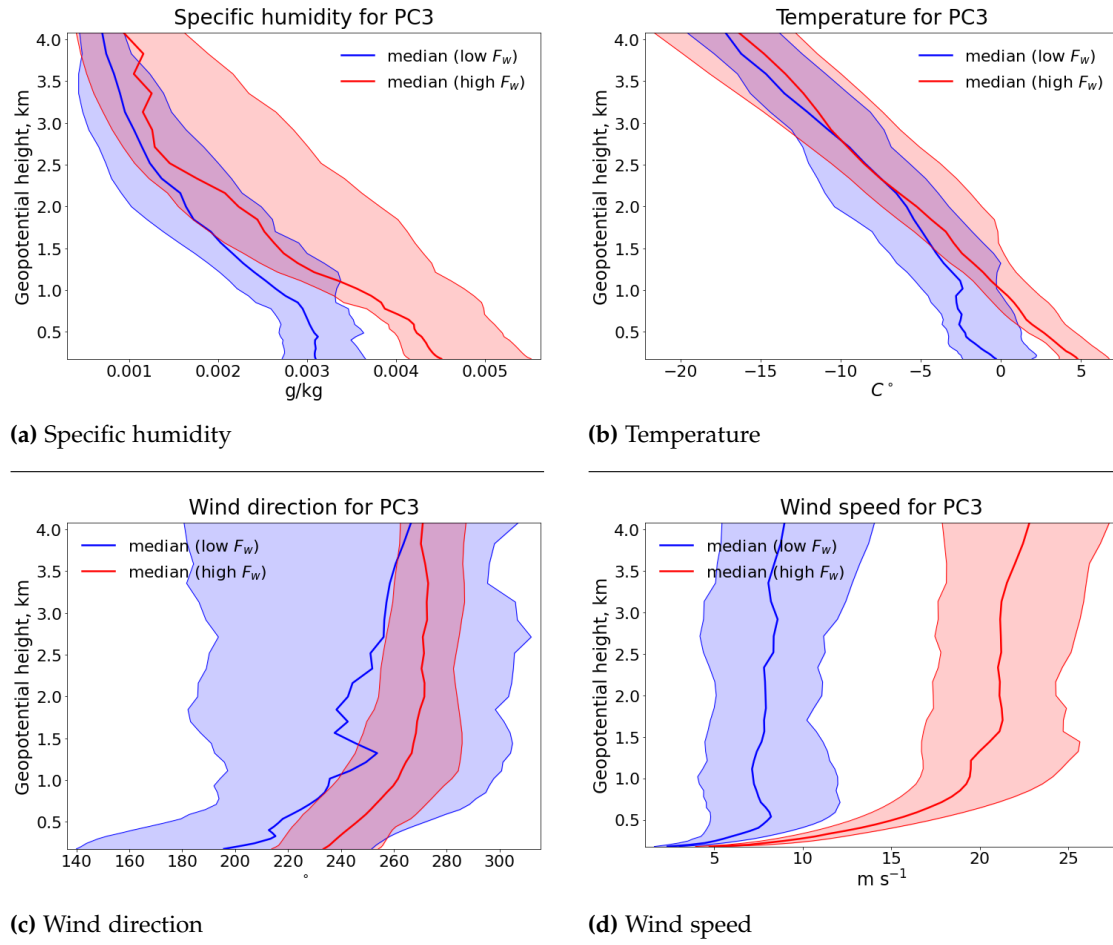


Figure 8.7: Vertical profiles for PC3 under high (red) and low (blue) Froude number regimes. Shading shows the 25th–75th percentile range.

Summary

The semi-idealised ICON experiments provide a controlled framework for testing how the large-scale environments associated with PC1–PC3 shape the orographic precipitation response. Under the colder and drier PC1 conditions, the model produces weak and localized precipitation, consistent with predominantly blocked flow and low F_w . By contrast, PC2 and PC3 generate stronger and more variable precipitation responses, reflecting warmer and moister conditions and a broader range of flow regimes.

The moist Froude number analysis shows that PC2 and PC3 each contain both lower- F_w and higher- F_w states, corresponding broadly to more blocked and less blocked flow

conditions. High- F_w subsets are associated with stronger winds, often higher humidity, and a more strongly ridge-centred positive COSMO–RADKLIM precipitation difference. Low- F_w subsets show weaker and more diffuse bias structures, but positive differences remain present in many cases.

Together, these findings support the interpretation that a substantial part of the Harz precipitation bias is structural and tied to how the model system responds to terrain-induced ascent across different flow regimes. They also provide a dynamical bridge between the PCA-based error modes and the case-study behaviour discussed in the preceding chapter.

8.3 Conclusion

This part of the thesis set out to determine whether precipitation errors over a low mountain range contain a systematic and physically interpretable component. The guiding idea is that recurring error structures are informative in two complementary ways. If similar large-scale environments repeatedly produce similar spatial mismatch patterns, then the error is not merely event-specific noise but reflects reproducible model behaviour. At the same time, linking these patterns to flow regime and stability turns verification into process understanding and provides a predictability perspective in which the type of error becomes conditionally predictable even when event details remain sensitive to synoptic evolution. To address this, precipitation differences between COSMO-DE-EPS and RADKLIM were decomposed with PCA to identify dominant spatial error structures. These modes were then related to the upstream environment using ERA5 composites, interpreted through representative ICON case studies and microphysics sensitivity experiments, and examined in semi-idealised simulations designed to isolate the terrain–flow response under controlled initial profiles.

The results can be summarised most directly by returning to the research questions posed in the Introduction (see Chapter 4). In the following, each question is restated and answered using the combined evidence from the PCA modes, the ERA5 regime composites, and the ICON case and semi-idealised experiments.

Do precipitation errors over the Harz exhibit a small number of recurring spatial patterns, rather than only case-specific variability?

The analysis indicates that winter precipitation errors over the Harz are organized around a small set of coherent spatial structures. The leading PCA modes capture ridge-centred amplification, synoptic-scale placement and spreading errors, and a local displacement-type pattern. These structures recur across many winter dates and account for a substantial share of the total error variability. In particular, PC1 explains about 43% of the variance, while PC2 and PC3 explain about 13% and about 6%, respectively, so that the first three modes together capture roughly two thirds of the winter precipitation-error variance. The ensemble perspective strengthens this

interpretation. When the dominant PC is identified separately for each of the 20 COSMO-DE-EPS members, all members agree on the same leading mode for about 60.8% of winter dates. This level of cross-member consistency indicates that the regime classification is often robust to initial-condition perturbations and that the dominant error structure is frequently controlled by systematic model behaviour rather than by member-to-member differences in the resolved flow. In practical terms, this means that the error field does not behave as an unstructured residual, but instead repeatedly aligns with a limited set of characteristic patterns that can be used as an empirical basis for diagnosing model behaviour over the Harz.

To what extent are these recurring error patterns linked to distinct large-scale flow regimes and stability conditions?

The analysis shows, that recurring spatial patterns are closely associated with distinct upstream environments. ERA5 composites show systematic differences between the modes in temperature, humidity, and wind profiles, which supports a regime interpretation in which large-scale forcing and stability shape how the terrain modifies ascent and precipitation distribution. The moist Froude number provides a compact dynamical summary of this regime dependence. Lower- F_w states correspond to more blocked or partially blocked flow, which is associated with weaker and more spatially diffuse precipitation differences. Higher- F_w states correspond to less blocked flow and stronger cross-barrier ascent, which is associated with more sharply ridge-centred positive COSMO–RADKLIM differences. Importantly, the analysis also suggests that a single PCA mode can contain dynamically distinct subsets, particularly for PC2 and PC3. Splitting these samples into low- and high- F_w subsets reveals consistent differences in both the upstream profiles and the corresponding precipitation-bias structure. This demonstrates that the PCA modes are physically anchored in flow-regime behaviour rather than being purely statistical descriptors of the error field.

What physical processes, including flow blocking, terrain-induced uplift, and micro-physical assumptions, help explain the dominant error patterns?

The combined case studies, ICON sensitivity experiments, and semi-idealised simulations support a process interpretation that separates error mechanisms dominated by terrain-locked amplification from those dominated by synoptic placement and organisation. In ICON, the sensitivity suite varied the microphysics formulation (one- versus two-moment), the prescribed CCN background (CCN_type=6 and CCN_type=8), and the inclusion of ice-phase processes through no-ice experiments, with an additional higher-resolution test for the PC1 regime.

The PC1 regime is characterised by ridge-centred overamplification that persists across contrasting thermodynamic environments. This persistence indicates that thermodynamic differences primarily modulate how strongly the bias is expressed, while the spatial anchoring to the ridge reflects a dynamical control through terrain-induced ascent. The ICON perturbations modify precipitation intensity and smoothness and, depending on the case, shift the relative contribution of warm-rain and ice-phase

pathways. However, they do not remove the ridge-locked tendency, which points to a coupled dynamics–microphysics over-response to uplift as a plausible explanation for this mode.

The PC₂ regime is dominated by the placement and spatial extent of a synoptic precipitation band. In the representative frontal case, COSMO and ICON reproduce the general band orientation but disagree with RADKLIM in organisation and areal coverage. The persistence of the spatial mismatch across one- and two-moment schemes, across CCN settings, and under no-ice conditions indicates that local microphysical conversion efficiency is not the primary cause of the error pattern. Instead, the dominant contribution is consistent with errors in large-scale advection and frontal evolution, which control where the main precipitation system interacts with the terrain.

The PC₃ regime highlights a local displacement-type behaviour near the Harz, consistent with comparatively small shifts in the onset and location of terrain-induced ascent. Here, microphysical changes again modulate intensity more than location, which points to low-level stratification, boundary-layer mixing, and subtle differences in flow direction as likely contributors to the remaining mismatch.

Predictability perspective.

A central implication of these results is that predictability can be framed not only in terms of event-specific precipitation amounts, but also in terms of regime-dependent error structures. While the exact precipitation maximum in an individual storm remains sensitive to details of synoptic evolution, the analysis indicates that the dominant type of precipitation error becomes more predictable once the large-scale regime is known. Ridge-centred overamplification is the characteristic signature in PC₁-type environments. Frontal placement and spreading errors dominate in PC₂-type environments. Local redistribution between windward and ridge sectors is most relevant in PC₃-type environments. The moist Froude number provides an explicit dynamical bridge between the statistical error modes and the terrain–flow interaction regime, and it explains why the spatial focusing and amplitude of the bias strengthen systematically from more blocked to less blocked flow states.

These findings suggest that bias correction and model development efforts benefit from being regime-aware. Microphysical modifications can alter precipitation intensity, but they are unlikely to correct spatial error structures that are controlled by synoptic placement or by the dynamical response to terrain. The regime framework developed here therefore provides both a physically interpretable diagnosis of model limitations over the Harz and a foundation for targeted follow-up experiments, which are pursued in the semi-idealised analyses that follow.

Part III

PRECIPITATION-BASED PROBABILISTIC HAIL PREDICTION

INTRODUCTION

Hail is among the most damaging convective hazards in central Europe, yet individual hailstorms remain rare events at a fixed location. For Germany, insurance statistics analysed by Kunz and Puskeiler (2010) show that hail contributes a substantial share of the insured building losses associated with convective storms, a result that is consistent with the broader European overview in Punge and Kunz (2016). Hailstorm-frequency studies for southwest Germany by Kunz, Sander, and Kottmeier (2009) and radar-based severe convective storm climatologies such as Fluck et al. (2021) show pronounced spatial variability and substantial year-to-year variability in hail occurrence. Mohr and Kunz (2013) related these variations to changes in large scale circulation patterns and the distribution of convective parameters such as CAPE and vertical wind shear. This combination of high impact and low local frequency creates a strong demand for reliable information on when and where hail is likely, both on the nowcasting time scale of minutes to hours and on climatological time scales of years to decades. In addition, a warming climate is expected to modify the frequency and intensity distribution of severe convective storms, including hail, through changes in thermodynamic and kinematic environments, although the sign and magnitude of projected hail changes remain uncertain and region dependent (Seneviratne et al., 2021; Kahraman et al., 2025).

Hail forms within deep convective clouds in an environment that combines strong instability, abundant low level moisture, and vertical wind shear that organises and sustains persistent updrafts (Markowski and Richardson, 2010). Hail embryos, for example graupel particles or frozen raindrops, develop in the mixed phase region above the 0°C isotherm where supercooled liquid water and ice coexist (Korolev et al., 2017; Barrett and Hoose, 2023). Subsequent growth occurs mainly by riming of supercooled droplets and can proceed both within updrafts and outside them, as long as particles encounter sufficient liquid water along their trajectories. Strong updrafts are nevertheless crucial because they increase residence time in the mixed phase region and enable repeated cycling through liquid water rich regions, which supports growth to large sizes. As hailstones grow, they eventually fall out of the storm once their terminal fall speed exceeds the local upward air motion and they may continue to accrete liquid water while descending through cloud. Depending on the balance between latent heat release from riming, ventilation, and ambient temperature, growth can proceed in a dry regime with a frozen surface or in a wet regime with a liquid surface layer, which strongly affects hailstone density, shape, and melting behaviour (Korolev et al., 2017). If the hailstones survive melting and sublimation in the sub cloud layer, they reach the surface as hail.

This close link between hail and storm dynamics makes hail prediction demanding, because it requires a realistic representation of both the convective circulation and the mixed-phase microphysics. At the same time, it also means that fields which describe the evolving structure of deep convection, such as surface precipitation rate or accumulation and simulated radar reflectivity, can contain useful information about hail occurrence even if the model does not predict hail explicitly. Modern kilometre-scale numerical weather prediction models resolve many aspects of convective initiation and cell-scale dynamics, but their representation of the dynamics and organisation of larger-scale convective systems remains an active area of research, and microphysical processes are still parameterized. Most operational systems therefore do not provide an explicit prediction of surface hailfall, for example hail size distributions or hail accumulation at the ground. Some research and test configurations include a separate hail category within more advanced multi-moment microphysics, but this approach raises practical and conceptual issues for operational forecasting, including computational cost and strong sensitivity to assumptions about the hydrometeor size distribution when the number of hailstones is small (Milbrandt and Yau, 2005).

Operational guidance thus represents hail only indirectly through proxy quantities, for example graupel or hail mass mixing ratios and simulated reflectivity, and through storm-environment diagnostics such as CAPE or composite severe weather indices (Zängl, Reinert, Rípodas, et al., 2015; Deutscher Wetterdienst, 2021). These quantities act as proxies for hail because they represent hydrometeor mass within the cloud column. Whether hail reaches the surface depends additionally on sedimentation and melting processes that are not uniquely determined by a mixing ratio field alone. This limitation is illustrated in the AROME configuration discussed by Seity et al. (2012), where an explicit hail variable was implemented but not used operationally, and hail risk was instead diagnosed from the vertically integrated graupel content. Insight into hail occurrence must therefore be obtained indirectly from such proxy quantities, either through subjective interpretation or through dedicated post-processing that converts model output into probabilistic hail information. To understand what signal such schemes need to reproduce, it is useful to first summarise the observed hail climatology over Germany.

Radar-based hail detection algorithms and hail-track catalogues provide detailed information on the space-time distribution of hail occurrence and, depending on the algorithm, can also offer severity proxies such as hail probability or estimated hailstone size (Handwerker, 2002; Junghänel et al., 2016; Puskeiler, Kunz, and Schmidberger, 2016; Fluck et al., 2021). In this thesis I use hail tracks derived from the TRACE3D cell-tracking algorithm applied to the German weather radar network (Deutscher Wetterdienst, 2026a). These tracks reveal pronounced hail corridors that are most active along the Alpine foreland in southern Germany and downstream of several low mountain ranges, in particular the Black Forest, the Swabian Jura, and the Thuringian Forest (Figs. 9.1a and 9.1b). Histograms of track direction, seasonality, length and

speed (Fig. 9.2) further show that most hail cells occur between June and August, that typical track lengths are only a few tens of kilometres, and that track directions are dominated by west–south–westerly to westerly motion. The mean annual number of hail days reaches values of one hail day per year in the most active 0.1° grid boxes, while many northern and eastern regions experience less than one hail day per year (Fig. 9.1a). Similar patterns have been reported in earlier radar based hail climatologies for Germany and Europe (Puskeiler, Kunz, and Schmidberger, 2016; Fluck et al., 2021; Punge and Kunz, 2016), which supports the robustness of the hail corridors identified here.

These climatological results provide a useful benchmark, but they also motivate a prediction question. Given that hail occurrence is tightly linked to particular convective structures and that hail tracks cluster in preferred corridors, the key issue is to what extent these properties can be anticipated from quantitative precipitation forecasts alone. In other words, how much information about hail is already contained in model precipitation fields, and how well can this information be extracted and translated into a probabilistic hail product.

These questions guide the analysis in this part of the thesis:

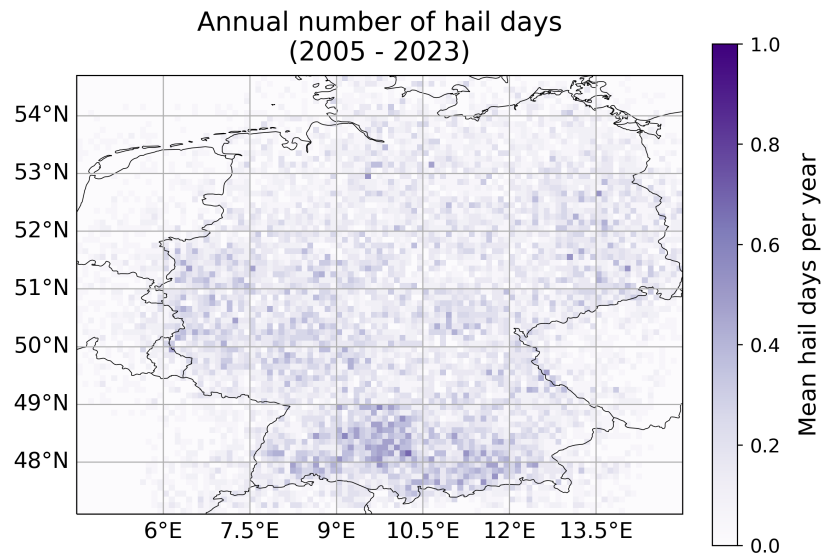
- How well can high-resolution radar precipitation fields alone separate hail-producing from non-hail-producing convection?
- How much information about hail occurrence is contained in precipitation fields alone, and how far can this signal be carried over to a convection-permitting ensemble forecast to construct probabilistic hail products?
- What is the predictability of hail-relevant convective structures in the ensemble, and how is this predictability linked to the large-scale environment?

The focus is on what can be inferred from the structure and intensity of the precipitation field itself, without relying on additional predictors such as CAPE or graupel mixing ratios.

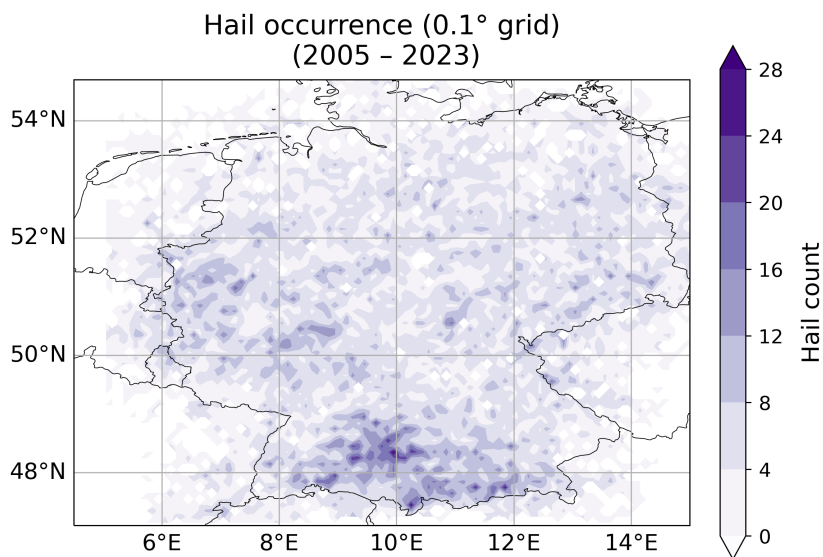
To address these questions, I link radar-based hail tracks to precipitation fields from first radar observations and then a limited-area ensemble forecast. I then use supervised machine-learning models to classify convective precipitation snapshots as hail-producing (1) or non-hail-producing (0) based solely on the structure and intensity of the precipitation field. Two complementary classifiers are considered: a Random Forest (RF) and a Convolutional Neural Network (CNN).

The required data preprocessing and snapshot construction are introduced first, starting with the 5-min RADKLIM fields and convective/stratiform separation, followed by the derivation of 1 h RADKLIM data and the harmonisation with the COSMO-DE-EPS precipitation output (Sections 10.1–10.3). The RF and CNN model setups are then described (Section 10.4). The results are organised to mirror this workflow. I first present the nowcasting performance based on 5 min RADKLIM precipitation snapshots

(Section 11), then assess RF performance on 1 h RADKLIM and COSMO snapshots (Sections 12). Finally, I evaluate COSMO-based hail predictions against observed hail tracks, including sensitivity to spatial and temporal matching choices and links to the large-scale environment and aerosol load (Sections 13).



(a) Mean annual number of hail days in Germany for 2005–2023 on 0.1° grid, derived from TRACE₃D hail tracks.



(b) Total number of hail occurrences per 0.1° grid box in Germany for 2005–2023.

Figure 9.1: TRACE₃D-based hail climatology for Germany (2005–2023). (a) Mean annual number of hail days. (b) Total hail occurrence per 0.1° grid box.

