Deriving probabilistic short-range forecasts from a deterministic high-resolution model

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Susanne Theis

aus

Köln

Bonn 2005

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Referent: Prof. Dr. Andreas Hense

2. Referentin: Prof. Dr. Jin-Song von Storch

3. Referent: Prof. Dr. Heini Wernli

Tag der Promotion: 24.10.2005

Diese Dissertation ist auf dem Hochschulschriftenserver der ULB Bonn
http://hss.ulb.uni-bonn.de/diss_online elektronisch publiziert.

Erscheinungsjahr: 2005

So eine Arbeit wird eigentlich nie fertig,
man muß sie für fertig erklären,
wenn man nach Zeit und Umständen
das Mögliche getan hat.

J. W. von Goethe, Italienische Reise, 16. März 1787

Abstract

In order to take full advantage of short-range forecasts from deterministic high-resolution NWP models, the direct model output must be addressed in a probabilistic framework. A promising approach is mesoscale ensemble prediction. However, its operational use is still hampered by conceptual deficiencies and large computational costs.

This study tackles two relevant issues: (1) the representation of model-related forecast uncertainty in mesoscale ensemble prediction systems and (2) the development of post-processing procedures that retrieve additional probabilistic information from a single model simulation. Special emphasis is laid on mesoscale forecast uncertainty of summer precipitation and 2m-temperature in Europe. Source of forecast guidance is the deterministic high-resolution model Lokal-Modell (LM) of the German Weather Service.

This study gains more insight into the effect and usefulness of *stochastic parametrisation schemes* in the representation of short-range forecast uncertainty. A stochastic parametrisation scheme is implemented into the LM in an attempt to simulate the stochastic effect of sub-grid scale processes. Experimental ensembles show that the scheme has a substantial effect on the forecast of precipitation amount. However, objective verification reveals that the ensemble does not attain better forecast goodness than a single LM simulation. Urgent issues for future research are identified.

In the context of statistical post-processing, two schemes are designed: the *neighbourhood method* and *wavelet smoothing*. Both approaches fall under the framework of estimating a large array of statistical parameters on the basis of a single realisation on each parameter.

The neighbourhood method is based on the notion of spatio-temporal ergodicity including explicit corrections for enhanced predictability from topographic forcing. The neighbourhood method derives estimates of quantiles, exceedance probabilities and expected values at each grid point of the LM. If the post-processed precipitation forecast is formulated in terms of probabilities or quantiles, it attains clear superiority in comparison to the raw model output.

Wavelet smoothing originates from the field of image denoising and includes concepts of multiresolution analysis and non-parametric regression. In this study, the method is used to produce estimates of the expected value, but it may be easily extended to the additional estimation of exceedance probabilities. Wavelet smoothing is not only computationally more efficient than the neighbourhood method, but automatically adapts the amount of spatial smoothing to local properties of the underlying data. The method apparently detects deterministically predictable temperature patterns on the basis of statistical guidance only.

Contents

1 Introduction

1.1 The need for probabilistic mesoscale forecasting

Prediction of small-scale surface weather

Predictability of the atmospheric state is of great societal and economic significance. Especially surface weather parameters are relevant for a wide range of man's activities and operations. For example, surface weather affects agricultural management, water resource management, and the transportation, energy and construction sector (BMBF, 2004). As many surface weather parameters are highly intermittent, it is also crucial to predict their timing and location as precisely as possible.

During the past decade, limited area models (LAM) have become a major source of forecast guidance on small-scale surface weather. The horizontal grid spacing of operational LAM has steadily been decreased. The model output contains more and more spatial details. Today, most operational LAM resolve the meso-β-scale and produce forecasts for a range of two or three days.

The core of a numerical weather prediction (NWP) model consists of a suitably discretised version of the laws which approximately govern the dynamics of atmospheric flow, augmented by so-called parametrisations of the processes that are unresolvable on the computational grid of the model.

Uncertainty in small-scale weather forecasts

In reality, neither model formulation nor input parameters are perfect, mainly due to the need of parametrising unresolved processes and due to operational constraints in the collection of observational data and their assimilation. Furthermore, the atmosphere behaves in a chaotic way (Lorenz, 1963) so that predicted states are extremely sensitive to small changes in model formulation and input parameters. There is substantial amplification of errors during the time integration of a NWP model. The inherent error growth is a consequence of nonlinearity and instability of atmospheric dynamics. Through the so-called inverse error cascade, errors in the smallest scales induce errors in successively larger scales. Eventually all scales are contaminated, in a systematic and/or random way.

Thus, the predicted state of interest can possess a substantial amount of intrinsic uncertainty, depending on forecast lead time, spatial scale, terrain, flow and forcing. With the advent of high-resolution NWP models this fact can not be neglected even in short-range forecasts since small-scale phenomena are particularly prone to forecast uncertainties. The relation between spatial scale and prediction skill has become evident in a multitude of theoretical and experimental studies (Lorenz, 1969; Zepeda-Arce et al., 2000; de Elía et al., 2002; Walser et al., 2004; Casati et al., 2004).

Consequently, the output of a deterministic NWP model must not be regarded as purely deterministic, but rather as an outcome of a random experiment that behaves according to a given probability density function (*pdf*). In other words, the output of a deterministic NWP model should be viewed as a *realisation of a random variable* and should be described by instruments of probability theory. Especially precipitation-related forecast products and information must be addressed in a probabilistic framework that accounts for their intrinsic uncertainty (Fritsch et al., 1998).

User's gain from a probabilistic forecast

The general public and specific users make many different decisions each day that depend at least in part on surface weather forecasts. Decisions may either be reached after a formal analysis or in an essentially intuitive manner. The connection between forecasts and decision making is self-evident. If a user consulted a forecast without even considering to use it within a decision making process, the purpose of a weather forecast would degenerate into mere entertainment.

In order to realise the full advantage of forecasts, regardless of the nature of the activity or of the individual's approach to decision making, the uncertainty inherent in the forecasts must be quantified and expressed in a concise and unambiguous manner (Murphy, 1979). Murphy (1977) has shown that the value of reliable—and even moderately reliable—probabilistic forecasts generally exceeds the value of traditional categorical forecasts. The superiority of a probabilistic forecast originates from its ability to suit different users and their different needs.

However, many users are not automatically able to properly interpret and use probabilistic forecasts. Similarly, subsequent prediction systems such as the prediction of air quality, water depth gauge, or road conditions are not immediately ready to cope with input from probabilistic weather forecasting. An operational program of probabilistic forecasting needs to be accompanied by a concerted effort to educate its users and to adapt subsequent prediction systems (Mylne, 2002).

Until such an effort takes effect, users and subsequent models still demand a categorical forecast. The direct NWP model output provides a categorical forecast, but it does not acknowledge the uncertainty in the prediction. An alternative to the direct NWP model output is the expected value of the random variable that underlies the prediction. Although the expected value is categorical, it can also be understood as probabilistic.

It is a function of the forecast *pdf* and takes the probabilistic nature of the direct NWP model output into account. A good estimate of the expected value contains less noise than a single realisation and is therefore a more reliable forecast.

1.2 Existing approaches to probabilistic prediction

Theoretical approach

In theory, probabilistic prediction could be achieved by solving equations that describe the time evolution of the *pdf* of the atmospheric state vector. Such evolution equations are known indeed: The Liouville equation and the Fokker-Planck equation (e. g. see Ehrendorfer, 1997, for a review). The Liouville equation is based on the notion that forecast uncertainties exclusively arise from an imperfectly known initial state and model nonlinearities. The Fokker-Planck equation additionally allows for model imperfections by the inclusion of stochastic elements. These equations have been successfully solved for low-dimensional systems, such as the Lorenz system (Lorenz, 1963).

In weather forecasting though, the solution of these equations is hampered by the extremely large dimensionality of the atmospheric equations. For example, a state-of-the-art mesoscale NWP model possesses a dimensionality of the order of 10^7. At present, using the Liouville or Fokker-Planck equation in operational NWP is out of reach. Their application is currently doomed to remain a purely theoretical approach and is not further mentioned in this study.

Ensemble prediction

The aim of ensemble prediction is to attain a Monte Carlo sampling of the phase space of the future atmospheric state. Ensemble prediction is a tool to investigate error growth and infer probabilistic properties of the future atmospheric state.

In ensemble prediction, a NWP system produces several simulations so as to derive multiple deterministic numerical weather predictions for the same forecast time. Each simulation or *ensemble member* forms another realisation of the random variable that underlies the prediction.

The simulations differ in the specification of one or several elements of the NWP system which are subject to uncertainty. Most commonly, initial condition uncertainty is addressed, because small differences in initial conditions can have a large impact on the prediction. An initial condition sample is produced and a deterministic NWP model is started from each sample element. This is the *perfect model scenario* which neglects uncertainties in model formulation and corresponds to the perspective of the Liouville equation.

An ideal ensemble design should involve perturbations of all elements of the NWP system which are subject to uncertainty (Balzer, 1998). This vision has not been put into practice yet, but remains as a guiding idea. Other sources of uncertainty are listed in Section 1.3.

Statistical methods based on historical data

Probabilistic prediction has also been achieved by purely statistical methods. The development of these methods requires a sufficiently large historical data base of NWP model output and/or observations.

A well-known example of statistical prediction is *model output statistics* (MOS) (Glahn and Lowry, 1972). Common MOS versions produce a categorical forecast, but probabilistic formulations also exist (cf. Applequist et al., 2002). MOS consists in linear regression conditional on NWP model output. In addition to linear regression, there is a multitude of other statistical approaches, for example logistic regression, neural networks, discriminant analysis, a classifier system (see Applequist et al., 2002, for an overview), quantile regression (Bremnes, 2004), Markov chains (Raible et al., 1999) or a Bayesian recursive method (Craig et al., 2001). Another example is the derivation of *statistical ensembles* which are used as a supplement to dynamical ensembles (Wilks, 2002; Roulston and Smith, 2003).

Statistical methods based on historical data have shown to be a valuable contribution to probabilistic forecasting. However, the need for a historical data set can be an obstacle. Most methods require a data base of historical model simulations that are representative of the model version currently in use. Due to frequent model refinements such a data base might not be at hand so that time-consuming reruns of the model become necessary before the statistical method can be developed.

In an attempt to limit the scope of this study, statistical methods based on historical data sets are not further mentioned here, despite their beneficial effect on probabilistic forecasting. Methodologies developed in this study do without historical data in order to extend their applicability to environments that do not have historical data sets at their disposal. Whenever such data should become available, it is recommended to use statistical methods based on historical data as a supplementary tool in probabilistic forecasting.

Forecaster's subjective judgment

At some forecast centres probabilistic forecasts are issued by the forecaster. A forecaster's probabilistic judgment is usually based on the deterministic output of one or several NWP model simulations, combined with the forecaster's experience and intuition.

The success of subjective probabilistic forecasts have only been proven in part (Hamill and Wilks, 1995; Krzysztofowicz and Sigrest, 1999). One problem of subjective judgment is its dependency on the intuitive process which may vary from forecaster to forecaster, and even from situation to situation for a particular forecaster. Another problem is that forecasters tend to formulate forecasts in a manner designed to suit a particular user's needs which cannot be representative of all users. For example, this can result in an overestimation of probabilities if forecasters tend to err on the safe side. Furthermore, the use of subjective judgment in operational high-resolution forecasting may cause large costs for the forecast centre, since the production of each local forecast involves the assistance of an experienced forecaster. Due to these reasons the subjective approach to probabilistic forecasting is not tackled in this study.

1.3 Current achievements and problems in ensemble prediction

Overview of mesoscale ensemble prediction

During the past decade, the simulation of initial condition related errors in *global* ensemble prediction systems has been well established (Toth and Kalnay, 1993; Molteni et al., 1996; Houtekamer et al., 1996). Operational use of *mesoscale* ensemble prediction has lagged far behind. It has been proposed that ensemble methods could also benefit short-range mesoscale forecasts (Brooks and Doswell III, 1993; Brooks et al., 1995). Though this is by no means guaranteed, results are generally positive (Du et al., 1997; Hamill and Colucci, 1998; Stensrud et al., 1999; Hou et al., 2001; Wandishin et al., 2001; Grimit and Mass, 2002). Mesoscale ensemble forecasting is on the agenda of many weather forecasting centres. Some ensemble prediction systems have already come into operational use, for example the short-range ensemble forecasting (SREF) system at the National Centers of Environmental Prediction (NCEP) in the United States and the limited-area ensemble prediction system (LEPS) of the Consortium for Small-scale Modeling (COSMO) in Europe. A recent overview of current research and issues of discussion in short- and medium-range ensemble forecasting can be found in (Hamill et al., 2005).

Experimental and operational limited-area ensemble prediction systems currently address various sources of forecast uncertainty:

- initial conditions (Du et al., 1997; Misra and Yau, 2001; Walser et al., 2004),
- lateral boundary conditions (Laprise et al., 2000; Marsigli et al., 2001; Frogner and Iversen, 2002),
- model physics (Arritt et al., 2001; Bright and Mullen, 2002),

- surface conditions (Hamill and Colucci, 1997; Wen et al., 2000; Mölders, 2001; Theis et al., 2002),
- and combinations thereof (Du et al., 1997; Stensrud et al., 2000; Hou et al., 2001; Grimit and Mass, 2002; Bright and Mullen, 2002).

Many questions remain concerning optimal design and potential benefits of mesoscale ensemble prediction. The best method of defining initial condition perturbations is unclear since error growth in limited-area mesoscale models is not a straightforward extension of error growth in global models. The presence of lateral boundary conditions introduces an added complexity. Large scales that are provided as nesting information at the lateral boundary are almost perfectly reproduced, whereas shorter scales behave chaotically and diverge as in an autonomous global model (de Elía et al., 2002). Ensemble dispersion may be inhibited even when perturbed lateral boundary conditions are applied (Nutter, 2003). In addition, the different role of physics and surface forcing (Anthes et al., 1985) requires special attention. In summary, it is still an open question how errors grow and organise on the mesoscale.

Horizontal resolution and ensemble size

A drawback of operational ensemble prediction is its large computational cost due to the need of multiple simulations. Together with the time-pressure in operational weather forecasting and the immense complexity of a state-of-the-art NWP model, this poses a challenge to computing environments of operational weather forecasting centres. The result is usually a trade-off between the horizontal grid box size of the model and the number of ensemble members. The optimal balance is still an issue of discussion.

An increase in grid box size is especially disadvantageous in mesoscale prediction. A realistic simulation of phenomena on the meso-β-scale and meso-γ-scale requires a grid box size of the order of 10 km and less. The model output of quantitative precipitation benefits from decreased grid spacing in mesoscale models, mainly because the realism of the results is improved (e. g. Mass and Kuo, 1998; Cacciamani et al., 2000; Colle et al., 2000; Riphagen et al., 2002; Mass et al., 2002).

As the number of ensemble members puts a limit to model grid box size, operational mesoscale prediction systems usually confine themselves to a relatively small ensemble size, typically a size of approximately ten members. However, such a small ensemble size can lead to considerable undersampling of the phase space of the future atmospheric state. The undersampling problem has lately been recognized as an important issue in the forecast quality assessment of ensemble systems (Richardson, 2001b; Mason, 2004).

Undersampling does not only arise out of practical limitations in the number of ensemble members, but is severely aggravated by the immense dimensionality of complex NWP models and the so-called *curse of dimensionality*. The curse of dimensionality expresses a major problem in direct sampling Monte Carlo simulation of high-dimensional sys-

tems. In order to maintain a given level of accuracy, the number of samples per variable must increase exponentially with the number of variables. Due to the large dimensionality of NWP models, the undersampling problem will remain omnipresent in ensemble prediction of weather and climate.

The undersampling problem becomes most prominent when a forecast centre puts maximum emphasis on small grid box size and refuses operational ensemble prediction as a whole. Then the operational forecast is based on a single deterministic simulation which provides only one realisation of the future atmospheric state. Due to the lack of an ensemble, forecast uncertainties are not quantified automatically. For example, such a situation arose at the German Weather Service (DWD) in 1999 when the non-hydrostatic Lokal-Modell (LM) was introduced, one of the world's leading operational NWP models in terms of small grid box size.

The problem becomes even more relevant since the model output of a high-resolution simulation explicitly contains small-scale phenomena. As smaller scales generally possess a shorter predictability limit, the direct model output is expected to contain a larger amount of randomness. This can substantially impair overall forecast quality and offset the benefits due to a smaller model grid box size if forecast uncertainties are not addressed explicitly. In fact, the objectively scored accuracy of a quantitative precipitation forecast does not necessarily improve compared to forecasts that lack spatial detail (Mass et al., 2002, and references within). The short predictability limit of small scales confines the gain of high-resolution models if the application aims to increase the accuracy of the prediction in intensity and phase at a given point (de Elía et al., 2002). In order to realise the full advantage of the small-scale simulation despite the lack of an ensemble or a sufficiently large ensemble, the model output needs statistical treatment.

Representation of model related uncertainty

The role of model related uncertainty has recently attracted increased attention in the field of ensemble prediction (Smith, 2000; Palmer, 2001; Toth and Vannitsem, 2001; Hansen, 2002). In principle, model related uncertainty arises from every part of the model which might be formulated in one or another way. Consequently, model uncertainty also refers to the uncertainty of model physics.

Model related uncertainty needs to be addressed explicitly, as results drawn from the perfect model scenario are of limited applicability. The perfect model scenario potentially leads to shortage of ensemble spread, i. e. underdispersiveness. Just as uncertainty in the initial condition severely limits the utility of a single forecast even in a perfect model, so model uncertainty severely limits attempts to obtain the forecast *pdf* (Smith, 2000; Orrell et al., 2001). Experiments with complex models indicate a high relevance of model related uncertainty. In short-range mesoscale prediction, model formulation appears to play a larger role than initial conditions when the large-scale forcing is weak (Stensrud et al., 2000).

A significant part of random model error in complex weather and climate models may be caused by neglect of sub-grid scale variability (Palmer, 2001; Lin and Neelin, 2002; Shutts, 2003). Because atmospheric models have finite grid sizes, they must parametrise sub-grid phenomena in terms of grid-scale variables. A measurement of the effect of the collective behaviour of unresolved processes must be some sort of average from an ensemble of fluctuating entities; it is inherently probabilistic. Parametrisations replace probability density functions in our physical description by conditional expectations that we then call deterministic bulk formula. Deterministic bulk representations of small-scale processes within a grid box only account for their mean or first-order effect, higher moments associated with sub-grid variability are not explicitly considered.

The amount of sub-grid scale variability can be substantial, as (Xu et al., 1992) have shown for convective processes in a cloud resolving model. Even if the sub-grid scale motions are energetically weak compared with the large-scale circulations, sub-grid scale variability can act as noise forcing that potentially shows up at the large scales. The upscale transformation from small-scale to large-scale error is a fundamental characteristic of error growth in weather and climate modelling. It is associated with an upscale energy transfer, also called *energy backscatter* (e. g. Shutts, 2004). Possible scenarios for an upscale cascade of model error are described in (Palmer, 2001). However, little is understood about the extent to which sub-grid scale variability affects large-scale motions stochastically in weather modelling.

A few years ago, most operational ensemble systems followed the perfect model scenario, i. e. they neglected uncertainties in the model formulation. Recently, more attempts have been made to account for model uncertainty. One approach is *stochastic parametrisation* (Buizza et al., 1999; Palmer, 2001; Lin and Neelin, 2002) which aims at the representation of sub-grid scale variability through introducing a stochastic element into current parametrisations. Another approach is the *multi-model* or *multi-scheme ensemble* (Palmer et al., 2000; Krishnamurti et al., 2000; Wandishin et al., 2001; Richardson, 2001a; Mylne et al., 2002) which pools several parametrisation schemes or models in one ensemble. Both approaches are unsatisfying, at least in their current state. Stochastic parametrisations are still in their infancy and multi-model or multi-scheme ensembles lack scientific underpinning. A more extensive overview of existing approaches follows in Section 3.1.

1.4 Aims of this study

This study aims to set the basis for further improvements of the ensemble approach and it aims to develop novel methods of probabilistic forecasting that improve the output of a single deterministic forecast and are suited for immediate operational use. The study focusses on mesoscale short-range forecasting of surface weather parameters in Europe. Forecast uncertainty encompasses random error only; systematic errors are not addressed in this study. Special focus is laid on small-scale forecast uncertainties and on

the warm season. Source of forecast guidance is the deterministic high-resolution model LM of DWD.

Isolation of certain aspects

This study looks into specific aspects of probabilistic forecasting *in isolation* instead of devising a fully-fledged probabilistic prediction system. The study tackles two problems: (1) stochastic representation of model related uncertainties in mesoscale ensemble forecasting and (2) probabilistic mesoscale forecasting when only one deterministic simulation is available, i. e. when the undersampling problem is most prevalent. The first problem is addressed through the implementation of a stochastic parametrisation scheme. The second problem is addressed through the development of statistical post-processing methods.

Stochastic parametrisation in the LM In terms of model related uncertainties, the study aims to gain more insight into the effect and usefulness of stochastic parametrisation schemes in short-range ensemble forecasting. In an attempt to simulate the stochastic effect of unresolved processes on the predictability of the resolved scales, this study designs a stochastic parametrisation scheme and implements it into the LM. The stochastic scheme introduces a random element into the existing parametrisation schemes of the LM. An ensemble of simulations is produced and each ensemble member is generated through a different realisation of the random element.

Statistical post-processing of LM output In terms of probabilistic forecasting on the basis of one deterministic simulation, the study aims at the development of a statistical post-processing scheme that accounts for forecast uncertainty in the operational LM output and does without ensemble simulations, historical error statistics, or the operational interaction of a forecaster. The statistical post-processing scheme attempts to *parametrise* forecast uncertainty, in the sense that it does not trace forecast uncertainty to its origins, but uses ancillary indicators for its estimation. In this study, two different schemes are developed: the so-called *neighbourhood method* and *wavelet smoothing*. If the respective post-processing products prove worthwhile, the schemes are envisaged to become operational at DWD.

The issue emerged from a bilateral project of DWD and the University of Bonn. In 1999 DWD introduced the high-resolution NWP model LM into operational use. The model has an exceptionally small grid box size and is used for high-resolution short-range forecasting. However, the benefits of small grid box size came with a price tag, as one simulation demanded such a large amount of computing power that an operational ensemble prediction system went out of reach. Due to limited predictability of small scales, DWD ran into problems in the interpretation of the forecast. DWD called for a pragmatic remedy that would require a minimum amount of computing power and be suitable for almost immediate operational use.

Pretensions of the approaches in this study

The stochastic parametrisation scheme and the statistical post-processing scheme are associated with very different intentions and expectations. In order to understand the achievements of this study, their respective pretensions is outlined in the following paragraphs.

The LM stochastic parametrisation scheme aims at gaining more insight into the stochastic simulation of model uncertainties in ensemble prediction. The field is relatively unexplored so that it is probably too early for downright answers in terms of an operational system. This study rather aims at defining relevant questions which must be tackled by future research. Furthermore, the stochastic parametrisation scheme could serve as a guideline in the development of the LM post-processing schemes. The experimental LM ensembles might provide insight into LM small-scale predictability and might be generalized in some way. Thus, the computationally less demanding post-processing method might be tuned to resemble the results of the LM ensemble and then be able to roughly mimic small-scale forecast uncertainties without actually producing an ensemble.

The statistical post-processing schemes need to satisfy the technical requirements of operational use; they must be application-oriented, ready-to-use and computing-efficient. The statistical post-processing schemes should attain noticeable improvements in probabilistic forecasting with a minimum of effort. The post-processing schemes do not aim at replacing the ensemble approach, but are designed as a supplementary tool in situations that temporarily do not allow for sufficient ensemble simulations. The value of the post-processed forecast is limited compared to that attainable through an ensemble of forecasts that can properly account for forecast uncertainty due to initial conditions and model related errors (cf. Toth et al., 2001). However, conceptual deficiencies and computational limitations still prevent the ensemble approach from reaching its full potential. The LM statistical post-processing schemes aim at a remedy for these shortcomings. Note that this should not discourage any simultaneous effort to improve the ensemble design itself.

Possible use in a wider context

Investigations of this study take place in very special situations. The experimental ensembles exclusively focus on the stochastic effect of sub-grid scale processes and neglect all other sources of forecast uncertainty; the statistical post-processing scheme pretends that no ensemble simulations are available at all.

The findings of this study may also be useful in various other contexts of probabilistic forecasting. The investigation of the stochastic parametrisation scheme can show whether such schemes might benefit existing ensembles that currently neglect the stochastic effect of sub-grid scale processes. The investigation of the statistical post-processing scheme can indicate whether statistical post-processing might be able to

enrich the statistical information derived from a small-sized mesoscale ensemble which severely suffers from undersampling.

Forecast products and variables

Recall that the LM output should be regarded as a realisation of a random variable. This study aims to estimate certain statistical properties of its underlying *pdf*. The estimates are issued at each grid point and output time of a LM simulation so that they can be coded in the same way as the operational LM output and used as supplementary information. The following statistical properties are estimated:

Exceedance Probability The exceedance probability is a function of a certain threshold value and denotes the probability that the observation will exceed the threshold. This probability is assumed to equal the probability that the forecast exceeds a certain threshold. The exceedance probability is interesting in terms of warnings, especially if the exceedance of a certain threshold is associated with the sudden onset of disastrous weather.

Quantiles Quantiles are the inverse of the exceedance probability. Quantiles denote certain exceedance values and are a function of a probability. If the quantile is correctly forecast, there will be a probability p that the corresponding quantile is greater than the observation. Again, quantiles are interesting in terms of warnings, especially if the forecast user wishes to control a certain amount of risk.

Expected Value In this study the estimate of the expected value is intended for those users who are not trained to cope with probabilistic formulations. Loosely speaking, the expected value is the value which would be attained on average if an infinite number of realisations of the same random variable were available. For example, the average of a perfect ensemble would be an unbiased estimator of the expected value of the forecast. The expected value is a categorical forecast which is reduced to features that are deterministically predictable. A good estimate of the expected value should be more reliable than the direct model output. The expected value can also be seen in the context of *signal* preservation and *noise* removal. The signal represents features that are deterministically predictable and the noise represents others. The field of the estimated expected values is then a *denoised* forecast.

The experimental ensemble derives estimates of the expected value and the exceedance probability, the neighbourhood method derives estimates of all three quantities and wavelet smoothing derives estimates of the expected value.

A selection of forecast variables is considered which bear special relevance to users: precipitation at the ground (1h-, 6h-, 12h-, 24h-accumulations), temperature in 2 m height (hourly, minimum, maximum), wind speed in 10 m height (mean and gusts), cloud cover and surface net radiative flux (solar and thermal). For the sake of brevity, this report

restricts itself to the presentation of selected variables. In terms of the experimental LM ensemble, the report shows investigations of the precipitation forecast only; in terms of the statistical post-processing scheme the report shows investigations of precipitation and hourly 2m-temperature (temperature in 2 m height above the ground).

1.5 Overview of this study

After a brief introduction to the model LM (Chapter 2), Chapter 3 deals with the stochastic parametrisation scheme which is developed in this study. The chapter gives an overview of existing approaches to the simulation of model uncertainties and outlines the benefits of stochastic parametrisation schemes. It then describes the design of the LM stochastic parametrisation scheme and draws conclusions from the corresponding LM ensemble experiments. These experiments are carried out for a case study of extreme summer precipitation (10 July 2002) and for 14 additional short-range forecasts which are investigated statistically.

Chapters 4 and 5 present the post-processing schemes that estimate the statistical properties of the underlying *pdf* from the data of a single deterministic LM simulation. Two methods are developed: The *neighbourhood method* and *wavelet smoothing*. The neighbourhood method (Chapter 4) is based on the notion of spatio-temporal ergodicity including explicit corrections for enhanced predictability from topographic forcing. It merges the estimation of quantiles, expected value and exceedance probabilities into a single concept. Wavelet smoothing (Chapter 5) originates from the field of image denoising and includes concepts of multiresolution analysis and non-parametric regression. In this study, wavelet smoothing estimates the expected value, but the method may be extended to the additional estimation of exceedance probabilities in a future study.

These chapters are organised as follows: First, the new post-processing scheme is put into the context of existing methodologies. Similarities and differences are revealed and the particular design of the new scheme is motivated. Underlying assumptions are discussed in detail and an optimal configuration is sought. The schemes are applied to the raw LM output of the case study above (10 July 2002). First conclusions are drawn from eye-ball inspection of the post-processed forecast fields.

Chapter 6 supplements the subjective assessment by objective verification of 15 short-range forecasts. The ensemble forecasts and the post-processed forecasts are compared to corresponding observational data and forecast goodness is evaluated. Forecast goodness is condensed into a few verification measures which allow for a direct comparison between the results of different forecast systems. Chapter 7 gives a summary of the whole study and compiles the main conclusions.

2 Lokal-Modell (LM)

2.1 Overview

The LM is a numerical model of the atmosphere, designed to model the meso-β- and meso-γ-scale. Its development was organised as an internal project of DWD in the 1990's. In 1999 the LM replaced the regional model Deutschland-Modell (DM) at DWD and the LM is the regional part of the DWD operational modelling system since then. Today LM is also the basic model component of the Consortium for Small-scale Modeling (COSMO), an international association of currently seven meteorological services in Europe. A major use of the model is its application to operational weather forecasting within the COSMO system.

The LM is a non-hydrostatic, elastic model that is designed to cover all resolutions from 50 km to 50 m. The development of the LM was motivated by the desire to decrease the grid spacing of operational weather forecasts below 10 km. It is an attempt to capture explicitly all small-scale, short-duration severe weather events and significant flow systems which are related to the non-hydrostatic scales of motion. The key issue is the numerical prediction of near-surface weather.

Using a model with a grid spacing below 10 km for operational weather forecasting broke new ground in science. The operational version of LM is run with a grid spacing of 7 km compared to a 14 km grid spacing of its predecessor DM.

2.2 Short description

A description of the LM and some examples of model predictions at the meso-γ-scale can be found in Steppeler et al. (2003). An extensive documentation of the LM (Doms and Schättler, 1999) is available in the Internet (*www.cosmo-model.org*). The LM undergoes constant refinements and new versions are frequently released. Updated information on DWD models are regularly presented in the "Quarterly Report of the German NWP System" (*www.dwd.de/en/FundE/Veroeffentlichung/Quarterly_new/Quarterly_new.htm*). This study uses versions 2.14 and 2.19 which were both released in the year 2002. In the following subsections, the basic features of the LM are described.

Model equations

The LM is a non-hydrostatic model based on the primitive equations. Prognostic variables comprise the three-dimensional wind vector, pressure, temperature, specific water vapour content, specific cloud water content, and (optional) specific cloud ice content. A multitude of further variables are calculated diagnostically, such as the total density, flux of rain and snow, cloud cover, 2m-temperature and wind gusts.

Initial and boundary conditions

Initial conditions are generated by LM assimilation runs using the nudging method (Schraff, 1997; Schraff and Hess, 2002). As the LM is a limited-area model, it also requires lateral boundary conditions during the simulation. They are usually provided by the simulation of a coarse-grid global model via one-way nesting using the relaxation technique by Davies (1976). In the operational setting of DWD, forecasts of their global model GME[1] are used. The top boundary condition of the LM is a rigid lid, supplemented by a Rayleigh damping layer.

Model grid

In the horizontal, spherical co-ordinates are used. The grid is rotated on the sphere, so that the model domain is almost uniform and avoids regions with strong convergence of the meridians. The equations are formulated on an Arakawa-C grid (Arakawa, 1966). In the vertical, a generalized terrain following co-ordinate is used. The thickness of the layers increases with distance to the model surface. In the operational setting of DWD, horizontal grid spacing is 7 km and the model atmosphere is divided into 35 layers. Vertical spacing is less than 100 m in the planetary boundary layer and almost 1 km in the upper troposphere. The highest vertical level is located at a pressure level of 20 hPa.

Numerical integration

The model equations are discretised in time and space and their solution is derived by numerical integration. The numerical integration scheme has to be computationally effective and should yield a sufficient degree of accuracy.

The LM integration scheme splits fast and slowly varying modes of motion. The time integration method is the leapfrog scheme (Klemp and Wilhelmson, 1978). Standard second order differences are used for the evaluation of the slow mode tendencies. For the fast modes the integration step is subdivided into a number of small time steps and

[1] The expression GME is a combination of its predecessors: GM (Global-Modell) and EM (Europa-Modell)

a split-explicit scheme is used. Horizontal motion of fast and slow modes is treated explicitly, whereas vertical motion is treated by the implicit Crank-Nicholson scheme (e. g. Großmann and Roos, 1994).

Parametrisations

A multitude of physical processes operates on small scales that are not resolved by the model grid. However, many sub-grid scale processes have a considerable effect on the resolved scales. Prominent examples are radiation, convection, cloud microphysics and turbulence. Their effect is covered by parametrisations that formulate the sub-grid scale forcing as deterministic functions of the resolved scales. An overview of LM parametrisations is given in Heise (2002).

The heating rate due to short-wave and long-wave radiation fluxes is calculated by the δ-two-stream approximation of the radiative transfer equation according to Ritter and Geleyn (1992). The parametrisation scheme for the formation of grid-scale clouds and precipitation is based on a Kessler-type bulk formulation (Kessler, 1969). Moist convection is parametrised by the mass flux scheme of Tiedtke (1989). Vertical diffusion due to turbulent transport is represented by a prognostic equation for turbulent kinetic energy (Raschendorfer, 2001) that is a level 2.5 closure scheme according to Mellor and Yamada (1974). The lowest model layer is assumed to be located within the Prandtl layer. A stability and roughness-length dependent surface flux formulation according to Louis (1979) is used.

For land surfaces, the soil-vegetation-atmosphere-transfer (SVAT) model TERRA provides the temperature and the specific humidity at the ground. Usually SVAT models are classified as parametrisations, too, although they mainly consist of explicit forecasts of temperature and humidity in several soil layers.

Additionally, parametrisations need external parameters, such as orographic height, surface type or roughness length. Due to their close relation to parametrisation schemes, their determination is also classified as a parametrisation.

Overall picture

The structure of the LM is made up of several modules. The sequence of actions within one slow mode time step (Fig. 2.1) is repeated again and again during a model simulation. First, the parametrisation schemes are invoked which calculate the parametrised forcings or tendencies. Then the parametrised forcings are added to the grid scale forcings and enter the numerical integration scheme that modifies the prognostic variables. An exception is the parametrisation of turbulence, as vertical motion is integrated implicitly, not explicitly. The turbulence parametrisation is directly entangled with the implicit integration scheme (not shown in Fig. 2.1). The relevant tangible output of the

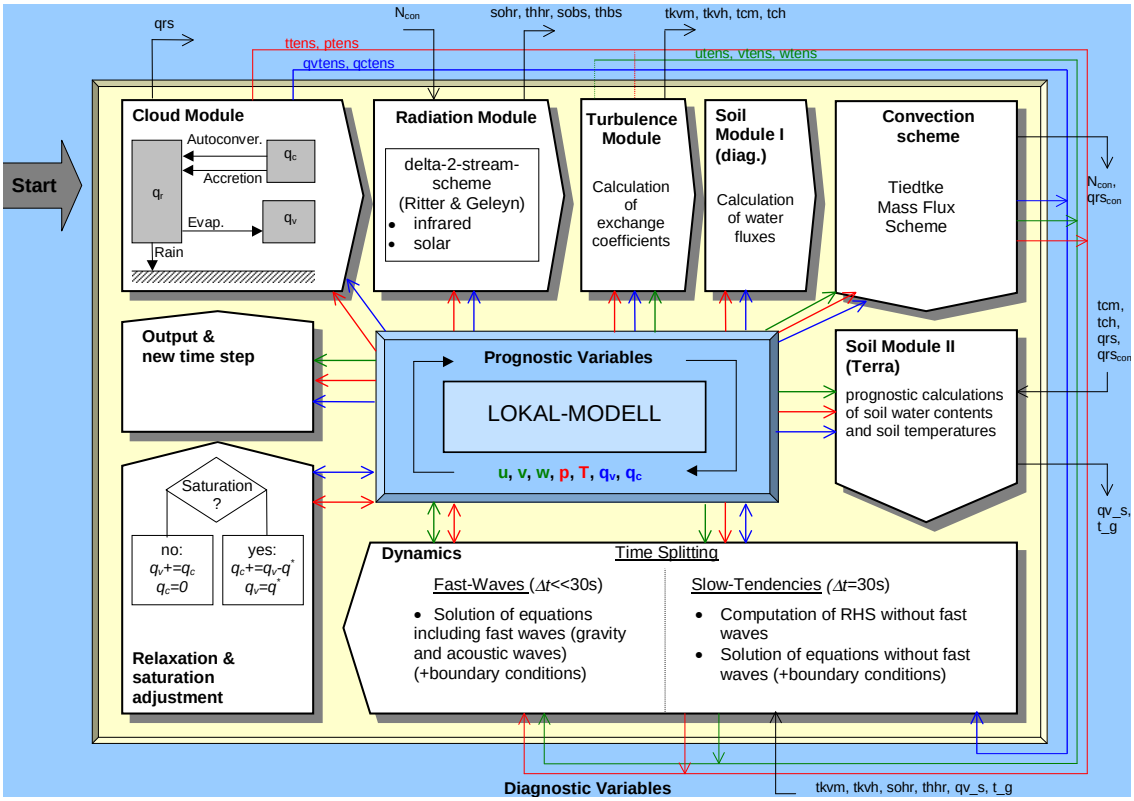

Figure 2.1: *Sequence of actions within one slow mode time step. The modules are invoked clockwise, starting at the upper left. The prognostic variables are in the center of the diagram and the diagnostic variables at the brim are used for communication between the modules (after Ament, 2001).*

turbulence scheme are the turbulent transfer rates that enter the integration scheme. Only a small part of the entire parametrised forcing of turbulence shows up explicitly.

Implementation

The implementation of the LM is based on features of FORTRAN 90 and is suitable for parallel computers using the communication standard *message passing interface* (MPI). The code performs on a variety of platforms and is designed to facilitate the co-operation of different model developers.

3 Stochastic parametrisation in the Lokal-Modell

In order to set the basis for improvements in mesoscale ensemble forecasting, a stochastic parametrisation scheme is implemented into the LM and experimental ensembles are produced. This endeavour aims to give insight into the effect and usefulness of stochastic parametrisation schemes in mesoscale forecasting. Additionally, the findings are envisaged to serve as a guideline in the development of the post-processing schemes in this study.

First, Section 3.1 gives an overview of existing approaches that deal with the simulation of random model error and motivates the use of a stochastic parametrisation scheme in more detail. Then the design of the stochastic parametrisation scheme is introduced in Section 3.2. Experimental ensembles are produced for a case study (Section 3.3) and then for a forecast data set covering two weeks (Section 3.4). The case study gives a subjective impression of the scheme's impact and is also used to explore sensitivities to the setup of the scheme. The more extensive data set serves as a basis for a statistically more robust analysis of simulated error growth. Furthermore, the data are used in Section 6.7 which assesses forecast goodness by comparison with observational data. The outcome of the ensemble experiments is discussed in Section 3.5. The experiments are juxtaposed with related research, possible deficiencies of the scheme are raised, and its use in the development of the statistical post-processing scheme is discussed.

3.1 Existing approaches to the representation of model uncertainties

Subsection 3.1.1 presents a brief overview of past research activities that have already dealt with stochastic approaches to the representation of model uncertainties. Subsection 3.1.2 juxtaposes the stochastic approaches with the multi-model and multi-scheme approaches. These approaches are more pragmatic and much more popular than the stochastic approaches, but lack scientific underpinning.

3.1.1 Stochastic approaches

Little is known about the dynamics of model error in complex systems, but some experiments have been carried out with low-order chaotic models (Smith et al., 1999; Judd and Smith, 2001; Nicolis, 2003). Recent results (Judd and Smith, 2001) establish that, given an imperfect nonlinear chaotic model of a deterministic system, better state estimation and better probability forecasts can almost certainly be obtained by using a stochastic model.

Since the landmark paper of Hasselmann (1976) the use of stochastic noise to represent unpredictable small-scale variability has become familiar in a number of geophysical models and in a number of different research fields. A review on stochastic modelling can be found in Penland (2003) or Imkeller and von Storch (2001).

In the field of complex NWP models and general circulation models, the use of stochastic noise has only recently gained attention (Buizza et al., 1999; Palmer, 2001; Lin and Neelin, 2002). There has been an increasing awareness that the neglect of sub-grid scale variability may be a cause of insufficient ensemble spread. Stochastic parametrisation schemes could improve ensemble outcomes by addressing another origin of uncertainty that has been left out so far. Its future implementation into existing prediction systems is currently on the agenda of numerous weather forecasting centres. Expectations are especially high in terms of short-range mesoscale forecasting since the perturbation of initial conditions and lateral boundary conditions have not led to a sufficient representation of forecast uncertainty so far.

One way of representing sub-grid scale variability is through introducing a stochastic element into current parametrisations. Papanicolaou and Kohler (1974) show that rapidly varying quantities with small but finite correlation times may be treated as stochastic noise. However, the best method of defining a stochastic parametrisation scheme is unclear. Only few experiments have been carried out with NWP models that are used operationally otherwise (Evans et al., 1998; Buizza et al., 1999; Bright and Mullen, 2002; Shutts, 2004) and very little experience has been gained in terms of high-resolution modelling. Many questions are still unanswered.

Since 1999 the European Centre of Medium-Range Weather Forecasts (ECMWF) has been applying a stochastic parametrisation scheme in the operational version of their global ensemble prediction system. One of its main achievements is an improvement of the short-range precipitation forecast (Buizza et al., 1999). The scheme is simplistic though and needs further research and development (Shutts, 2004).

3.1.2 Multi-model and multi-scheme approaches

Obviously, the application of stochastic methods in operational NWP models is still in its infancy. Nevertheless, a simulation of model related uncertainties is urgently needed.

This discrepancy led to the development of multi-model and multi-scheme approaches which are much more common in operational weather forecasting today.

In a multi-scheme ensemble the model uses different parametrisation schemes in each ensemble member while the multi-model approach simply combines different models in one ensemble. Multi-model approaches offer a way to address model uncertainties with a minimum of effort in ensemble system development. Multi-scheme approaches involve somewhat more effort, as the schemes have to be tuned to the respective model first.

First experiments have shown that, despite the larger ensemble size, the quality of an initial condition ensemble is not superior to that of the multi-model ensemble (Ziehmann, 2000). Multi-model ensembles have been successfully applied to a variety of time scales, ranging from short-range forecasting to climate change scenarios (Palmer et al., 2000; Krishnamurti et al., 2000; Wandishin et al., 2001; Richardson, 2001a; Mylne et al., 2002).

However, neither multi-model nor multi-scheme ensembles explicitly address the neglect of sub-grid scale variability, because each individual scheme or model still uses deterministic bulk formulae. The probability density function of sub-grid scale processes is assumed to be sharp around the mean and the ensemble accounts for the fact that the mean itself is not well known. Each individual model or scheme still aims at simulating the most probable state. As a practical consequence, the multi-model or multi-scheme ensemble may yield poor results in terms of estimating the risk of extreme events.

Additionally, a multi-model or multi-scheme ensemble can pose problems in the interpretation of its direct output. The individual schemes or models may possess very different error characteristics which can be flow-dependent or entirely unknown. The ensemble forecasts cluster strongly by scheme or model. It may be necessary to assign weights to the individual simulations, but their definition is unclear.

As long as the stochastic approach is still in its infancy, multi-model and multi-scheme ensembles are a valuable approach to the representation of model uncertainties. However, the quick success of multi-model ensembles should not hide the fact that the stochastic approach is theoretically more sound. Stochastic model formulation needs continued research effort so that the approach may eventually mature and provide convincing answers to ensemble prediction at some point in the future.

3.2 Design of the stochastic parametrisation scheme

In order to explore the use of stochastic parametrisation schemes in high-resolution modelling, this study implements such a scheme into the LM. The scheme is similar to the operational scheme at ECMWF, but adapts to the smaller grid size and shorter time step of the LM, contains several upgrades such as an energetically consistent perturbation of

the turbulence scheme, a perturbation of roughness length and a correction for energetic inconsistency.

3.2.1 Notation and overview

As most other climate or weather prediction models, the LM subdivides the vertical and horizontal extent of the atmosphere into a finite number of layers and grid boxes (cf. Chapter 2). At each grid box, a vector $\mathbf{e}$ is defined that contains the seven prognostic variables (cf. Chapter 2). The model possesses a dimensionality of 10^7.

The operational version of the LM does not contain any stochastic elements. Its parametrisation schemes are based on deterministic bulk formulae $\mathbf{P}\big(\mathbf{e}(t);\alpha\big)$ that are functions of the prognostic variables $\mathbf{e}$ on the model grid and functions of fixed parameters α that need to be set empirically and are independent of the simulation.

Prognostic variables of a future point of time are calculated by numerical schemes which use a time step Δt. The prognostic variables at time $t + \Delta t$ are a function of the non-parametrised processes $\mathbf{A}$ and the parametrised processes $\mathbf{P}$.

$$\mathbf{e}(t + \Delta t) = \mathbf{e}(t) + \Delta t \cdot \Big\{\mathbf{A}\big(\mathbf{e}(t)\big) + \mathbf{P}\big(\mathbf{e}(t);\alpha\big)\Big\}. \tag{3.1}$$

As the choice of the particular scheme is not in the focus of interest here, Eq. 3.1 is formulated in terms of the simple Euler forward scheme, although the LM applies more complex schemes (cf. Chapter 2). In the context of this study, a model simulation can be seen as the successive application of Eq. 3.1.

There are about as many different approaches to stochastic parametrisations as there are investigations in this field. The perturbations can enter into different parts of the parametrisations $\mathbf{P}\big(\mathbf{e}(t);\alpha\big)$. Some examples are listed here.

Input fields e: Lin and Neelin (2000) and Bright and Mullen (2002) perturb input fields $\mathbf{e}$ before they enter the parametrisation schemes. Lin and Neelin (2000) perturb a measure of convective available potential energy (CAPE) before its use in the convective bulk formula and Bright and Mullen (2002) perturb input fields $\mathbf{e}$ of the convective trigger function. Aside from their use in parametrisation schemes, the prognostic variables $\mathbf{e}$ remain unperturbed.

Tunable parameters α: In addition to the perturbation of input fields $\mathbf{e}$, Bright and Mullen (2002) perturb a tunable parameter α, the critical bulk Richardson number in the turbulence parametrisation scheme.

Parametrised tendency $\mathbf{P}(\mathbf{e}(t);\alpha)$: The operational ECMWF ensemble prediction system directly perturbs the parametrised tendency $\mathbf{P}(\mathbf{e}(t);\alpha)$ which represents the net diabatic forcing from all parametrisation schemes (Buizza et al., 1999).

Recently, the research department at ECMWF has started to perturb the non-parametrised tendency as well. Shutts (2004) developed a stochastic backscatter forcing which introduces vorticity perturbations into the model flow.

The LM stochastic parametrisation scheme of this study mainly follows the scheme of the operational ECMWF system (Buizza et al., 1999) and transfers the approach from a global model to a limited-area high-resolution model. In addition to the ECMWF scheme, the LM stochastic scheme contains several upgrades such as an energetically consistent perturbation of the turbulence scheme, a perturbation of roughness length and a correction for energetic inconsistency.