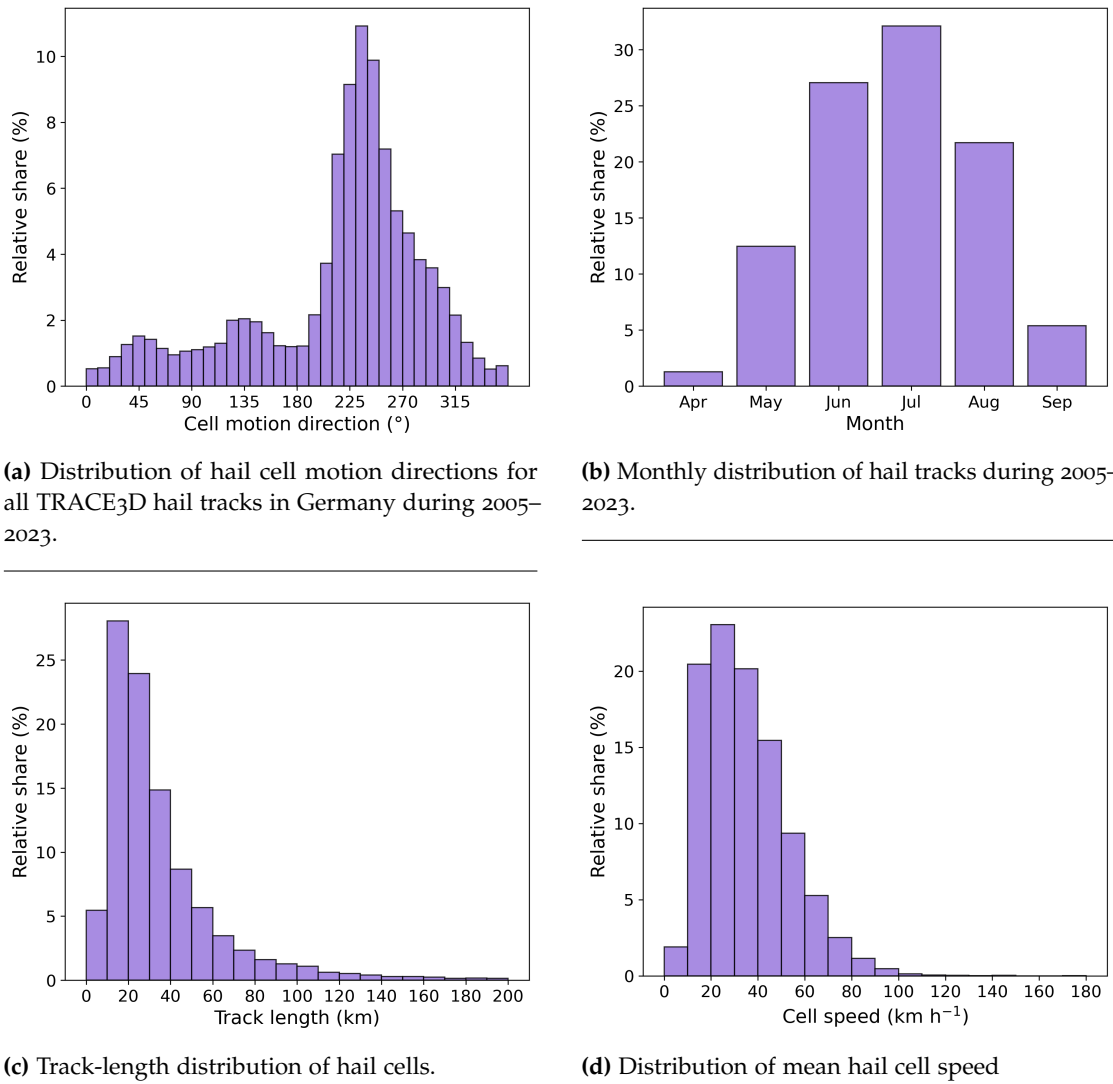


Figure 9.2: Statistical properties of TRACE3D hail tracks over Germany for 2005–2023. Panels show (a) cell motion direction, (b) monthly frequency of tracks, (c) track length and (d) mean cell speed. All histograms are normalised to give relative frequency in percent.

In this chapter I describe the preprocessing of two RADKLIM radar precipitation products, the 5 min fields and the 1 h accumulations. The 5 min fields are used to identify convective elements and to build labelled, event-centred snapshots aligned with hail tracks. The 1 h data are used to construct an hourly RADKLIM dataset on a grid comparable to COSMO-DE-EPS output, so that radar-based and model-based inputs can be evaluated consistently. After outlining these preprocessing steps (including temporal/spatial harmonization and snapshot extraction), I describe the learning setups for the random forest (RF) and the convolutional neural network (CNN).

10.1 5 min RADKLIM data

Surface precipitation observations were taken from the RADKLIM dataset (see Section 3.1). RADKLIM offers high spatial and temporal resolution over a multi-year period, making it suitable for climatological analyses and for comparison with the hail-track dataset. In this section, the 5 min RADKLIM fields are used for nowcasting purposes based on radar precipitation snapshots and are not intended for direct comparison with the coarser-temporal-resolution model forecasts. This provides a high-resolution benchmark for the precipitation-based hail signal and illustrates the potential value of higher-frequency output from future convection-permitting ensemble systems. A consistency check of the 5 min RADKLIM archive showed that the warm-season dataset is essentially complete. For April–September 2005–2023, 1 001 376 expected 5 min timestamps were examined, of which none were missing and only four were identified as corrupted, corresponding to about 0.0004%. These few invalid time steps were excluded from the analysis.

For each analysis time the RADKLIM fields provide the accumulated amount of precipitation over the preceding time interval but no information about the precipitation type. To enable a distinction between stratiform and convective precipitation, where the latter is typically associated with hail, I applied the objective convective–stratiform partitioning of Steiner, Houze, and Yuter (1995).

10.1.1 Convective/Stratiform Classification

RADKLIM precipitation type is classified by adapting the Steiner–Houze–Yuter (1995) (in the following referred to as SHY95) convective/stratiform scheme, which was devel-

oped for 2D radar reflectivity fields, to the 2D rain rate fields available in RADKLIM. In SHY95, convective pixels are identified based on (i) intensity and (ii) the local “peaked-ness” of reflectivity relative to a surrounding background, reflecting typical differences between convective and stratiform precipitation patterns (Steiner, Houze, and Yuter, 1995).

Five-minute accumulations are converted to rain rate R (mm h^{-1}). Since the thresholds in the SHY95 algorithm are specified in reflectivity (dBZ), I temporarily convert RADKLIM rain rate to an equivalent reflectivity using the following Z – R relation (Weigl, 2008):

$$Z = \begin{cases} 200 R^{1.6}, & Z \leq 44 \text{ dBZ}, \\ 77 R^{1.9}, & Z > 44 \text{ dBZ}. \end{cases}$$

This conversion is used only for translating the SHY95 thresholds for application to RADKLIM rain rate data.

A local background reflectivity Z_{bg} is estimated from a circular neighbourhood with radius 11 km (corresponding to ≈ 11 RADKLIM grid cells). A pixel is flagged *convective* if (i) $Z \geq 40$ dBZ, or (ii) $Z \geq Z_{\text{bg}} + \Delta Z$, where

$$\Delta Z = \begin{cases} 10 \text{ dB}, & Z_{\text{bg}} < 0 \text{ dBZ}, \\ 10 - \frac{Z_{\text{bg}}^2}{180} \text{ dB}, & 0 \leq Z_{\text{bg}} < 42.43 \text{ dBZ}, \\ 0 \text{ dB}, & Z_{\text{bg}} \geq 42.43 \text{ dBZ}. \end{cases}$$

To capture the immediate surrounding area around each detected convective pixel, I expand the convective mask by marking as convective all pixels within a disk of radius r grid cells around each seed (i.e., morphological dilation). The radius increases with the local background reflectivity Z_{bg} : $r = 1, 2, 3, 4, 5$ for $Z_{\text{bg}} < 20$, $20 \leq Z_{\text{bg}} < 25$, $25 \leq Z_{\text{bg}} < 30$, $30 \leq Z_{\text{bg}} < 35$, and $Z_{\text{bg}} \geq 35$ dBZ, respectively. After this growth step, any remaining pixels not included in the convective mask are labeled stratiform. As a result, I obtain a convective element mask every 5 min, that can be aligned with the 15 min hail-track data.

Next, I aggregate adjacent convective pixels into coherent objects using connected-component labeling. Specifically, I apply `scipy.ndimage.label` (8-neighbour connectivity), which assigns a unique integer ID to each contiguous area and returns the total object count.

Figure 10.1 summarizes the structure and intensity of the convective objects identified by the SHY95-based classification. The area distribution spans roughly two orders of magnitude (from a few to $\gtrsim 10^3 \text{ km}^2$), with most cases in the 10^1 – 10^3 km^2 range. The prevalence of larger objects can indicate organised systems rather than small isolated

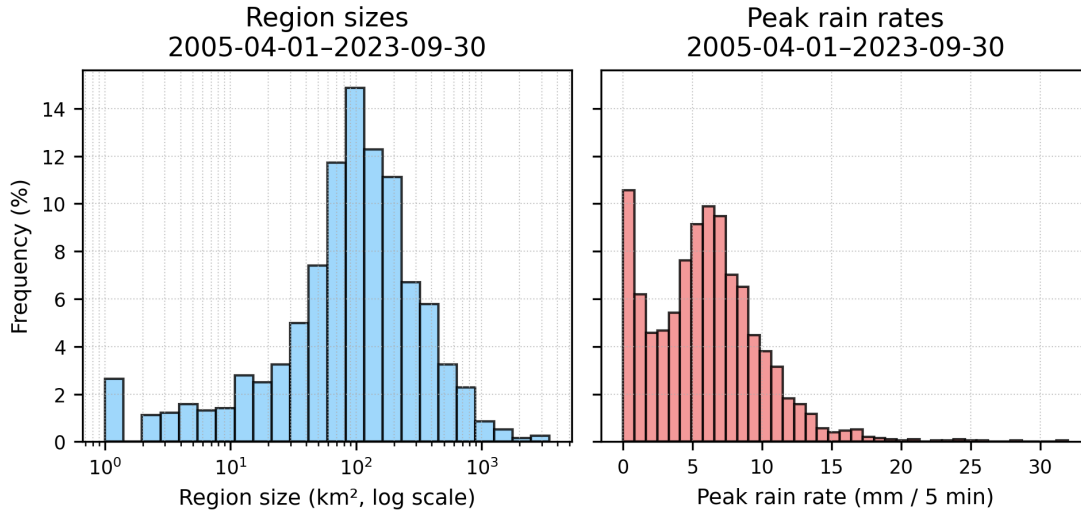


Figure 10.1: Distributions of (left) convective-object area (km^2 , log scale) and (right) peak 5-min rain rate ($\text{mm}/5 \text{ min}$) for convective cells collocated with hail tracks from 2005-04-01 to 2023-09-30.

cells. The peak rain-rate distribution is right-skewed. Here, peak rain rate is defined as the maximum 5 min rain rate within each convective object. Most peaks fall in the moderate range (about 3–10 mm per 5 min), with a smaller number of very intense events reaching 15–25 mm per 5 min. Together these histograms illustrate the range of convective-element sizes and intensities captured by the detection method. In later analyses, I keep only convective objects larger than 3 km^2 , because smaller patches are usually artefacts of the detection method.

10.1.2 Snapshot extraction and labeling

I construct a dataset for supervised learning from RADKLIM by pairing precipitation snapshots with hail/no-hail labels. Each sample consists of a native-grid $20 \times 20 \text{ km}$ precipitation patch (input) and a binary target indicating whether hail was observed in the patch at that time (label). Inputs are (i) coherent-object files that provide the location of convective regions and (ii) the hail-track dataset described in 3.4. For each analysis time I:

1. Select convective regions with area $\geq 3 \text{ km}^2$.
2. For each region, locate the precipitation maximum and extract a native-grid $20 \times 20 \text{ km}$ snapshot of the RADKLIM field centred on this point.
3. Assign the target label $y=1$ if the snapshot overlaps a hail track at that time, otherwise $y=0$. I define “overlap” as a spatial intersection between the snapshot and the hail-

track positions at the corresponding time stamp (15 min track times are matched to the nearest 5 min RADKLIM analysis time). If multiple hail-track points fall within one snapshot, the label remains $y=1$.

To ensure class balance, I randomly downsample the non-hail ($y=0$) snapshots to match the count of hail ($y=1$) snapshots. This produces a structured set of positive and negative examples suitable for training hail classifiers, including a convolutional neural network (CNN) and a random forest (RF), that use only radar precipitation snapshots as input.

10.2 Hourly RADKLIM data

10.2.1 Temporal and spatial harmonisation of RADKLIM data

To enable a precipitation-only hail classifier to be applied to COSMO-DE-EPS output, the radar-based training data must be made compatible with the model snapshots in both space and time. RADKLIM therefore needs to be transformed to the same nominal temporal resolution and to a COSMO-aligned grid-point representation. RADKLIM provides 5 min precipitation estimates on an approximately 1 km grid, whereas COSMO-DE-EPS supplies hourly totals on a ~ 2.8 km grid. The aim of this section is therefore to construct labelled radar snapshots that represent 1 h accumulations and can be mapped to fixed 9×9 patches on the COSMO grid. On the COSMO-DE-EPS grid, such a patch corresponds to a physical area of roughly 25×25 km. This is slightly larger than the native-grid 20×20 km RADKLIM snapshots used in the 5 min nowcasting experiment, but it still represents the same storm-scale context, including the convective core and its immediate surroundings, while ensuring direct compatibility with the model grid. The processing is carried out for the COSMO-DE-EPS period (April–September 2011–2017 and 1 April 2018 to 15 May 2018) and is restricted to times for which hail-track information is available.

Hourly accumulations from 5 min RADKLIM fields

I first convert the 5 min RADKLIM precipitation data into 1 h accumulated fields. For each analysis time t with hail information, a 1 h accumulation is constructed by summing the previous twelve 5 min frames, including the current one. This corresponds to the period $[t - 55 \text{ min}, t]$ and yields a rolling 1 h total that reflects the precipitation over the past hour. Pixels with missing-value codes in the RADKLIM files (i.e., fill values used to mark undefined pixels, which occur in particular near the domain boundary) are treated as missing when computing the sum. Any hour with at least one missing 5 min frame is discarded. The resulting 1 h accumulation is stored with the time stamp t , so that the RADKLIM and COSMO fields refer to the same valid time. In

this way, every valid hail-mask time within April–September is paired with a consistent hourly radar accumulation field that can be used for snapshot extraction.

Convective region-based snapshot sampling and labelling

At each hourly analysis time t , convective regions are identified using the labelled objects from the RADKLIM coherent-area objects (see Section 10.1.2). Only regions larger than a minimum area of 3 km^2 are retained, so that very small and noisy objects are excluded. For each remaining region, the corresponding 1 h accumulation field at time t is used to locate the grid point with the highest 1 h precipitation inside the region. This point marks the most intense part of the convective region in terms of accumulated rainfall.

Around this maximum, a fixed native-grid RADKLIM cut-out is extracted. Cut-outs that would extend beyond the RADKLIM domain are discarded. Each retained cut-out is then resampled by bilinear interpolation to a 9×9 COSMO-aligned grid. In this harmonised representation, each snapshot corresponds to a storm-centred patch of roughly $25 \times 25 \text{ km}$. The same transformation is applied to the corresponding longitude and latitude tiles to preserve location information.

The snapshot size is chosen to capture the storm core together with its immediate surroundings. A hail-producing convective cell typically exhibits a compact precipitation maximum but is embedded in a broader precipitation structure. Native-grid RADKLIM snapshots of $20 \times 20 \text{ km}$ provide the local storm-scale context used in the 5 min now-casting setup. Their harmonised 9×9 COSMO-aligned counterparts represent roughly $25 \times 25 \text{ km}$ and retain a similar local spatial context while ensuring compatibility with the model grid. Because the snapshots are centred on the local precipitation maximum within the convective region, advection between consecutive analysis times mainly affects the internal structure within the window rather than shifting the system out of the patch.

To assign labels, the convective region is tested against the hail-track dataset at time t . If at least one hail-track point falls within the convective region (set of pixels) at that time, the snapshot is labelled as hail ($y = 1$). If no hail-track point is present within the convective region, the snapshot is labelled as no hail ($y = 0$). Hail-track times at 15 min resolution are matched to the nearest RADKLIM analysis time t .

This procedure is applied to all convective regions at all analysis times in the period of interest, regardless of whether hail is present. As a result, it produces a large and diverse pool of both hail and no-hail examples. Because hail is relatively rare, the raw dataset is strongly imbalanced. To avoid this during model training, I balance the classes separately for each month between April and September by randomly downsampling the no-hail snapshots within each month so that the number of no-hail snapshots matches the number of hail snapshots. The resulting balanced set of 9×9 hourly snapshots, together with their labels and metadata (time stamp, convective

region identifier, and location), is used as radar-based training data for the RF and CNN.

10.2 HOURLY RADKLIM DATA

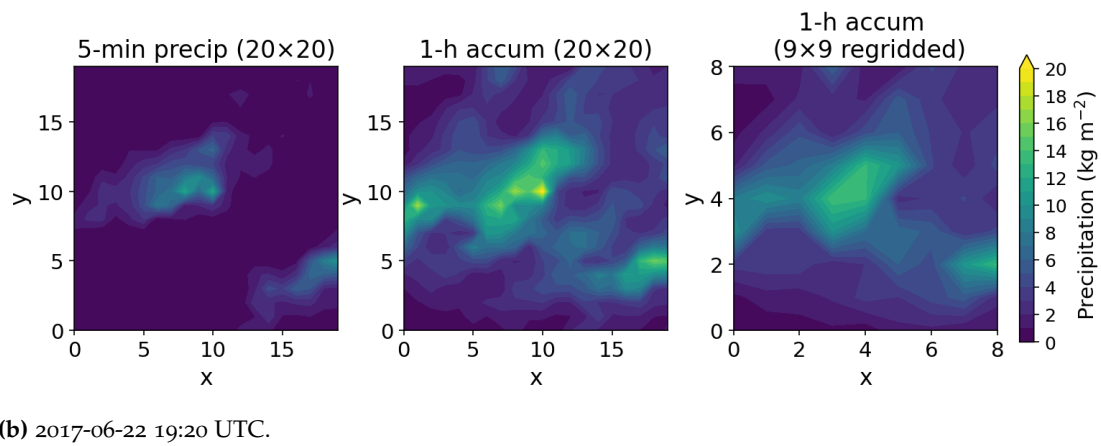
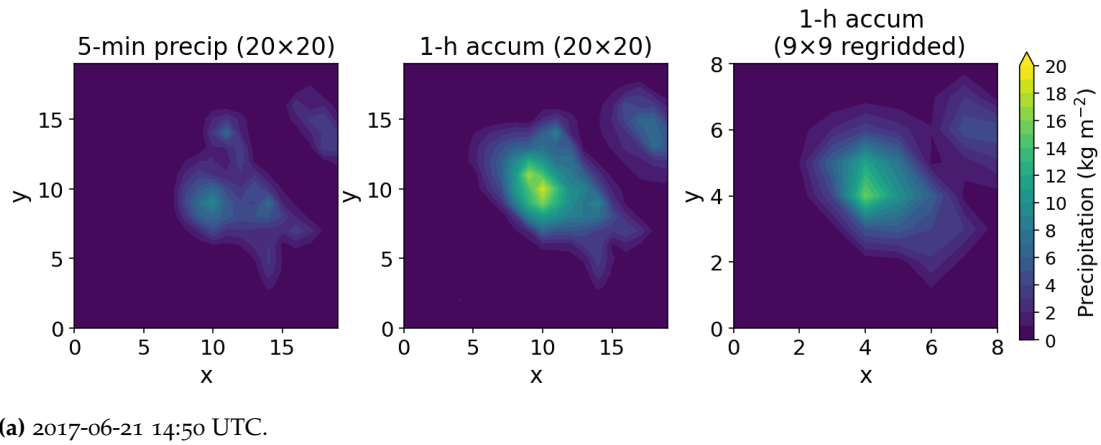


Figure 10.2: Examples of temporal aggregation and regridding for (a) 2017-06-21 14:50 UTC and (b) 2017-06-22 19:20 UTC. Columns show 5 min RADKLIM (left), 1 h RADKLIM accumulation (middle), and the regrided 1 h snapshot on the COSMO-aligned 9×9 grid (right).

The effect of the temporal and spatial harmonisation is illustrated in Fig. 10.2. It shows, for two example cases, the 5 min RADKLIM precipitation snapshot used to define the convective region (left), the corresponding 1 h RADKLIM accumulation snapshot on the native grid (middle), and the 9×9 snapshot after regridding to the COSMO-aligned resolution (right). This visual comparison highlights that the resampling preserves the location of the precipitation maximum and the main spatial gradients while reducing small-scale variability at the native RADKLIM resolution.

10.3 COSMO-DE data processing

10.3.1 Hourly COSMO-DE precipitation fields

In addition to the radar-based snapshots, I also use precipitation from the COSMO-DE-EPS ensemble in Section 12. In the operational archive, COSMO-DE-EPS total precipitation is stored as accumulated fields at hourly forecast times (up to a maximum lead time of +21 h) for each ensemble member rather than as hourly totals. I therefore reconstruct a continuous series of hourly precipitation fields by taking differences between successive accumulations. This hourly COSMO series then defines the temporal resolution to which RADKLIM is aggregated in Section 10.2, so that RADKLIM and model can be compared on a common 1 h grid.

The reconstruction is carried out for April–September 2011–15 May 2018, that is the convective season within the COSMO-DE-EPS period available here. For each day and ensemble member, I use accumulated precipitation at lead times 18, 19, 20, and 21 hours. Hourly totals are obtained as differences between these accumulations, which yields three consecutive 1 h precipitation fields per forecast run on the native ~ 2.8 km COSMO grid. This procedure is repeated for all available days and ensemble members, so that the selected period is covered by a set of hourly COSMO precipitation fields at the chosen valid times. Across this period, 20 808 forecast times were available for each lead time. Of these, 359 times at 18 h lead time, 344 times at 19 h, 346 times at 20 h, and 361 times at 21 h, corresponding to about 1.7% in each case, were identified as broken across all ensemble members. Missing or incomplete input files are explicitly flagged and propagated as missing values in the hourly fields rather than being implicitly set to zero.

The resulting dataset is a consistent archive of hourly COSMO-DE-EPS precipitation for all ensemble members on the native model grid. It is directly comparable to the 1 h RADKLIM fields described in Section 10.2. These hourly fields are then used to construct the convective region-centred COSMO-DE snapshots described in the next section 10.3.2.

10.3.2 COSMO-DE snapshots

To drive the radar-trained RF with model data, I extract COSMO-DE-EPS precipitation snapshots in a way that mirrors the convective cell-centred sampling used for RADKLIM. COSMO-DE-EPS provides hourly accumulated precipitation on a ~ 2.8 km grid for 20 ensemble members. For each year between 2011 and 2017, I use the convective season from April to September. For 2018, the period is limited to April and May, consistent with the available COSMO-DE-EPS archive.

For each month in this period, the hourly COSMO datasets described in Section 10.3.1 are used. To define candidate precipitation objects for snapshot extraction, I use a simple intensity threshold rather than the SHY95-based convective/stratiform separation applied to RADKLIM. This is because the COSMO sampling step only requires a robust identification of intense precipitation areas to centre the fields, while the subsequent hail discrimination is performed by the machine-learning model trained on the RADKLIM data. Accordingly, intense precipitation areas are identified using a precipitation threshold of 1 mm h^{-1} .

For each ensemble member and each hourly time step, I form a binary mask in which grid points with precipitation $\geq 1 \text{ mm h}^{-1}$ are set to one and all others to zero. Connected components in this mask are then labeled as individual intense precipitation areas using a standard connectivity algorithm. Regions with no threshold-exceeding pixels are skipped. For each remaining region, the corresponding hourly precipitation field is used to find the grid point with maximum intensity inside the region, which is treated as the region centre in COSMO.

Around each region centre, a fixed 9×9 patch is extracted from the hourly precipitation field. This patch is centred on the maximum and represents a $\sim 25 \times 25$ km neighbourhood on the COSMO grid. If the region is close to the domain boundary, parts of the 9×9 window fall outside the COSMO domain and are filled with missing values in the patch. The same extraction is applied to the latitude and longitude fields, so that every precipitation patch has collocated coordinate information. No hail labels are used at this stage. The aim is to collect a pool of COSMO precipitation snapshots that can later be classified as hail-producing or non-hail-producing by the radar-trained model.