The LM contains four individual parametrisation modules: cloud microphysics, turbulence, convection and radiation scheme. Stochastic forcing enters the cloud microphysics, convection and radiation scheme via the perturbation of their net diabatic forcing (Section 3.2.2). For the sake of simplicity, a separate treatment of these parametrisation schemes is omitted.

Due to the specific design of the LM, the tendencies of the turbulence scheme are not included in the perturbation of the net diabatic forcing, as the respective tendencies do not explicitly appear in the code of the model. Thus, stochastic forcing enters the turbulence scheme via the perturbation of input parameters (Section 3.2.3).

The choice of input noise is discussed in Section 3.2.4. An attempt is made to restore energetic consistency with respect to radiative forcing (Section 3.2.5). Additionally, stochastic forcing enters the turbulence scheme via the perturbation of a tunable parameter: the roughness length (Section 3.2.6). By perturbing the roughness length, the LM ensemble demonstrates how the perturbation of tunable parameters can be generally included in a stochastic parametrisation scheme.

3.2.2 Cloud microphysics, convection and radiation scheme

The perturbations of the cloud microphysics scheme, the convection scheme, and the radiation scheme follow the approach of Buizza et al. (1999). The perturbation directly accesses the net diabatic forcing $\mathbf{P}_{\mathrm{m/c/r}}$ of these schemes. At every grid point r and at every time t the parametrised tendency $\mathbf{P}(\mathbf{e}(t);\alpha)$ is multiplied by a random number $\langle x(r,t)\rangle_{D,T}$:

$$\mathbf{P'}_{m/c/r}\big(\mathbf{e}(t)\big) \equiv \langle x(r,t)\rangle_{D,T} \cdot \mathbf{P}_{\mathrm{m/c/r}}\big(\mathbf{e}(t);\alpha\big). \tag{3.2}$$

The indices D, T imply some spatio-temporal auto-correlation of the random numbers which are specified in Section 3.2.4. The multiplication is carried out for each component of the state vector $\mathbf{e}$ and for all vertical levels. The random number varies in time and horizontal directions, but the same number is used for different prognostic variables within the atmospheric state vector and for different vertical layers.

In the experimental LM ensemble, each ensemble member results from the iterative

application of the *perturbed* equations:

$$\mathbf{e}_j(t+\Delta t) = \Delta t \cdot \Big\{ \mathbf{A}\big(\mathbf{e}_j(t)\big) + \mathbf{P'}_{m/c/r}\big(\mathbf{e}_j(t);\alpha\big) \Big\}. \tag{3.3}$$

The perturbed equations arise from a series of random numbers $\big\langle x_j(r,t)\big\rangle_{D,T}$ that is unique to each ensemble member j. The random numbers are distributed according to a specific probability distribution function that is chosen *a priori*. Note that the perturbed equations (Eq. 3.3) have a discretised form of the Langevin equation (Gardiner, 1997) which is the prototypical stochastic differential equation.

3.2.3 Turbulence scheme

Due to the specific design of the LM, the perturbation of the net diabatic effect (Eq. 3.2) does not include the tendency from the turbulence parametrisation scheme. As the LM integrates the turbulence part of the parametrised tendencies implicitly, the tendency $\mathbf{P}_{\mathrm{turb}}\big(\mathbf{e}(t);\alpha\big)$ due to turbulence is not directly accessible in the LM source code.

Instead of perturbing the tendency itself, the stochastic scheme perturbs the diffusion coefficients for turbulent vertical transport of momentum and heat. The diffusion coefficients directly relate to the turbulence tendency via the closure approximation (cf. e. g. Stull (1988)). As the diffusion coefficients are calculated from a combination of grid-scale fields and fixed parameters, the perturbation implicitly addresses both input fields $\mathbf{e}$ and tunable parameters α.

The perturbation of turbulence is directly connected with the perturbation of the other three parametrisation schemes, because the stochastic scheme multiplies the diffusion coefficients by the same random number as the net diabatic forcing $\mathbf{P'}_{\mathrm{m/c/r}}$ (Eq. 3.2). In contrast to the perturbation of tendencies, the perturbation of the diffusion coefficients has the positive side-effect of conserving the energetic consistency of parametrised tendencies with associated surface fluxes.

3.2.4 Choice of random numbers

Determining the specific *pdf* of the input noise is one of the key issues in the development of a stochastic parametrisation scheme. The calibration of stochastic parametrisation schemes is still subject of current discussion and research. As this is the very first approach to implementing stochasticity in the LM, the choice of input noise follows the design of the operational scheme at ECMWF (Buizza et al., 1999) and adapts the approach to the smaller grid size and shorter time step of the LM.

The distribution function

Following the ECMWF scheme (Buizza et al., 1999), the *pdf* of the random numbers is chosen to be uniform. The random number $\langle x_j(r,t)\rangle_{D,T}$ is sampled *uniformly* from the interval $[1-a, 1+a]$ where a is a fixed parameter of noise amplitude ($0 < a \leq 1$).

Strictly speaking, there are more scientific grounds for a Gaussian distribution, because the central limit theorem states that a multi-scale dynamical system may be approximated as depending on the realisations of a white-noise process under certain conditions (cf. e. g. Papanicolaou and Kohler, 1974; Penland, 2003). However, the uniform distribution is also justifiable, because it is still superior to the delta-function which is implicitly applied in the deterministic parametrisation schemes of the operational LM version.

Noise amplitude and spatio-temporal auto-correlation

Once the shape of the distribution has been chosen, the values of the noise amplitude a and the spatio-temporal noise auto-correlation (D, T) have to be selected. In terms of the auto-correlation function, the approach of Buizza et al. (1999) is followed: The random numbers are identical in a spatial region of $D \times D$ grid boxes and during a time interval of T time steps of the model integration. Apart from the constancy within boxes and time intervals, the auto-correlation is zero.

Operational values of amplitude a and auto-correlation D and T can be determined through an empirical optimisation or tuning process. Buizza et al. (1999) pick several configurations of amplitude a and auto-correlation D and T, carry out fairly extensive test simulations for each configuration and apply the one giving best verification results in future operational use. However, this procedure is only feasible in generously equipped computing environments like that at ECMWF.

As a fully-fledged tuning process is associated with prohibitive computational costs, this study has to choose the noise properties *a priori*. The *a priori* choice of noise properties is relatively unexplored. Perturbations in the parametrisation increments generated by the stochastic scheme are meant to represent statistical fluctuations about the ensemble-mean value. In case of convective parametrisation one could relate these fluctuations to the low convective cloud population density so that the number of clouds within a grid box fluctuates substantially (Shutts, 2004). Either observational data or simulations from a model with a finer horizontal mesh size might serve as a guidance. Bright and Mullen (2002), for example, base their estimate of temporal auto-correlation on radar-derived information about the lifetime of individual convective cells, but also admit that their estimate is somewhat arbitrary. Especially the discrepancy between the size of individual convective cells and a model grid box is a problem. Principally, problems of this kind are always expected to come up, when applying observational findings to model grid box variables.

As a first step, this study carries out a sensitivity study (cf. Section 3.3). The study

Table 3.1: *The parameter settings of the four experimental ensembles with stochastic forcing. Δx denotes the grid spacing of the model and Δt denotes the time step in the model integration.*

	Amplitude a	Spatial Auto-Correlation D	Temporal Auto-Correlation T
LLL	0.25	$5 \cdot \Delta x = 35\,\mathrm{km}$	$4 \cdot \Delta t = 160\,\mathrm{s}$
MLL	0.50	$5 \cdot \Delta x = 35\,\mathrm{km}$	$4 \cdot \Delta t = 160\,\mathrm{s}$
HLL	1.00	$5 \cdot \Delta x = 35\,\mathrm{km}$	$4 \cdot \Delta t = 160\,\mathrm{s}$
HHH	1.00	$10 \cdot \Delta x = 70\,\mathrm{km}$	$16 \cdot \Delta t = 640\,\mathrm{s}$

investigates the impact of the injected noise on the ensemble forecast by applying four different configurations of noise amplitude and auto-correlation (see Tab. 3.1). In relation to the grid box size and time step of the respective model, the values of D and T are equivalent to those of the ECMWF setup (cf. Buizza et al., 1999). In an absolute sense, the values D and T are considerably smaller here, since the grid box size of the LM is much finer. The configurations are termed *LLL*, *MLL*, *HLL* and *HHH*, where L denotes "low", M denotes "medium" and H denotes "high". The denotation is specified in order of amplitude, spatial auto-correlation and temporal auto-correlation.

Results of the sensitivity study are presented in Section 3.3. The crude spatio-temporal correlation of the random numbers is refined in Section 3.4. Details are given there.

Random number generator

Before concluding this subsection a brief remark should be made about the random number generator that is applied throughout the investigations. The need for a careful choice of an appropriate generator should not be underestimated (Press et al., 1992; Penland, 2003). A fairly efficient, but simple random number generator is used: the basic generator of the NAG Fortran Library Mark 19. It applies a multiplicative congruential algorithm and its period is 2^{57}. The random number generator is used to produce a series of random numbers during LM simulations which have a maximum lead time of 48 hours and a maximum of 325×325 grid boxes. As the time step of the integration is $\Delta t = 40$ seconds, the generator is invoked 4320 consecutive times on a maximum of 105625 grid boxes each. In this application the period of the generator is still ample and the use of the generator appears justified.

Another remark should be made on the implementation of the random number generator in the parallel computing environment of the LM. All random numbers are generated by a single process and are then distributed to the remaining processes. This approach does not take full advantage of the available computing resources, but ensures that the random numbers satisfy the statistical properties that are guaranteed by the generator. Parallelisation of the random number generator would run the risk of altering these properties. In order to make sure that the statistical properties are still satisfied to a

sufficient extent, the resulting numbers would have to undergo an additional statistical test.

3.2.5 Restoring energetic consistency in surface heat fluxes

Introductory Remarks

The stochastic perturbation of the parametrised tendencies in Eq. 3.2 is not energetically consistent with associated surface fluxes of heat and momentum. This fact is simply neglected in the ECMWF scheme. In order to regain energetic consistency the perturbation of parametrised tendencies should be accompanied by a surface flux correction.

This study takes a first step towards an energetically consistent scheme by addressing the surface heat flux due to solar radiation. The heat flux is adjusted in a way as to be energetically consistent with the stochastic perturbation of temperature tendencies from the radiation scheme. A *virtual* solar radiation flux is defined which cumulatively accounts for energetic discrepancies due to stochastic perturbation and compensates for them by replacing the solar heat flux of the unperturbed simulation at the surface.

Methodology

The radiation scheme of the LM (Chapter 2) separates the total radiative flux Q according to short-wave radiation fluxes Q_{sol} and long-wave radiation fluxes Q_{th}. Both fluxes are calculated by a δ-two-stream approximation of the radiative transfer equation (Ritter and Geleyn, 1992) at each vertical model level. Due to energy preservation the vertical derivative of the solar radiation flux Q_{sol} yields the solar heating rate at each model level and is proportional to the temperature tendency due to solar radiation:

$$\frac{\partial Q_{\mathrm{sol}}}{\partial z} \propto \frac{\partial T_{\mathrm{sol}}}{\partial t}. \tag{3.4}$$

The surface heat flux due to solar radiation is simply the solar radiation flux Q_{sol} at the lowest model level.

Perturbing the radiative tendency of temperature in the LM ensemble destroys the energy balance between the radiative heating rate and the vertical derivative of radiation flux Q. A *virtual* radiation flux Q^v_{sol} is defined which is energetically consistent with the perturbed simulation at each model level:

$$\frac{\partial Q^v_{\mathrm{sol}}}{\partial z} = \big\langle x(r,t) \big\rangle \cdot \frac{\partial Q_{\mathrm{sol}}}{\partial z} \tag{3.5}$$

where $\big\langle x(r,t) \big\rangle$ is the random number in the tendency perturbation (Eq. 3.2).

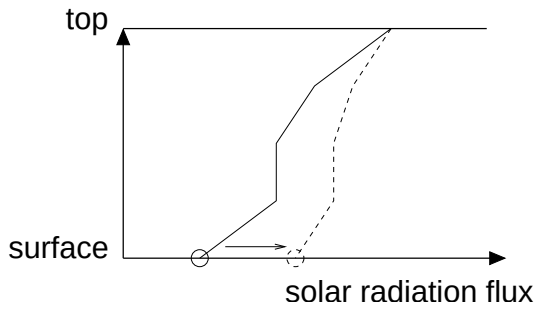

Figure 3.1: *Vertical profile of solar radiation fluxes. Schematic view. Solid line: Solar radiation flux Q_{sol} in an unperturbed LM simulation. Dashed line: Virtual radiation flux Q^v_{sol} in a perturbed LM simulation. The virtual radiation flux is consistent with the perturbed temperature tendency from the radiation scheme. In this case the random number $\langle x(r,t) \rangle$ has the value 0.5. The circles at the surface mark the corresponding surface heat fluxes due to solar radiation. The surface heat flux of the unperturbed simulation is replaced by the virtual surface heat flux so as to restore energetic consistency in terms of solar radiation fluxes.*

Let us assume that the perturbation does not affect the solar radiation flux at the top of the model atmosphere:

$$Q^v_{\mathrm{sol}} = Q_{\mathrm{sol}} \quad \text{at the top of the model atmosphere.}$$

Then the virtual radiation flux Q^v_{sol} can be inferred at any vertical model level by piecewise downward integration. In particular, a virtual surface heat flux due to solar radiation becomes available (cf. Fig. 3.1). It cumulatively accounts for energetic discrepancies due to the stochastic perturbation of $\frac{\partial T_{\mathrm{sol}}}{\partial t}$. The surface heat flux of the unperturbed simulation is replaced by the virtual surface heat flux so as to restore energetic consistency in terms of solar radiation fluxes.

A correction of the thermal radiative heat flux has not been implemented so far, but is strongly recommended for future work. Corrections of other heat fluxes would be desirable as well.

3.2.6 Perturbation of tunable parameters

Motivation

Amongst the multitude of tunable parameters α within LM parametrisation schemes, this study focusses on surface parameters and selects two examples: orographic elevation and roughness length. The surface roughness length determines the turbulent transfer rates of heat and momentum at the surface (cf. e. g. Stull, 1988). Over land surfaces the LM uses roughness length values that are constant with time.

At first glance, the idea of perturbing surface parameters might be startling, as one might argue that they are measurable quantities and not subject to uncertainty. However, an atmospheric model does not simply contain the measurable surface parameters, but a grid box mean which represents their mean effect on the atmosphere best. The parameter values are optimised with regard to their best performance and do not represent related uncertainty or sub-grid scale variability in a deterministic model simulation. Especially the optimisation of roughness length is associated with a considerable statistical spread (Duynkerke, 1992; Verhoef, 1997; Mölder, 1998). The spread is due to shortcomings of the theoretical concept, to the neglect of relevant effects such as the seasonal or diurnal cycle, and to measurement uncertainty. In addition, many spatial fields of surface parameters exhibit self-similarity. A very typical example is orographic elevation. The self-similarity makes the determination of the grid box mean difficult, no matter how small the grid box size is chosen. A small lateral shift of the entire model grid might induce large changes in the surface parameter fields.

Houtekamer et al. (1996) already implement perturbations of orographic elevation and roughness length into the assimilation cycle and forecast simulations of the global ensemble prediction system at the Canadian Meteorological Centre. Here the idea is carried over to a mesoscale forecast system.

The following subsection describes the methodology of perturbing roughness length. Experiments with perturbed orography are not followed up further, because the perturbation causes the position of vertical levels to re-adjust and thereby changes the entire physiognomy of the LM.

Perturbation methodology

As to simulate the effect of roughness length uncertainty on forecast uncertainty in the LM ensemble, perturbed fields of roughness length are produced and a different field is prescribed in each ensemble member except for the control simulation. The roughness length perturbation can be coded according to the notation of the previous subsections, as it adds a stochastic element to one of the fixed parameters α within the turbulence parametrisation:

$$\mathbf{P'}_{\text{rough}}\big(\mathbf{e}(t)\big) \equiv \mathbf{P}_{\text{rough}}\big(\mathbf{e}(t); \{\alpha \cdot \langle y(r)\rangle\}\big).$$

In this study the probability density function of the random numbers $\langle y(r)\rangle$ depends on location r only. The random numbers do *not* vary with forecast time, so it is not a stochastic scheme in the classical sense.

How are the perturbed fields of land surface roughness length derived in the LM ensemble? A probability density function of roughness length is postulated at each grid box and used to draw realisations of roughness length for each ensemble member. The differences between the realisations should ideally represent the uncertainty of roughness length or the stochastic effect of sub-grid scale variations of roughness length. For the

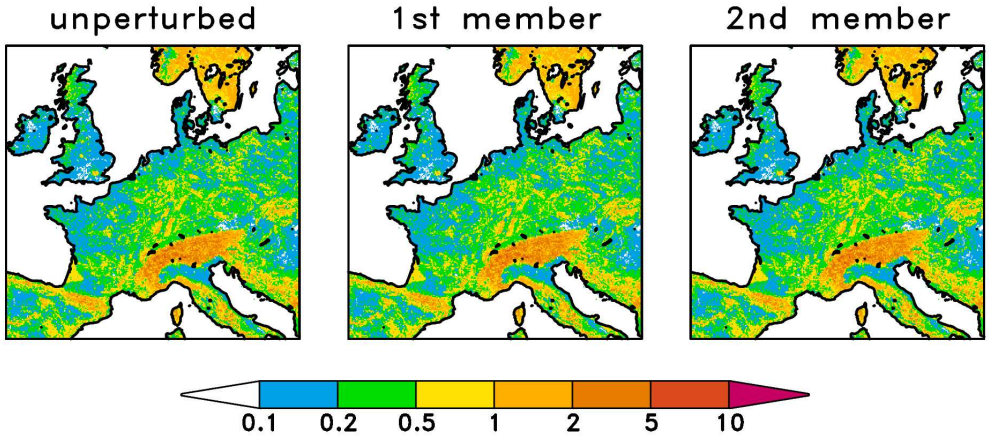

Figure 3.2: *Left: The unperturbed field of LM roughness length. Middle and right: Examples of perturbed roughness length fields in the LM ensemble. Unit: meter. Large-scale characteristics such as big cities or mountain ranges are identical in each perturbed field.*

sake of simplicity, the perturbations are not auto-correlated in space. Such a refinement remains for future research.

The LM contains a roughness length for momentum exchange and a roughness length for heat exchange. In this study, the roughness length for momentum exchange is explicitly perturbed. This also affects the roughness length for heat exchange, because the two quantities are closely related in the model design of the LM. If the roughness length for momentum exchange is smaller than 0.1 m the LM applies this value to heat exchange as well, otherwise a constant value of 0.1 m is chosen for heat exchange.

It would be straightforward to derive the probability density function directly from the optimisation procedure of roughness length. However, details of the optimisation procedure are not available, so several assumptions are made instead:

- The amplitudes of the perturbations depend on the local spatial variability of the model roughness length on the resolved scale.
- They are small enough to preserve large-scale features, such as big cities or mountain ranges.
- The median of the perturbed roughness lengths at a certain location is equal to the unperturbed roughness length at that location.

Roughness length uncertainty is deduced from grid-scale information that is not directly related to roughness length uncertainty in a physical sense, i. e. roughness length uncertainty is basically *parametrised*. The parametrisation of roughness length variability is developed empirically and checked by statistical and eye-ball inspection of the resulting fields. Examples of perturbed fields are shown in Fig. 3.2. As intended, large-scale characteristics such as big cities or mountain ranges are identical in each perturbed field. The

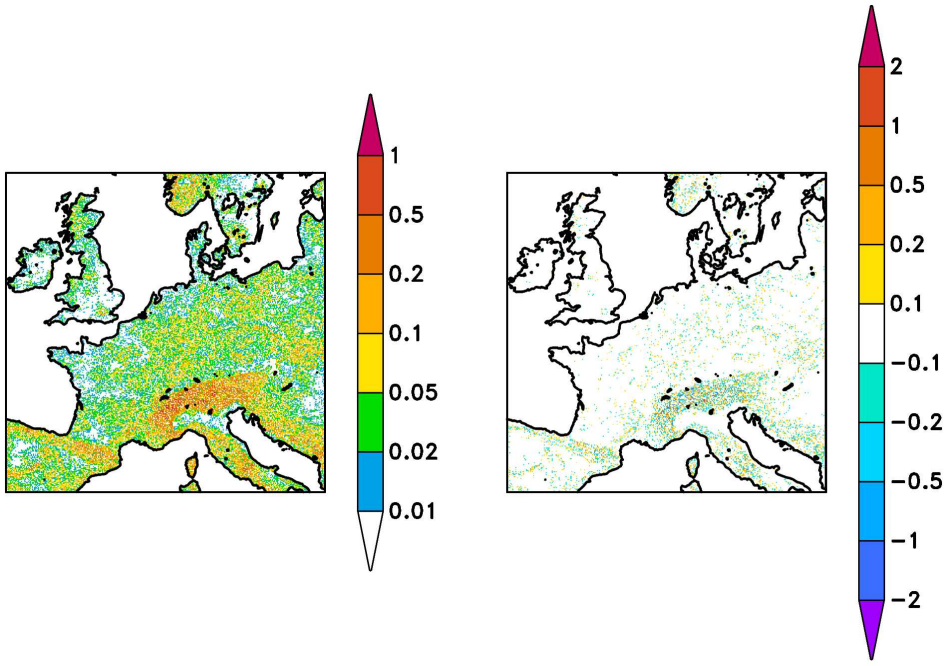

Figure 3.3: *Left: Empirical standard deviation between the ten roughness length fields which are prescribed to the ensemble members, respectively. Unit: meter. Right: Difference between the empirical mean of the ten roughness length fields and the unperturbed roughness length field. Unit: meter.*

actual amount of perturbation can be seen in the standard deviation between the ten roughness length fields which are prescribed to the ensemble members, respectively (left of Fig. 3.3). In many areas the perturbation amplitude turns out to be roughly proportional to the unperturbed roughness length and amounts to 5%–10% of the unperturbed values. The perturbation amount is still conservative compared to the uncertainties encountered in the derivation of roughness length values (Duynkerke, 1992; Verhoef, 1997; Mölder, 1998).

As the perturbation methodology preserves the median instead of the mean, the difference between the empirical mean of the ten roughness length fields and the unperturbed roughness length field is checked in addition (right of Fig. 3.3). The difference turns out to be sufficiently small.

3.3 Ensemble experiments for the case study *Berlin*

A variety of ensemble experiments is carried out. First experiments of this study were confined to the perturbation of roughness length. Five case studies of severe precipitation were investigated: 4 July 1994; 5 May 2001; 10 July 2002; 21 May 1999; 25 Dec 1999. The experiments showed that the perturbation of roughness length has a considerable effect on the small-scale quantitative precipitation forecast.

More recent experiments test the stochastic parametrisation scheme. Two case studies are investigated: 4 July 1994 and 10 July 2002. Both case studies are characterised by intense convective precipitation. In this section and the following, the case study of 10 July 2002 is presented as an example. It is named *Berlin*, because strong wind gusts led to severe damages and casualties in the city of Berlin on that day.

First, the experimental setup is described in Subsection 3.3.1. Subsection 3.3.2 shows the effect of the stochastic parametrisation scheme and Subsection 3.3.3 compares it to the impact of initial conditions. As the multiple ensemble forecasts demand a large amount of computer capacities, these investigations are confined to one case study. A more extensive data set is produced in Section 3.4 which also forms the basis for a comparison to observational data in Section 6.7.

3.3.1 Configuration of the ensemble experiments

Starting time of the ensemble forecast is 00 UTC, July 10, 2002. Each ensemble comprises 10 members, consisting of 9 perturbed simulations and the control simulation. The lateral boundary conditions are provided by LM analyses. The model domain is reduced to 177×177 grid points and is centred over Germany. Version 2.14 of the LM is used. Details of the model such as the grid box size are presented in Chapter 2.