The extraction loop runs over all ensembles and all time steps in the selected months. For each detected region, one 9×9 patch of hourly COSMO precipitation is stored, together with the matching latitude patch, longitude patch, valid time, and ensemble-member index. In this way, the method builds up a large sample of region-centred COSMO snapshots that are structurally consistent with the radar-based snapshots, but reflect the coarser model resolution and the spread across ensemble members.

10.4 Random forest (RF) and convolutional neural network (CNN)

In this section I introduce the two classifiers used in the hail analysis and describe how they are configured and trained. Both models are trained on convective RADKLIM snapshots (Sections 10.1.2 and 10.2.1): the input is a precipitation-only snapshot extracted around a convective region, and the target label (hail / no hail) is derived from the TRACE_{3D} hail-track dataset (Section 3.4). I use two forms of radar input in this framework. For the nowcasting experiment, the models operate on 5 min snapshots on the native 1 km RADKLIM grid (Section 10.1.2). For the model-transfer experiment, they use 1 h snapshots that are spatially and temporally harmonised with COSMO-DE-EPS (Section 10.2.1).

The first classifier is a random forest (RF) that operates on vectorised snapshots. The second is a compact convolutional neural network (CNN) that ingests the two-dimensional precipitation fields directly. The RF and CNN architectures, input representations, and training procedures are described in Sections 10.4.1 and 10.4.2. The evaluation of these models on 5 min radar snapshots is presented later in the nowcasting results (Section 11). The subsequent application of the RF to COSMO-DE precipitation snapshots is described in Section 12. The CNN is used only in the radar-based nowcasting experiment and is not applied to COSMO, because the model-transfer results are presented for the RF, which provides the stronger baseline performance in this study.

10.4.1 Random forest (RF) setup

As a baseline method, I employ a random forest (RF) to classify hail occurrence from the precipitation snapshots. A random forest is an ensemble of decision trees, each trained on a bootstrap sample of the data and using random subsets of features at each split (Breiman, 2001). By averaging the predictions of many such trees, the RF can capture non linear relationships and interactions while remaining relatively robust to noise and irrelevant predictors. It also requires little input preprocessing, which makes it a convenient choice for snapshot based precipitation fields once they are flattened into feature vectors. In addition, RFs provide class probabilities and simple diagnostics such as feature importance, which are useful for interpreting a precipitation only hail classifier. The RF outputs a probability $p \in [0, 1]$ for the hail class. When a binary hail decision is required, I apply a fixed decision threshold p_0 , so that $p \geq p_0$ is counted as “hail” and $p < p_0$ as “no hail”. Unless stated otherwise, I use $p_0 = 0.5$, which corresponds to the default decision rule in `scikit-learn` for binary classification.

For the RF, each RADKLIM snapshot is treated as a small image that is flattened into a feature vector. If a snapshot has dimensions $n_y \times n_x$, the corresponding feature vector has $P = n_y n_x$ elements. Stacking all snapshots gives a design matrix $X \in \mathbb{R}^{N \times P}$, where

N is the number of examples, and the target variable is a binary hail label $y \in \{0, 1\}$. The label indicates whether the convective region associated with the snapshot overlaps the hail-track data (hail) or not (no hail), as described in Sections 10.1.2 and 10.2.1. The same RF configuration is used for both 5 min and 1 h RADKLIM snapshots. Only the underlying snapshots differ in spatial and temporal resolution (Sections 10.1.2 and 10.2.1).

I split the snapshot dataset into training and test subsets using an 80/20 split with a fixed random seed, so that RF and CNN see exactly the same examples in each split. The RF is implemented with the `scikit-learn` library (Pedregosa et al., 2011). To avoid committing to an arbitrary set of hyperparameters, I tune the RF using a grid search on the training data and keep the test data strictly held out for later evaluation. The grid search tests the following values:

- number of trees $n_{\text{estimators}} \in \{50, 100, 200, 300\}$
- maximum tree depth $\text{max_depth} \in \{\text{None}, 10, 20, 30\}$
- minimum samples to split a node $\text{min_samples_split} \in \{2, 5, 10\}$
- minimum samples in a leaf $\text{min_samples_leaf} \in \{1, 2, 4\}$
- feature subsampling $\text{max_features} \in \{\text{sqrt}, \text{log2}, \text{None}\}$
- splitting criterion $\text{criterion} \in \{\text{gini}, \text{entropy}\}$

Configurations are ranked by classification accuracy on the training folds, and the best performing configuration is then evaluated on the held-out test set.

The best performing configuration is a 300-tree RF with unlimited tree depth. Each split considers $\sqrt{P}$ randomly chosen features and uses the Gini impurity as the splitting criterion. Nodes are allowed to split as long as they contain at least two samples ($\text{min_samples_split}=2$) and terminal leaves may contain a single sample ($\text{min_samples_leaf}=1$). Throughout, I use $\text{class_weight}='balanced'$ so that hail and no-hail snapshots contribute equally in expectation. This tuned RF is then refitted on the full training set and used as the reference precipitation-based classifier in the nowcasting and COSMO implementation experiments.

10.4.2 CNN setup

As a complementary approach to the RF, I also use a convolutional neural network (CNN) to predict hail from the RADKLIM precipitation snapshots. CNNs exploit the spatial structure of gridded data by applying local convolutional filters across the input field (LeCun et al., 1998). This makes them well suited to detecting patterns such as compact high-intensity cores and surrounding gradients in two-dimensional precipitation fields.

The CNN is primarily applied to 5 min RADKLIM snapshots on the native 1 km grid (Section 10.1.2), represented as single-channel images on fixed native-grid 20×20 km windows, corresponding to 20×20 pixels. Each input snapshot is therefore a 20×20 array of precipitation values, and the network outputs a single probability $p \in [0, 1]$

that the snapshot contains hail. The target label is the same binary hail or no-hail indicator used for the RF, derived from the overlap between the associated convective region and the TRACE_{3D} hail-track dataset (Section 3.4). In additional experiments I also apply the CNN to the 1 h harmonised radar snapshots (Section 10.2.1), and the relative performance of RF and CNN on these inputs is discussed in Section 12.

Before entering the network, the precipitation field is transformed using $\log(1 + P)$ to compress the heavy upper tail, and then standardised using the mean and standard deviation over all snapshots. Inputs are precipitation only, without additional predictors, so the CNN learns directly from the shape and intensity pattern of the precipitation footprint.

I tested a small set of compact CNN configurations that varied the number of convolutional layers (two vs three), the use of batch normalisation, and the use of data augmentation during training. The configuration used in this thesis consists of three convolutional layers with 3×3 kernels. The first two layers (with 32 and 64 feature maps) are each followed by batch normalisation and a ReLU activation, and 2×2 max pooling reduces the spatial resolution from 20×20 to 5×5 . A third convolutional layer with 128 feature maps and batch normalisation refines the representation at this coarser scale. The resulting feature maps are flattened and passed through a fully connected layer with 256 units and ReLU activation, followed by dropout with a rate of 0.3, and a final output layer that produces logits for the two classes. A softmax activation is used at evaluation time to obtain hail probabilities.

The dataset is stratified by class and split into training and test sets with an 80/20 proportion, using a fixed random seed so that the RF and CNN are trained and tested on identical snapshot sets. Within the training portion, a further 10% is held out for validation in order to monitor generalisation during training. The CNN is trained with a cross-entropy loss and the Adam optimiser. A learning-rate scheduler reduces the learning rate when the validation loss stops improving, and early stopping is applied based on the validation loss in order to avoid overfitting once the model has converged.

To improve generalisation and reduce sensitivity to the exact orientation of the precipitation footprint, I apply lightweight geometric data augmentation during training. Each training snapshot can be randomly rotated by up to 15° and randomly flipped in the horizontal direction. These operations preserve precipitation intensities but expose the network to a wider variety of spatial patterns.

This CNN configuration is used only for the radar-based nowcasting experiment. Its performance, alongside that of the RF, is evaluated on 5 min RADKLIM snapshots in Section 11.

I begin by evaluating the two snapshot-based classifiers, a compact convolutional neural network (CNN) and a random forest (RF). The dataset contains 37,056 labelled 20×20 km snapshots. Of these, 80% are used for training and validation and 20% (7,412 snapshots) are held out for testing, with the split stratified by class so that both models see identical training and test sets. Both classifiers are trained on 5 min RADKLIM snapshots. For details on the model configurations, input preprocessing, and training procedure, the reader is referred to Sections 10.4.1 and 10.4.2.

This chapter evaluates the models on the held-out test set using accuracy, confusion matrices, diagnostic composites, and probabilistic calibration. Here, accuracy is defined as the fraction of correctly classified snapshots, $ACC = (TP + TN) / (TP + TN + FP + FN)$. Across multiple runs, the best test accuracies are very similar, with 85.91% for the CNN and 85.71% for the RF. The different runs use the same train test split but differ only because parts of the training procedure are stochastic. The details are described in Sections 10.4.1 and 10.4.2. This suggests that both methods extract comparable predictive information from the radar snapshots.

To see how this skill is distributed across the two classes, I first look at the confusion matrices (Fig. 11.1). Most cases lie on the diagonal for both models, which confirms that the high accuracies are not dominated by one class alone. The main differences appear in the off-diagonal elements. The RF produces slightly fewer false negatives but somewhat more false positives than the CNN. In Fig. 11.1, the RF yields $FN = 531$ and $FP = 528$, while the CNN yields $FN = 566$ and $FP = 478$. Compared to the CNN, the RF reduces false negatives by 35 cases but increases false positives by 50 cases. The RF is therefore more inclined to flag hail. The CNN, in contrast, is a little more conservative and misses a few more hail cases while keeping the number of false alarms lower. Since the overall accuracies are almost identical, the choice between the two models is mainly a question of whether missed events or false alarms are more critical for the intended application.

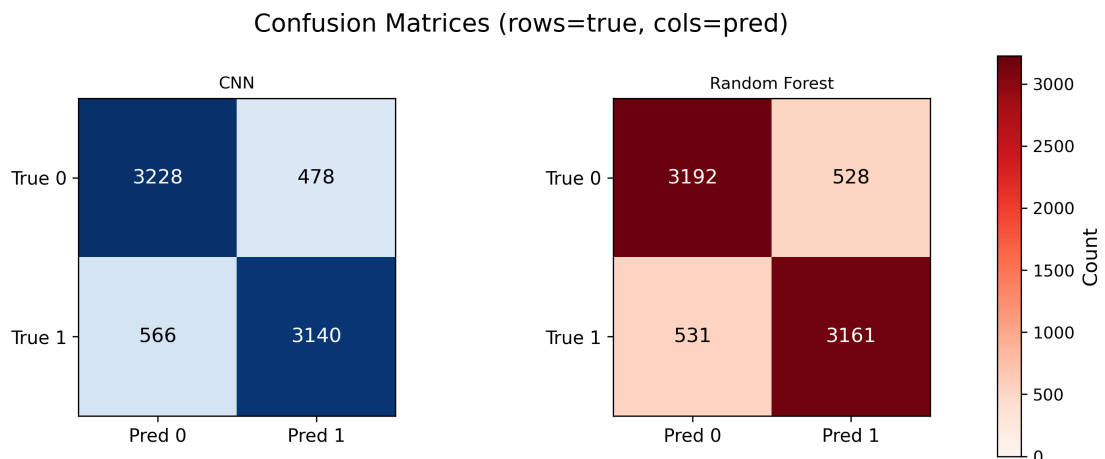


Figure 11.1: Confusion matrices for the CNN (left) and the RF (right) on the 5 min test set (rows = true class, columns = predicted class).

The confusion matrices quantify how many errors of each type occur, but they do not show what those scenes look like in terms of precipitation structure. To gain physical intuition, I therefore composite the original 5 min RADKLIM precipitation fields separately for true positives (TP), false positives (FP), and false negatives (FN) (Fig. 11.2). For both models, the TP composites (CNN: Fig. 11.2a, RF: Fig. 11.2d) show a strong, compact precipitation maximum near the centre of the window with a smooth decay towards the edges of the snapshot. This structure is consistent with the expectation for hail-labeled snapshots in this setup, since the hail label is defined using the overlap between convective regions and the hail mask. Hail-labeled snapshots therefore often coincide with intense, localised precipitation cores in the radar field.

The FP composites (CNN: Fig. 11.2b, RF: Fig. 11.2e) exhibit weaker peak precipitation and broader, more diffuse patterns than the TPs. At the same time, some snapshots labelled as non-hail still display compact and locally enhanced precipitation features that resemble the TP structures. This highlights an intrinsic ambiguity when using precipitation as the only predictor, since intense convective rain can produce patterns that are similar to hail-labeled scenes in the radar field.

The FN composites (CNN: Fig. 11.2c, RF: Fig. 11.2f) have the lowest mean 5 min precipitation values, with only a small and localised enhancement near the centre of the patch. Compared with the TP composites (Fig. 11.2a,d), the missed hail cases therefore tend to be associated with weaker precipitation signatures in the radar field. A secondary contribution may come from variability in where the precipitation maximum occurs within the fixed 20×20 km window, since averaging fields that are not perfectly co-located tends to smooth the composite and reduce the peak amplitude.

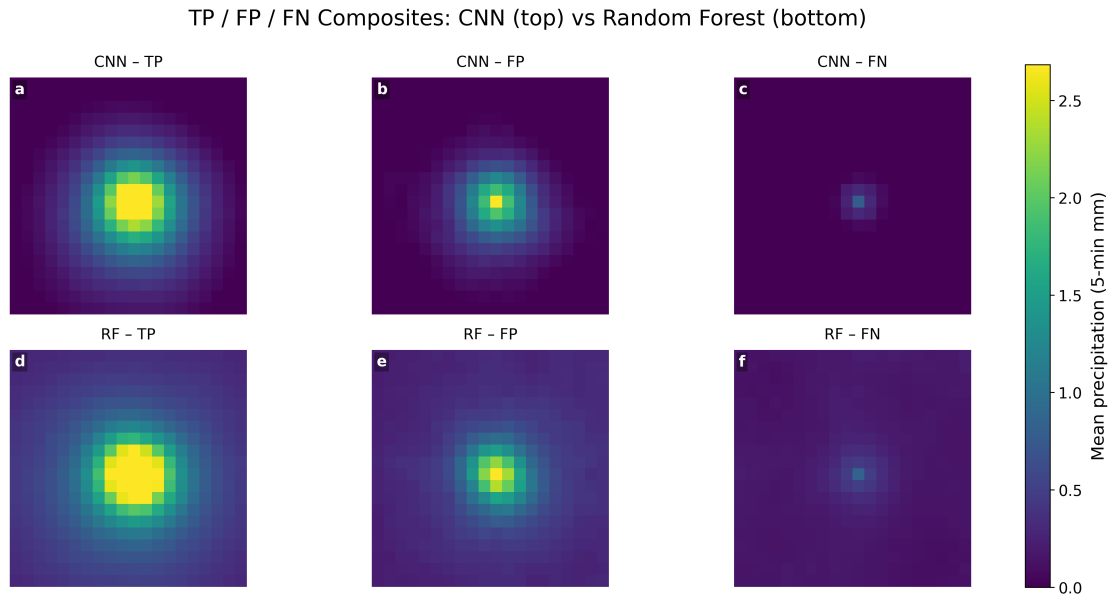


Figure 11.2: Composites of mean 5 min RADKLIM precipitation snapshots grouped by true positives (TP), false positives (FP), and false negatives (FN) for the CNN (a–c) and the RF (d–f). Panels show TP (a,d), FP (b,e), and FN (c,f). Colours indicate mean precipitation (mm per 5 min).

Overall, the CNN and RF composites are broadly similar across all categories (Fig. 11.2a–f), with only subtle differences in the sharpness of the central maximum and the surrounding gradients, in line with the small differences seen in the confusion matrices (Fig. 11.1).

Because both models provide probabilistic outputs, I now assess whether these probabilities are reliable enough to be used directly as hail risk estimates. I examine calibration and sharpness (Fig. 11.3). Calibration describes whether predicted probabilities correspond to observed event frequencies, while sharpness describes how often the forecasts are confident, with probabilities concentrated near 0 or 1 rather than spread across intermediate values. The reliability diagram in the top panel compares predicted probabilities with the observed hail frequency in each probability bin. Both curves lie close to the one-to-one line, with only modest deviations in the mid range, so neither model shows strong miscalibration. The histograms conditioned on the true class in the middle and bottom panels show that non-hail cases cluster near 0 and hail cases near 1 for both models, which indicates good discrimination.

Probability Calibration: CNN vs Random Forest

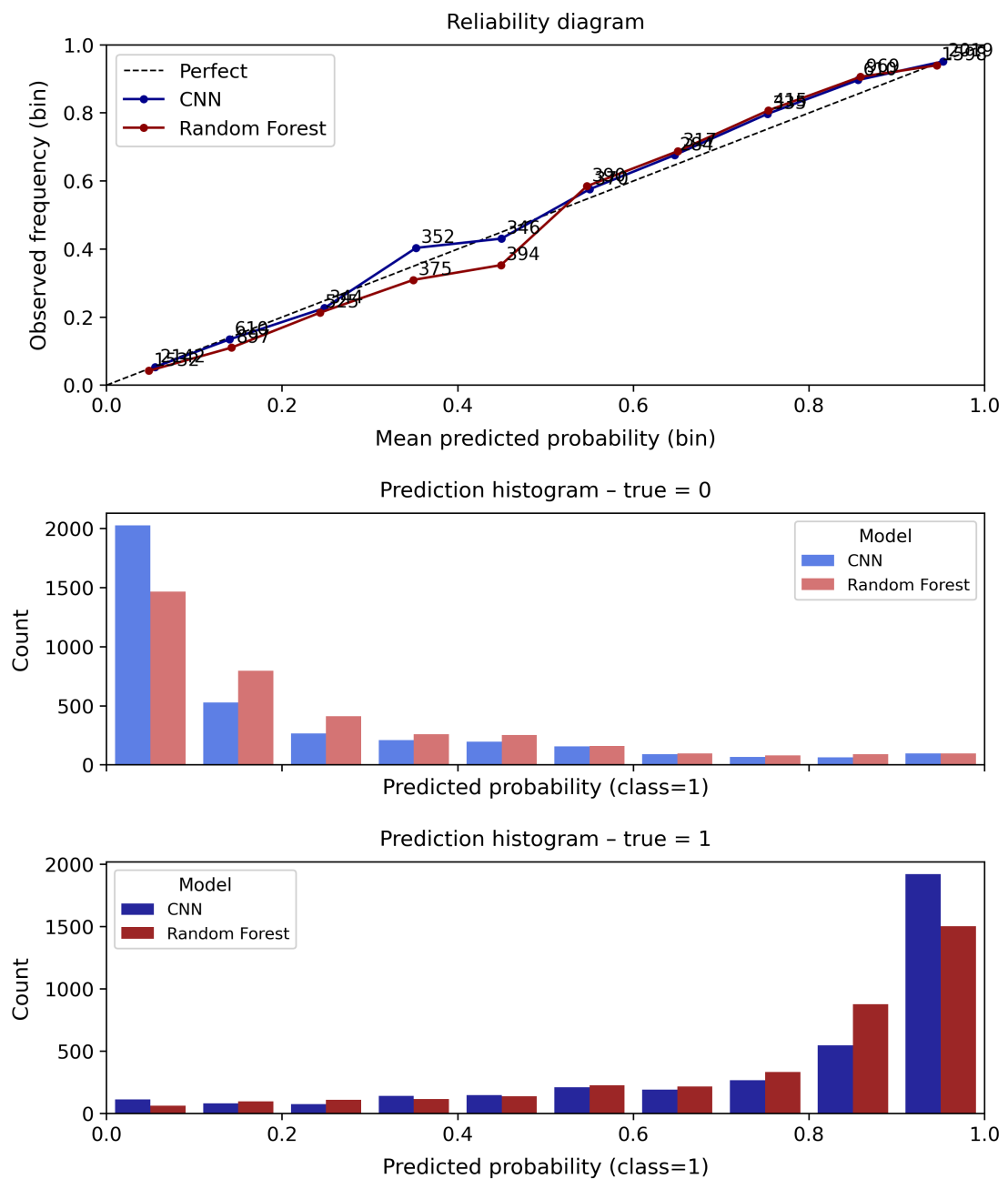


Figure 11.3: Calibration and sharpness for the CNN (blue) and RF (red) on the 5 min test set. Top: reliability diagram with the one-to-one reference (dashed), where numbers denote the sample size in each probability bin. Middle: histogram of predicted probabilities for non-hail cases (true = 0). Bottom: histogram of predicted probabilities for hail cases (true = 1).

The RF places slightly more probability mass at the extremes, while the CNN assigns relatively more probability to intermediate bins. This behaviour is consistent with the RF being slightly more willing to issue very high hail probabilities and with the CNN retaining a somewhat more conservative approach. For example, a forecast of $p = 0.8$ can be interpreted as an estimated 80% chance of hail for that snapshot, so a user could trigger an alert only when $p \geq 0.7$ to reduce false alarms, or use a lower threshold such as $p \geq 0.3$ when missed events are more costly.

Operational warning systems already use radar-based products that provide hail probabilities for decision-making. For example, the WSR-88D Hail Detection Algorithm derives cell-based probabilities of severe hail (POSH) from radar reflectivity and environmental data, which are then thresholded to support warning decisions (Witt et al., 1998). In Europe, C-band radar products providing probability of hail and probability of severe hail have been developed and evaluated for Belgium and Switzerland (Lukach et al., 2017; Kopp et al., 2024), and recent work has started to employ machine-learning and CNN-based probabilistic hail forecasts from radar and NWP data for near-real-time decision support (Gagne et al., 2017; Forcadell et al., 2024; Rombeek, Leinonen, and Hamann, 2024).

Within this broader context, the CNN and RF developed here can be viewed as compact, RADKLIM-based hail-probability generators for the German radar domain. Compared with many existing operational and research products, the approach differs in three main aspects. First, the predictors are restricted to precipitation snapshots only, which isolates the information content of the radar precipitation field and yields a model that can be applied wherever RADKLIM-type precipitation estimates are available. Second, the models operate on fixed 20×20 km windows at 5 min resolution, which directly targets cell-scale structure rather than cell tracking attributes. Third, performance is quantified using both deterministic metrics and probabilistic verification, so that the output probabilities can be interpreted as risk-style estimates rather than only as categorical labels.

At 5 min resolution, the output field $p(x, y, t) \in [0, 1]$ can be displayed as a hail-risk map, which helps forecasters see where hail-producing cells are most likely and how the risk evolves from one snapshot to the next. The same probabilities can also be turned into simple decision rules by applying a threshold $p \geq \tau$ to trigger very short-term warnings, temporarily protect exposed assets, or prioritise regions that should be examined in more detail using radar and model fields, with τ tuned to balance missed events against false alarms. Because the CNN is slightly more conservative and the RF is slightly more inclined to flag hail, they can be offered as complementary products, with the RF providing a more sensitive hail-alert layer and the CNN offering a more cautious, low false-alarm alternative.

Radar-based systems are limited to the radar footprint, and their lead time is inherently short. In practice, lead times beyond the current radar scan are obtained through radar-signal nowcasting methods that extrapolate observed precipitation patterns for-

ward in time, often providing guidance up to about 1–2 h depending on the situation. Such approaches are used operationally, for example based on optical-flow radar extrapolation at DWD (Deutscher Wetterdienst, 2024b), and they are also a building block within DWD’s seamless forecasting initiative SINFONY (Deutscher Wetterdienst, 2024a) that aims to integrate nowcasting and NWP on short lead times. The probability outputs developed here can be used within these short-range workflows, either by applying the classifiers to extrapolated radar precipitation fields or by tracking and propagating the derived hail-probability patterns.

For applications that require longer lead times or guidance independent of radar availability, a next step is to investigate whether the same hail-prediction logic can be transferred to numerical model output. In the next section, I therefore apply the radar-trained RF to precipitation snapshots from COSMO-DE-EPS and assess how well hail occurrence can be recovered directly from the model fields.

In the previous section, I evaluated the RF and CNN classifiers trained on 5 min RADKLIM snapshots on the native 1 km radar grid and discussed how their probabilistic output could be used for near-real-time hail nowcasting over Germany. In the following, I adapt this framework to precipitation fields from the operational COSMO-DE-EPS, which provides 20 ensemble members at coarser spatial (about 2.8 km) and temporal (hourly) resolution. The aim is to drive the radar-trained RF with model precipitation snapshots so that hail probabilities can be obtained from numerical forecasts.

To achieve this, I use the harmonised COSMO and RADKLIM snapshots described in Chapter 10. Briefly, hourly COSMO precipitation from all 20 ensemble members is processed into 9×9 gridpoint patches centred on local hourly precipitation maxima. The 5 min RADKLIM fields are aggregated to hourly totals (Section 10.2) and regridded onto the COSMO grid so that radar and model snapshots share the same spatial and temporal resolution (Section 10.3.2). Even after this harmonisation, the precipitation intensity distributions can differ systematically between COSMO and RADKLIM, and this matters because the RF was trained on radar-based intensities. If uncorrected, such a bias can shift the RF decision boundary and lead to systematically overconfident or underconfident hail probabilities when the model fields are used as input. On this basis, the next subsection quantifies systematic intensity differences between COSMO and RADKLIM, after which the RF is applied to the COSMO snapshots and the resulting hail predictions are evaluated.

12.1 Bias correction

To obtain hail probabilities from numerical forecasts, I transfer the RF trained on RADKLIM precipitation snapshots to COSMO precipitation fields. This provides hail-probability guidance with lead times beyond radar-based monitoring and short-range extrapolation. Training an RF directly on COSMO is not straightforward because COSMO does not predict hail explicitly and therefore does not provide a native hail label that is comparable to the radar-based hail mask used in this thesis. Using the radar-trained RF and applying it to COSMO precipitation snapshots is therefore a practical way to obtain hail probabilities from model output while keeping the label definition consistent.

A prerequisite for this transfer is that the RF should see precipitation intensities that are comparable to those used during training. However, even after regridding and temporal aggregation, COSMO and RADKLIM precipitation can differ systematically

in their intensity distributions. If uncorrected, this mismatch shifts the RF input distribution and can lead to biased hail probabilities. I therefore quantify these differences and apply a bias-correction procedure before applying the RF on COSMO snapshots. Figure 12.1 illustrates the systematic intensity differences and motivates the correction described below.

For the bias-correction step, I use a RADKLIM snapshot set that is constructed specifically to be directly comparable to the COSMO snapshots on the harmonised hourly grid. For consistency, this RADKLIM sample is not identical to the labelled hourly RADKLIM snapshot set described in Section 10.2.1, which is based on SHY95 coherent-area objects and is used for RF and CNN training and evaluation. Instead, the bias-correction sample is defined with the same threshold-based sampling applied to COSMO (Section 10.3), so that the two datasets are treated symmetrically for distribution matching and the correction primarily reflects systematic precipitation intensity differences. The implementation of the bias correction procedure is summarised below.

COSMO hourly precipitation fields from all 20 ensemble members are thresholded at 1 mm h^{-1} . Connected components in this mask are identified, and a fixed 9×9 patch centred on the local precipitation maximum is extracted, as described in Section 10.3. RADKLIM is first aggregated to 1 h accumulations and regridded to the COSMO grid. The same 1 mm h^{-1} threshold and snapshot extraction are then applied to the hourly radar fields. This yields two comparable collections of hourly convective snapshots from COSMO and RADKLIM that are constructed without reference to hail labels and are used to diagnose and correct systematic precipitation intensity differences.

The resulting sample comprises $N_{\text{rad}} = 5,403,564$ RADKLIM snapshots and $N_{\text{cos}} = 15,770,762$ COSMO snapshots. The ratio is not expected to be 20, even though COSMO-DE-EPS has 20 ensemble members, because the snapshot archive is object-based rather than field-based. One snapshot is stored for each threshold-exceeding precipitation object, so the total count depends on both the number of available member-hour fields and the mean number of detected objects per field. In addition, the COSMO archive used here provides hourly precipitation only for a limited set of forecast lead times (18–21 h), which yields three consecutive 1 h fields per forecast run (Section 10.3.1). Moreover, if COSMO precipitation is systematically wetter, areas exceeding the 1 mm h^{-1} threshold are more likely to be spatially connected, so nearby cores merge into fewer but larger connected exceedance regions.

Using all grid-point values from the COSMO and RAD snapshots, I compare the one-dimensional distributions of hourly precipitation in COSMO and RADKLIM (Fig. 12.1a). Although the COSMO snapshot archive is larger, the histograms are normalised, so the comparison reflects relative frequencies rather than absolute counts. The larger sample size therefore mainly reduces sampling noise and yields a smoother curve, while any systematic offset between the curves reflects genuine intensity differences between the datasets.

The overall shapes of the distributions are similar, but COSMO is clearly wetter. Its histogram is shifted toward higher values and shows a much heavier upper tail. As an additional diagnostic, the same wet shift is also apparent when the snapshots are stratified into hail and no-hail subsets using the radar-based labels. This stratification is used only for illustration and does not enter the bias-correction mapping itself. The wet shift is consistent with the COSMO wet bias discussed in Part II. Because the RF is trained on radar-derived precipitation, it has learned to associate certain intensity ranges with hail. If those ranges are filled with systematically higher COSMO values, many situations that would be non-hail in the radar data start to look hail-like to the model, which leads to overprediction. This is confirmed by a preliminary run of the radar-trained RF on raw COSMO snapshots. Using a default decision threshold of $p \geq 0.5$, the predicted class frequencies are

Class 0 (no hail): 12,479,404 samples (79.13%), Class 1 (hail): 3,291,358 samples (20.87%).

In other words, the RF would imply hail in roughly one-fifth of all COSMO snapshots, which is unrealistically high compared with the much rarer hail occurrence seen in the radar record and in typical hail climatologies.

To mitigate this, I tested three simple bias-correction strategies applied to COSMO precipitation before RF inference: (i) a single global multiplicative factor based on the COSMO-to-radar mean ratio, (ii) bin-wise scaling based on ratios of the radar and COSMO histograms, and (iii) quantile mapping (QM).

Quantile mapping performed best and behaved most robustly across subsets and months. In QM, each COSMO value is transformed according to its empirical quantile in the COSMO distribution and mapped to the corresponding quantile of the RADKLIM distribution (Thiemeßl, Gobiet, and Leuprecht, 2011; Gudmundsson et al., 2012). Equivalently, a COSMO value at the 90th percentile of the COSMO snapshot-value distribution is mapped to the 90th-percentile value of the corresponding radar distribution. This aligns the full distribution of COSMO values with that of radar rather than adjusting only a single moment such as the mean or the variance.

QM has several advantages. It can handle non-linear, intensity-dependent biases, it corrects both the bulk and the tail of the distribution, and it is relatively easy to implement once a sufficiently large sample from both datasets on a common grid and time resolution is available. At the same time, it assumes that the statistical relationship between COSMO and RADKLIM is stationary over time and may be less reliable in the very far tail, where sample sizes are small. QM also cannot correct errors in the timing or spatial placement of precipitation. It only adjusts the local intensity distribution.

In this application, these assumptions are acceptable because the COSMO–RADKLIM relationship is estimated from a large multi-year sample on the harmonised hourly grid, and the dominant discrepancy is a smooth wet shift in intensity rather than structural errors in placement. Moreover, the RF is trained purely on radar snapshots, so the bias-correction step does not use any hail labels and does not alter the training of

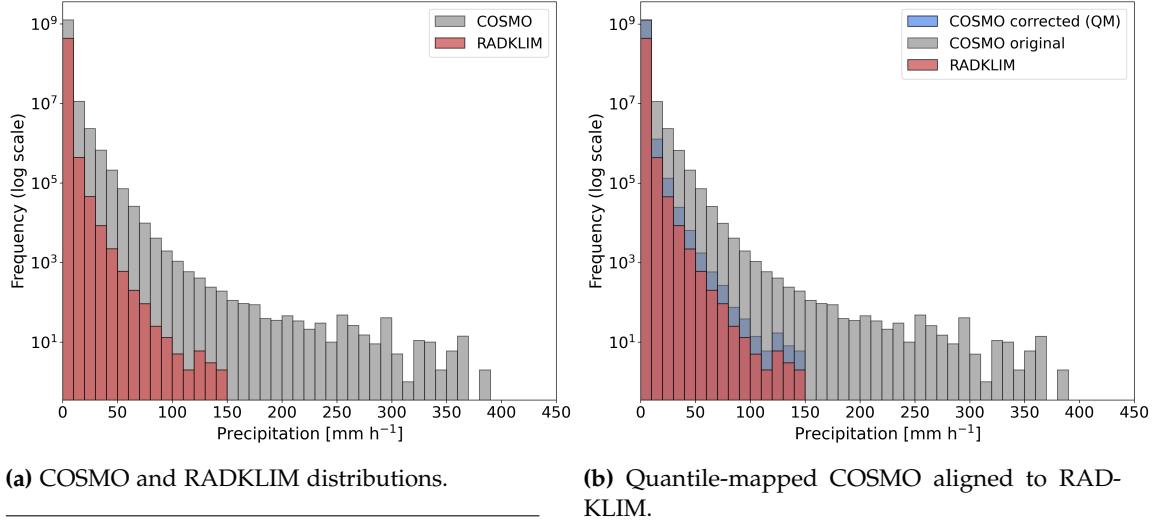


Figure 12.1: COSMO and RADKLIM hourly snapshot-value distributions on the harmonised grid (a) and the corresponding quantile-mapped COSMO distribution aligned to RADKLIM (b).

the classifier itself. I therefore compute a single quantile-mapping function from the full multi-year COSMO and RADKLIM snapshot-value distributions and apply this mapping to the entire COSMO dataset before running the RF. As shown in Fig. 12.1b, this correction brings the COSMO distribution much closer to RADKLIM in both the bulk and the upper tail.

The practical impact on RF inference is illustrated by repeating the preliminary RF run using quantile-mapped COSMO snapshots. Using the same default decision threshold of $p \geq 0.5$, the predicted class frequencies become

Class 0 (no hail): 15,097,652 samples (95.73%), Class 1 (hail): 673,110 samples (4.27%).

Compared with the uncorrected COSMO input, the implied hail frequency is substantially reduced and becomes more consistent with the rarity of hail in the radar record.

Unless stated otherwise, all COSMO-based results presented in the following sections use quantile-mapped (bias-corrected) COSMO precipitation snapshots as RF input.

12.2 Random forest benchmark on 1 h RADKLIM snapshots

To obtain a prediction model for hail occurrence at COSMO-like resolution, I benchmark the two supervised classifiers introduced in Sections 10.4.1 and 10.4.2 on hourly RADKLIM snapshots. Both models are trained and evaluated on the same hourly snapshot dataset, so their performance can be compared directly. Across the CNN configurations

discussed in Section 10.4.2, the hourly-snapshot test accuracy ranges from about 73% up to a best value of 76.78%. A plausible reason for this lower CNN skill is that temporal accumulation and spatial resampling smooth the compact precipitation cores and sharp gradients that are prominent in the 5 min fields and that appear to be particularly informative for the CNN in the nowcasting benchmark (Section 11).

In contrast, the RF achieves substantially higher accuracy on the same hourly dataset. I therefore adopt the RF as the main classifier for the COSMO-DE implementation and focus on its behaviour in the following.

The RADKLIM snapshots are constructed as described in Section 10.2. In brief, twelve 5 min frames are accumulated into 1 h totals, convective regions are identified from the 5 min convective masks, and storm-centred windows are extracted around local precipitation maxima and resampled to 9×9 COSMO-aligned patches representing roughly 25×25 km. On this 1 h radar dataset, the RF setup from Section 10.4.1 is used without modification. The snapshot pool contains $N_{\text{rad},1\text{h}} = 37,194$ labelled hourly snapshots. I split these into an 80% training set (29,755 samples) and a 20% test set (7,439 samples) with a fixed random seed. The RF is trained on the 1 h RADKLIM snapshots only and evaluated on the held-out hourly radar test set.

On this 1 h radar benchmark, the RF achieves a test accuracy of 85.11%. The confusion matrix on the test set is

$$\begin{pmatrix} \text{TN} & \text{FP} \\ \text{FN} & \text{TP} \end{pmatrix} = \begin{pmatrix} 3098 & 614 \\ 494 & 3233 \end{pmatrix}.$$

The RF correctly classifies most hourly snapshots in both classes, with somewhat more false alarms (614) than missed events (494). Compared with the 5 min nowcasting experiment, the overall accuracy is slightly lower, which is consistent with hourly accumulation and the coarser 9×9 representation reducing the sharpness and compactness of convective precipitation maxima. This can weaken class separation when precipitation structure is the only predictor. At the same time, the RF still provides a strong radar-based reference at COSMO-like resolution.

12.3 Random forest results on COSMO snapshots

After training and benchmarking the RF on 1 h RADKLIM snapshots, I next apply the same classifier to the bias-corrected COSMO-DE-EPS snapshots. All convective COSMO snapshots in the 2011–2018 period are collected on the harmonised hourly grid and passed through the RF, which returns a hail probability for each 9×9 patch. As discussed in Section 12.1, the full COSMO snapshot sample is split into 15,097,652 no-hail cases (95.73%) and 673,110 hail cases (4.27%).

As a simple structural check, Fig. 12.2 shows mean hourly COSMO snapshots composited over all hail-labelled cases (left, label 1) and all no-hail cases (right, label 0).

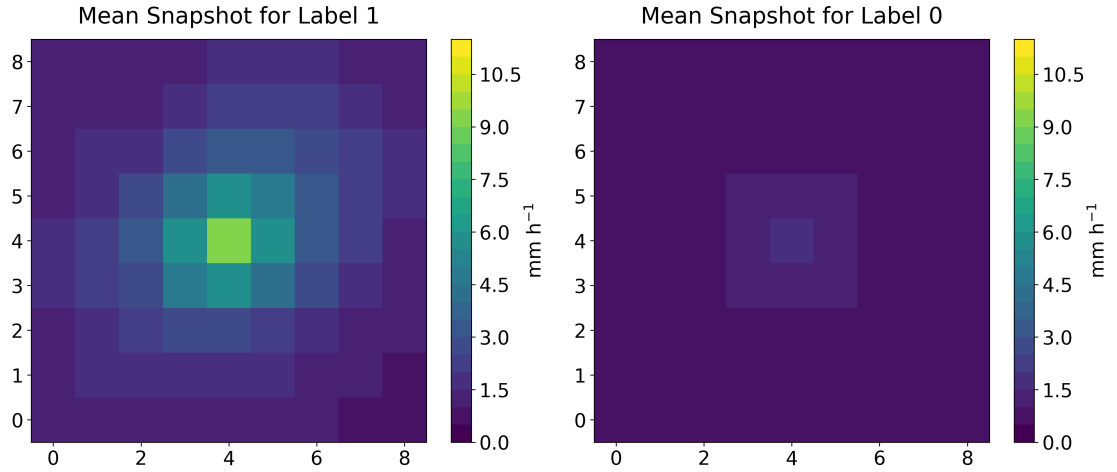


Figure 12.2: Mean hourly COSMO-DE-EPS snapshots for hail cases (left, label 1) and no-hail cases (right, label 0) based on RF-derived labels. Colours show mean precipitation (mm h^{-1}).

Hail cases exhibit a compact and intense central maximum with smoothly decreasing precipitation toward the edges, whereas the no-hail composite has a much weaker and more diffuse signal. A peaked structure is expected in this setup because the RF was trained on snapshots centred on local precipitation maxima within convective regions, and hail-labelled radar regions typically coincide with intense, localised convective precipitation cores. The similarity to the radar-based composites from the nowcasting experiment suggests that the RF is responding to consistent storm-scale structures in the COSMO fields rather than to artefacts of the bias correction.

In the next section, I evaluate the practical predictability of hail in the post-processed COSMO-DE-EPS dataset by comparing the COSMO-based hail probabilities with the hail tracks and testing the sensitivity to the spatial matching radius and the allowed time offset between model and observations.

In this part of the hail analysis, I verify the COSMO-based RF hail predictions against the observed hail tracks. The aim is to assess how often and where predicted hail coincides with observed hail and to identify the dominant sources of mismatch. The chapter addresses two questions. First, how sensitive are the verification results to the choice of spatial matching radius and temporal matching window. Second, how does the resulting skill vary across Germany and from year to year.

A strict point-by-point comparison is not appropriate for hail because the tracks are highly localised in space and time and both observations and forecasts contain displacement uncertainty. Even small spatial offsets or timing errors can turn a physically reasonable prediction into a miss when verification is performed at a single grid point and a single valid time. I therefore use neighbourhood verification and allow a temporal tolerance window. By varying the neighbourhood radius and the allowed time offset, I quantify how much of the apparent disagreement is consistent with displacement and timing differences rather than a complete failure to predict hail.

COSMO-DE-EPS provides 20 ensemble members. I apply the RF to each member separately, which yields a hail probability for every snapshot. For verification, these probabilities are converted into a binary hail / no-hail decision using the same fixed threshold as in the RF setup ($p_0 = 0.5$, unless stated otherwise). The ensemble forecast is then defined using an “at least one member” rule, meaning that hail is considered predicted if any of the 20 members indicates hail at a given location and time.

The verification setup is then defined by the ensemble event rule and by the spatial and temporal matching tolerances used to pair forecasts with tracks. The temporal window Δt specifies the maximum allowed time difference between an observed hail-track time and a forecast time for the two to be considered the same event. The radius R specifies the maximum horizontal distance between a hail-track location and the nearest forecast hail grid point for a match. A forecast is counted as a hit if a predicted hail grid point occurs within radius R of a hail-track location and within the temporal tolerance window Δt . If no such forecast can be matched, the event is counted as a miss. Predicted hail events that cannot be matched to any track within $(R, \Delta t)$ are counted as false alarms.

In the first sensitivity experiment (Section 13.1), I fix $\Delta t = 30$ min and vary the spatial radius $R \in \{20, 25, 30, 35, 45, 60, 80, 100\}$ km. In the second experiment (Section 13.2), I fix the spatial radius to $R \in \{20, 100\}$ km and vary the temporal matching window $\Delta t \in \{30, 60, 90, 120, 150, 180\}$ min, representing a strict and a very tolerant spatial match, respectively.

All comparisons between COSMO predictions and hail tracks are summarised in contingency tables that count hits (TP), misses (FN), false alarms (FP), and true negatives (TN) within the chosen spatial neighbourhood and time window. From these counts I compute several standard scores:

- Accuracy (ACC): $ACC = (TP + TN) / (TP + FP + FN + TN)$. Values range from 0 to 1, with 1 indicating that all events are classified correctly.
- Probability of detection (POD): $POD = TP / (TP + FN)$. POD ranges from 0 to 1 and measures how often observed hail is detected (POD = 1 means no misses, POD = 0 means all hail events are missed).
- False alarm ratio (FAR): $FAR = FP / (TP + FP)$. FAR also ranges from 0 to 1 and measures how many predicted hail events do not correspond to an observed hail track. FAR = 0 means no false alarms, FAR = 1 means that all hail forecasts are false.
- Critical success index (CSI): $CSI = TP / (TP + FP + FN)$. CSI ranges from 0 to 1, combines hits, misses, and false alarms, and is widely used for rare events. CSI = 1 indicates perfect forecasts, whereas CSI = 0 corresponds to no correctly predicted hail events.

13.1 Sensitivity to spatial radius ($\Delta t = 30$ min)

To keep the discussion focused, I show only two illustrative radii in the main text. For hourly COSMO precipitation and point-based hail tracks, a neighbourhood radius of order 45–60 km together with $\Delta t = 30$ min is a physically reasonable compromise. It allows for typical spatial displacement and small timing differences while still requiring event-level proximity. The sensitivity analysis below brackets this reference choice by comparing a tight criterion ($R = 20$ km) with a very permissive one ($R = 100$ km).

For the reference setup ($\Delta t = 30$ min, $R = 60$ km), the rare-event skill remains modest. Across the domain, CSI typically ranges between 0 and about 0.2, while FAR is generally high (roughly 0.7–1).

Figure 13.2 shows accuracy (ACC) and the false alarm ratio (FAR) for $R = 20$ and $R = 100$ km. ACC is uniformly high (typically ≥ 0.9) across Germany at both radii. This behaviour is expected because TN counts dominate the contingency table. As a result, ACC is not very informative for hail verification, whereas FAR provides more insight. At $R = 20$ km, FAR is close to one over large parts of the domain, which means that many hail forecasts cannot be matched to a track within this tight neighbourhood. Increasing the radius to $R = 100$ km reduces FAR in several regions because the larger neighbourhood tolerates displacement and converts some near misses into hits.

13.1 SENSITIVITY TO SPATIAL RADIUS ($\Delta t = 30$ min)

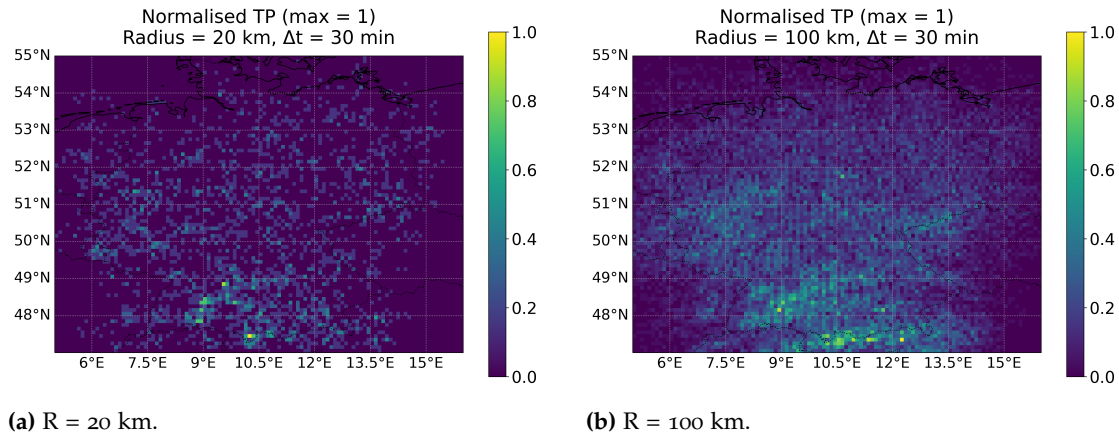


Figure 13.1: Normalized true-positive (TP) density for $R = 20$ km (a) and $R = 100$ km (b).