The stochastic parametrisation scheme described in Section 3.2 is used to add a stochastic element into the deterministic model simulations. The ensemble forecast is produced four times. Each ensemble employs a different parameter setting of the stochastic parametrisation scheme (Tab. 3.1). In addition to the perturbation of diabatic forcing, the roughness length is perturbed. The perturbation methodology is the same in each of the four ensembles and is described in the previous section.

Ensembles of this experimental setting are presented in the following Subsection 3.3.2. Results reveal the overall impact of the stochastic parametrisation scheme and its sensitivity to the configuration of the scheme.

Subsection 3.3.3 compares the impact of stochastic parametrisation to the impact of initial conditions. A "poor man's method" of perturbing initial conditions is developed and further ensemble experiments are produced.

3.3.2 Impact of the stochastic parametrisation scheme

Fig. 3.4 shows the precipitation forecast of the control simulation. The precipitation amount is a combination of convective and grid scale precipitation. With the exception of precipitation amounts of approximately 5 mm/h and higher, the coherent precipitation field in northern Germany and southern Denmark is entirely caused by grid scale precipitation. The patterns are moving northeastwards. The western part of the precip-

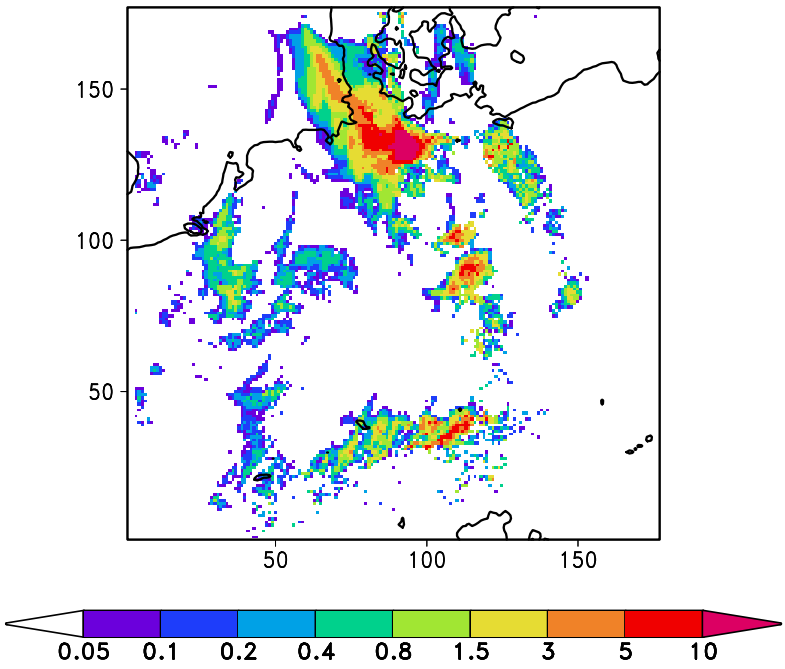

Figure 3.4: *1h-accumulations of precipitation (mm/h) at 17–18 UTC on July 10, 2002. Precipitation forecast of the unperturbed simulation. Starting time of the forecast is 00 UTC, July 10, 2002.*

itation near the Alps is also due to grid scale precipitation. All remaining precipitation fields are generated by the convection scheme.

Results of the *LLL* and the *HHH* configurations are presented. The standard deviation between the ensemble members estimates the impact of the perturbations on the precipitation forecast (Fig. 3.5). The standard deviation is also called *ensemble spread*. In the *LLL* ensemble the impact of the perturbations is fairly low. The standard deviation reaches approximately 30% of the precipitation amounts in the unperturbed forecast (Fig. 3.4). The perturbations only affect the smallest spatial scales of the forecast. Occurrence and non-occurrence of precipitation are hardly affected.

The *HHH* ensemble (right of Fig. 3.5) exhibits a much larger impact of the stochastic parametrisation scheme. The ensemble spread is often as high as the precipitation amount in the unperturbed forecast (Fig. 3.4). Compared to the *LLL* ensemble the perturbations affect slightly larger scales of the forecast. In some locations the structure of convective precipitation patterns is changed, but the broad pattern of precipitating and non-precipitating regions is still hardly affected. The enhanced impact of the perturbations results both from increasing the amplitude and from increasing the spatio-temporal auto-correlation.

In summary, the impact of stochastic parametrisation on the LM precipitation forecast can be substantial, but strongly depends on the amplitude and the spatio-temporal auto-

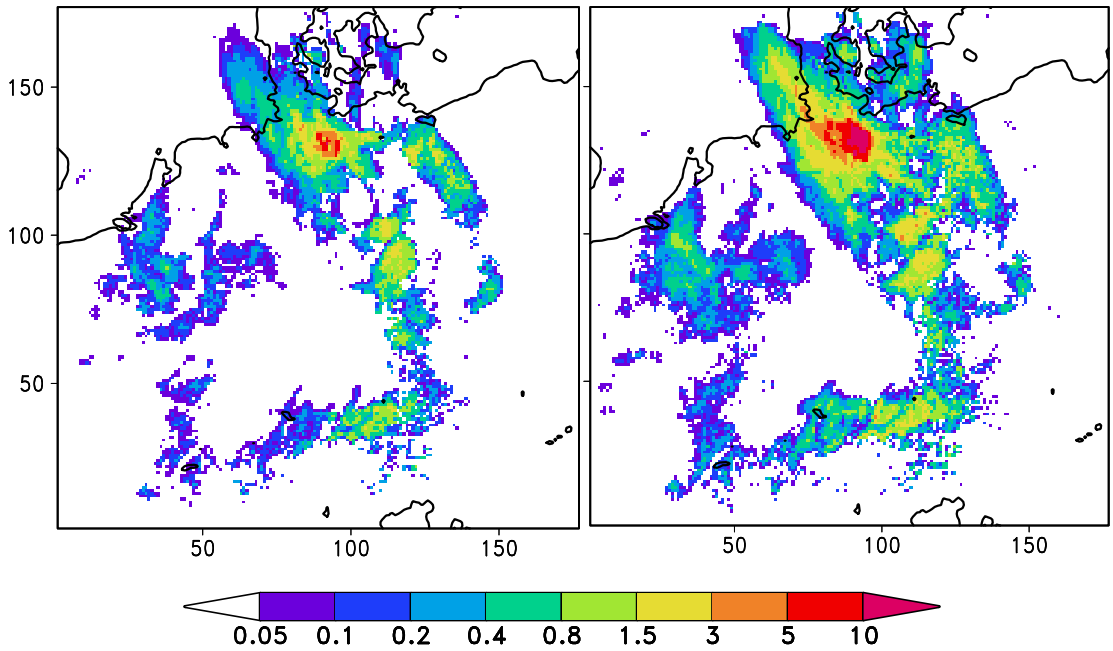

Figure 3.5: *1h-accumulations of precipitation (mm/h) at 17–18 UTC on July 10, 2002. Starting time of the forecast is 00 UTC, July 10, 2002. The figure shows the ensemble spread, i. e. the standard deviation between the precipitation forecasts of the ensemble members. The two ensembles differ by their configurations of stochastic parametrisation (cf. Tab. 3.1). Left:* LLL *configuration. Right:* HHH *configuration.*

correlation of the random numbers. However, all configurations show that the effect is restricted to small scales of the forecast field and to quantitative precipitation. The stochastic parametrisation scheme barely influences occurrences and non-occurrences of precipitation. Therefore, it must be suspected that the experimental ensemble captures only a small portion of the total error in a precipitation forecast. This suspicion is reaffirmed by a comparison to observational data in Section 6.7.

3.3.3 Impact of initial conditions

In order to assess the relative impact of stochastic parametrisation compared to the impact of initial conditions, a "poor man's method" of perturbing initial conditions is developed. The "poor man's method" is expected to give only a crude insight into the impact of initial conditions on the forecast.

The LM simulation is initialised by an LM-analysis of a slightly shifted time. Such an analysis is available from the nudged LM assimilation run for any point of time. At present, the LM analysis of 01 UTC and the analysis of 23 UTC of the previous day are used, while the LM simulation is actually started at 00 UTC. The various analyses are

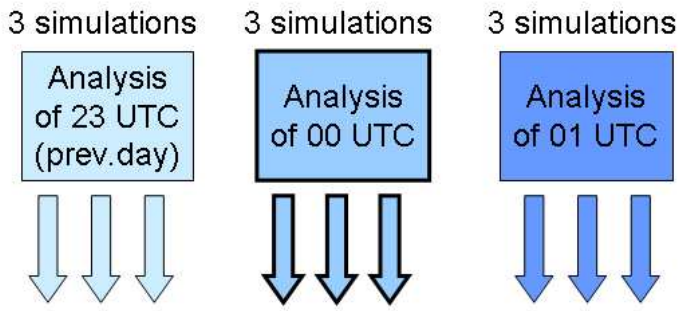

Figure 3.6: *Setup of the ensemble including perturbations of initial conditions. Three simulations per initial condition are started, differing by the stochastic parametrisation and roughness length. Together with the control simulation, the ensemble consists of 10 members.*

considered equally likely initial conditions at starting time 00 UTC. Together with the analysis of 00 UTC, the LM analyses of 23 UTC and 01 UTC constitute three different initial conditions. The time difference of one hour is selected for convenience, as these analyses are readily available from the data base. A somewhat shorter time step may be more reasonable and remains as an issue for future work. Likewise, the number of different initial condition fields should be raised above the small number of three. A simple way to achieve this might consist in averaging or *mirroring* the initial conditions.

The new ensemble setup combines the variation of initial conditions with the previous ensemble setup. An outline of the methodology is given in Fig. 3.6. From each of the three initial conditions, three simulations are started. They differ by the stochastic parametrisation and the roughness length. In the stochastic parametrisation scheme the *LLL* configuration is used (see Tab. 3.1). Altogether the ensemble comprises 9 perturbed simulations plus the control simulation.

Fig. 3.7 presents the ensemble spread of the precipitation forecast. The ensemble spread is about as large as in the *HHH* ensemble (right of Fig. 3.5), but the effect of initial condition perturbations is very different from the effect of the stochastic parametrisation scheme. The ensemble forecasts cluster strongly by the choice of initial conditions (not shown). The clustering causes the three maxima in the field of ensemble spread (Fig. 3.7). The initial conditions of 23 UTC result in a southward shift of the original maximum (Fig. 3.5) by 100 km, while the initial conditions of 01 UTC result in a northward shift of the maximum by 100 km. Whereas the stochastic parametrisation scheme mainly affects scales close to the grid box size, initial conditions affect scales up to 200 km. This becomes obvious through a scale-dependent calculation of ensemble spread (not shown) or simply by looking at individual ensemble members (not shown).

In summary, the magnitude of the ensemble spread suggests that initial conditions and the stochastic parametrisation scheme have a similar effect when the stochastic parametrisation scheme applies a large noise amplitude and large spatio-temporal cor-

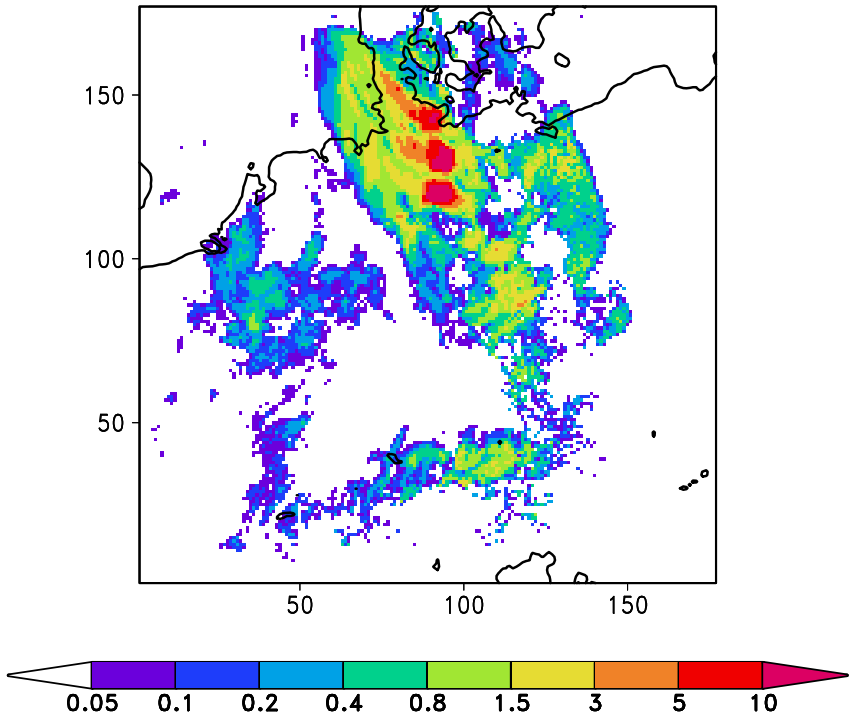

Figure 3.7: *1h-accumulations of precipitation (mm/h) at 17–18 UTC on July 10, 2002. Starting time of the forecast is 00 UTC, July 10, 2002. The figure shows the ensemble spread, i. e. the standard deviation between the precipitation forecasts of the ensemble members. In addition to perturbations of roughness length and stochastic parametrisation, the initial conditions are varied according to Fig. 3.6.*

relations. However, the initial conditions affect much larger scales of the forecast than the stochastic parametrisation scheme. As opposed to the stochastic parametrisation scheme, the perturbation of initial conditions is able to shift patterns of precipitation and it affects the occurrence of precipitation instead of precipitation amount only. Therefore, the initial condition perturbations are suspected to capture a substantially larger portion of total forecast uncertainty.

Note, however, that this investigation is only expected to give a very crude insight into the relative effects of the stochastic scheme and of initial condition perturbations. A direct comparison between the two effects may be unfair, because the intensities of the respective perturbations may not agree at all. The initial condition perturbations probably possess a much larger magnitude and a much larger spatio-temporal autocorrelation than the perturbations within the stochastic parametrisation scheme.

3.4 Ensemble experiments with more extensive data

The validity of the findings in the previous section are confined to just one case study. In an attempt to carry out a statistically more robust analysis and to produce a forecast data set suitable for verification in Section 6.7, consecutive ensemble forecasts are produced for a period of two weeks. First, the experimental setup of the ensemble is described in Subsection 3.4.1 and then the effect of the stochastic parametrisation scheme is analysed in Subsection 3.4.2.

3.4.1 Configuration of the ensemble experiments

General design

On each day of the time period July 10–24, 2002, an ensemble forecast is started at 00 UTC. The 15 ensemble forecasts comprise 10 members each, consisting of 9 perturbed simulations and one control simulation. All forecasts have a lead time of 48 hours.

The configuration of the experimental ensemble is similar to Subsection 3.3.2. The perturbation of ensemble members consists in the stochastic parametrisation scheme and a perturbed roughness length. Perturbations of initial conditions are omitted here, so as to look at the effect of the stochastic scheme in isolation. Again, the lateral boundary conditions are provided by LM analyses.

In contrast to the experiments in Subsection 3.3.2, a more recent version of the LM (version 2.19) is used and the operational model domain of 325×325 grid points is chosen. Thus, the experimental configuration is closer to the operational setting of DWD so that the results bear more implications for operational forecasting at DWD. Using the operational domain increases the computational cost of the simulations substantially, but appears necessary, as changing the domain size can modulate the intrinsic unstable dynamics substantially (Chomé et al., 2002).

Furthermore, the specification of random numbers in the stochastic parametrisation scheme differs from the configuration in Subsection 3.3.2. The random numbers used so far have a crude spatio-temporal auto-correlation through their constancy within boxes and time intervals. An attempt is made to establish a more sophisticated auto-correlation. As the simulation of 15 ensembles comprising 10 members each demands a considerable amount of computing time, the ensemble forecasts are produced only once. One specific configuration of input noise must be chosen which is described in the following paragraph.

Amplitude and spatio-temporal auto-correlation of input noise

For the sake of simplicity, it is decided to retain the uniform distribution. The amplitude a is chosen to be 0.25. In terms of turbulence, the order-of-magnitude is in some way supported by Högström (1988) who estimate observational uncertainties of the dimensionless wind shear and lapse rate. In terms of radiation, some support may be drawn from Scheirer and Macke (2001) who investigate radiative flux uncertainties in a high-resolution model, arising from the simplification of cloud geometries as plane-parallel homogeneous layers.

The simple spatio-temporal auto-correlation function of Subsection 3.2.4 is refined. The spatio-temporal auto-correlation function in Subsection 3.2.4 assumes that a unique spatial scale D and temporal scale T act as separators of perfect and zero auto-correlation between sub-grid scale effects. Such scales are difficult to determine, because their existence is far from reality. The approach is simplistic and needs refinement. This is also an issue of current research at ECMWF. Palmer (2001) proposes more dynamically based techniques such as stochastic forcing in singular vector space or a cellular automaton model. Shutts (2004) has recently used a cellular automaton to generate a forcing field in his stochastic kinetic energy backscatter algorithm.

In this study, the simplistic auto-correlation function is replaced by an exponentially decaying function. Bright and Mullen (2002) use a similar approach for their simulation of convective sub-grid scale effects. As a first step, the auto-correlation is confined to the time direction. Just like Bright and Mullen (2002), this study provisionally omits the spatial auto-correlation of stochastic noise and leaves its implementation to future research.

The refined auto-correlation function results from a first-order auto-regression model (Wilks, 1995):

$$y_t = \eta \cdot y_{t-1} + \epsilon_t, \quad \text{with} \quad y_0 = \epsilon_0,$$

where $t = 1, \ldots, n$ counts the number of time steps during the simulation, ϵ_t is a realisation of a Gaussian white noise process and $0 < \eta < 1$. Consequently, the random number y_t is also Gaussian and the random variables y_t and $y_{t-\tau}$ at time steps t and $t - \tau$ are correlated by η^τ. The e-folding time of the auto-correlation is $T = -\frac{1}{\ln\eta}$. Its value is chosen to be $T = 5$ minutes so as to be of the same order-of-magnitude than the time auto-correlation in the sensitivity experiments of the previous section.

In order to regain a uniform distribution of the random numbers, the Gaussian random number y_t is transformed to a random number x_t, before it enters the stochastic parametrisation scheme. The random number x_t is uniformly distributed in the interval $[1 - a, 1 + a]$ $(0 < a \leq 1)$. An excerpt of the random number series applied in the LM stochastic parametrisation scheme is shown in Fig. 3.8.

The random variable x_t is derived by the inverse of a logit transformation. Generally, the inverse of a logit transformation can be used to convert a Gaussian sequence of random numbers into a sequence of realisations that are approximately distributed according

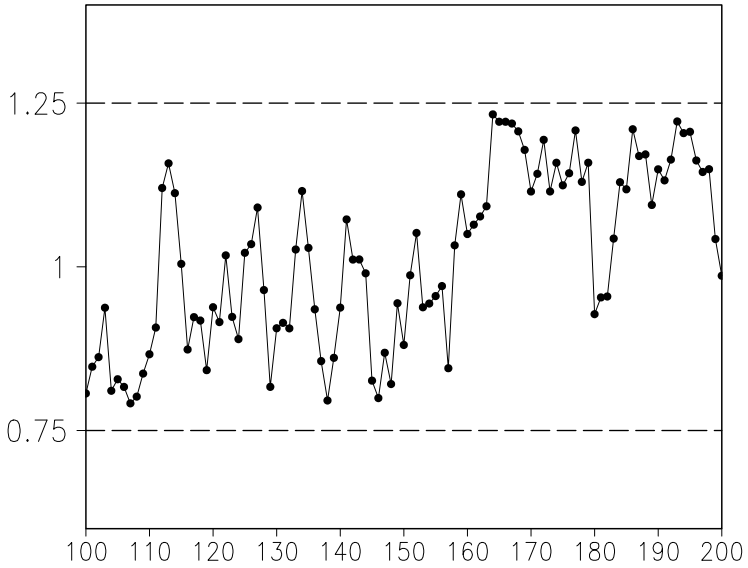

Figure 3.8: *Excerpt of the random number series applied in the LM stochastic parametrisation scheme. Abscissa: Numbering of the time steps in the model integration.*

to a beta-distribution. This property is also useful with respect to the generation of uniformly distributed numbers, because the uniform distribution is a special case of the beta-distribution.

However, the transformation results in a distortion of the correlation in x_t. The correlation between realisations with a value close to $(1 + a)$ or $(1 - a)$ is larger than the correlation between realisations with a value close to 1. A more sophisticated transformation should avoid this side-effect. Alternatively, a future study could circumvent the problem by letting the random numbers follow a Gaussian distribution. Initially, the uniform distribution was motivated by its simplicity, but now it turns out that the uniform distribution introduces more complexity than the Gaussian distribution.

Case study

Before handing over to the statistical analysis of the entire forecast data set, let us have a look at a single case, as the previous paragraph introduces some alterations in the configuration of the stochastic scheme compared to the sensitivity study described in Subsection 3.3. Fig. 3.9 shows the ensemble mean and the ensemble spread. The effect of the stochastic parametrisation scheme is similar to the effect of the LLL configuration (Fig. 3.5). Note that a straight comparison is not possible, because there are also changes in model version, model domain and the starting time of the forecast which already cause differences in the respective unperturbed forecasts. These differences become obvious when comparing the original forecast of the previous setup (Fig. 3.4) with the ensemble

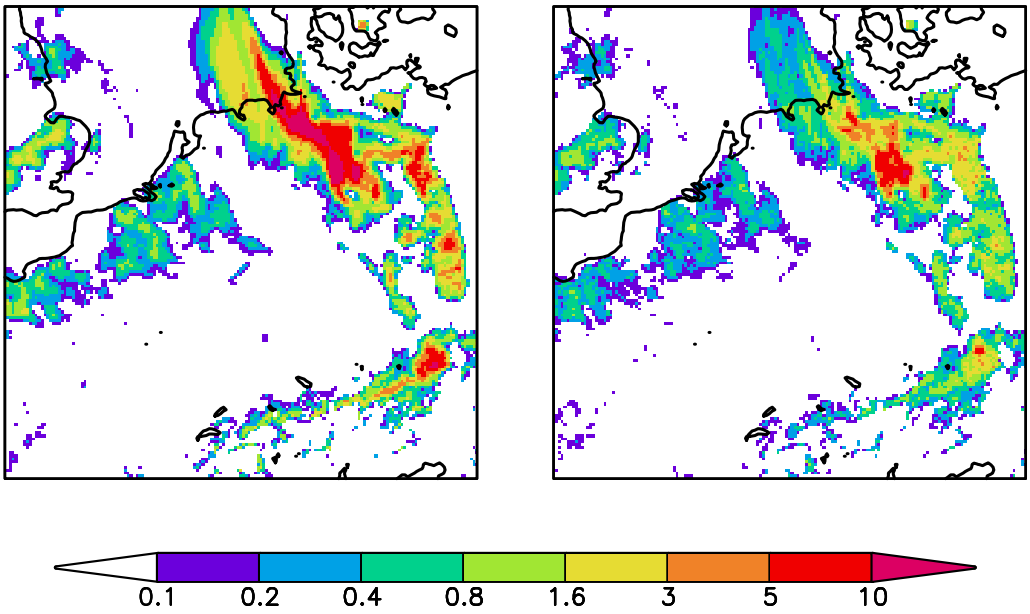

Figure 3.9: *1h-accumulations of precipitation (mm/h) at 17–18 UTC on July 10, 2002. Starting time of the forecast is 00 UTC, July 9, 2002. A section of 165 × 165 grid points is shown. Left: Ensemble mean. Right: Ensemble spread, i. e. standard deviation between the precipitation forecasts of the ensemble members. As opposed to Fig. 3.5 and 3.7, a new ensemble configuration is used which is described in Subsection 3.4.1.*

mean of the new setup (left of Fig. 3.9).