Given the strong class imbalance, I focus on the critical success index (CSI) and the probability of detection (POD). CSI combines hits, misses, and false alarms and is widely used for rare events, while POD isolates the detection component and helps explain spatial variations in CSI. CSI is low and patchy at $R = 20$ km, typically below 0.05 (Figure 13.3). At $R = 100$ km CSI becomes more uniform and increases to about 0.10–0.18 over extended areas. In this verification framework, CSI values in this range indicate modest but meaningful event-level overlap between predicted and observed hail on the coarse grid. The score remains far from one because hail is rare and the contingency table penalises both missed events and false alarms accumulated over many time steps. The POD map at $R = 20$ km shows small, localised areas with relatively high detection surrounded by much larger regions with poor detection. Together with the high FAR, this explains why CSI remains modest when matching is restricted to a small radius.

Figure 13.1 represents the true-positive (TP) density for $R = 20$ km and $R = 100$ km. Despite the low CSI at small radii, the TP maps remain physically consistent with the observed hail pattern (Figure 9.1b). TP hotspots cluster along the same hail-active corridors that appear in the track climatology. This indicates that the COSMO-based RF captures the spatial distribution of hail and tends to place hail in the right regions, even if individual events are often displaced by more than 20 km when using hourly accumulations. In summary, increasing the neighbourhood radius improves the rare-event scores by relaxing spatial displacement errors, and the TP fields confirm that the forecasts target the correct hail-prone areas climatologically.

13.1 SENSITIVITY TO SPATIAL RADIUS ($\Delta t = 30$ min)

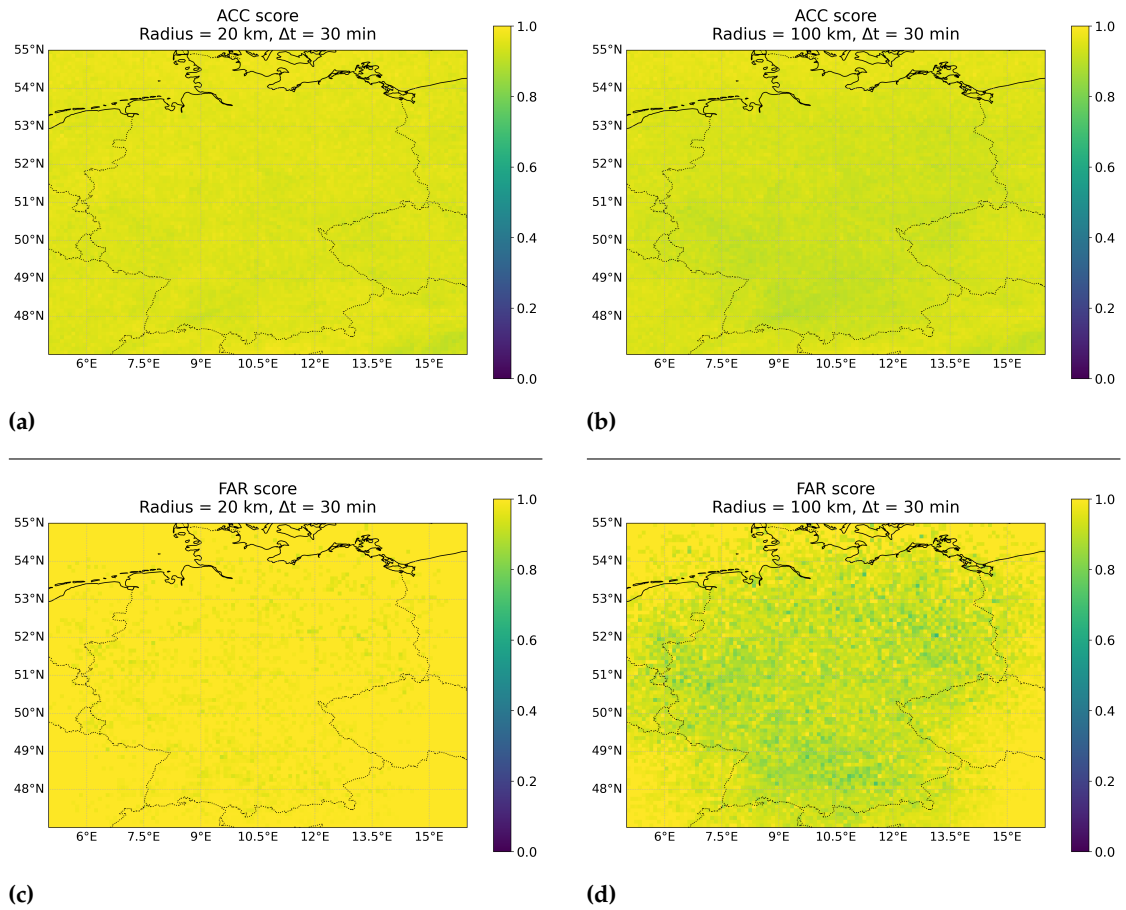


Figure 13.2: Spatial maps of accuracy (ACC; (a),(b)) and false-alarm ratio (FAR; (c),(d)) for $R = 20$ km (left) and $R = 100$ km (right), verified with a 30 min window.

13.1 SENSITIVITY TO SPATIAL RADIUS ($\Delta t = 30$ min)

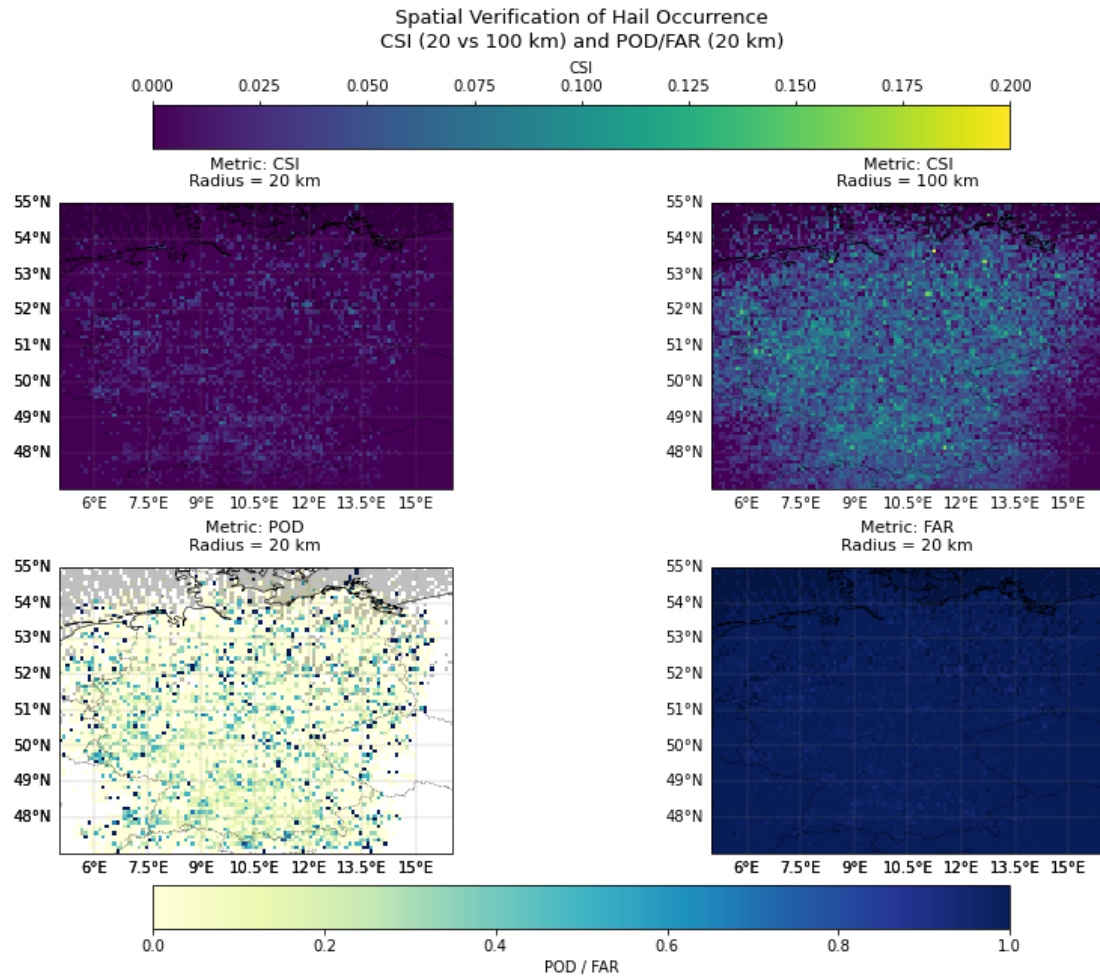


Figure 13.3: Skill dashboard for the neighbourhood verification. Top: CSI for $R = 20$ km and $R = 100$ km (common colour scale). Bottom: POD and FAR for $R = 20$ km.

13.2 Sensitivity to temporal window Δt

To complement the spatial analysis, I next examine how the verification depends on the temporal matching window Δt . Using the same neighbourhood matching and gridded TP, FP, FN, and TN counts, I evaluate $\Delta t \in \{30, 60, 90, 120, 150, 180\}$ min for two fixed spatial radii, $R = 20$ km and $R = 100$ km, representing a strict and a very tolerant spatial match, respectively. Figure 13.4 summarises the results. The top row shows domain-mean skill scores (POD, FAR, CSI, ACC) as a function of Δt , and the bottom row shows the corresponding summed event counts (TP, FP, FN, TN, log scale).

At $R = 20$ km the spatial match is very strict, so the main limitation is that forecasts and tracks often do not line up close enough in space. In this situation, enlarging Δt only has a modest effect. As the window grows from 30 to 180 minutes, more hail hours find a matching forecast within the tight radius, so the number of true positives increases and POD rises slightly. False alarms increase much more slowly, so FAR stays almost flat and CSI remains nearly unchanged, with at most a small improvement at the longest windows. Accuracy stays very high (about 0.95–0.99), which is expected.

At $R = 100$ km the picture changes. Here the spatial match is already generous, so many forecasts and tracks are close enough in space. Extending Δt now mainly adds extra opportunities to count a forecast as a hit even if it occurs well before or after the observed hail. Both TP and FP increase with Δt , but FP grows more rapidly than at 20 km. POD is similar or only slightly higher, while FAR increases and CSI becomes flat or even a little lower for the longest windows. Thus, the marginal gain in detection obtained by enlarging the time window is accompanied by a disproportionate increase in false alarms.

Taken together, these results show that relaxing the temporal tolerance does help to recover some missed events, but it also inflates the number of false alarms, especially when the spatial radius is already large. To keep this under control, I adopt $\Delta t = 30$ min as the default time window for the main analyses. This choice keeps temporal overlap conservative, while the separate experiments with $R = 20$ km and $R = 100$ km describe the range of outcomes associated with stricter and more tolerant spatial matching.

13.2 SENSITIVITY TO TEMPORAL WINDOW Δt

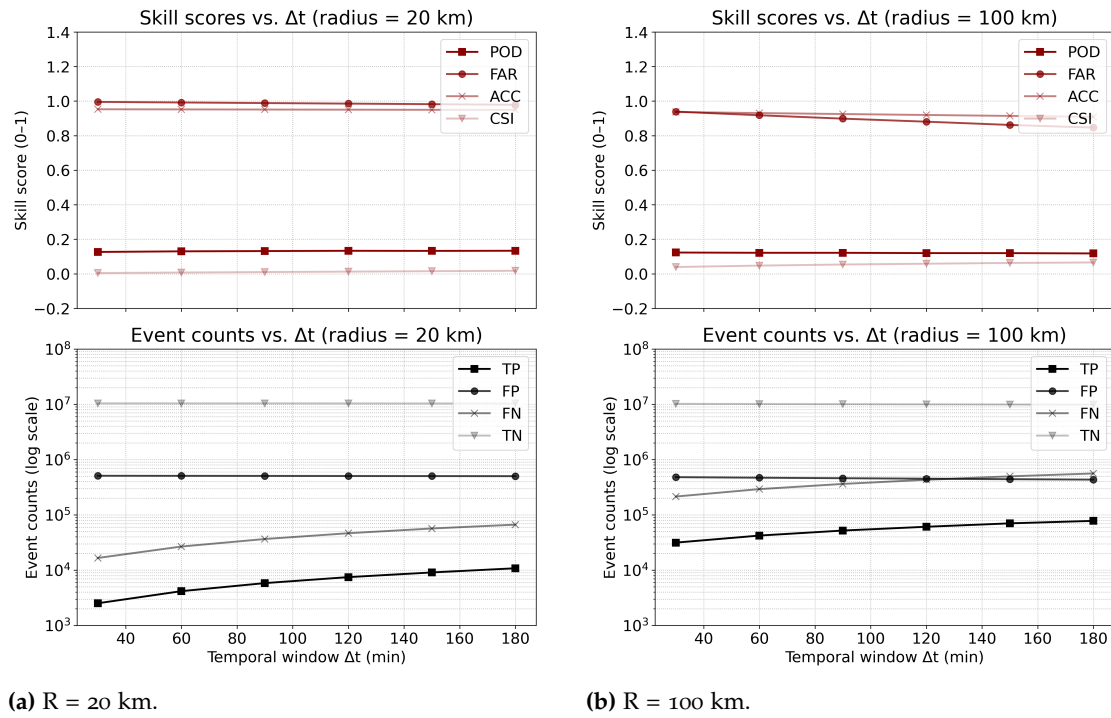


Figure 13.4: Temporal-window sensitivity at fixed radius. (a): $R = 20$ km. (b): $R = 100$ km. Top: domain-mean skill scores (POD, FAR, CSI, ACC) versus Δt . Bottom: total counts (TP, FP, FN, TN, log scale).

Year-by-year summary of hail counts

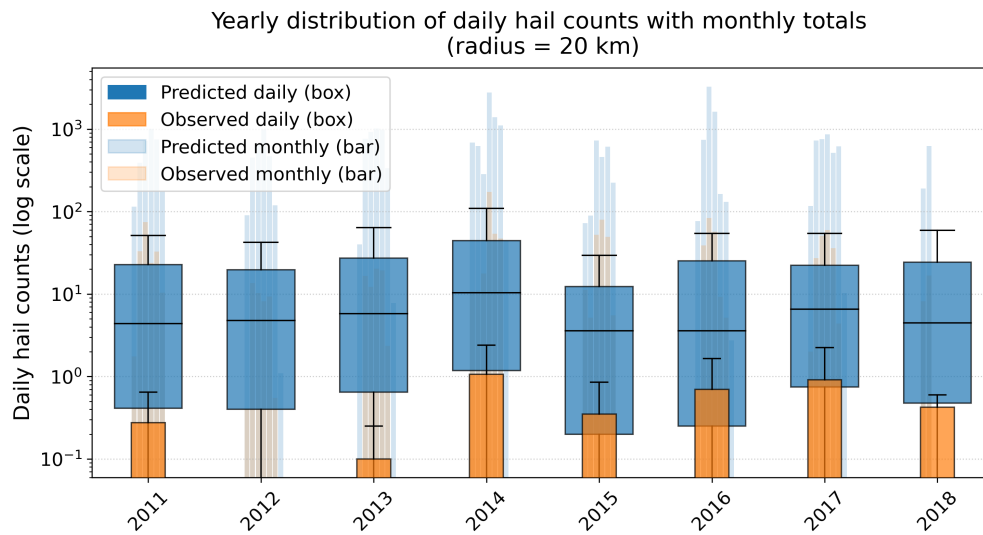
Beyond pointwise scores, it is useful to analyse whether the RF based predictions could reproduce the year-to-year variability over Germany. To do this, I aggregate the COSMO RF predictions and the hail tracks to daily and monthly domain-total hail tracks counts and compare them for each year from 2011 to 2018. Figure 13.5 shows the resulting distributions for two spatial matching radii. The blue boxes represent the distribution of predicted daily hail counts, and the orange boxes show the observed daily counts from the observed track dataset. Light bars in the background indicate the corresponding monthly totals. The upper panel uses a strict matching radius of $R = 20$ km, while the lower panel applies a more tolerant radius of $R = 100$ km. The vertical axis is logarithmic, so differences of one unit correspond to an order of magnitude in daily hail counts.

For $R = 20$ km, the RF tends to produce more hail than is observed. In most years the predicted daily distributions sit clearly above the observed ones, with higher medians and a wider spread toward large daily counts. The monthly totals show a similar

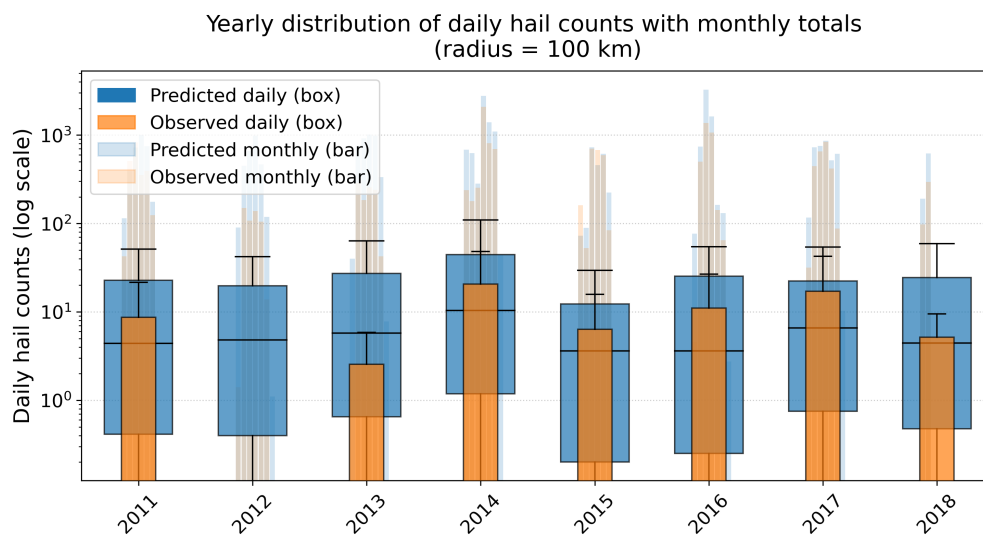
behaviour, with COSMO RF generally predicting more hail days and larger monthly hail counts than indicated by the observed track dataset. Despite this positive bias in amplitude, the relative ranking of years is captured reasonably well. Years that are active in the observations, such as 2013, 2014, 2016, and 2017, also show elevated predicted counts, whereas the years, such as 2011 and 2015, appear as relatively low in both datasets. This suggests that the model and RF respond to the same large-scale variability, even though strict matching at 20 km leaves many events unmatched.

When the radius is increased to $R = 100$ km, the observed hail counts rise substantially, because more storms are counted as “nearby” the tracks. The predicted counts also increase, but the gap between predictions and observations becomes smaller. In several years the predicted and observed daily distributions overlap more strongly, and the monthly totals are closer in magnitude. The inter annual pattern remains similar, with active and quiet years lining up in both datasets, but the overall bias is reduced. This behaviour is consistent with the spatial-radius analysis in Section 12.3: a larger radius compensates partly for displacement errors in space and allows more of the COSMO hail cases to be matched to the observed tracks.

Taken together, the year-by-year aggregates indicate that the COSMO RF system overestimates hail frequency, but it reproduces the main interannual variations in hail activity over Germany. Using a larger matching radius reduces the amplitude bias and supports the view that a substantial part of the discrepancy arises from spatial displacement rather than from a complete failure to represent hail-prone conditions.



(a) R = 20 km.



(b) R = 100 km.

Figure 13.5: Year-by-year summary of hail counts from predictions (blue) and observations (orange) using two spatial aggregation radii ((a): 20 km; (b): 100 km).

LARGE-SCALE CONDITIONS ASSOCIATED WITH DETECTED AND MISSED HAIL TRACKS

The COSMO–RF hail probabilities are inferred only from model precipitation snapshots, whereas hail production is governed more directly by storm-scale vertical motions and microphysical growth processes. These storm-scale processes are, in turn, modulated by the larger-scale thermodynamic and kinematic environment and potentially by aerosol and dust conditions. To test whether the COSMO–RF behaves differently under different environments, I compare ERA5 vertical profiles for hail tracks that are detected by the COSMO–RF with those that are missed. I distinguish between profiles sampled at the track time and profiles sampled one hour earlier to represent a pre-storm environment.

Tracks are classified as detected or missed using the COSMO–RF predictions. For each hourly COSMO snapshot, I define a latitude–longitude bounding box from the footprint of the 9×9 patch and check whether any hail-track point at the corresponding valid hour lies inside a box where the RF indicates hail. If at least one point of a track overlaps a predicted-hail patch in this sense, the track is flagged as detected. Tracks with no such overlap are flagged as missed. This track-level definition differs from the snapshot-based verification in Chapter 13, which evaluates forecast and observed events on the snapshot grid.

COSMO-DE-EPS provides 20 ensemble members. For this environmental analysis, the member-level hail probabilities are first reduced to an ensemble event using the same “at least one member” rule as in the verification, so that a track is treated as detected if any member produces a hail-classified snapshot that overlaps the track point in space and time.

For each hail-track time and location, ERA5 temperature, specific humidity, wind speed, and wind direction profiles are extracted and vertically interpolated to a common height coordinate. Composites are then formed separately for detected and missed tracks, both at the track time and one hour earlier (pre-storm). Figure 14.1 shows median profiles with the interquartile range (25th–75th percentiles) for the four combinations (detected vs. missed, track time vs. pre-storm).

The temperature and specific-humidity profiles (top row) are very similar for detected and missed tracks. Differences are small compared with the within-group spread at both sampling times, suggesting that missed events do not occur in a systematically different large-scale thermodynamic environment than detected ones.

The wind profiles (bottom row) show slightly more structure. Detected tracks tend to be associated with somewhat stronger winds through parts of the mid-troposphere, which is consistent with modestly enhanced deep-layer shear and more organized

convection. Wind-direction profiles are broadly comparable between detected and missed cases. These tendencies are subtle relative to the variability within each group, but they align with the expectation that hail from better-organized storms is more likely to be captured by a precipitation-only classifier.