3.4.2 Statistical analysis of error growth

An interesting issue in ensemble forecasting are the characteristics of error growth. Do forecast perturbations keep growing with time or do they reach a level of saturation? Is there an upscale transformation from small-scale to large-scale error?

Fig. 3.10 shows the normalised ensemble spread as a function of forecast lead time. The normalised ensemble spread is the ratio of ensemble spread and ensemble mean. The spread is normalised because the spread itself strongly depends on the values of the mean. It tends to be larger for higher precipitation values, because the probability distribution of precipitation is skewed towards higher values. Normalisation reduces this dependency, but the diurnal cycle of the mean is still visible in the normalised spread. Note that the decay of normalised spread during the afternoon does not indicate a recovery of predictability, but is simply a result of the diurnal cycle of the mean.

In Fig. 3.10 only cases are taken into account with an ensemble mean greater than 0.01 mm. The spread is averaged over the domain shown in Fig. 3.9 and then averaged

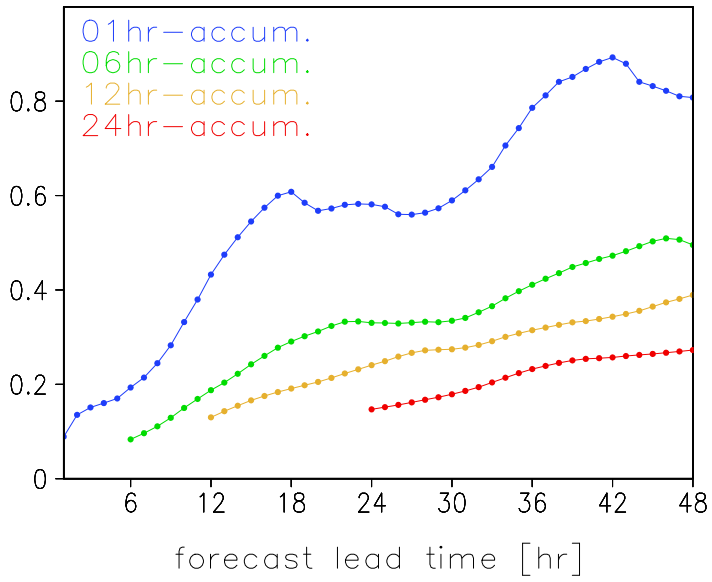

Figure 3.10: *Time series of the ensemble spread, normalised by the ensemble mean. The abscissa shows the lead time of the forecast in hours. Only cases are taken into account with an ensemble mean greater than 0.01 mm. The spread is averaged over the area shown in Fig. 3.9 and then averaged over all 15 ensemble forecasts within the time period 10–24 July 2002.*

over all 15 ensembles. The ensemble spread increases throughout the simulation and does not appear to reach a saturation level. This does not only become obvious in the averaged data, but also in each of the 15 individual ensemble simulations (not shown).

The differences between a perturbed and the original simulation are not only analysed in terms of their amplitude, but also in terms of their spatio-temporal coherence. Fig. 3.11 shows an example of the difference between a perturbed and the original simulation. Some spatial auto-correlation of the differences are already noticeable by eye-ball inspection.

A statistical analysis of spatio-temporal error correlation is carried out and its results are shown in Fig. 3.12. Forecasts with an ensemble mean greater than 0.01 mm and with a lead time of 25–48 hours enter the calculation. Data of the domain shown in Fig. 3.11 is used, from all 15 forecasts with a starting date within the period of 10–24 July, 2002 and from the 9 perturbed simulations within the ensemble. The results reveal that spatial auto-correlation drops down to zero for spatial distances larger than 30 km. Temporal auto-correlation is still existent for a time lag of one hour. A time lag of one hour is quite large compared to the temporal auto-correlation function of the prescribed noise which possesses an e-folding time of only 5 minutes. This might indeed indicate an upscale error propagation in terms of time scales.

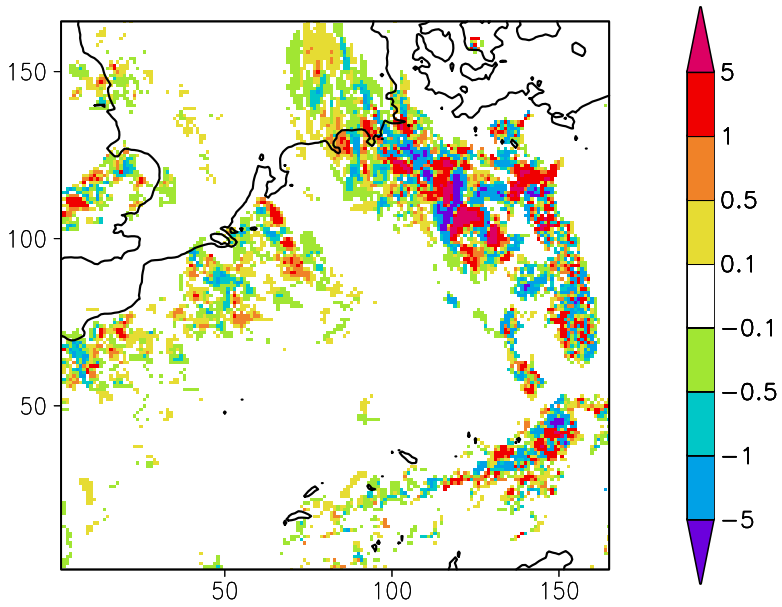

Figure 3.11: *1h-accumulations of precipitation (mm/h) at 17–18 UTC on July 10, 2002. Differences between the unperturbed simulation and one of the perturbed simulations.*

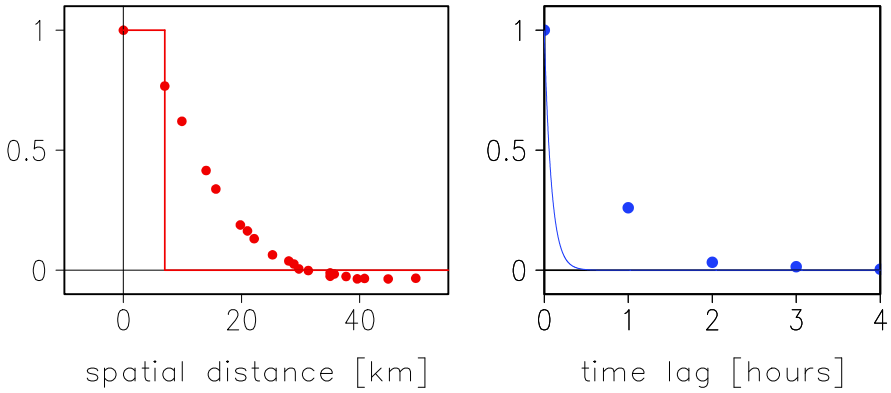

Figure 3.12: *The closed circles indicate the spatial and temporal auto-correlation of the differences between the original simulations and perturbed simulations (cf. Fig. 3.11). Left: Auto-correlation in space. Right: Auto-correlation in time. The auto-correlation is shown as a function of horizontal distance and time lag, respectively. For comparison, the auto-correlation of the input noise is visualised by the solid lines (red and blue). The closed circles are estimated from all 15 ensemble forecasts and from the 9 perturbed simulations within each ensemble forecast. Forecasts with an ensemble mean greater than 0.01 mm and with a lead time of 25–48 hours enter the calculation. Data of the domain shown in Fig. 3.11 are used.*

3.5 Discussion and summary

3.5.1 Possible explanations of the small effect

The previous sections have shown that the LM stochastic parametrisation scheme has only little effect on LM forecasts when compared to the total uncertainty of the forecast. There is a considerable influence on the predicted precipitation amount on small scales close to grid box size, but the stochastic parametrisation scheme does not change the occurrence of precipitation and it does not shift the position of precipitation patterns. The experimental ensemble misses a substantial portion of total forecast uncertainty. This is reaffirmed by objective verification which attests to extreme underdispersiveness of the experimental ensemble (Section 6.7).

Two explanations are conceivable:

- Sub-grid scale variability is not an important player in forecast uncertainty.
- The stochastic parametrisation scheme does not produce a realistic simulation of sub-grid scale variability.

If we trust the LM stochastic parametrisation scheme, the first conclusion is valid. Similar investigations show that their parametrisation perturbations also have extremely weak effects (Anderson et al., 2002; Bright and Mullen, 2002), so it can be argued that they support this conclusion.

However, some doubts about the abilities of current stochastic parametrisation schemes might be justified. A multitude of reasons may be held responsible for an unrealistic simulation of sub-grid scale processes, ranging from implementation-related problems to very fundamental problems, such as the applicability of the central limit theorem (Penland, 2003) or the ability of the model to represent certain states of the atmosphere realistically (Smith, 2000, 2003). In the following subsection, problems related to the implementation of stochastic parametrisation schemes are discussed. Two key issues are highlighted: properties of the injected noise and the role of the integration scheme.

3.5.2 Possible problems in the implementation

Properties of the prescribed noise

The appropriate choice of perturbation amplitude and correlation remain an open question. As only little is known about sub-grid scale variability, the choice of noise properties (Section 3.2.4) is somewhat arbitrary. At the same time, Section 3.3 points out that the simulated effect is very sensitive to the chosen noise properties, so the outcome of stochastic schemes may be somewhat arbitrary as well.

Larger noise amplitudes and larger noise correlations attain larger effects on the forecast. Thus, a diffident attitude towards amplitude and correlation might dilute the inherent power of the LM stochastic parametrisation scheme. For comparison, Bright and Mullen (2002) choose an e-folding time of $T = 164$ minutes for their time auto-correlation of convective sub-grid scale effects as opposed to $T = 5$ minutes in the auto-correlation of the LM scheme. A somewhat shorter value does seem to be justified for the LM, as the area of a LM grid box is about 25 times smaller than in their model. Nevertheless, the choice of $T = 5$ minutes may still be on the short side. Shutts (2004) argues that small correlation scales tend to generate high frequency gravity waves whose impact on weather system evolution is only slight.

One way out of this problem is tuning, given that the computational environment allows for the operating expenses. The tuning procedure alters the configuration of the stochastic parametrisation scheme until the experimental ensemble matches observed forecast uncertainties best (cf. Buizza et al., 1999). As tuning is certainly of immediate practical use, it would be very interesting to find a way of minimising its technical expenses so as to widen its applicability. However, tuning also has some drawbacks. The more tuning parameters are added to a simulation system, the more shaky its theoretical foundation gets. Caution has to be taken not to end up in headless action. Improvements might be achieved for wrong theoretical reasons. Subsequent alterations of a different part of the simulation system might actually degrade the results, although they really point into the right direction.

There is yet another possibility to circumvent the need for an *a priori* selection of a particular spatio-temporal auto-correlation of the input noise. Along the lines of Palmer (2001), the general design of the stochastic scheme could move away from grid point forcing towards a dynamical approach. Then the coupling between the stochastic system and the resolved processes extends over a range of scales. These scales are determined by certain modes that result from a dynamical model. Palmer (2001) suggests two examples: the application of singular vector modes and the cellular automaton model.

Numerical integration scheme

There is an important aspect in stochastic parametrisation that has not been touched upon in this study so far: the numerical integration scheme. The operational version of the LM does not contain any stochastic elements. Thus, its numerical scheme is designed to solve deterministic differential equations with a sufficient degree of accuracy and a minimum of computational cost. By introducing stochastic elements into the LM parametrisation schemes the model equations are no longer deterministic, but stochastic. A standard algorithm for integrating deterministic differential equations is not properly sanctioned for the integration of stochastic differential equations. Simply adding a stochastic term to existing computer code may yield results that look reasonable, and yet be completely spurious. Ewald et al. (2003) state that "the choice of integration technique can greatly affect the outcome of a stochastically forced numerical

model, and the variety of results possible in different implementations of a single model is frightening".

The impact of the numerical integration scheme is only starting to be recognised in stochastic modeling of weather and climate. Although Kloeden and Platen (1992) present a collection of well-researched stochastic integration schemes, most meteorological implementations of stochasticity have not applied them yet (e. g. Buizza et al., 1999; Lin and Neelin, 2000; Bright and Mullen, 2002), as the adequate rewriting of complex numerical models is an extensive task. Only recently, a numerical scheme has been developed by which stochastic forcing can be introduced into one of the simplest meteorologically relevant numerical models, the barotropic vorticity model (see Ewald et al., 2003, and references within).

As the numerical integration scheme appears to have such a tremendous impact, current experiments with complex stochastic models should be extremely careful in their interpretation, if not pause for a moment. It may be appropriate to take a step back and invest in research activities that aim at adjusting the numerical scheme first.

3.5.3 Summary

This study is one of the first to systematically analyse the benefits of stochastic parametrisation in operational short-range forecasting. The implementation of stochasticity in short-range ensemble forecasting is on the agenda of several operational forecast centres and high expectations are raised. In contrast, only little experience has been gained using complex weather forecasting models. Thus, this study provides a significant advancement of current research.

Initially, this study envisaged the experimental LM ensemble to serve as a guidance in the development of statistical post-processing procedures. However, the ensemble requires a specific configuration of input perturbations, which has a considerable impact on the ensemble results. Therefore, this configuration must be justified in a theoretical or empirical way. In the absence of such justification, the choice of input perturbations remains largely open leading to results of the experimental LM ensemble that remain arbitrary to a large extent. As a consequence, it is elected to shelve this initial plan until the ensemble can settle for a specific configuration of input perturbations.

Apart from the context of statistical post-processing, the study provides useful guidance for the future development of stochastic parametrisation schemes in ensemble prediction systems. The study raises relevant questions which need to be tackled first before stochastic parametrisation may produce more credible results and provide better insight into small-scale predictability. Future fields of research should not only include the specific configuration of input perturbations, but also a reconsideration of the numerical integration scheme.

4 Neighbourhood methodology

While the previous chapter deals with the ensemble approach, this chapter presents a different approach to probabilistic forecasting: statistical post-processing of a *single* deterministic LM simulation. In this study, statistical post-processing offers a low-budget approach to probabilistic forecasting that is suitable for almost immediate operational use and requires only little computing effort compared to ensemble forecasting. A brief overview of the method and an outline of this chapter are given below.

4.1 Introductory remarks

4.1.1 Overview

The statistical post-processing scheme attempts to *parametrise* forecast uncertainty, in the sense that it does not trace forecast uncertainty to its origins, but uses ancillary indicators for its estimation. In other words, forecast dependence on initial conditions and on model physics is not taken into account. Instead, the model output is perceived as a function of space and time only. The method relies on the spatio-temporal structure of the forecast which may give some information on deterministic predictability.

It is certainly a bold attempt to base predictability estimates on the spatio-temporal structure of one deterministic simulation. However, a few generic characteristics of deterministically predictable structures are vaguely known, for example the scale-dependency of predictability or the enhancement of predictability through topographic forcing. The post-processing scheme attempts to take advantage of such knowledge.

The notion of a so-called *neighbourhood* is central to the post-processing methodology in this chapter. The post-processing scheme looks in the spatio-temporal neighbourhood of a point to get a set of forecasts and assumes that the set is a sample of the forecast at the central point of the neighbourhood. The neighbourhood sample is then used to derive a probabilistic forecast at this grid point. The procedure is repeated for each grid point of the LM so that an entire field of probabilistic forecasts is generated.

The neighbourhood sample is also called *surrogate sample*, implying that it is a surrogate for a dynamically produced ensemble. The surrogate ensemble serves as a starting point for the estimation of the expected value, quantiles and exceedance probabilities. The estimation of exceedance probabilities and of the expected value simply applies the

respective maximum-likelihood estimators. Due to the risk of a large variance error in quantile estimates, quantile estimation uses a more complex methodology which consists in a kernel regression of an empirical quantile function.

The estimation procedures of this study are non-parametric. The term *non-parametric* delimits such procedures from *parametric* approaches. Parametric estimation makes a rigid pre-assumption which confines the solution space of the resulting *pdf* to a certain family of functions. Within such a family, functions only differ by the specification of a few parameters. Parametric methods are reduced to the direct estimation of these parameters. Estimates of the *pdf* or certain statistical quantities are then inferred from the estimated parameters. In non-parametric estimation the resulting estimate is more freely guided by the data. Certain pre-assumptions about a *pdf* are still allowed, but they do not pre-determine the estimate to such a large extent. This is an advantage here, because the shape of the *pdf* is often not known. Thus, parametric methodologies can be inappropriate, as they may force the estimate to go awry.

4.1.2 Why another methodology?

Existing non-parametric methodologies

The neighbourhood method estimates statistical parameters at each point of the model grid on the basis of a single model simulation. Thus, the approach falls under the framework of estimating a large array of statistical parameters on the basis of a single realisation on each parameter, a recently recurring theme in geostatistics. Approaches to non-parametric estimation already exist and are briefly mentioned here. A more detailed presentation follows in Section 4.3.

The neighbourhood concept bears similarities to two prominent methodologies: *kernel smoothing* and *kriging*. In the last 40 years, kernel smoothing has been a standard methodology in non-parametric density estimation and in non-parametric regression (Silverman, 1986; Härdle, 1990; Simonoff, 1996). Kriging is the central methodology in geostatistics, a field that emerged in the 1960s with the work of Georges Matheron (Wackernagel, 1998; Chilès and Delfiner, 1999).

Kernel smoothing and kriging are able to derive estimates of the expected value and the exceedance probability. A quick overview is given by Fig. 4.1 and 4.2. More extensive explanations follow in Section 4.3.

In summary, kriging and kernel smoothing mainly consist in weighted averaging of realisations that are available at different locations. The key element of the methods is the determination of appropriate spatio-temporal weights. In order to estimate the exceedance probability, the realisations can be transformed to *indicator data* before averaging. The transformation function is valued 1 if the realisation is larger than the exceedance threshold, and valued 0 if it is not.

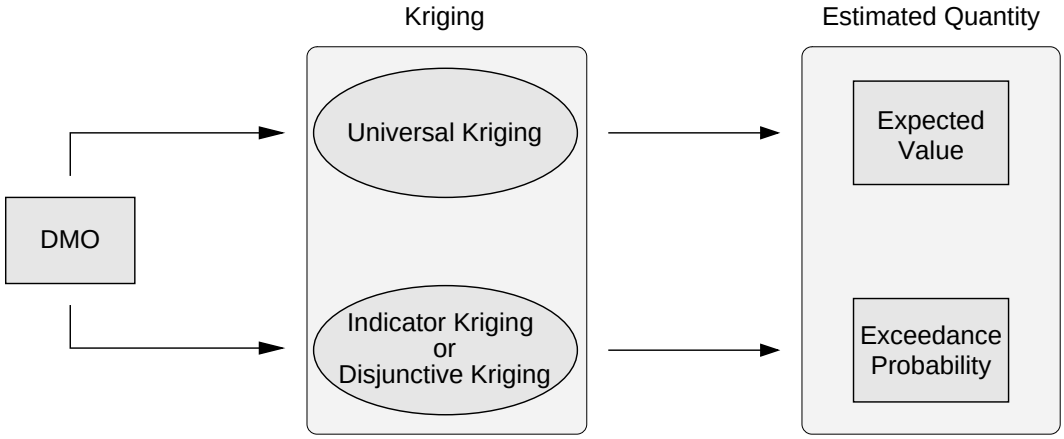

Figure 4.1: *Kriging is an existing non-parametric methodology that produces estimates of the expected value and the exceedance probability on the basis of single realisations from several locations. When applied to the LM output, it starts out from the direct model output (DMO) of a single LM simulation and is able to derive estimates of the expected value and exceedance probabilities at each point of the model grid. Estimates of the expected value can be derived via* universal kriging *(Chilès and Delfiner, 1999). Estimates of the exceedance probability can be derived via* indicator kriging *or* disjunctive kriging.

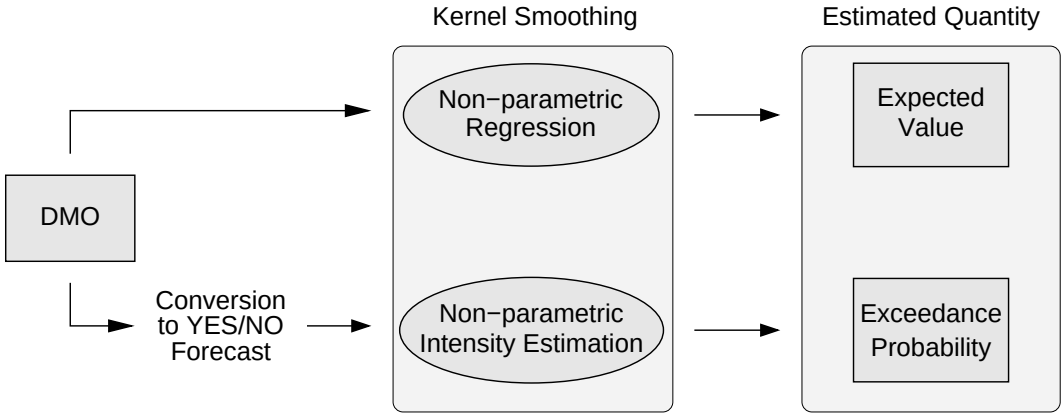

Figure 4.2: *Kernel smoothing is an existing non-parametric methodology that produces a large array of statistical parameters on the basis of a single realisation on each parameter. When applied to the LM output, it starts out from the direct model output (DMO) of a single LM simulation and is able to derive estimates of the expected value and exceedance probabilities at each point of the model grid. The kernel estimate of the expected value arises from the framework of* non-parametric regression *of a function (Härdle, 1990; Simonoff, 1996). The kernel estimate of the exceedance probability arises from the framework of estimating the* intensity function *of a* point process *(Cressie, 1991).*

Kernel smoothing and kriging are similar to the neighbourhood methodology, because they are non-parametric and because they assume that the statistical quantity possesses some degree of smoothness. Kernel smoothing and kriging are based on the notion that many variables exhibit some form of structure, in an average sense, reflecting the fact that points close in space tend to assume close values. Just like the neighbourhood method, they result in spatio-temporal smoothing of the noisy field.

The need for a surrogate sample

In terms of expected value and exceedance probability, kernel smoothing and kriging are useful concepts of non-parametric estimation. They would provide a sufficient starting point for the development of an adequate post-processing procedure in this study, if the aim of LM post-processing was limited to the estimation of the expected value and the exceedance probability. However, LM post-processing is required to achieve quantile estimates in addition. Quantiles are interesting in terms of warnings, especially if the forecast user wishes to control a certain amount of risk.

Whereas the estimation of the expected value and exceedance probabilities can be mastered by weighted averaging of the realisations, quantile estimation at a specific point demands a rough survey of the probability function at that point. As opposed to the estimation of exceedance probabilities, such a survey cannot be circumvented by prior transformation of the realisations. Non-parametric quantile estimation is inseparably linked to the operations of ordering and sorting the realisations by their value. Therefore, the availability of a sample is indispensable.

As the underlying data only consist of single realisations at several locations, the surrogate sample must be generated at each point of estimation. As kernel smoothing and kriging only provide weights instead of a sample, quantile estimation requires the development of an additional methodology in this study: the *neighbourhood methodology*.

When the neighbourhood methodology succeeds to generate the sample, it is only a tiny step to derive estimates of the expected value and the exceedance probabilities as well. Thus, the neighbourhood methodology is able to combine the estimation of quantiles, expected value and exceedance probabilities into a single concept (cf. Fig. 4.3).

Although kernel smoothing and kriging are not explicitly applied in this study, parts of their concepts certainly enter into the development of the neighbourhood methodology. In fact, it is shown that the basic version of the neighbourhood methodology can be reformulated as a special case of kernel smoothing and kriging when estimates of the expected value and exceedance probability are derived. Details follow in the course of this chapter.