Comparing pre-storm and track-time profiles is useful because storms can modify the local environment through precipitation evaporation, melting, cold-pool formation, and vertical transport of heat, moisture, and momentum. In these composites, however, the median profiles change only weakly over the one-hour offset for all variables. This indicates that the inferred differences between detected and missed cases are not dominated by storm-induced local modification at the track time.

Overall, the ERA5 composites suggest that detected and missed hail tracks occur in broadly similar large-scale environments. Small tendencies toward stronger mid-level winds for detected tracks hint that shear and storm organization may play a role, but there is no evidence for fundamentally different synoptic regimes. This supports the interpretation that many remaining errors arise from displacement and timing mismatches, storm-scale microphysical uncertainty, or limitations of the snapshot-based hail proxy rather than from systematically different large-scale environmental conditions.

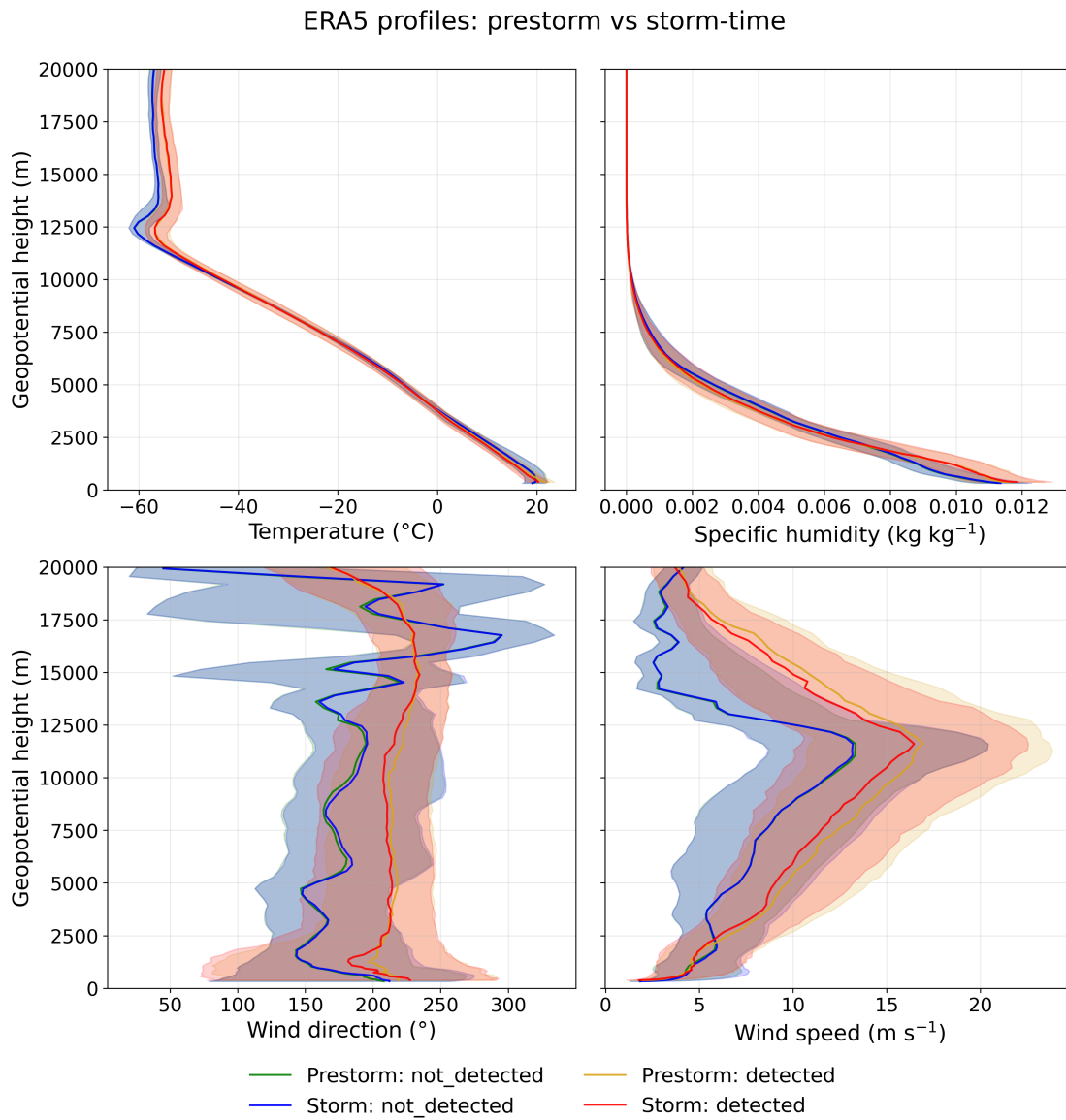


Figure 14.1: ERA5 environmental profiles for hail tracks that are detected (yellow/red) and missed (blue/green) by the COSMO-RF. Top row: temperature T and specific humidity q . Bottom row: wind speed and wind direction. Solid lines show median profiles and shaded bands indicate the interquartile range (25th–75th percentiles). “Pre-storm” refers to profiles sampled one hour before the track time, and “track time” refers to profiles sampled at the hail-track time.

14.1 Aerosol load in predicted and observed hail cases

The previous subsection focused on the large-scale thermodynamic and kinematic environment of detected and missed hail tracks. As a complementary step, I now wonder whether the COSMO–RF hail predictions tend to occur under particular aerosol conditions, in particular elevated mineral dust.

Mineral dust can act as both cloud-condensation nuclei and ice-nucleating particles and can therefore influence mixed-phase microphysics, deep convection, and hail growth (DeMott et al., 2010). At the same time, the CAMS dust analyses used here are subject to uncertainties in emissions, transport, and data assimilation, with documented region- and season-dependent biases over Europe (Marécal et al., 2015; Inness et al., 2019). The results below should therefore be interpreted as statistical relationships rather than a definitive attribution of causality. They also do not imply that dust information is used by the RF itself, which in this work is driven only by COSMO precipitation snapshots.

For each hourly COSMO snapshot I extract the CAMS dust fields at the nearest valid time and sample all model grid cells within the 9×9 snapshot bounding box. Each grid-cell sample is tagged using the COSMO–RF output and the hail-track verification into three categories: `pred_and_obs` (hit), `pred_only` (false alarm), and `obs_only` (miss). I analyse the vertically integrated dust mass in three particle-size modes (0.03–0.55 μm , 0.55–0.9 μm , and 0.9–20 μm). For each mode I show log-count histograms together with normalised histograms (fraction of samples per bin), so that both the frequency of high-dust situations and the shapes of the distributions can be compared fairly across categories.

Figures 14.2–14.4 summarise these distributions. The distributions of vertically integrated dust mass in the three particle-size modes are strongly right-skewed, with most samples at low values and a long tail towards high dust loadings. A consistent pattern nevertheless emerges across modes. Categories that involve a COSMO–RF hail prediction (`pred_only` and, to a lesser extent, `pred_and_obs`) show a heavier upper tail than `obs_only`. In other words, large dust loads occur more frequently when the RF marks a snapshot as hail-producing than when hail is observed without a corresponding RF prediction. This contrast is clearest in the fine and mid-size modes and remains visible, though weaker, in the coarse mode.

One interpretation is that elevated dust co-occurs with synoptic and boundary-layer regimes in which COSMO generates intense and spatially compact precipitation structures. Since the RF is driven only by precipitation intensity and texture, these situations are more likely to be labelled as hail-producing by the classifier. The aerosol histograms then indicate that the subset of the most intense hourly precipitation snapshots, which the RF is most prone to flag as hail-producing, tends to coincide with higher CAMS dust loads.

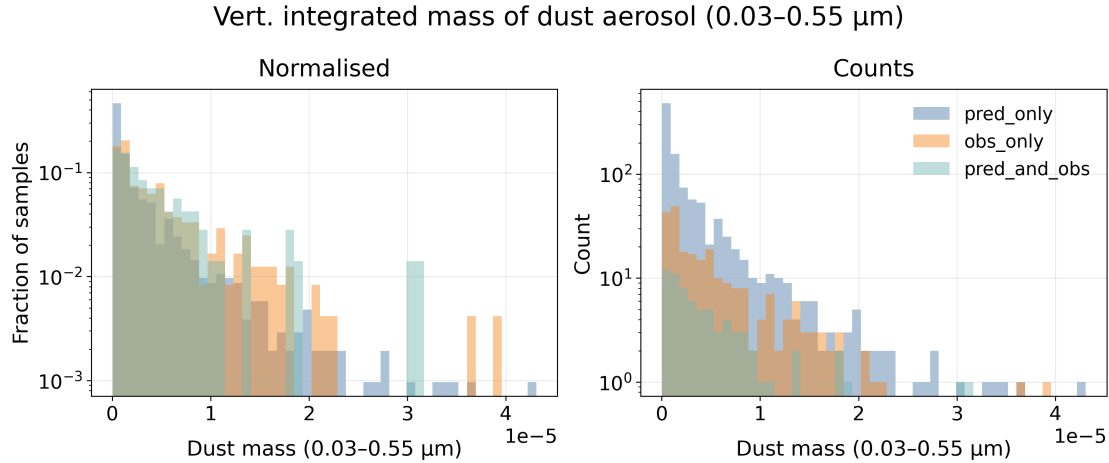


Figure 14.2: Distributions of vertically integrated CAMS mineral dust mass for the 0.03–0.55 μm mode, sampled over all grid cells within the 9×9 COSMO snapshot window. The left panel shows the normalized histogram (fraction of samples per bin), and the right panel shows the corresponding histogram counts on a logarithmic scale. Colours denote *pred_only*, *obs_only*, and *pred_and_obs*.

At the same time, the heavy-dust tail is shared by both hits (*pred_and_obs*) and false alarms (*pred_only*). High dust therefore resembles a background condition that increases the prevalence of hail predictions rather than a factor that cleanly separates correct from incorrect forecasts. Moreover, *obs_only* cases show comparatively few events at very large dust loadings. A plausible hypothesis consistent with this contrast is that high dust may be associated with processes that reduce the probability of observed hail, for example through aerosol impacts on cloud microphysics that could inhibit efficient hail growth. This interpretation remains tentative because CAMS dust analyses carry uncertainties and because the present analysis is purely statistical. In particular, the RF does not ingest dust information, so the relationship identified here should be viewed as a contextual co-variability between dust loading and the likelihood of a COSMO–RF hail prediction rather than as evidence for a causal control of prediction skill.

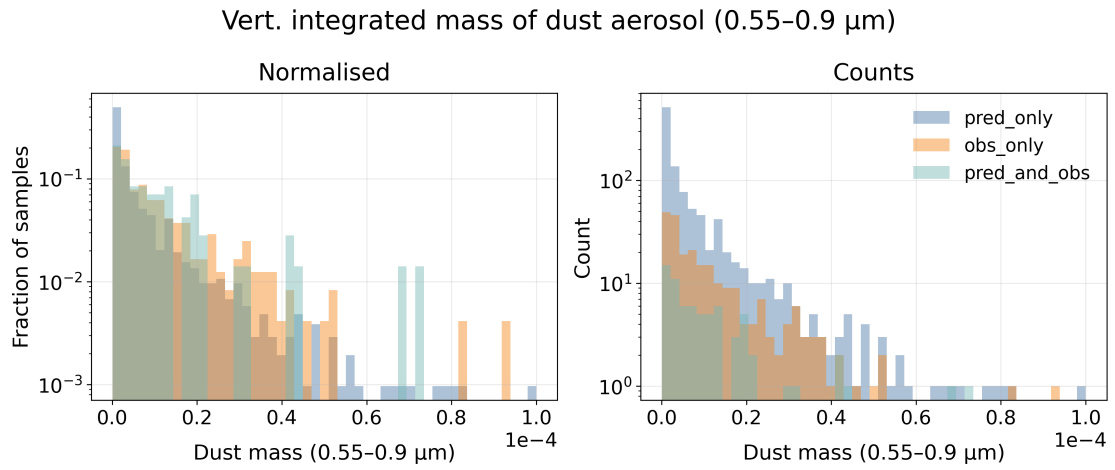


Figure 14.3: Same as Fig. 14.2, but for the 0.55–0.9 μm mode.

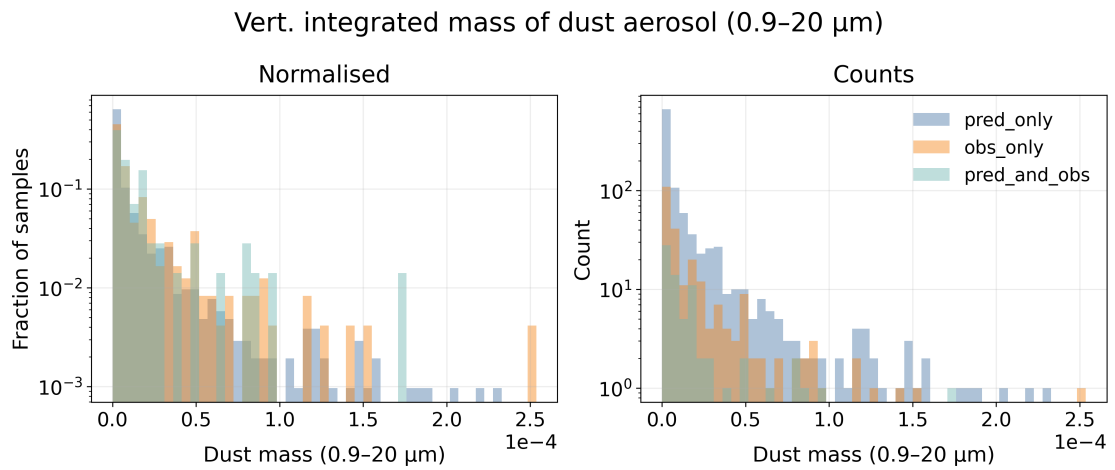


Figure 14.4: Same as Fig. 14.2, but for the 0.9–20 μm mode.

SUMMARY AND CONCLUSIONS

Large hail is a high-impact but relatively rare phenomenon that arises mainly in strongly organised deep convection with active mixed-phase microphysics. Operational numerical weather prediction models such as ICON-D2 (Zängl, Reinert, Rípodas, et al., 2015) and ICON-EU (Deutscher Wetterdienst, 2022) do not provide an explicit forecast of surface hail. In practice, hail is therefore diagnosed indirectly from model output, for example through thresholds applied to proxy variables (such as simulated reflectivity, graupel-related fields, or updraft diagnostics) or through post-processing schemes that combine precipitation and environmental information in composite indices or machine-learning frameworks (Rasmussen and Blanchard, 1998; Gagne et al., 2017; Cintineo et al., 2020). A central question in this context is how strongly hail occurrence is reflected in the structure and intensity of precipitation fields and to what extent this signal can be extracted.

This part of the thesis addressed that question through three related research questions (p. 77). First, I tested whether high-resolution radar precipitation snapshots alone contain enough information to separate hail-producing from non-hail-producing convection. Second, I examined how far a classifier trained on RADKLIM data can be transferred to COSMO-DE-EPS precipitation fields to construct a probabilistic hail product. Third, I assessed the predictability of hail-relevant convective structures in the ensemble and related remaining forecast errors to environmental conditions.

Using 5 min RADKLIM snapshots on a 20×20 km window, I trained a random forest (RF) and a compact convolutional neural network (CNN) to distinguish hail from non-hail events based solely on the radar precipitation pattern. Across multiple runs, the best test accuracies are very similar, with 85.91% for the CNN and 85.71% for the RF. Both models also learned a consistent hail signature in the radar data, with a compact and intense precipitation core in true positives and weaker or more diffuse structures in false positives.

The probability calibration is adequate for probabilistic use. The reliability diagrams for both the CNN and RF remain close to the one-to-one line across most probability bins, with the largest departures occurring in the mid-probability range (approximately $p \sim 0.3-0.5$). At high forecast probabilities, both models again lie close to the diagonal, although these bins contain fewer samples (Fig. 11.3; Section 11). The accompanying prediction histograms also show useful sharpness, with non-hail cases concentrated toward low predicted probabilities and hail cases concentrated toward high predicted probabilities.

These results show that precipitation fields alone can support skilful probabilistic discrimination between hail and non-hail snapshots within the radar domain, even though the models use only precipitation patterns and no explicit environmental predictors such as CAPE, vertical wind shear, or freezing-level height.

In a second step, I adapted this approach to the COSMO-DE-EPS model. RADKLIM was aggregated to hourly accumulations and resampled to fixed 9×9 snapshots on the COSMO grid, and the same RF configuration was trained and evaluated on these hourly radar snapshots. COSMO-DE hourly precipitation was then processed into analogous 9×9 convective snapshots. A central issue that emerged at this stage was the systematic wet bias in COSMO compared to RADKLIM, consistent with the wet-bias behaviour discussed in Part II. I therefore tested several bias-correction strategies and found that quantile mapping of COSMO precipitation to the RADKLIM distribution most effectively reduced the mismatch across both the bulk and the upper tail. All subsequent COSMO-based predictions in this part use quantile-mapped precipitation fields as RF input.

After bias correction, the radar-trained RF was applied to the COSMO snapshots to produce model-based hail probabilities over Germany. Applying the RF to raw COSMO snapshots led to unrealistically frequent hail predictions (20.87% of COSMO snapshots classified as hail at the RF decision threshold), while after quantile mapping the predicted hail fraction decreased to 4.27%, with 95.73% classified as no hail. This confirms that intensity bias correction is a necessary prerequisite for a physically plausible transfer of the radar-trained classifier to COSMO precipitation fields.

Verification against TRACE3D hail tracks showed that strict matching at hourly resolution is too restrictive and yields low rare-event skill. For a reference setup ($\Delta t = 30$ min and $R \approx 45$ to 60 km), CSI remains modest (typically of order 0.1, with spatial values in this study generally below about 0.2) and FAR remains high (roughly 0.7 to 1, depending on region and matching settings). As the spatial radius is increased, more near misses are converted into hits, CSI improves, and FAR decreases at intermediate radii. Maps of true positives show that these hits cluster in the main hail-active regions of Germany, which indicates that the COSMO–RF system captures the broad spatial hail climatology even when individual events are displaced. Sensitivity tests with the temporal matching window show similar behaviour. A moderate increase in the allowed time offset improves detection, whereas very long windows mainly increase permissiveness and reduce the diagnostic value of the verification.

These skill levels should be interpreted in the context of the intentionally restricted predictor design used here. Unlike many hail diagnostics and post-processing approaches that combine multiple predictors such as instability, vertical wind shear, storm proxies, radar signatures, or other environmental information (Rasmussen and Blanchard, 1998; Gagne et al., 2017; Cintineo et al., 2020), the COSMO–RF method in this study uses hourly precipitation snapshots alone. In addition, the verification target is based on point-like hail-track observations, whereas the forecast signal is derived

from precipitation patches and evaluated with neighbourhood tolerances in space and time. A direct score-to-score comparison with other hail-forecasting tools is therefore not straightforward because predictor sets, event definitions, lead times, spatial scales, and verification methods differ. Within this constrained setup, the results nevertheless show that COSMO-DE-EPS precipitation fields contain useful hail-related information, especially for reproducing broad spatial patterns and interannual variability.

On a climatological scale, comparisons of yearly hail counts show that the COSMO-RF system reproduces the interannual variability of hail activity. Years with enhanced hail occurrence in the track data tend to be active in the model-based predictions as well, especially when a moderate spatial tolerance is used. Under tight radii, the system still tends to overestimate hail occurrence, but this bias is reduced when some displacement tolerance is allowed. These findings suggest that the COSMO-RF combination responds to the same large-scale hail-favouring regimes as the observations, even though event-by-event agreement in space and time remains limited.

The last research question for this part focused on whether differences between detected and missed hail events can be related to the larger-scale environment and aerosol conditions. The environmental analysis using ERA5 profiles and the aerosol analysis provide context for this question. Detected and missed tracks occur in broadly similar thermodynamic environments, with only subtle indications of stronger mid-level wind speeds for detected events. Elevated dust loads occur more often in situations that the RF marks as hail-producing than in cases where hail is observed without a corresponding prediction. However, the high-dust tail is shared by both hits and false alarms. Dust therefore appears here as a background condition that co-varies with hail predictions rather than as a factor that clearly separates correct from incorrect forecasts in the present setup.

Overall, the results in this part of the thesis show that a relatively simple precipitation-based RF can extract reproducible information about hail occurrence from both radar and COSMO-DE-EPS precipitation fields once systematic intensity biases are corrected. The method captures key aspects of the spatial and interannual hail climatology and provides a quantitative probabilistic hail product over Germany from model precipitation alone. At the same time, it does not resolve individual hail events with high positional accuracy, and the use of hourly accumulations and precipitation-only predictors limits sharpness at event scale. These limitations point to natural extensions, including additional environmental predictors, higher-frequency model output, and explicit hail-related diagnostics. A useful next step in future work is also a direct benchmark against established operational hail diagnostics and post-processing tools, in order to place the skill of the precipitation-based RF approach in a broader forecasting context.

Part IV

CONCLUSIONS & PERSPECTIVES

CONCLUSIONS

High-impact precipitation remains one of the most difficult forecast problems in operational numerical weather prediction because forecast uncertainty arises from both rapid scale-dependent error growth and persistent model deficiencies. This challenge is particularly pronounced in convection-permitting systems, where forecasts resolve more of the relevant mesoscale structure but remain sensitive to initial conditions, boundary conditions, and process representation. At the same time, long time series of operational ensemble forecasts provide a rich basis for diagnosing how uncertainty and systematic error depend on flow regime, scale, and hazard type.

The overarching goal of this thesis is to explore what information on systematic model errors and initial-condition uncertainty can be extracted from a long time series of operational high-resolution ensemble forecasts. In particular, this question is examined for two contrasting precipitation types: orographic precipitation over complex terrain and hail associated with deep convection over Germany. The aim is to identify what is practically predictable, meaning what can be anticipated reliably once the large-scale environment, forecast uncertainty, and model limitations are taken into account.

Four research questions posed in the introduction (Chapter 1) guide the synthesis of the results.

- **RQ1: To what extent do systematic precipitation errors over complex terrain recur under similar large-scale flow conditions, and can they be described by a small number of robust spatial patterns linked to specific regimes?**

The results in Chapter 6 show that a substantial fraction of wintertime COSMO precipitation error variability projects onto a small number of recurring spatial structures. The leading modes identified by the PCA provide a compact description of systematic model–observation differences and are associated with distinct large-scale environments rather than being purely case-specific noise. This supports the central premise that part of the precipitation error is flow conditioned and therefore predictable in a statistical sense.

The use of ensemble information further strengthens this interpretation. On a substantial subset of winter dates, all ensemble members project most strongly onto the same dominant error mode, which indicates that the pattern is robust to perturbations in the initial conditions. On other dates, members split between modes, which suggests sensitivity to flow differences near regime boundaries. This separation helps distinguish robust flow-conditioned error structures from variability associated with forecast uncertainty in the large-scale setup.

Taken together, the results show that systematic wintertime precipitation error over the Harz is genuinely regime dependent and can be represented in a low-dimensional way. In practical terms, once the large-scale flow and stability are known, the likely spatial structure of model error is constrained, which provides a basis for regime-aware diagnostics and post-processing.