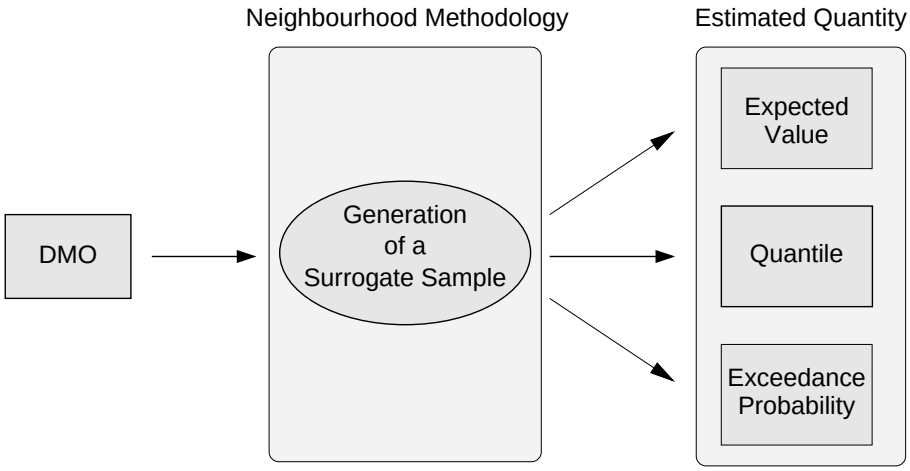

Figure 4.3: *When applied to the LM output, the neighbourhood methodology starts out from the direct model output (DMO) of a single LM simulation and is able to derive estimates of quantiles, the expected value and exceedance probabilities at each point of the model grid. The crucial step of the methodology consists in the generation of a surrogate sample at each point of estimation.*

4.1.3 Organisation of the chapter

A simple, preliminary version of neighbourhood sample generation is described and discussed in Section 4.2. The section contains a definition of the simple version of the neighbourhood, some examples, and a discussion of key assumptions. Section 4.3 works out the differences and analogies between the neighbourhood method, kriging and kernel smoothing. The comparison provides new insight into the underlying assumptions of the neighbourhood methodology and into the errors in resulting estimates. Section 4.4 develops the final definition of the neighbourhood sample by introducing so-called explanatory variables. Explanatory variables represent sources of deterministic small-scale predictability such as orography. Section 4.5 describes the estimation of quantiles in more detail. An empirical quantile function is determined and then transformed to the final quantile estimate via kernel regression.

Throughout this chapter the resulting estimates are assessed by eye-ball inspection. Several examples of the post-processed forecast are presented for the case study of 10 July 2002. The subjective assessment is supplemented by objective verification in Chapter 6. A more extensive set of post-processed forecast data is compared to observational data and forecast goodness is assessed by various verification measures.

4.2 General concept of a neighbourhood

4.2.1 Definition and notation of the surrogate sample

First, some notation is introduced. The LM output of interest consists of one or several surface predictands on a three-dimensional grid comprising two dimensions in space and one dimension in time. Thus, the model output may be seen as a function which relates all points of the discrete model domain $\mathcal{N} = \{1, \ldots, N_x\} \times \{1, \ldots, N_y\} \times \{1, \ldots, N_t\}$ to values of the predictands. The number of grid points is given by $N_x \in \mathbb{N}$, $N_y \in \mathbb{N}$, $N_t \in \mathbb{N}$ and the model grid spacing is given by Δx, Δy, Δt. The size of a grid box is determined by Δx and Δy. The time step between successive model output times is denoted Δt. In an operational LM simulation the grid spacing is $\Delta x = \Delta y = 7\,\mathrm{km}$ and $\Delta t = 1\,\mathrm{h}$.

When post-processing the LM output at a given spatio-temporal location $(i_a, j_a, k_a) \in \mathcal{N}$ of the model grid, a *neighbourhood* around this grid point is defined. The neighbourhood is the set of points (i, j, k) that lie within a certain distance from the point (i_a, j_a, k_a):

$$\mathcal{U}_{i_a j_a k_a} = \left\{ (i,j,k) \in \mathcal{N} \middle| \left(\frac{i_a - i}{\rho_x}\right)^2 + \left(\frac{j_a - j}{\rho_y}\right)^2 + \left(\frac{k_a - k}{\rho_t}\right)^2 \leq 1 \right\}. \tag{4.1}$$

The distances $(i_a - i)$, $(j_a - j)$, $(k_a - k)$ can be phrased in terms of indices here, because a Cartesian grid is used. The size of the neighbourhood is determined by $\rho_x \in \mathbb{R}^+$, $\rho_y \in \mathbb{R}^+$ and $\rho_t \in \mathbb{R}^+$.

The simplest version of the neighbourhood is isotropic in space so that the neighbourhood sizes in the two spatial co-ordinate directions are equal:

$$\rho_{xy} \equiv \rho_x = \rho_y.$$

Fig. 4.4 shows a schematic view of the neighbourhood in the (i, j)-plane and in the (i, k)-plane. Shaded grid boxes belong to the neighbourhood. If the point (i_a, j_a, k_a) is near a boundary of the discrete model domain $\mathcal{N}$ the neighbourhood is simply cut off at the boundary. This happens when i_a is less than ρ_{xy} or greater than $(N_x - \rho_{xy})$. Similarly, the case occurs if $j_a < \rho_{xy}$, $j_a > (N_y - \rho_{xy})$ or $k_a < \rho_t$, $k_a > (N_t - \rho_t)$.

The set of forecasts within the neighbourhood $\mathcal{U}_{i_a j_a k_a}$ is denoted

$$\mathcal{S}_{i_a j_a k_a} = \left\{ z_{ijk} \middle| (i,j,k) \in \mathcal{U}_{i_a j_a k_a} \right\} \tag{4.2}$$

where z_{ijk} are the direct LM output data of the predictand. In order to estimate statistical properties of the predictand at the point (i_a, j_a, k_a), the set $\mathcal{S}_{i_a j_a k_a}$ is taken to be the surrogate sample of the predictand at point (i_a, j_a, k_a).

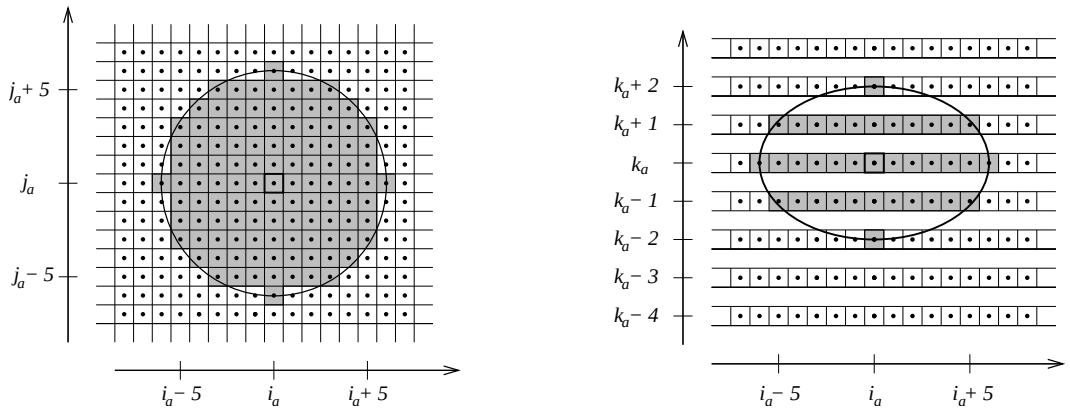

Figure 4.4: *Example of a spatio-temporal neighbourhood of a given grid point at location* $(i_a, j_a, k_a) \in \mathcal{N}$. *Left: The spatial neighbourhood in the* (i, j)*-plane. Right: The spatio-temporal neighbourhood in the* (i, k)*-plane. Shaded grid boxes belong to the neighbourhood.*

4.2.2 Neighbourhood in action: Simple examples

Once the sample $\mathcal{S}_{i_a j_a k_a}$ has been successfully determined, all sorts of statistical properties of the random variable $Z_{i_a j_a k_a}$ may be estimated. This study derives estimates of the expected value, quantiles for various probabilities and exceedance probabilities for various thresholds. Definitions and examples of the estimators of expected value and exceedance probability are presented in the following paragraphs. The estimation of quantiles is tackled in Section 4.5.

Expected value

The neighbourhood $\mathcal{U}_{ijk}$ provides a sample $\mathcal{S}_{ijk} = \{z_1, \ldots, z_N\}_{ijk}$ of the forecast at each arbitrarily chosen location (i, j) on the model grid and at each model output time k. The estimator $\hat{\mu}_{ijk}$ of the expected value $E[Z]$ at location (i, j) and time k is the maximum-likelihood estimator

$$\hat{\mu}_{ijk} = \frac{1}{N} \sum_{l=1}^{N} z_l \quad \text{with} \quad z_l \in \mathcal{S}_{ijk}. \tag{4.3}$$

Fig. 4.6 shows an example of the estimated expected value of 2m-temperature for the case study of 10 July 2002. It is a post-processed version of the respective LM forecast in Fig. 4.5.

Exceedance probability

The exceedance probability $\pi^c = \Pr[Z > c]$ is closely related to the distribution function F of the predictand: $\pi^c = 1 - F(c)$. The exceedance probability can be derived for

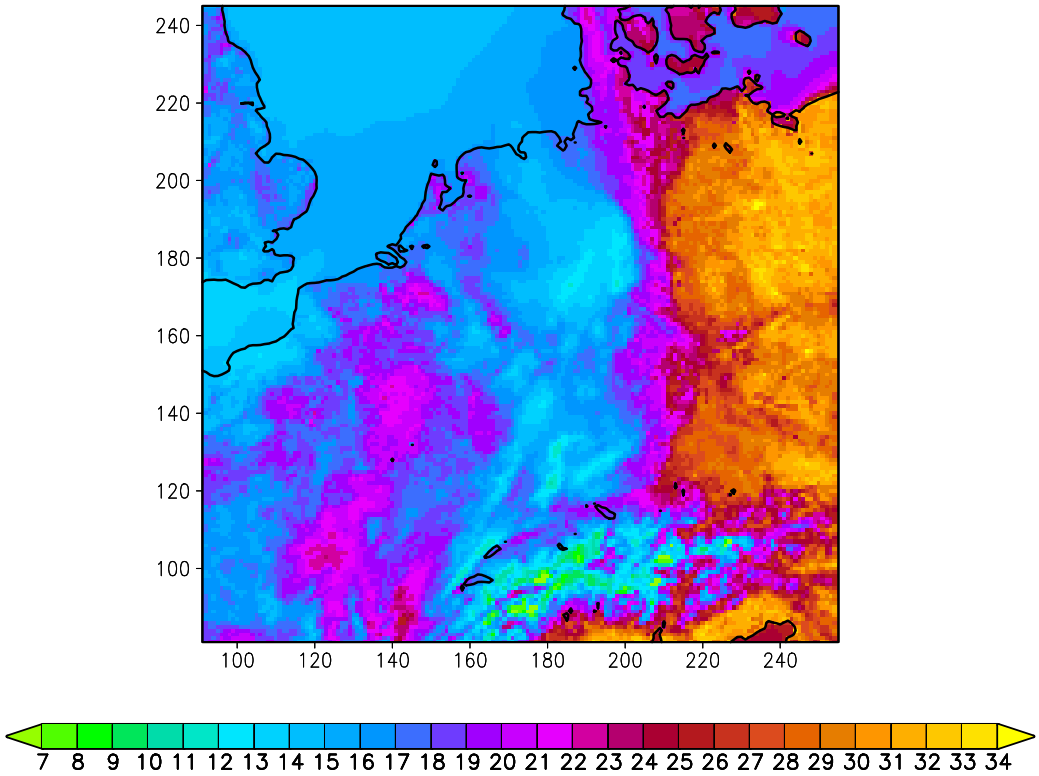

Figure 4.5: *Original LM forecast of 2m-temperature on July 10, 2002, 14 UTC. Unit: Degree Celsius. A section of* 165×165 *grid points is shown.*

arbitrarily chosen exceedance thresholds $c \in \mathbb{R}_0^+$. The function

$$\tilde{\delta}^c(z) \equiv \begin{cases} 1 & : \quad z > c \\ 0 & : \quad z \leq c \end{cases} \tag{4.4}$$

uses the threshold c as a criterion for relating the model output variable z to the values zero and one. Fig. 4.8 shows an example of a field $\tilde{\delta}^c(z_{ij})$ with threshold $c = 0.01\,\mathrm{mm}$. It is derived from the LM precipitation forecast in Fig. 4.7.

Similarly, a set $\mathcal{S} = \{z_1, \ldots, z_N\}$ can be transformed into a set $\mathcal{S}^c = \{\tilde{\delta}^c(z_1), \ldots, \tilde{\delta}^c(z_N)\}$. The set $\mathcal{S}^c$ can be viewed as a sample of the discrete random variable Z^c defined on the mutually exclusive events $\mathcal{A}^c = \{1\}$ and $\bar{\mathcal{A}}^c = \{0\}$. The neighbourhood methodology estimates the probability π^c of $\mathcal{A}^c$ at point (i, j, k) from the maximum-likelihood estimator

$$\hat{\pi}^c_{ijk} = \frac{1}{N} \sum_{l=1}^{N} \tilde{\delta}^c(z_l) \quad \text{with} \quad z_l \in \mathcal{S}_{ijk}. \tag{4.5}$$

The probability π^c is the exceedance probability with exceedance threshold c. Fig. 4.9 shows an example of the estimate. It is a post-processed version of the LM precipitation forecast in Fig. 4.8 and the exceedance threshold is chosen to be $c = 0.01\,\mathrm{mm}$.

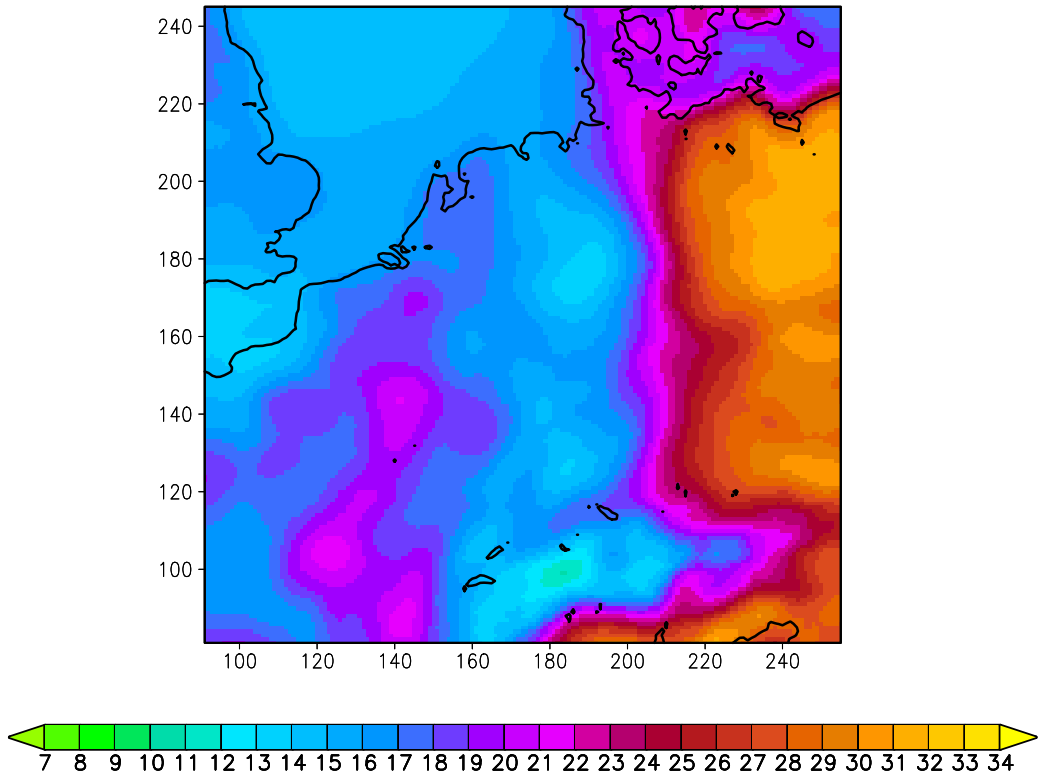

Figure 4.6: *Post-processed version of the 2m-temperature forecast in Fig. 4.5. Unit: Degree Celsius. The estimate of expected value is shown. Post-processing procedure: maximum likelihood estimator (Eq. 4.3) from the simple neighbourhood sample $\mathcal{S}_{ijk}$ (Eq. 4.2).*

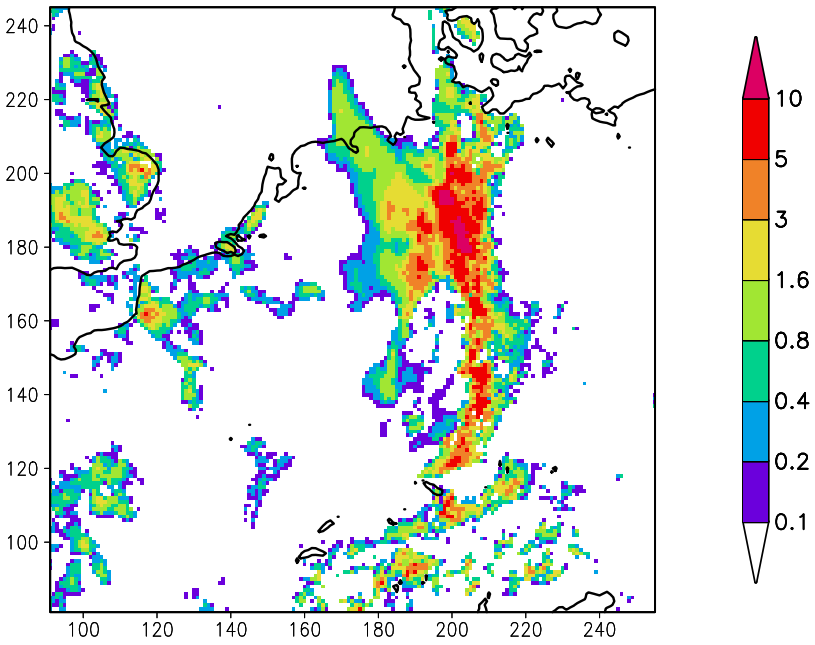

Figure 4.7: *Original LM forecast of 1h-accumulations of precipitation on July 10, 2002, 13–14 UTC. Unit: mm. A section of* 165×165 *grid points is shown.*

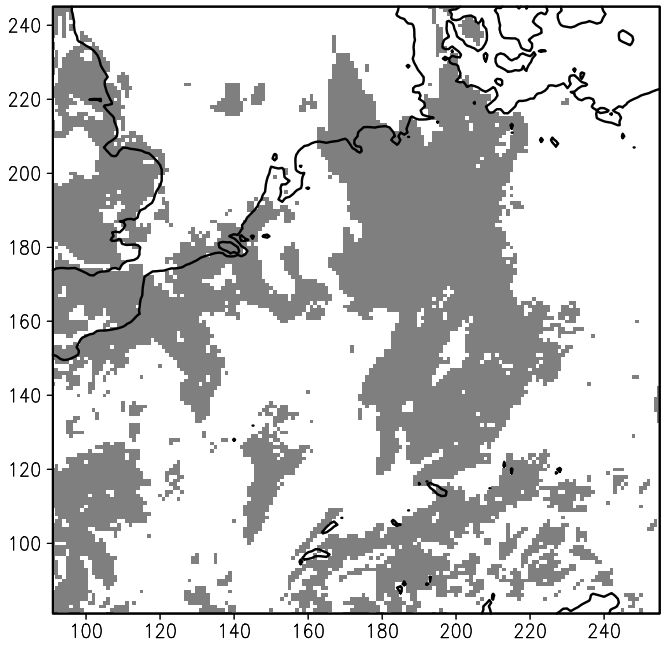

Figure 4.8: *Transformation of the LM precipitation forecast in Fig. 4.7 to values of zero and one. The transformation follows Eq. 4.4 with a threshold* $c = 0.01$ *mm. Values of one are grey and values of zero are white.*

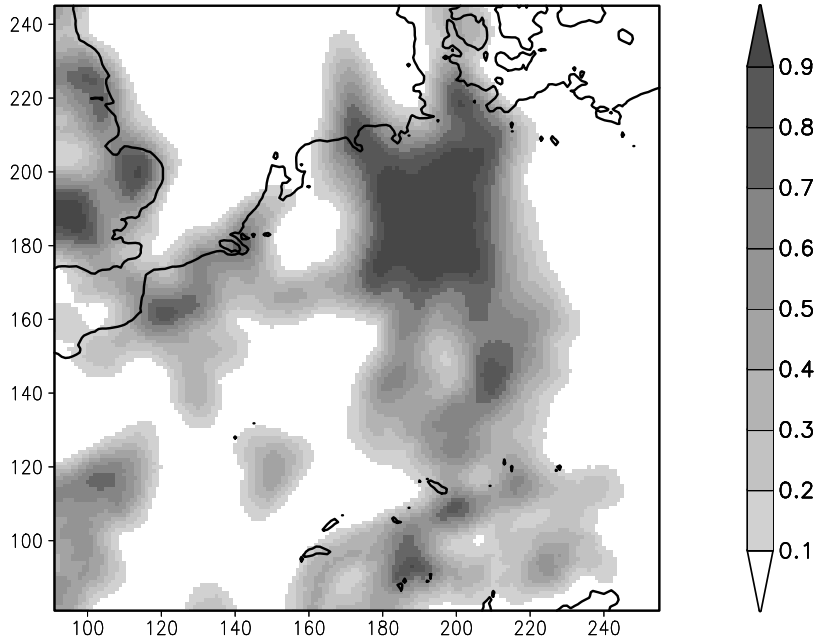

Figure 4.9: *Post-processed version of the precipitation forecast in Fig. 4.7. Estimate of exceedance probability with exceedance threshold 0.01 mm. Post-processing procedure: maximum likelihood estimator (Eq. 4.5) from the simple neighbourhood sample $\mathcal{S}_{ijk}$ (Eq. 4.2).*

4.2.3 The underlying assumption: Ergodicity

What is the underlying assumption of the neighbourhood concept?

The neighbourhood concept assumes that the values $z_{ijk} \in \mathcal{S}_{i_a j_a k_a}$ are independent and identically distributed (*iid*) according to the distribution of the random variable $Z_{i_a j_a k_a}$. Further, it postulates a strictly stationary process which is even *ergodic*, i. e. the spatial and temporal space is assumed to be equivalent to probability space. Ergodicity is confined to small spatio-temporal scales of up to ρ_{xy} and ρ_{tt} (Eq. 4.1) in size.

The ergodic assumption may be illustrated in terms of a probabilistic precipitation forecast. Ergodicity implies the deliberate mixing of two different terms: (1) the spatio-temporal distribution of the deterministic precipitation forecast and (2) a probabilistic precipitation forecast at a specific point in space and time on the model grid. Both are single event probabilities, but they refer to different reference classes. The first type of probability could be used to state that it will rain tomorrow in, say, 30% of the area and time; the second type of probability could be used to state that it will rain on, say, 30% of the days like tomorrow. Reference classes are often accidently confused, as they are not always specified (Gigerenzer and Edwards, 2003). This study does not intend to encourage such erroneous confusion, but deliberately uses the spatio-temporal distribution as an estimator of the probabilistic precipitation forecast at a specific point

in space and time on the model grid.

The underlying assumption of the post-processing methodology implies that patterns within the neighbourhood are unpredictable and entirely due to noise, whereas the large-scale model forecast is assumed to be perfect. The size of the neighbourhood defines a universal spatio-temporal scale which acts as a hypothetical separation between predictable scales and unpredictable scales. Forecast uncertainty is assumed to be solely associated with spatio-temporal small-scale variability of the forecast itself. Small location and timing errors are supposedly the only source of the overall forecast error. Systematic error and large-scale uncertainties are neglected. Furthermore, forecasts at neighbouring grid points are assumed to be independent of each other.