- **RQ2: Which physical mechanisms are associated with the recurring systematic model errors in the orographic precipitation case, and is there evidence for a unique dominant mechanism behind the model failures?**

The combined evidence from environmental composites, targeted ICON case studies, and semi-idealised experiments indicates that the recurring error structures are physically interpretable, but not attributable to a single universal mechanism. The dominant modes correspond to different balances between synoptic forcing, flow blocking, terrain-induced ascent, and the microphysical response. This links the statistical error patterns identified in the PCA to physically distinct terrain-interaction regimes.

The sensitivity experiments further show that changes in microphysical configuration affect the amplitude and secondary details of the precipitation response, while key structural tendencies, including overly ridge-centred enhancement in some regimes, persist. The semi-idealised experiments reproduce important aspects of the systematic bias even without time-varying synoptic variability and without external lateral-boundary uncertainty during the analysed period, which points to structural properties of the model dynamics–physics coupling over complex terrain rather than to a single parameter choice.

The results therefore suggest that systematic model failures in the orographic case emerge from an interaction between flow regime, terrain forcing, and microphysical representation. This is an important conclusion for predictability because it implies that physically meaningful regime descriptors are more useful than a single globally applied correction or a single-mechanism explanation.

- **RQ3: How much information about hail occurrence is contained in precipitation fields alone, and how far can this signal be used to construct probabilistic hail products from convection-permitting model precipitation?**

The results in Chapter 11 show that precipitation fields alone contain a reproducible signal related to hail occurrence that is strong enough to support skilful probabilistic discrimination between hail-producing and non-hail-producing convective snapshots in radar data. Both the RF and the compact CNN learned consistent precipitation signatures associated with hail, and the resulting probability forecasts were reasonably calibrated in the radar-only setting. This demonstrates that a meaningful hazard-relevant signal is embedded in precipitation structure at high temporal and spatial resolution, despite the complexity of hail growth and storm evolution.

Transferring this approach to a convection-permitting ensemble forecast required explicit treatment of systematic differences between modelled and radar-derived precipitation (see Chapter 12). Bias correction of COSMO precipitation, especially by quantile mapping, provided the most effective reduction of the systematic wet bias across the distribution of model precipitation. After this step, precipitation-based RF predictions applied to COSMO yielded a model-based hail probability product that captures major features of the spatial hail climatology and reproduces aspects of interannual variability. This indicates that the post-processed ensemble responds to hail-prone large-scale regimes in a way that is broadly consistent with observations at aggregated scales relevant for practical risk assessment.

At event scale, however, clear predictability limitations remain. Strict pointwise matching at hourly resolution yields low skill because it penalises displacement errors that are intrinsic to convective forecasts and to the intermittent nature of hail. Neighbourhood verification with spatial and temporal tolerances provides a more meaningful assessment and improves performance, but remaining false alarms and misses confirm that precipitation-only predictors and hourly accumulations limit positional accuracy and sharpness.

- **RQ4: Do hail forecast successes and failures cluster under specific large-scale meteorological conditions, and are forecast outcomes linked to different aerosol (mineral dust) loadings?**

The environmental and aerosol analyses show that forecast outcomes do exhibit structure in environmental space, but not in the form of a single dominant discriminator that cleanly separates hits, misses, and false alarms in the current setup (Chapter 14). Instead, the results indicate partial clustering and co-variability with larger-scale conditions, including thermodynamic, kinematic, and aerosol-related differences, while substantial overlap remains between outcome classes.

This has two implications for practical predictability. First, part of the remaining forecast error likely reflects unresolved or weakly represented controls on hail occurrence at the ground, including storm-scale kinematics, melting depth, and microphysical pathway variability. Second, the absence of a single dominant discriminator is itself informative, because it indicates that residual forecast error arises from multiple interacting sources rather than from one missing predictor. In predictability terms, this is consistent with a system in which some hazard-relevant information is recoverable from precipitation patterns and large-scale context, but event-scale outcomes remain strongly affected by short-scale uncertainty and process complexity.

Across the two contrasting precipitation types, the thesis highlights complementary routes to practical predictability in high-resolution ensemble forecasting. In the orographic case, useful predictability emerges through low-dimensional, flow-conditioned structure in systematic model error. In the hail case, useful predictability emerges

through a detectable hazard signal in precipitation patterns that can be translated into calibrated probabilities, even when the hazard is not explicitly forecast by the model. In both settings, ensembles are essential because they quantify flow-dependent uncertainty, reveal robustness or sensitivity of forecast behaviour to perturbations, and enable probabilistic products. Together, the results support a unified perspective in which predictability is not only a property of the atmospheric state, but also of the forecast system, the hazard definition, and the verification and post-processing framework used to extract actionable information.

DISCUSSION AND PERSPECTIVES

The results provide a coherent perspective on precipitation predictability in convection-permitting systems, where uncertainty arises both from intrinsic scale-dependent error growth and from systematic, flow-dependent model behaviour. The wintertime analysis over the Harz indicates that terrain-related precipitation biases can be described by a small set of recurring structures that depend on the large-scale environment and the terrain-interaction regime. This is encouraging for the predictability of systematic error. It does not imply that individual events are always predictable in detail, but it does suggest that the dominant error type is constrained once the flow regime is known. A practical consequence is that model evaluation and post-processing should condition explicitly on regime descriptors rather than aggregating over all situations. Such conditioning can be based on flow orientation and stability measures and complemented by nondimensional metrics, such as the moist Froude number, that summarise blocking likelihood. This regime-aware view helps separate amplitude corrections from structural error modes and reduces the risk that a single global correction masks compensating errors across different regimes.

The convective hazard analysis further illustrates that predictability depends strongly on the spatial and temporal scale at which skill is assessed. At the snapshot level, precipitation patterns contain discriminative information that can be mapped to probabilistic hail occurrence, which supports the idea that precipitation fields encode a transferable hazard signal even without explicitly forecasting hail. At the event scale, however, displacement uncertainty dominates verification, making neighbourhood approaches essential for interpreting forecast quality. At climatological scales, the post-processed ensemble guidance captures spatial hail corridors and year-to-year activity variations more robustly than it captures the exact location and timing of individual events. This behaviour is consistent with the multi-scale nature of convective predictability, where the large-scale environment is often more predictable than the evolution of cell-scale features. Practically, the results suggest that precipitation-only hail products are most useful as calibrated risk guidance that is interpreted with spatial tolerance, rather than as deterministic event forecasts.

Several limitations frame the interpretation and also clarify where the remaining errors likely originate. For the terrain-focused analysis, separating flow-related structural bias from observational uncertainty remains challenging, particularly in complex terrain where radar-based precipitation products can have regime-dependent errors. For the hail application, precipitation alone cannot represent important controls on hail survival to the surface, such as melting depth and variability in microphysical pathways,

and the use of hourly accumulations reduces information about storm evolution and peak intensity. These constraints likely explain part of the remaining false alarms and misses even after bias correction and neighbourhood verification.

Two extensions follow naturally from the predictability framing developed in Chapter 2. First, regime-aware precipitation bias correction over mountains could be developed by conditioning corrections on flow and stability metrics, rather than applying a single global mapping across all situations. Second, probabilistic hail products could be improved by adding predictors that represent the thermodynamic and kinematic environment and hail-relevant process proxies, while retaining the precipitation-pattern information that was shown here to be valuable. From a methodological perspective, higher-frequency model output and verification designs aligned with expected spatial and temporal uncertainty would likely yield sharper, more actionable probabilistic guidance.

Overall, the thesis indicates that useful predictability of hazardous precipitation can be extracted in two complementary ways. One is by identifying low-dimensional, flow-conditioned structure in systematic model error over complex terrain. The other is by translating precipitation-pattern information into calibrated probabilities for a hazard that is not explicitly forecast by the model. Both perspectives emphasise that forecast value depends on matching evaluation and post-processing strategies to the dominant uncertainty sources at the relevant scale.

Part V

APPENDIX

ICON SENSITIVITY EXPERIMENTS

This appendix provides additional ICON sensitivity experiments for four representative PCA cases. Each figure compares simulated 21-h accumulated precipitation fields from the 1-moment configuration and selected 2-moment configurations with varying CCN_type and ice-phase settings.

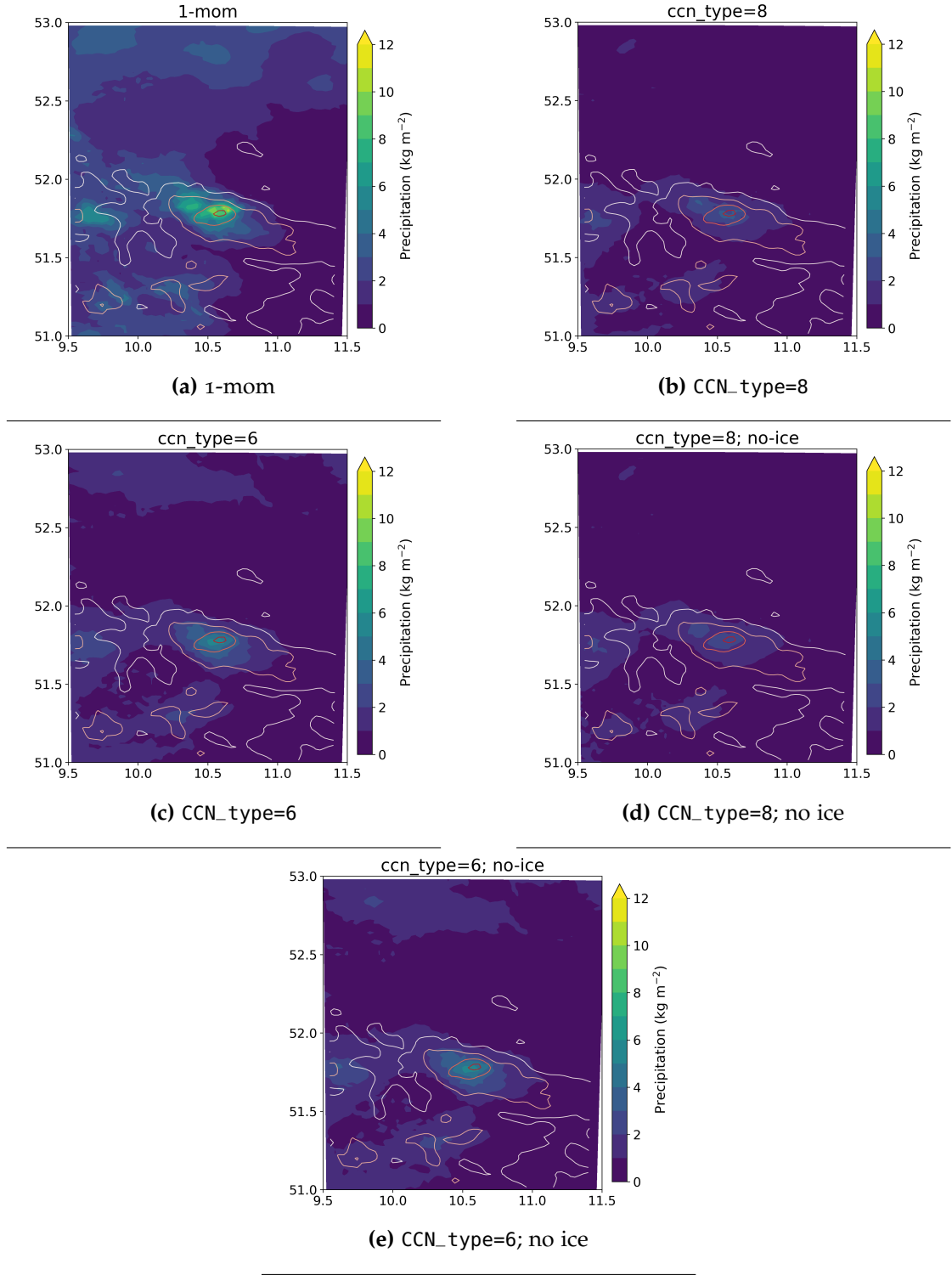


Figure A.1: ICON sensitivity experiments for the PC1 case (13.12.2011 06:00 UTC). Simulated 21-h accumulated precipitation for (a) 1-moment, (b) 2-moment CCN8, (c) 2-moment CCN6, (d) 2-moment CCN8 no ice, and (e) 2-moment CCN6 no ice.

ICON SENSITIVITY EXPERIMENTS

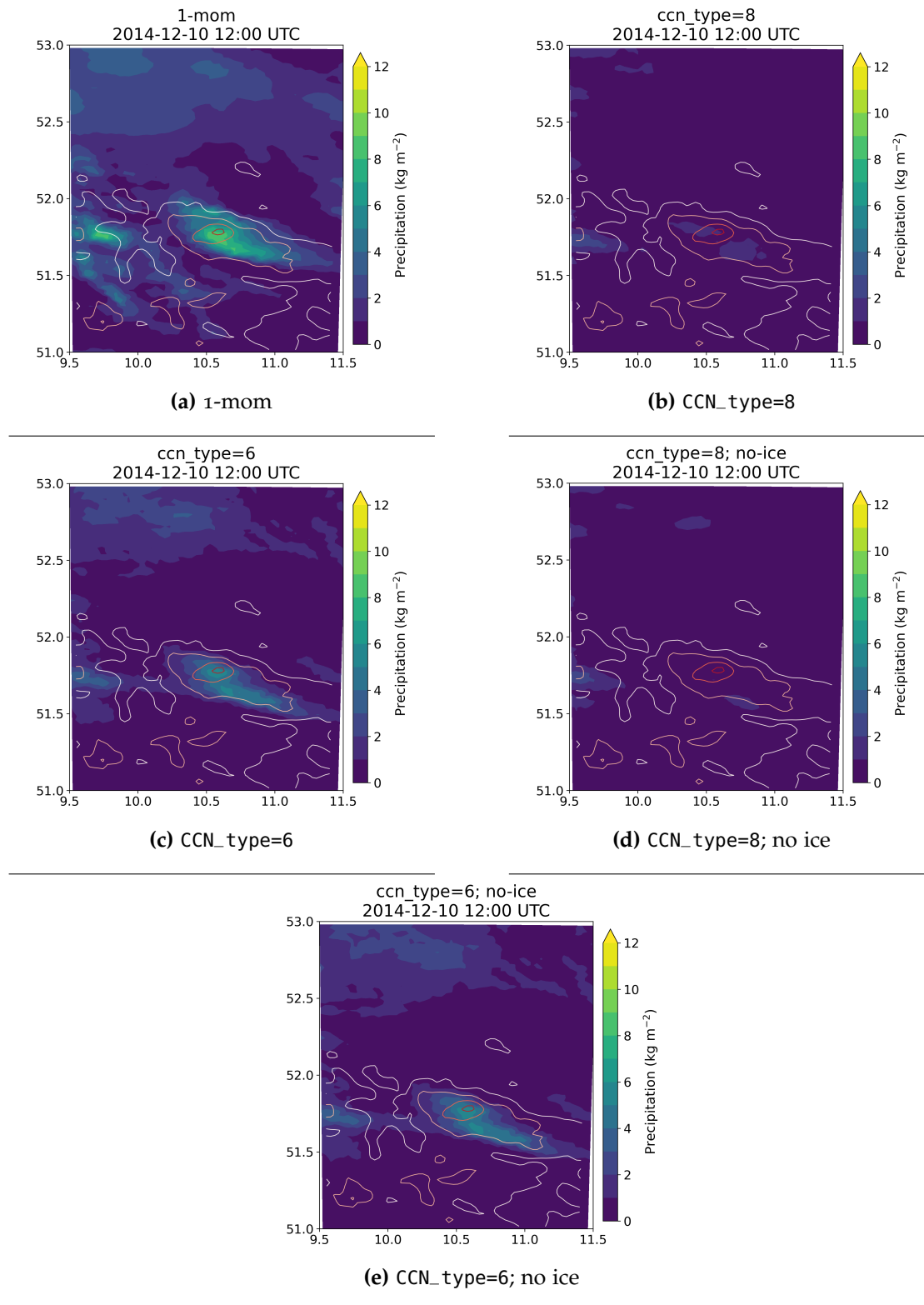


Figure A.2: ICON sensitivity experiments for the PC1 case (10.12.2014 12:00 UTC). Simulated 21-h accumulated precipitation for (a) 1-moment, (b) 2-moment CCN8, (c) 2-moment CCN6, (d) 2-moment CCN8 no ice, and (e) 2-moment CCN6 no ice.

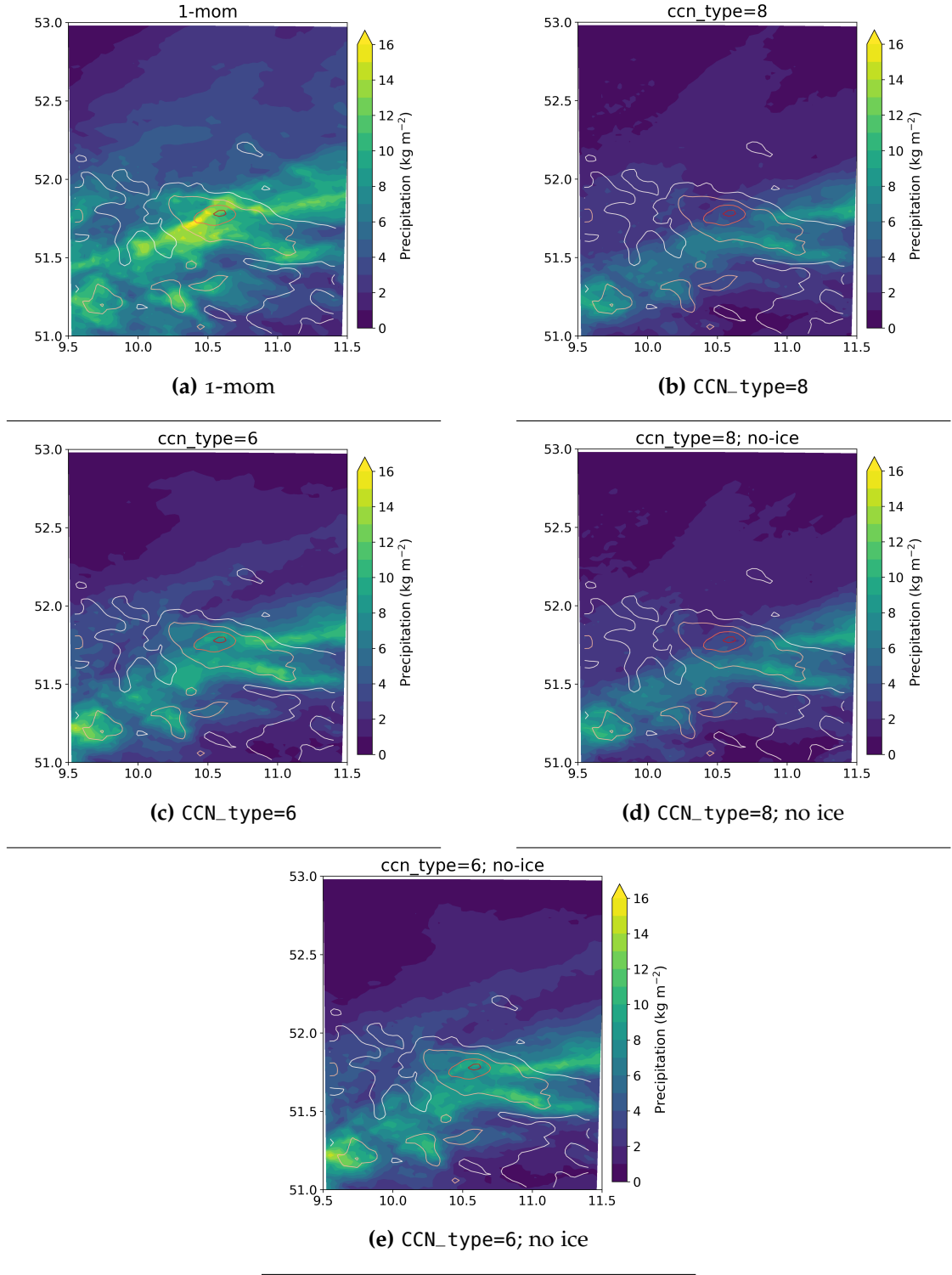


Figure A.3: ICON sensitivity experiments for the PC2 case (28.12.2013 21:00 UTC). Simulated 21-h accumulated precipitation for (a) 1-moment, (b) 2-moment CCN8, (c) 2-moment CCN6, (d) 2-moment CCN8 no ice, and (e) 2-moment CCN6 no ice.

ICON SENSITIVITY EXPERIMENTS

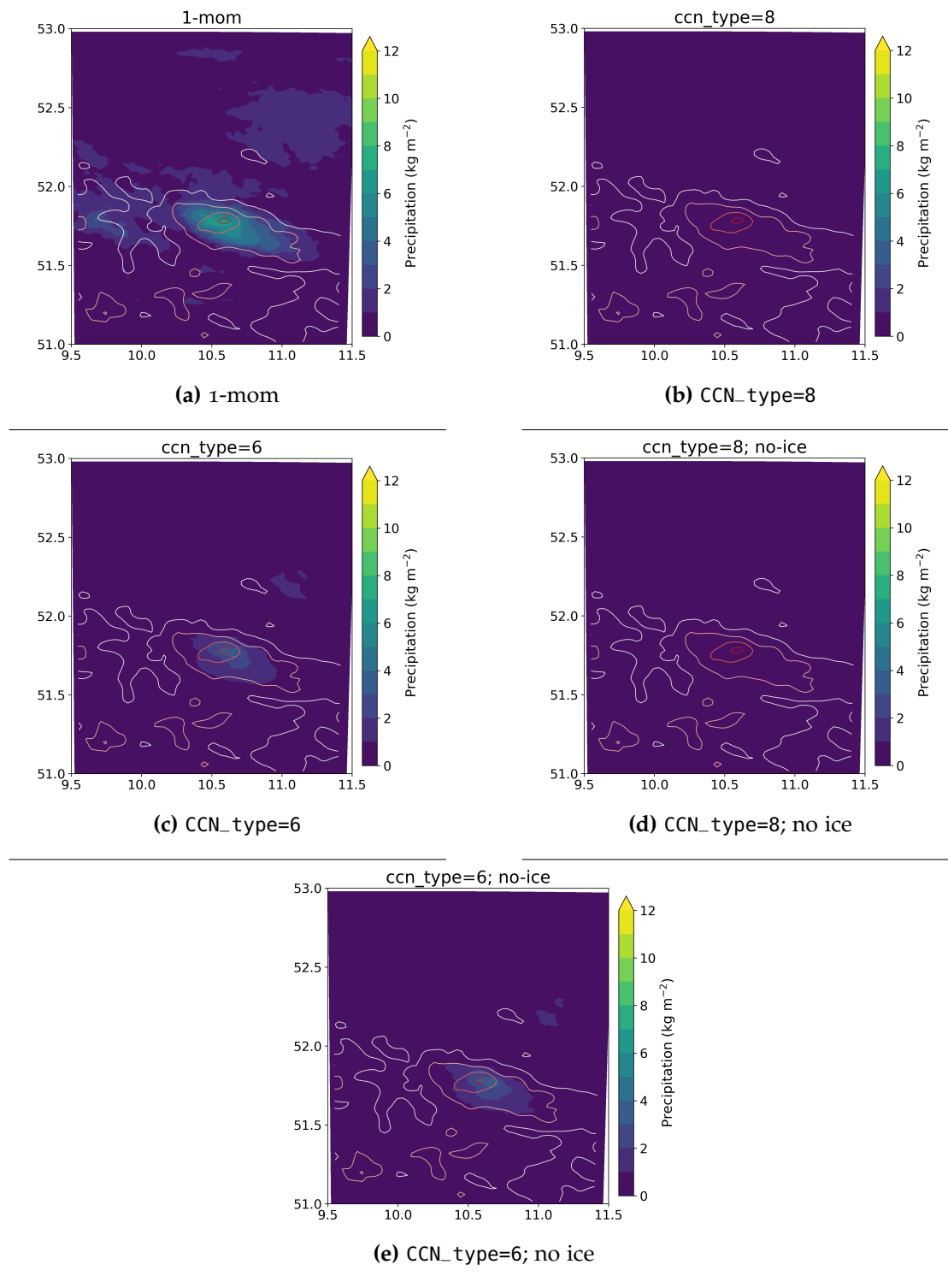


Figure A.4: ICON sensitivity experiments for the PC₃ case (04.01.2012 21:00 UTC). Simulated 21-h accumulated precipitation for (a) 1-moment, (b) 2-moment CCN8, (c) 2-moment CCN6, (d) 2-moment CCN8 no ice, and (e) 2-moment CCN6 no ice.