Arguments against the neighbourhood concept

Plenty of cases can be imagined which do not meet these conditions even crudely. For example, variations of predictability with terrain, flow, forcing and forecast lead time are not taken into account at all, despite their large impact.

Furthermore, the assumption of small-scale spatio-temporal stationarity is unrealistic, especially since the post-processing procedure still allows random variables $Z_{i_a j_a k_a}$ and $Z_{i_b j_b k_b}$ at neighbouring grid points to differ. If a value z_{ijk} is contained in both neighbourhoods $\mathcal{U}_{i_a j_a k_a}$ and $\mathcal{U}_{i_b j_b k_b}$, one and the same value z_{ijk} is used as a realisation of differing random variables. It might be more appropriate to soften the stationarity assumption and replace it by an assumption of spatio-temporal smoothness. This is discussed in the following Section 4.3.

Last but not least, the fact that meteorological fields are spatio-temporally coherent and that atmospheric models apply numerical smoothing contradicts the assumption of small-scale independence and zero covariance in the realisations z_{ijk}. Relevant meteorological structures often pertain to scales that are substantially larger than grid box size. As a consequence, random variables Z_{ijk} at neighbouring grid points might not just be correlated, but the elements z_{ijk} of a small-scale sample $\mathcal{S}_{i_a j_a k_a}$ most probably depend on each other, as they belong to the same realisation of a large-scale structure.

Arguments in favour of the neighbourhood concept

On the other hand, some results of verification and predictability studies provide basic arguments in favour of the neighbourhood concept. Theory (Lorenz, 1969) as well as investigation of forecast data (Zepeda-Arce et al., 2000; Casati et al., 2004) point to a relation between spatial scale and prediction skill, thus giving rise to the hope that scale-dependent smoothing could be an appropriate tool in this study. Precipitation forecast errors are largest on the smallest spatial scales (Zepeda-Arce et al., 2000; Mass et al., 2002; Casati et al., 2004) and predictability of small spatial scales is shorter than that of larger scales (Laprise et al., 2000; de Elía et al., 2002; Walser et al., 2004).

Location and timing errors are indeed a dominant source of the total precipitation forecast error (Hoffmann et al., 1995; Ebert and McBride, 2000). Often a simulated convective storm is a few hours off, or a few tens of kilometers away from the observed one (Bernardet et al., 2000). Kumar and Foufoula-Georgiou (1993a) argue that large-scale features of rainfall capture the morphological organisation of a storm which should be modelled deterministically, whereas small-scale features of rainfall exhibit scaling laws and merely contain statistical information or noise. These studies support the notion that forecast uncertainty is mostly due to small-scale uncertainty in space and time.

Some advanced verification strategies like entity-based or object-based approaches implicitly take this fact into account. In addition, Atger (2001) develops a simple verification procedure which resembles the concept of the neighbourhood even more. It consists of taking into account all precipitation amounts that are predicted in the vicinity of an observation in order to produce a probabilistic forecast from a single model run. Atger (2001) argues that forecasters usually consider the whole model output and take advantage of its spatial distribution in order to predict its local probability. Harold E. Brooks (NSSL, Oklahoma) presented a similar approach at the 1998 Workshop on Mesoscale Model Verification (Davis and Carr, 2000), confined to a spatial neighbourhood.

Before concluding this subsection, it should be mentioned that there are also numerical arguments which state that the grid spacing of a model is not synonymous with its ability to represent accurately a given feature (Pielke, 1991; Grasso, 2000). Even in the absence of predictability limits it would be inappropriate to infer the forecast at a given point by considering the deterministic model prediction at the corresponding model grid point only.

4.2.4 Neighbourhood size

A crucial issue is the determination of an optimal size and shape of the neighbourhood. Some indication of the order of magnitude can be inferred from verification and predictability studies, especially from those using scale decomposition (e. g. Casati et al., 2004).

Three different sizes are tested in this study. They are termed “small”, “medium” and “large” and are listed in Tab. 4.1. Fig. 4.4 depicts the spatial and temporal extension of the “medium” size and Fig. 4.10 shows the spatial extension of all three configurations.

The ratio between spatial and temporal extension is kept relatively small, because small-scale variations in terrain are suspected to limit spatial representativeness of surface weather parameters. Therefore, the ratio does not correspond to the expected speed of mesoscale precipitation and temperature patterns. An expected speed of 10 m/s would call for larger spatial extensions so as to match the respective temporal extensions.

In the current version of the post-processing procedure the size of the neighbourhood is held fixed on the entire model domain and at all lead times of the simulation. In

Table 4.1: *The three different configurations of the post-processing procedure which are applied and evaluated. They differ in the size of their spatio-temporal neighbourhood.*

Configuration	Spatial Diameter of Neighbourhood	Temporal Diameter of Neighbourhood	Size ρ_{xy} (Eq. 4.1)	Size $\rho_{t\!t}$ (Eq. 4.1)
"Small"	$6 \cdot \Delta x = 42\,\mathrm{km}$	$3 \cdot \Delta t\!t = 3\,\mathrm{h}$	$\rho_{xy} = 3$	$\rho_{t\!t} = 1.5$
"Medium"	$12 \cdot \Delta x = 84\,\mathrm{km}$	$4 \cdot \Delta t\!t = 4\,\mathrm{h}$	$\rho_{xy} = 6$	$\rho_{t\!t} = 2$
"Large"	$20 \cdot \Delta x = 140\,\mathrm{km}$	$6 \cdot \Delta t\!t = 6\,\mathrm{h}$	$\rho_{xy} = 10$	$\rho_{t\!t} = 3$

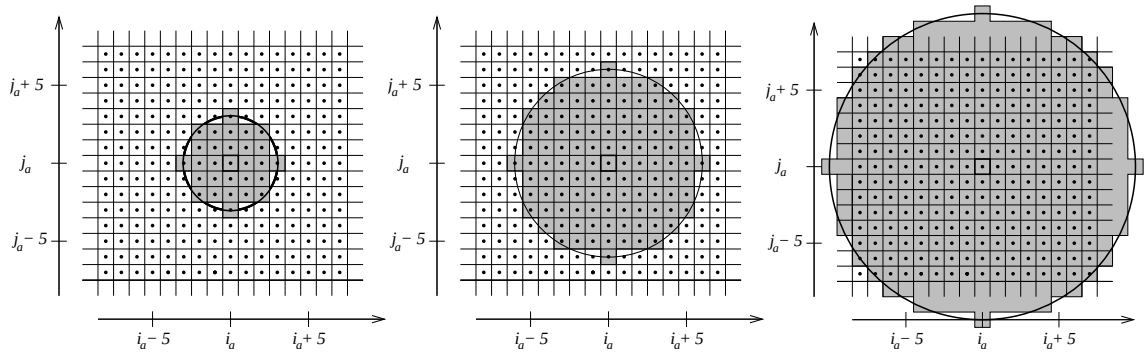

Figure 4.10: *The three different sizes of the neighbourhood as described in Tab. 4.1. Only the spatial extension is shown.*

a more sophisticated study, a variable neighbourhood size could be considered, for example a neighbourhood size that grows with forecast lead time. A spatially varying neighbourhood shape is already developed in this study (Section 4.4.3).

Neighbourhood size does not only determine the assumed scales of deterministic predictability, but it is also highly relevant in terms of computational efficiency: The smaller the neighbourhood, the less computing time is required in the execution of the post-processing scheme. The large neighbourhood of Tab. 4.1 already entails a severely increased computing effort which may be difficult to justify in an operational environment.

4.2.5 Consistency between precipitation and 2m-temperature

The current methodology deals with precipitation and 2m-temperature separately. When the user looks at both parameters simultaneously, consistency between them becomes an issue. Surface precipitation amount and 2m-temperature are expected to be correlated in a physical and statistical way. For example, a negative statistical correlation becomes obvious in the forecast on 4 July 1994 (not shown), a situation characterised by heavy, convective summer precipitation. So as to achieve a realistic post-processed forecast, the post-processing procedure should ideally retain this correlation.

A Bayesian method is elaborated that explicitly takes this correlation into account. First, an *unconditional* probabilistic precipitation forecast is formulated. Then a probabilistic temperature forecast is derived which is *conditional* on the hypothetical outcome of precipitation. For example, the procedure could result in three *unconditional* precipitation quantiles for the probabilities 25%, 50% and 75% and another three *conditional* temperature quantiles for each of the three precipitation quantiles. Obviously, the forecast then consists of nine different probabilistic forecast values.

The ordering of the procedure can also be reversed, i. e. an unconditional probabilistic temperature forecast may be derived first and then the probabilistic precipitation forecast conditional on temperature may be added. In general, the respective results are different.

From a standard user's perspective, a number of nine forecast values for two parameters is likely to be impractical. Thus, it is recommended to abandon the Bayesian methodology and to simply use identical shapes and sizes of the neighbourhood for each parameter instead. This way exactly the same points contribute to the derivation of the precipitation forecast and to the derivation of the 2m-temperature forecast. This is a much simpler solution which roughly preserves the correlation as well. However, it comes with a price tag, as it prevents any attempts of tailoring the neighbourhood to individual characteristics of different parameters.

4.3 Relation to existing non-parametric methodologies

As outlined in Section 4.1.2, the neighbourhood methodology bears similarities to two existing methodologies in non-parametric statistics: kernel smoothing and kriging. This section works out the differences and analogies between the neighbourhood method, kriging and kernel smoothing. A direct comparison is only possible for the estimation of the expected value and the exceedance probability, because kernel smoothing and kriging do not derive estimates of quantiles (cf. Fig. 4.1 and 4.2). The comparison allows to view the neighbourhood in the light of existing research. Furthermore, the underlying assumptions of the neighbourhood methodology and errors in the resulting estimates are reconceived.

In terms of expected value estimation, an overview of the differences and analogies between the three methods is shown in Fig. 4.11. Apart from the estimation procedure itself, the methods also differ in their underlying statistical models and in their assumptions about the raw data. Additionally, they offer different explanations of the error in their estimates. Explaining the error in the estimate is not only crucial for the interpretation and evaluation of the post-processed forecast, but also for optimising the tuning parameters of a method.

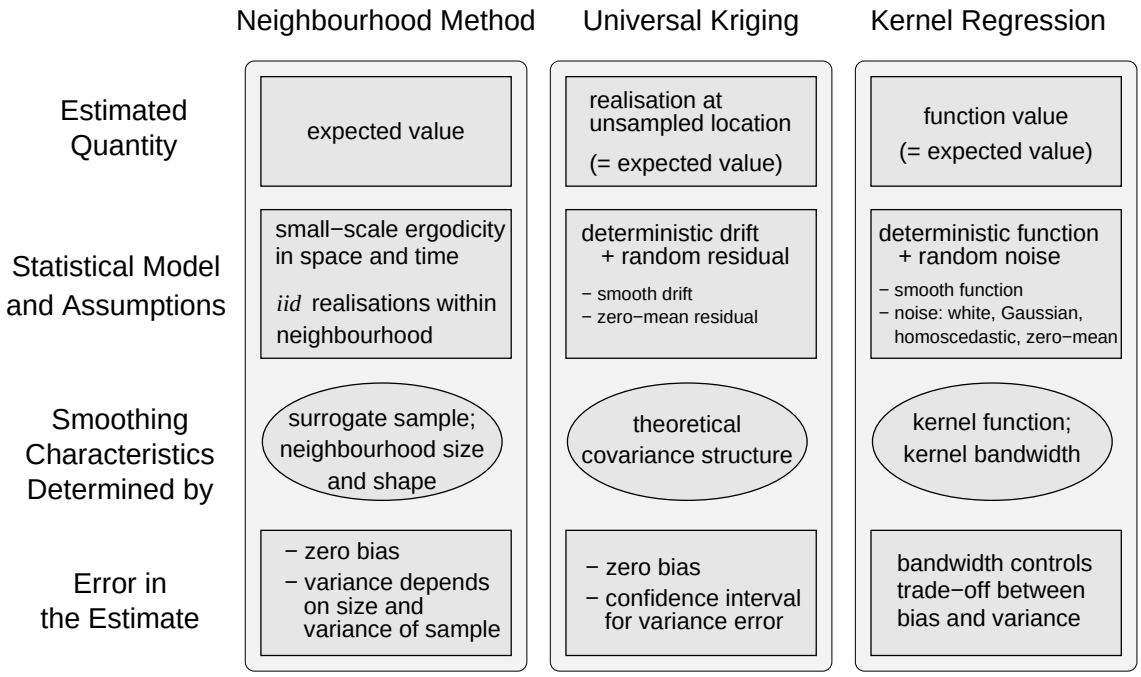

Figure 4.11: *Differences and analogies between the neighbourhood method, kernel smoothing and kriging with respect to the estimation of the expected value.*

Kriging and kernel smoothing are briefly introduced in Subsection 4.3.1. Emphasis is laid on the description of the estimation procedure, its applicability to LM post-processing and a discussion of their underlying statistical models. Subsection 4.3.2 shows that the neighbourhood estimators of expected value and exceedance probability can even be reformulated as a special case of kernel estimators and kriging estimators. This insight sheds new light on the neighbourhood concept. The different explanations of errors in the resulting estimates are discussed in Subsection 4.3.3.

4.3.1 Brief introduction to kriging and kernel smoothing

Similarly to the neighbourhood methodology, kriging and kernel smoothing perceive the LM model output as a realisation of some random process. The nature of the random process is stated in a statistical model. Depending on the respective methodology and the estimated quantity, different statistical models are postulated. It turns out that the underlying assumptions of kriging and kernel smoothing are less rigid than the ergodic assumption in the neighbourhood methodology.

Kriging estimator of the expected value

Kriging has originally been designed to estimate the value of a realisation at an unsampled location on the basis of some realisations at nearby locations (Cressie, 1991;

Wackernagel, 1998; Chilès and Delfiner, 1999). The available realisations are usually located at irregularly spaced points.

Kriging offers a multitude of specific procedures: simple kriging, ordinary kriging, universal kriging, cokriging, quantile kriging, indicator kriging, probability kriging and disjunctive kriging, to name only some of them. Cokriging, for example, offers the possibility to preserve the consistency between several variables so that it could be very useful in LM post-processing (cf. Subsection 4.2.5).

In general, the kriging estimate is a weighted average of the available realisations at the sampled locations. The weights are determined by some theoretical covariance structure between the random variables at different locations. This covariance structure is a relevant part of the underlying statistical model in kriging.

In LM post-processing, the initial situation is slightly different than the standard kriging problem. The available realisations are located on a regular grid. Instead of an additional realisation, an estimate of the expected value is sought. A realisation is already available at the location under consideration.

Kriging is still useful in this situation, because the universal kriging estimator of a realisation is unbiased, i. e. the expected value of the estimator is equal to the expected value of the underlying random variable itself. Furthermore, the kriging estimator of the expected value is a better estimator than the realisation which is already available at that point, because the kriging estimator theoretically has a lower variance error. The regularity of the grid does not pose any additional problems.

In universal kriging the realisations are believed to be the sum of a random residual and a deterministic function, the so-called *drift*. The drift is assumed to be a smooth function in space and the residuals are assumed to have zero mean. The smoothness of the drift expresses the notion that deterministically predictable features at neighbouring locations tend to assume close values.

Due to the notion of a drift, the statistical model of kriging is far less rigid than the ergodic assumption of the neighbourhood methodology. Whereas the neighbourhood method postulates that forecasts within a small-scale spatio-temporal region must be realisations of exactly the same random variable, the kriging methodology allows the random variables to differ at different locations. The differences in their first moment must be small though, because the drift is assumed to be smooth. Furthermore, the kriging concept is able to explicitly acknowledge the spatio-temporal coherence of meteorological fields whereas the neighbourhood concept is confined to the simplistic assumption of independence.

Kernel estimator of the expected value

When estimating the expected value, kernel smoothing can be applied in the context of non-parametric regression of a function (Härdle, 1990; Simonoff, 1996). Statistical

regression aims at recovering an unknown continuous function from noisy data that are available at discrete locations, mostly on a regular grid. The motivation is twofold: (1) interpolation of the function at unsampled locations and (2) denoising the available data, i. e. the replacement of the noisy data by new estimates of the function values that possess a reduced variance error. In the context of LM post-processing, the second motivation is determining, as the estimates are derived at grid point locations.

Similarly to the kriging estimator, the kernel estimator is a weighted average of the available data at different locations. As opposed to kriging, the weights are not determined by the theoretical covariance structure, but by a pre-defined kernel function that usually depends on spatial distance only.

Kernel smoothing of a function is based on the statistical regression model which makes assumptions on the uncertainty of the data and the roughness of the unknown function. The model is similar to the kriging model, but contains more restrictions. In the following paragraphs, the regression model is introduced in detail, because it is picked up again when the wavelet estimator of the expected value is introduced in the following Chapter 5.

When applied to LM post-processing, the regression model regards the LM model output as a three-dimensional function of space and time that is contaminated by noise. The noisy data is given by the LM output data z_{ijk} at discrete grid locations (i, j) and output times k. N_x, N_y are the number of discrete model grid points in the x- and y-direction, respectively, and N_t denotes the number of discrete model output times.

The simplest paradigmatic regression model is

$$Z_{ijk} = \zeta(x_i, y_j, t_k) + \varepsilon_{ijk}, \quad i = 1, \ldots, N_x \quad j = 1, \ldots, N_y \quad k = 1, \ldots, N_t. \tag{4.6}$$

The direct model output z_{ijk} is a realisation of Z_{ijk} and the random variables Z_{ijk} refer to the predictand at grid points $(x_i, y_j, t_k) = (i\Delta x, j\Delta y, k\Delta t) \in \mathbb{R}^3$. The random variables ε_{ijk} are *iid* and follow the normal distribution $N(0, \sigma^2)$, i. e. the noise is assumed to be white, Gaussian and homoscedastic with zero mean.

As the noise ε_{ijk} has zero mean, the function value $\zeta(x_i, y_j, t_k)$ is also the expected value of the random variable Z_{ijk}. Therefore, it is the estimate of $\zeta(x_i, y_j, t_k)$ which is sought in LM post-processing.

In LM post-processing, the random variables ε_{ijk} represent forecast error and the function value $\zeta(x_i, y_j, t_k)$ represents the "true" or deterministically predictable state of the predictand. Again, it is assumed that the underlying function $\zeta(x_i, y_j, t_k)$ possesses some degree of smoothness.

Similarly to the kriging model, the regression model is far less rigid than the ergodic assumption, because it allows the first moments of the random variables to differ at different locations. Still, the statistical regression model contains some simplifications that are generally not in accordance with the real world. For example, the assumption of homoscedastic noise is not justified, because forecast errors mostly depend on location,

time and the atmospheric state. Furthermore, spatio-temporal correlation in forecast error is neglected and systematic forecast error is not addressed either.

Before concluding this subsection, a general remark should be made on the "true" state of the atmosphere. So far, all of the statistical models in this study express the true state by a parameter value. In reality, the true state is not perfectly known even when observations are available, because it is an estimated mean of a grid box *area* and a time *period*. Strictly speaking, the true state should be treated as a random variable. However, it is still common practice to neglect this fact. This study simply conforms to this practice, because this problem is outside the scope of this study.

Kriging estimator of the exceedance probability

The simplest kriging estimate of the exceedance probability is provided by so-called *indicator kriging* (Chilès and Delfiner, 1999). The concept is very similar to the kriging estimator of the expected value (cf. Fig. 4.11), but replaces the initial data z by *indicator data* $\tilde{\delta}^c(z)$ (Eq. 4.4) first.

Analogously to the neighbourhood estimation of exceedance probability, a function transforms the realisations z of the continuous random variable Z into realisations of a random variable with a discrete distribution having the two possible outcomes zero and one (cf. Eq. 4.4). The transformation function is also known as the *indicator function*. The expected value of the discrete random variable is then the exceedance probability associated with the exceedance threshold c.

The replacement by indicator data constitutes a clear loss of information, because it takes into account the position of a value relative to the threshold but not its proximity. Therefore, more sophisticated kriging procedures have been developed: probability kriging, indicator cokriging and disjunctive kriging (cf. Chilès and Delfiner, 1999).

The neighbourhood methodology does not take advantage of such advanced concepts, because they are specific to the estimation of the exceedance probability and because the neighbourhood methodology is required to provide a uniform concept for the simultaneous estimation of exceedance probability, expected value and quantiles. Therefore, the neighbourhood estimator of exceedance probability is less refined than other existing estimators. This is the price tag when requiring a single methodology that covers the estimation of three statistical quantities.

Kernel estimator of the exceedance probability

Kernel estimation of the exceedance probability bears many technical similarities to kernel estimation of the expected value (cf. Fig. 4.11), but the underlying statistical model is entirely different. Whereas the estimation of the expected value is based on function regression, the estimation of the exceedance probability can be viewed as the

estimation of the intensity function of a spatio-temporal point process (cf. Diggle, 1985; Reiss, 1993).

Informally speaking, point process theory is a mathematical tool that describes the occurrence of isolated events in time and/or space. After transforming the realisations to indicator data with values 0 and 1, they can be viewed as an outcome of a spatio-temporal point process. Locations that exhibit an indicator value of 1 are interpreted as the outcome of the spatio-temporal point process.

In the context of precipitation fields, the application of point process theory is already familiar in meteorology. Models for spatio-temporal precipitation based on stochastic point processes go back at least as far as the fundamental work of Cam (1961). This approach developed rapidly in the 1980s through a series of papers (e. g. Waymire et al., 1984; Rodriguez-Iturbe and Eagleson, 1987) and is still subject of current research (e. g. Salsón and Garcia-Bartual, 2003). However, the majority of these papers does not focus on weather forecasting, but on rainfall generators and hydrological applications.

A point process is partly characterised by its intensity function. In LM post-processing, the intensity function of a point process is of special interest, because it is directly linked to the probability of the event under consideration, i. e. the event of exceeding a threshold (cf. Cressie, 1991). A local estimate of the intensity function can be derived by the kernel method (Diggle, 1985). Kernel intensity estimation bears many analogies with kernel density estimation which is extensively described in Silverman (1986). The kernel estimate consists in a weighted average of the indicator values and the weights are determined by the kernel function. The unknown intensity function is assumed to be smooth, again reflecting the notion that deterministically predictable features at neighbouring locations tend to assume close values.

The notion of a spatio-temporal point process requires a different statistical perspective than in the previous subsections. So far, this study has perceived the model output as the realisation of a random variable Z_{ijk} at some point (i, j, k) on the model grid. Forecast uncertainty has been expressed via the uncertainty in the amount z_{ijk} while the point (i, j, k) has been taken as fixed. Now a new approach is pursued which takes the amount as fixed and treats the point (i, j, k) as a realisation of a multivariate random variable (I, J, K). The random variable (I, J, K) follows a discrete distribution confined to the discrete model grid locations (i, j) and the discrete output times k.

The approach bears similarities to the neighbourhood method, as it treats forecast uncertainty as a result of spatio-temporal uncertainty. However, the formalisation of this notion is obviously different here. As opposed to the small-scale ergodic assumption in the neighbourhood methodology, the kernel estimator merely requires the validity of the point process concept and a certain degree of smoothness in the unknown intensity function.

Before concluding this subsection, some remarks should be made on the discrimination between discrete and continuous point processes. In the LM output the realisations of the point process are confined to the discrete model grid points, so model output

data must be described as a discrete point process. However, standard kernel density estimation or kernel intensity estimation starts from the assumption that the function to be estimated is continuous. Therefore, special care must be taken when applying the kernel estimator in the discrete case. So-called discrete kernel functions (Rajagopalan and Lall, 1995) could be able to cope with this special situation.

4.3.2 Neighbourhood methodology as a special case of kernel smoothing and kriging

In terms of expected value and exceedance probability estimation, the basic version of the neighbourhood methodology can be reformulated as a special case of kernel smoothing and kriging. The reformulation allows to discover definite parallels between the various methods and it reveals that the neighbourhood method may still be justified even if the data do not satisfy the rigid prerequisite of small-scale ergodicity. This fact may alleviate the uneasiness about the ergodic assumption in the neighbourhood concept. Furthermore, the comparison also allows a different view of errors in the resulting neighbourhood estimates (Subsection 4.3.3).