BIBLIOGRAPHY

- Baldauf, M. et al. (2011). **Operational Convective-Scale Numerical Weather Prediction with the COSMO Model: Description and Sensitivities**. *Monthly Weather Review* **139**, 3887–3905.
- Barrett, A. I. and C. Hoose (2023). **Microphysical Pathways Active Within Thunderstorms and Their Sensitivity to CCN Concentration and Wind Shear**. *Journal of Geophysical Research: Atmospheres* **128**, e2022JD036965.
- Bauer, P., A. Thorpe, and G. Brunet (2015). **The Quiet Revolution of Numerical Weather Prediction**. *Nature* **525**, 47–55.
- Bi, K. et al. (2023). **Accurate medium-range global weather forecasting with 3D neural networks**. *Nature* **619**, 533–538.
- Breiman, L. (2001). **Random Forests**. *Machine Learning* **45**, 5–32.
- Bretherton, C. S., C. Smith, and J. M. Wallace (1992). **An intercomparison of methods for finding coupled patterns in climate data**. *Journal of Climate* **5**, 541–560.
- Buizza, R. (2018). **Ensemble Forecasting and the Need for Calibration**. *Statistical Post-processing of Ensemble Forecasts*. Ed. by S. Vannitsem, D. S. Wilks, and J. W. Messner. Elsevier, 15–48.
- Cintineo, J. et al. (2020). **NOAA ProbSevere v2.0—ProbHail, ProbWind, and ProbTor**. *Weather and Forecasting* **35**, 1523–1543.
- Clark, P. et al. (2016). **Convection-Permitting Models: A Step-Change in Rainfall Forecasting**. *Meteorological Applications* **23**, 165–181.
- Copernicus Atmosphere Data Store (2020). *CAMS Global Reanalysis (EAC4): Dataset Overview*. <https://ads.atmosphere.copernicus.eu/datasets/cams-global-reanalysis-eac4?tab=overview>. Global grid 0.75°, 3-hourly resolution.
- Copernicus Climate Change Service (2023). *ERA5: Fifth Generation of ECMWF Atmospheric Reanalyses of the Global Climate*. <https://cds.climate.copernicus.eu/>. Accessed 2025-11-12.
- Dawson, A. (2016). **eofs: A Library for EOF Analysis of Meteorological, Oceanographic, and Climate Data**. *Journal of Open Research Software* **4**.
- DeMott, P. J. et al. (2010). **Predicting global atmospheric ice nuclei distributions and their impacts on climate**. *Proceedings of the National Academy of Sciences* **107**, 11217–11222.
- Deser, C. et al. (2010). **Sea surface temperature variability: Patterns and mechanisms**. *Annual Review of Marine Science* **2**, 115–143.
- Deutscher Wetterdienst (2021). *Replacement of COSMO-D2 / COSMO-D2-EPS with ICON-D2 / ICON-D2-EPS*. Operational NWP system change notice. Accessed 2025-11-14. Deutscher Wetterdienst.

- Deutscher Wetterdienst (2022). *ICON Database Reference Manual*. Tech. rep. Accessed 2026-03-03. Offenbach, Germany: Deutscher Wetterdienst.
- (2024a). *SINFONY – Seamless INtegrated FOrecastING sYstem*. https://www.dwd.de/DE/forschung/forschungsprogramme/sinfony_iafe/sinfony_start.html. Accessed 2026-02-18.
 - (2024b). *SINFONY Pilot Project: Operational precipitation nowcasting using optical flow*. https://www.dwd.de/DE/forschung/forschungsprogramme/sinfony_iafe/sinfony_pilotprojekt/sinfony_pilotprojekt_1_02.html. Accessed 2026-02-18.
 - (2025). *Klimatafel Brocken/Harz (Klimanormalperiode 1961–1990)*. https://www.dwd.de/DWD/klima/beratung/ak/ak_104530_kt.pdf. Accessed 2025-11-17.
 - (2026a). *Radar Products: Weather Radar Network of the DWD*. https://www.dwd.de/EN/ourservices/radar_products/radar_products.html. Accessed 2026-02-04.
 - (2026b). *Regional Model COSMO-DE*. https://www.dwd.de/EN/research/weatherforecasting/num_modelling/01_num_weather_prediction_modells/regional_model_cosmo_de.html. Accessed 2026-02-05.
- Doms, G. et al. (2011). *A Description of the Nonhydrostatic Regional COSMO Model. Part II: Physical Parameterization*. LM F90, Version 4.20, September 2011. Consortium for Small-scale Modelling (COSMO). Offenbach, Germany.
- Fabig, I. (2007). Wandel der Niederschlagsverhältnisse im Lee des Harzes – Indikatoren eines regionalen Klimawandels? *Hercynia, Neue Folge* **40**. Open access PDF available, 33–39.
- Fluck, E. et al. (2021). *Radar-based assessment of hail frequency in Europe*. *Natural Hazards and Earth System Sciences* **21**, 683–701.
- Forcadell, V. et al. (2024). *Severe-Hail Detection with C-Band Dual-Polarisation Radars Using Convolutional Neural Networks*. *Atmospheric Measurement Techniques* **17**, 6707–6734.
- Gagne, D. J. et al. (2017). *Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles*. *Weather and Forecasting* **32**, 1819–1840.
- Galewsky, J. and A. H. Sobel (2005). *Moist Dynamics and Orographic Precipitation in Northern and Central California during the New Year’s Flood of 1997*. *Monthly Weather Review* **133**, 1594–1612.
- Gebhardt, C. et al. (2011). *Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries*. *Atmospheric Research* **100**, 168–177.
- Gudmundsson, L. et al. (2012). *Downscaling RCM precipitation to the station scale using statistical transformations – a comparison of methods*. *Hydrology and Earth System Sciences* **16**, 3383–3390.
- Handwerker, J. (2002). *Cell tracking with TRACE3D – A new algorithm*. *Atmospheric Research* **61**, 15–34.

- Hannachi, A., I. T. Jolliffe, and D. B. Stephenson (2007). **Empirical orthogonal functions and related techniques in atmospheric science: A review**. *International Journal of Climatology* **27**, 1119–1152.
- Hassell, D. et al. (2017). **A data model of the Climate and Forecast metadata conventions (CF-1.6) with a software implementation (cf-python v2.1)**. *Geoscientific Model Development* **10**, 4619–4646.
- Hersbach, H. et al. (2020). **The ERA5 Global Reanalysis**. *Quarterly Journal of the Royal Meteorological Society* **146**, 1999–2049.
- Hurrell, J. W. (1995). **Decadal trends in the North Atlantic Oscillation: Regional temperatures and precipitation**. *Science* **269**, 676–679.
- Inness, A. et al. (2019). **The CAMS reanalysis of atmospheric composition**. *Atmospheric Chemistry and Physics* **19**, 3515–3556.
- Junghänel, T. et al. (2016). **Towards a radar- and observation-based hail climatology for Germany**. *Meteorologische Zeitschrift* **25**, 435–445.
- Kahraman, A. et al. (2025). **Future changes in severe hail across Europe, including regional emergence of warm-type thunderstorms**. *Nature Communications* **16**, 8438.
- Kalnay, E. (2003). **Atmospheric Modeling, Data Assimilation and Predictability**. Cambridge, UK: Cambridge University Press.
- Kopp, J. et al. (2024). **Verification of Weather-Radar-Based Hail Metrics with Crowd-sourced Observations from Switzerland**. *Atmospheric Measurement Techniques* **17**, 4529–4552.
- Korolev, A. et al. (2017). **Mixed-Phase Clouds: Progress and Challenges**. *Meteorological Monographs*. Vol. 58. American Meteorological Society, 5.1–5.50.
- Kunz, M. and M. Puskeiler (2010). **High-resolution assessment of the hail hazard over complex terrain from radar and insurance data**. *Meteorologische Zeitschrift* **19**, 427–439.
- Kunz, M., J. Sander, and C. Kottmeier (2009). **Recent trends of thunderstorm and hailstorm frequency and their relation to atmospheric characteristics in southwest Germany**. *International Journal of Climatology* **29**, 2283–2297.
- LeCun, Y. et al. (1998). **Gradient-Based Learning Applied to Document Recognition**. *Proceedings of the IEEE* **86**, 2278–2324.
- Lengfeld, K. et al. (2019). **Characteristic spatial extent of hourly and daily precipitation events in Germany derived from 16 years of radar data**. *Meteorologische Zeitschrift* **28**, 363–378.
- Leutbecher, M. and T. N. Palmer (2008). **Ensemble Forecasting**. *Journal of Computational Physics* **227**, 3515–3539.
- Lorenz, E. N. (1963). **Deterministic Nonperiodic Flow**. *Journal of the Atmospheric Sciences* **20**, 130–141.
- (1969). **The predictability of a flow which possesses many scales of motion**. *Tellus* **21**, 289–307.

- Lukach, M. et al. (2017). **Estimating the Occurrence and Severity of Hail Based on 10 Years of Observations from Weather Radar in Belgium.** *Meteorological Applications* **24**, 250–259.
- Marécal, V. et al. (2015). **A regional air quality forecasting system over Europe: the MACC-II daily ensemble production.** *Geoscientific Model Development* **8**, 2777–2813.
- Markowski, P. and Y. Richardson (2010). **Mesoscale Meteorology in Midlatitudes.** English (US). John Wiley and Sons.
- Mellor, G. L. and T. Yamada (1982). **Development of a Turbulence Closure Model for Geophysical Fluid Problems.** *Reviews of Geophysics* **20**, 851–875.
- Milbrandt, J. A. and M. K. Yau (2005). **A Multimoment Bulk Microphysics Parameterization. Part I: Analysis of the Scheme and Comparison with Observations.** *Journal of the Atmospheric Sciences* **62**, 3056–3081.
- Mohr, S. and M. Kunz (2013). **Recent trends and variabilities of convective parameters relevant for hail events in Germany and Europe.** *Atmospheric Research* **123**, 211–228.
- Mohr, S., M. Kunz, and M. Schmidberger (2024). **An Updated 3D Radar-Based Hail Statistic for Germany (2005–2023).** Poster presentation.
- Morales, A., H. Morrison, and D. J. Posselt (2018). **Orographic precipitation response to microphysical parameter perturbations for idealized moist nearly neutral flow.** *Journal of the Atmospheric Sciences* **75**, 1933–1953.
- Morrison, H. et al. (2020). **Confronting the Challenge of Modeling Cloud and Precipitation Microphysics: The Promise of Machine Learning.** *Journal of Advances in Modeling Earth Systems* **12**, e2019MS001689.
- Palmer, T. N. (2000). **Predicting Uncertainty in Forecasts of Weather and Climate.** *Reports on Progress in Physics* **63**, 71–116.
- Pathak, J. et al. (2022). **FourCastNet: A Global Data-Driven High-Resolution Weather Model Using Adaptive Fourier Neural Operators.** *arXiv preprint arXiv:2202.11214*. arXiv: 2202.11214 [physics.ao-ph].
- Pedregosa, F. et al. (2011). **Scikit-learn: Machine Learning in Python.** *Journal of Machine Learning Research* **12**, 2825–2830.
- Pham, T. V. et al. (2021). **ICON in Climate Limited-area Mode (ICON release version 2.6.1): a new regional climate model.** *Geoscientific Model Development* **14**, 985–1005.
- Photiadou, C. et al. (2016). **Incorporating circulation statistics in bias correction of GCM ensembles: hydrological application for the Rhine basin.** *Climate Dynamics* **46**, 187–203.
- Prein, A. F. et al. (2015). **A Review on Regional Convection-Permitting Climate Modeling: Demonstrations, Prospects, and Challenges.** *Reviews of Geophysics* **53**, 323–361.
- Punge, H. J. and M. Kunz (2016). **Hail observations and hailstorm characteristics in Europe: A review.** *Atmospheric Research* **176**, 159–184.
- Puskeiler, M., M. Kunz, and M. Schmidberger (2016). **Hail statistics for Germany derived from single-polarization radar data.** *Atmospheric Research* **178–179**, 459–470.

- Raible, C. C. et al. (2020). Extratropical cyclones in the present and future climate. *WIREs Climate Change* **11**, e620.
- Rasch, P. J. and J. E. Kristjánsson (1998). A Comparison of the CCM₃ Model Climate Using Diagnosed and Predicted Condensate Parameterizations. *Journal of Climate* **11**, 1587–1614.
- Rasmussen, E. N. and D. O. Blanchard (1998). A Baseline Climatology of Sounding-Derived Supercell and Tornado Forecast Parameters. *Weather and Forecasting* **13**, 1148–1164.
- Rauthe, M. et al. (2013). A Central European precipitation climatology – Part I: Generation and validation of a high resolution gridded daily data set (HYRAS). *Meteorologische Zeitschrift* **22**, 235–256.
- Roe, G. H. (2005). Orographic precipitation. *Annual Review of Earth and Planetary Sciences* **33**, 645–671.
- Rombeek, N., J. Leinonen, and U. Hamann (2024). Exploiting Radar Polarimetry for Nowcasting Thunderstorm Hazards Using Deep Learning. *Natural Hazards and Earth System Sciences* **24**, 133–144.
- Rotunno, R. and R. A. Houze (2007). Lessons on orographic precipitation from the Mesoscale Alpine Programme. *Quarterly Journal of the Royal Meteorological Society* **133**, 811–830.
- Schneider, L. and C. Barthlott (2018). The precipitation response to variable terrain forcing over low-mountain ranges in different weather regimes. *Quarterly Journal of the Royal Meteorological Society* **144**, 970–989.
- Schulz, B. and S. Lerch (2022). Machine Learning Methods for Postprocessing Ensemble Forecasts of Wind Gusts: A Systematic Comparison. *Monthly Weather Review* **150**, 235–257.
- Seity, Y. et al. (2012). Cloud and microphysical schemes in ARPEGE and AROME models. Tech. rep. ECMWF Workshop on Parametrization of Clouds and Precipitation, 5–8 November 2012. ECMWF.
- Seneviratne, S. I. et al. (2021). Weather and Climate Extreme Events in a Changing Climate. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by V. Masson-Delmotte et al. Cambridge, UK and New York, NY, USA: Cambridge University Press. Chap. 11, 1513–1765.
- Singh, S. and N. Kalthoff (2022). Process Studies of the Impact of Land-Surface Resolution on Convective Precipitation Based on High-Resolution ICON Simulations. *Meteorology* **1**, 288–310.
- Slingo, J. and T. N. Palmer (2011). Uncertainty in weather and climate prediction. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **369**, 4751–4767.
- Smith, R. B. (1979). The influence of mountains on the atmosphere. *Advances in Geophysics* **21**, 87–230.

- Steiner, M., R. A. J. Houze, and S. E. Yuter (1995). Climatological Characterization of Three-Dimensional Storm Structure from Operational Radar and Rain Gauge Data. *Journal of Applied Meteorology* **34**, 1978–2007.
- Stephens, G. L. et al. (2010). Dreary State of Precipitation in Global Models. *Journal of Geophysical Research: Atmospheres* **115**, D24211.
- Theis, S., C. Gebhardt, and Z. Ben Bouallègue (2013). COSMO-DE-EPS: A new way predicting severe convection. https://presentations.copernicus.org/EMS2013/EMS2013-544_presentation.pdf. Accessed 2026-02-05.
- (2015). Beschreibung des COSMO-DE-EPS und seiner Ausgabe in die Datenbanken des DWD. https://www.dwd.de/SharedDocs/downloads/DE/modelldokumentationen/nwv/cosmo_de_eps/cosmo_de_eps_dbbeschr_20150922.pdf?__blob=publicationFile&v=4. Accessed 2026-02-05.
- Themeßl, M. J., A. Gobiet, and A. Leuprecht (2011). Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *International Journal of Climatology* **31**, 1530–1544.
- Thomas, J., A. I. Barrett, and C. Hoose (2023). Temperature and cloud condensation nuclei (CCN) sensitivity of orographic precipitation enhanced by a mixed-phase seeder–feeder mechanism: A case study for the 2015 Cumbria flood. *Atmospheric Chemistry and Physics* **23**, 1987–2002.
- Tiedtke, M. (1989). A Comprehensive Mass Flux Scheme for Cumulus Parameterization in Large-Scale Models. *Monthly Weather Review* **117**, 1779–1800.
- Trigo, R. M. et al. (2004). Winter blocking episodes in the European–Atlantic sector: climate impacts and associated physical mechanisms in the reanalysis. *Climate Dynamics* **23**, 17–28.
- Ulbrich, U., G. C. Leckebusch, and J. G. Pinto (2009). Extra-tropical cyclones in the present and future climate: A review. *Theoretical and Applied Climatology* **96**, 117–131.
- Von Storch, H. and F. W. Zwiers (2002). Statistical analysis in climate research. Cambridge university press.
- Wallace, J. M. and D. S. Gutzler (1981). Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Monthly Weather Review* **109**, 784–812.
- Weigl, E. (2008). Radarniederschlag: Prinzip der Niederschlagsbestimmung mit Radar inkl. Umrechnung der Radarreflektivitäten in Momentanwerte des Niederschlages. https://www.dwd.de/DE/fachnutzer/wasserwirtschaft/unsere_leistungen/radarniederschlagsprodukte/niederschlagsbestimmung_pdf.pdf?__blob=publicationFile&v=2. Version 1.0, Stand 12.03.2008. Accessed 2026-02-04.
- Wilks, D. S. (2011). Statistical methods in the atmospheric sciences. Elsevier Academic Press.
- Winterrath, T. et al. (2018). An Overview of the New Radar-Based Precipitation Climatology of the Deutscher Wetterdienst: Data, Methods, Products. Contribution to the 11th International Workshop on Precipitation in Urban Areas, Pontresina, Switzerland.

BIBLIOGRAPHY

- Witt, A. et al. (1998). *An Enhanced Hail Detection Algorithm for the WSR-88D*. *Weather and Forecasting* **13**, 286–303.
- Wulfmeyer, V. et al. (2008). *The Convective and Orographically-induced Precipitation Study: A Research and Development Project of the World Weather Research Program for improving quantitative precipitation forecasting in low-mountain regions*. *Bulletin of the American Meteorological Society* **89**, 1477–1486.
- Zängl, G., D. Reinert, and F. Prill (2022). *Grid Refinement in ICON v2.6.4*. *Geoscientific Model Development* **15**, 7153–7176.
- Zängl, G., D. Reinert, P. Rípodas, et al. (2015). *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.
- Zhang, F., C. Snyder, and R. Rotunno (2003). *Effects of Moist Convection on Mesoscale Predictability*. *Journal of the Atmospheric Sciences* **60**, 1173–1185.



Curriculum Vitae

Personal Information

Name: Yuliya Kazachkova
Address: Hintere Bleiche 18, 55116 Mainz
Nationality: Russian
Date of birth: 21 January 1997
Place of birth: Murmansk, Russia

Educational Background

since Feb 2022 **Doctoral studies in Atmospheric Physics**
Johannes Gutenberg University Mainz, Germany
Dissertation: *Predictability of Precipitation-Related Hazards: Insights from Ensemble Archives with Advanced Statistical Methods*

Sep 2019 – Oct 2021 **M.Sc. in Atmospheric Science**
University of Hamburg, Germany
Master thesis: *Impact of urban green distribution on urban temperature*

Sep 2015 – Jun 2019 **B.Sc. in Climatology and Meteorology**
St. Petersburg State University, Russia
Bachelor thesis: *Meteorological support of sustainable operation of wind turbines in Russia*

Academic Experience

Feb 2022 – present **Doctoral Researcher**
Johannes Gutenberg University Mainz, Germany
Research in precipitation predictability, radar-based hail detection, ensemble forecasting, and statistical and machine-learning methods for atmospheric data analysis.

Feb 2023 – Nov 2025 **Tutor**
Johannes Gutenberg University Mainz, Germany
Teaching support and supervision of students in the field of cloud microphysics and aerosols.

Jun 2018 – Jul 2018 **Research Intern**
Nansen Center
Processing of ERA-Interim reanalysis data and analysis of the surface radiation budget over the Irminger, Labrador, and Greenland Seas.

Jul 2018 **Forecasting Intern**
Federal Hydrometeorological Service
Preparation of synoptic charts and support of operational short-range weather forecasting under supervision.

Jun 2016 **Field Training in Microclimatology**
Microclimatic observations including air temperature, soil temperature, wind, and actinometric measurements.

Conferences and Training

Jul 2024	International Conference on Clouds and Precipitation (ICCP) Participant
Jun 2024	CISM Advanced School: Microphysics of Atmospheric Clouds Participant
Apr 2024	EGU General Assembly Participant
Jun 2023	SCALES Conference Participant
Mar 2022	Introduction to Cluster Computing (HPC-Intro) Training course
Mar 2022	Introduction to Shell Scripting (Bash) Training course

Languages

Russian	Native speaker
English	Fluent
German	B1

Technical Skills

Programming analysis	and	Python, MATLAB, Fortran, Bash, Linux, Git
Scientific tools		NumPy, SciPy, pandas, xarray, netCDF4, Matplotlib, PyTorch, scikit-learn
Computing environment	environ-	HPC, Slurm

Reference

Jun.-Prof. Dr. Annette Miltenberger	Johannes Gutenberg University Mainz Institute of Atmospheric Physics Supervisor of the doctoral thesis Email: amiltenb@uni-mainz.de
--	--

Mainz, March 3, 2026