Preparatory remarks

Just like the kriging and kernel estimators of expected value and exceedance probability, the corresponding neighbourhood estimators (Eq. 4.3 and 4.5) can be expressed as a weighted average of the direct or transformed model output, respectively:

$$\hat{\mu}_{i'j'k'} = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \sum_{k=1}^{N_{\#}} z_{ijk} \cdot \omega_{ijk}(i', j', k') \tag{4.7}$$

and

$$\hat{\pi}^c_{i'j'k'} = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \sum_{k=1}^{N_{\#}} \tilde{\delta}^c(z_{ijk}) \cdot \omega_{ijk}(i', j', k'). \tag{4.8}$$

Inside the neighbourhood the weights possess a constant value greater than zero, outside the neighbourhood they are zero and their overall sum is unity:

$$\omega_{ijk}(i', j', k') = \begin{cases} \frac{1}{N} & : \quad z_{ijk} \in \mathcal{S}_{i'j'k'} \\ 0 & : \quad z_{ijk} \notin \mathcal{S}_{i'j'k'} \end{cases} \tag{4.9}$$

where N is the number of elements in the neighbourhood sample $\mathcal{S}_{i'j'k'}$ (Eq. 4.2).

Kriging estimators

In order to reformulate the neighbourhood estimate as a special case of kriging, the covariance structure of the data must yield the same kriging weights as the neighbourhood

does. In other words, when the neighbourhood estimator is interpreted in the context of kriging, the neighbourhood implicitly postulates an extremely simple covariance structure. Realisations within the neighbourhood are supposed to have unit covariance matrix and realisations outside the neighbourhood are supposed to have zero covariance. This covariance structure does not necessarily correspond to the true spatial structure of the signal.

The neighbourhood method gives away one of the main advantages of kriging, i. e. the neighbourhood does not let the data itself determine the degree and structure of spatial smoothing within the estimation process. Although kriging is a much more flexible method than the neighbourhood, this study chooses to refrain from kriging methods, because they are not apt for quantile estimation. However, some existing ideas from kriging find their way into the neighbourhood method when its refinements are introduced in Section 4.4.

Kernel estimators

In order to reformulate the neighbourhood estimate as a special case of kernel smoothing, a kernel function must be determined that yields the same kernel weights as the neighbourhood does.

Let the point (i', j', k') lie within a sufficient distance from the boundaries of the model domain so that the lateral boundaries of the LM do not interfere with the neighbourhood $\mathcal{U}_{i'j'k'}$. Then the corresponding kernel function represents a spherical ball with radius one and volume one:

$$K_{\text{ball}}(u,v,w) = \begin{cases} V_{\text{ball}}^{-1} & : \quad u^2+v^2+w^2 \leq 1 \\ 0 & : \quad u^2+v^2+w^2 > 1 \end{cases} \qquad \text{with} \quad u,v,w \in \mathbb{R} \tag{4.10}$$

where the inverse of volume $V_{\text{ball}} = \frac{4}{3}\pi$ ensures that $\int\int\int K_{\text{ball}}(u,v,w)\,dw\,dv\,du = 1$. The function satisfies the standard requirements of a kernel function (cf. Silverman, 1986).

If a kernel estimate is sought at some grid point (i_a, j_a, k_a), the kernel weights at a point (i,j,k) are given by

$$\omega_{ijk}(i_a,j_a,k_a) = \frac{1}{h_x h_y h_{\text{t}}} \cdot \frac{1}{\Delta x \Delta y \Delta \text{t}} \cdot \int_{\left[(i-\frac{1}{2})\Delta x\right]}^{\left[(i+\frac{1}{2})\Delta x\right]} \int_{\left[(j-\frac{1}{2})\Delta y\right]}^{\left[(j+\frac{1}{2})\Delta y\right]} \int_{\left[(k-\frac{1}{2})\Delta \text{t}\right]}^{\left[(k+\frac{1}{2})\Delta \text{t}\right]} K_{\text{ball}}\Big(\frac{u-x_a\Delta x}{h_x\Delta x}, \frac{v-y_a\Delta y}{h_y\Delta y}, \frac{w-\text{t}_a\Delta \text{t}}{h_{\text{t}}\Delta \text{t}}\Big)\,dw\,dv\,du \tag{4.11}$$

where h_x, h_y, $h_{\text{t}} \in \mathbb{R}^+$ are the corresponding kernel bandwidths. As to match the kernel methodology with the neighbourhood methodology, the values of the bandwidths are

simply chosen according to the neighbourhood size defined in Eq. 4.1: $h_x = h_y = \rho_{xy}$, $h_t = \rho_t$.

When the neighbourhood estimator is interpreted in the context of kernel smoothing, the neighbourhood implicitly postulates a very simple shape of the kernel function. This is not a severe constraint in kernel smoothing, because the kernel shape theoretically does not have a great impact on the error of the estimate, at least when compared to the impact of kernel bandwidth. This is proven theoretically (Silverman, 1986) and reaffirmed by sensitivity studies with LM forecasts in this study (not shown). Subjective assessment of a case study reveals only a minor impact of kernel shape and objective verification does not indicate any influence of kernel shape on forecast quality. Thus, there should be no reservations to view the neighbourhood estimates of expected value and exceedance probability as fully-fledged kernel estimates.

Before concluding this subsection, it should be mentioned that the kernel weights above do not exactly correspond to the neighbourhood weights. In the neighbourhood methodology, all weights within the neighbourhood are equal. However, a kernel weight near the boundary of the neighbourhood is smaller than the interior kernel weights if the kernel function drops down to zero within the grid box. The correspondence between the neighbourhood methodology and the kernel methodology is only exact if the kernel weights ω_{ijk} (Eq. 4.11) are formulated in terms of a discrete kernel function as defined in Rajagopalan and Lall (1995).

4.3.3 Different explanations of errors in the estimate

The errors in the neighbourhood, the kriging and the kernel estimates are expected to possess certain statistical properties (cf. Fig. 4.11). Statistical models that explain the error in the resulting estimate are not only crucial for the interpretation and evaluation of the post-processed forecast, but also for optimising the tuning parameters of a method.

Expected statistical properties of the error are closely linked to the underlying statistical model of the respective methodology. Only if the assumptions of the model are satisfied by the data, the statistical properties of the error in the resulting estimate are satisfied as well. If the underlying model is not satisfied, the method runs the risk of introducing spurious errors that cannot be explained by the statistical model anymore.

As the neighbourhood, kriging and kernel methodologies are based on different statistical models, the errors of the resulting estimates are expected to possess different statistical properties as well. For example, the kernel methodology is the only one that allows for a bias (cf. Fig. 4.11). This might seem like a paradoxon after having shown that the neighbourhood estimates can be rephrased as a special case of kriging and kernel smoothing. Different viewpoints should certainly not change the quality of one and the same estimate. However, different viewpoints may well differ in their ability to *explain* the quality of one and the same estimate. The analogies between the neighbourhood

method, kriging and kernel smoothing allow for a better insight into the expected quality of the neighbourhood estimates.

Errors in the estimate are generally composed of a bias term and a variance term. Whereas the neighbourhood estimator and the kriging estimator are theoretically unbiased, the underlying model of kernel smoothing expects a non-zero bias term in the kernel estimate. If at all, the neighbourhood concept and the kriging concept only allow for a qualitative discussion of a possible bias in their estimates. One of the main arguments in such a discussion would be the violation of small-scale ergodicity or the violation of the theoretical covariance structure, respectively.

The kernel concept reveals a direct link between kernel bandwidth and the error in the resulting estimate. In the context of kernel smoothing, the error in the estimate can be phrased as a function of the roughness of the unknown function, the variance error of the available data, kernel shape and kernel bandwidth. The choice of bandwidth is a trade-off between bias and variance in the estimator. A larger bandwidth increases the bias and reduces the variance so that there exists an optimal bandwidth which maximises the overall accuracy in the estimate. As kernel bandwidth and neighbourhood size are directly linked, the kernel concept provides a very important insight into the quantitative effect of neighbourhood size.

4.4 Neighbourhood concept revisited: Explanatory variables

4.4.1 General remarks

Motivation

Small-scale 2m-temperature variations are often associated with small-scale variations of topographic forcing. In such a case, 2m-temperature variations within a neighbourhood are not entirely due to chance, but partly determined by the deterministic effect of topographic forcing. There are many variables that can enhance small-scale deterministic predictability of 2m-temperature and/or precipitation. Some examples are orographic elevation, orographic slope, land-sea mask, snow cover, leaf area index, albedo, time of day (i. e. position of the sun).

The existence of small-scale deterministic effects severely conflicts with the assumptions of the neighbourhood method and is one of the key problems of the neighbourhood methodology. Simple scale separation alone is not always sufficient to differentiate between deterministically predictable features and deterministically unpredictable features. If topographic forcing is not taken into account, the neighbourhood method erroneously assumes that all small-scale variations of 2m-temperature are entirely unpredictable. In

other words, the plain ergodic assumption of small scales (Section 4.2.3) is not always valid and needs some refinement.

The existence of deterministic effects has already received much attention in geostatistics. The causative variables are described as *explanatory variables* (e. g. Diggle et al., 1998) and their deterministic effect is called *external drift* (e. g. Chilès and Delfiner, 1999).

Aims

Accounting for explanatory variables in LM post-processing is a complex problem, especially when ensemble simulations are not available. The problem is probably not fully solvable, so this study merely aims at an alleviation of the unwanted effect of explanatory variables in statistical post-processing, but does not hope for a complete removal of the problem. Furthermore, the consideration of explanatory variables must be computationally cheap so that the low-budget character of post-processing is retained. One has to resist the temptation of expanding the LM post-processing procedure into another extensive model.

Overview of the method

Geostatistics have already elaborated methodologies that deal with explanatory variables. The kriging methodology narrows the problem down to a more complex model of the underlying covariance structure and suggests multivariate methods such as cokriging (e. g. Chilès and Delfiner, 1999).

This study refrains from the laborious procedure of cokriging, because LM post-processing starts from a more favourable situation than the standard kriging problem. The values of explanatory variables are known at each grid point and there is some rudimentary external knowledge about their effect.

In this study, the effect is formalised by the statistical model

$$Z_{ijk} = \breve{Z}_{ijk} + g_{ijk}(d^1_{ijk}, \ldots, d^{N_d}_{ijk}). \tag{4.12}$$

The random variable Z_{ijk} represents the raw model output realisations z_{ijk}. They are composed of a random variable $\breve{Z}_{ijk}$ and a term $g_{ijk}(d^1_{ijk}, \ldots, d^{N_d}_{ijk})$ which represents the deterministic effect of explanatory variables. The variables $(d^1_{ijk}, \ldots, d^{N_d}_{ijk})$ represent explanatory variables at point (i, j, k) and the function g_{ijk} relates them to their deterministic effect. The function is allowed to vary with location and time. Note that the statistical model addresses the *deterministic* effect of explanatory variables, since the term $g_{ijk}(d^1_{ijk}, \ldots, d^{N_d}_{ijk})$ is a parameter value instead of a random variable.

As the sample $\mathcal{S}$ (Eq. 4.2) contains realisations of Z_{ijk}, it captures the full small-scale variability of the raw model output, including the deterministic variability caused by explanatory variables $\boldsymbol{d}_{ijk}$. The variation within the sample $\mathcal{S}$ is not suited to represent

forecast uncertainty, as it also encompasses a substantial part that is deterministically predictable. Therefore, a refined version of $\mathcal{S}$ is sought whose elements are less influenced by explanatory variables.

This study alleviates the unwanted, small-scale effect of explanatory variables in two ways:

1. The raw model output is modified so as to reduce the deterministic effect of explanatory variables within the neighbourhood sample $\mathcal{S}$ (Subsection 4.4.2).
2. The shape of the neighbourhood is altered so as to reduce the variation of the explanatory variables within the neighbourhood $\mathcal{U}$ (Subsection 4.4.3).

4.4.2 Linear modification of the model output

Parametric assumption about the deterministic effect

First, a parametric assumption is made about the function g_{ijk} (Eq. 4.12):

$$g_{ijk}(\boldsymbol{d}_{ijk}) = \beta^{\circ}_{ijk} + (\boldsymbol{d}_{ijk})^T \boldsymbol{\beta}_{ijk} \tag{4.13}$$

where the vector $\boldsymbol{d}_{ijk}$ contains the explanatory variables $(d^1_{ijk}, \ldots, d^{N_d}_{ijk})$, the parameter β°_{ijk} is an offset parameter at $\boldsymbol{d}_{ijk} = \boldsymbol{0}$ and $\boldsymbol{\beta}_{ijk}$ is a parameter vector with N_d elements. The parametric assumption resembles the external drift model known from geostatistics (cf. Chilès and Delfiner, 1999) which provides a shortcut to bypass the more complex approach of cokriging.

The refined neighbourhood sample

The simple neighbourhood sample $\mathcal{S}$ (Eq. 4.2) is now upgraded to a refined version whose elements are less influenced by explanatory variables:

$$\breve{\mathcal{S}}_{ijk} \equiv \left\{ \left(z_{i'j'k'} - (\boldsymbol{d}_{i'j'k'})^T \boldsymbol{\beta}_{ijk} + (\boldsymbol{d}_{ijk})^T \boldsymbol{\beta}_{ijk} \right) \middle| (i', j', k') \in \mathcal{U}_{ijk} \right\}. \tag{4.14}$$

At each point $(i', j', k') \in \mathcal{U}_{ijk}$ the deterministic effect is subtracted from $z_{i'j'k'}$ and replaced by the deterministic effect at the central point (i, j, k) of the neighbourhood (cf. Fig. 4.12).

In contrast to the simple neighbourhood sample $\mathcal{S}_{ijk}$ the variation of the elements in $\breve{\mathcal{S}}_{ijk}$ is not influenced by the unwanted effect of explanatory variables $\boldsymbol{d}_{i'j'k'} \neq \boldsymbol{d}_{ijk}$. Thus, the refined version $\breve{\mathcal{S}}_{ijk}$ is more suitable to form a sample of the predictand at point (i, j, k) than the original set $\mathcal{S}_{ijk}$. This refinement of the neighbourhood is called *linear modification* throughout this study.

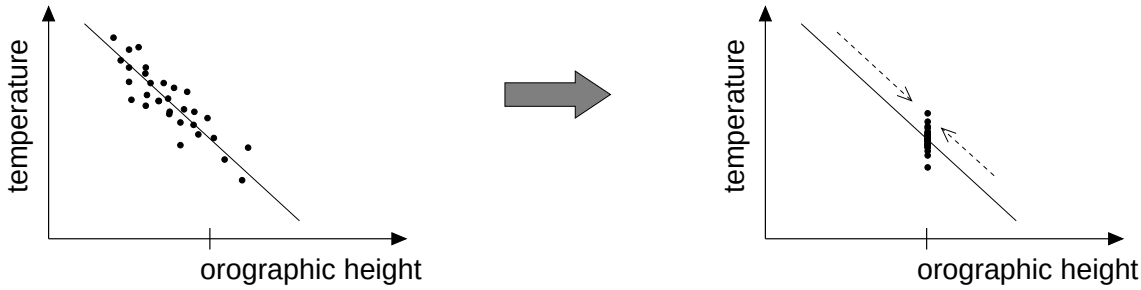

Figure 4.12: *Linear modification of the forecast values within the neighbourhood. Left: Estimation of a linear relation between the LM 2m-temperature forecast and orographic elevation. Right: The LM 2m-temperature values within the neighbourhood are modified according to the estimated regression line and the orographic elevation at the central point of the neighbourhood (Eq. 4.14).*

Estimating the effect of explanatory variables

In order to derive the modified sample $\breve{\mathcal{S}}_{ijk}$, the value of the parameter vector $\boldsymbol{\beta}_{ijk}$ is needed. This study refrains from prescribing a fixed value of $\boldsymbol{\beta}_{ijk}$, because the effect of explanatory variables may vary with the current weather situation and because external knowledge about the deterministic effect might not even be available. In order to capture the current effect of explanatory variables, the value of the parameter vector $\boldsymbol{\beta}_{ijk}$ is estimated from the model output data itself. The parameter is allowed to vary with location and time, so new estimates are derived at each grid point and output time.

For example, a flat estimate of the association between orographic elevation and 2m-temperature would be $-0.0065K/m$. This is the gradient of the standard atmosphere near the ground. In reality, however, the gradient varies with the weather situation, for example a temperature inversion exhibits quite a different gradient. Furthermore, the orographic influence on precipitation is even harder to capture (Prudhomme and Reed, 1999). The relation often depends on the prevailing wind direction. However, additional variables such as wind direction are deliberately excluded from the parametric assumption (Eq. 4.13), so as to keep the post-processing procedure simple. The influence of wind direction might also be covered by the spatio-temporal variability of $\boldsymbol{\beta}_{ijk}$.

Combining equations 4.12 and 4.13 and assuming the regression model 4.6 for the random variable $\breve{Z}_{ijk}$ yields the following statistical model of the raw model output:

$$Z_{ijk} = \beta^{\circ}_{ijk} + (\boldsymbol{d}_{ijk})^T \boldsymbol{\beta}_{ijk} + \zeta(x_i, y_j, t_k) + \varepsilon_{ijk}.$$

The assumptions about ε_{ijk} (Eq. 4.6) and a certain degree of spatial smoothness in ζ and $\boldsymbol{\beta}_{ijk}$ set the theoretical basis for the estimation of the parameter vector $\boldsymbol{\beta}_{ijk}$. The vector is estimated by a spatially weighted least squares fit between the raw model output and explanatory variables.

Choice of weights

When estimating the parameter $\boldsymbol{\beta}_{ijk}$ at location (i, j), more weight is given to those data $z_{i'j'}$ that are close to the point (i, j) and vice versa. The corresponding weights $\omega^{ij}_{i'j'}$ are chosen to follow a symmetric, truncated and normalised Gaussian function centered at point (i, j):

$$\omega^{ij}_{i'j'} \propto \begin{cases} e^{-\frac{1}{2}(u^2+v^2)} & : \quad |u| \leq B \wedge |v| \leq B \\ 0 & : \quad |u| > B \vee |v| > B \end{cases} \tag{4.15}$$

with

$$u = \frac{(i'-i)}{\gamma_{\text{expl}}} \quad \text{and} \quad v = \frac{(j'-j)}{\gamma_{\text{expl}}}$$

and

$$\sum_{j'=1}^{N_y} \sum_{i'=1}^{N_x} \omega^{ij}_{i'j'} = 1.$$

The smoothing parameter γ_{expl} determines to what degree a point (i', j') contributes to the estimate of $\boldsymbol{\beta}_{ijk}$ at point (i, j). Its value is chosen to be equal to the neighbourhood size ρ_{xy}. The truncation parameter $B = 2\gamma_{\text{expl}}$ guarantees that the weighting function has compact support so that the procedure is computationally efficient.

The smoothing parameter γ_{expl} is a key figure, as it controls the trade-off between bias and variance error in the estimator $\hat{\boldsymbol{\beta}}$. On the one hand, the smoothing parameter must be small enough to capture essential spatial variations of $\boldsymbol{\beta}$. If the smoothing parameter is too large, the resulting estimate $\hat{\boldsymbol{\beta}}$ becomes unrealistically smooth and runs the risk of having a large bias. On the other hand, the value of γ_{expl} must be large enough so as to avoid an over-fit of the data due to small sample size. An over-fit results in a large variance in the estimate.

As many explanatory variables are connected to land surface characteristics with a constant value over the sea, the estimation of the parameter $\boldsymbol{\beta}_{ijk}$ and the derivation of the modified sample $\breve{\mathcal{S}}_{ijk}$ (Eq. 4.14) are exclusively carried out at land points (i, j).

In addition, the weights $\omega^{ij}_{i'j'}$ adapt to surface characteristics. For example, the LM data show that the relation between 2m-temperature and snow cover changes abruptly at the transition between areas that are covered with snow and areas that are free of snow. Thus, weights $\omega^{ij}_{i'j'}$ are set to zero if the surface at point (i', j') differs from that at point (i, j) in such a respect.

Similarly, the weights $\omega^{ij}_{i'j'}$ also adapt to characteristics of the forecast field itself. In case of precipitation post-processing, linear modification is only applied to precipitation amounts greater than zero. At locations without precipitation the weights $\omega^{ij}_{i'j'}$ are set to zero.

If many weights are set to zero, the estimate of $\boldsymbol{\beta}_{ijk}$ is based on only few data $z_{i'j'k}$ and the risk of obtaining a bad estimate is larger. For example, such a case occurs in

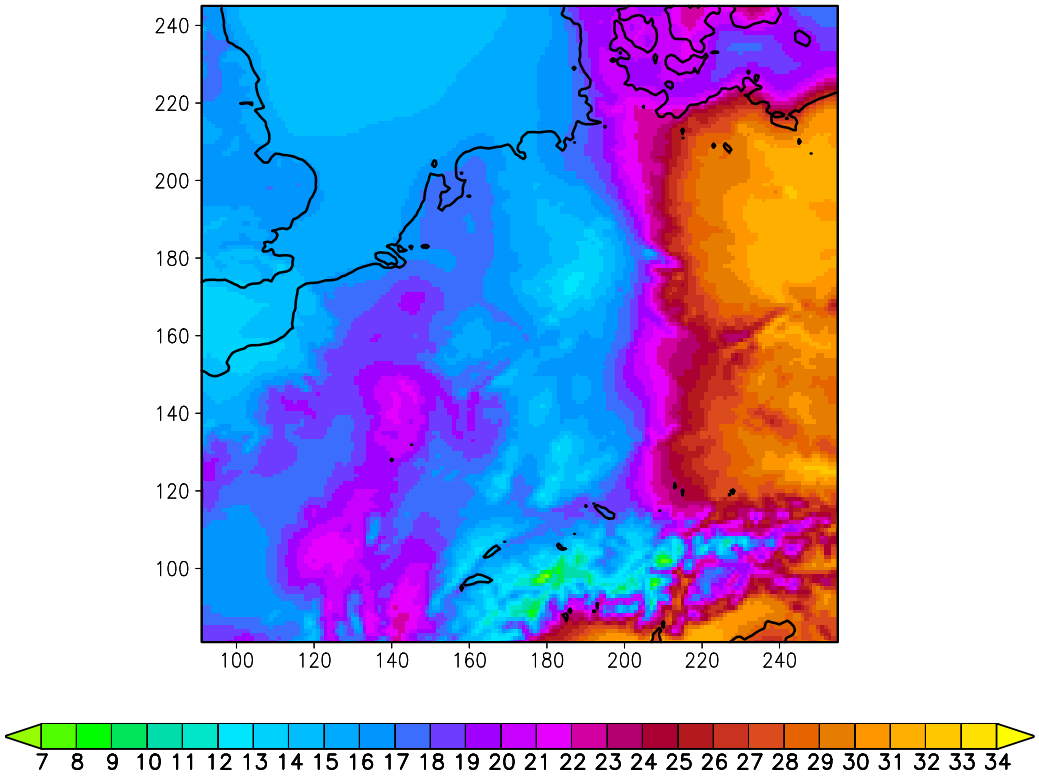

Figure 4.13: *Post-processed version of the 2m-temperature forecast in Fig. 4.5. Unit: Degree Celsius. Estimate of expected value. Post-processing procedure: maximum likelihood estimator (Eq. 4.3) from the refined neighbourhood sample $\breve{\mathcal{S}}_{ijk}$ with linear modification (Eq. 4.14).*

temperature post-processing if the point (i, j) is located within a very small region with snow cover. Therefore, the number of relevant data $z_{i'j'k}$ in the estimation procedure is checked. If the total number is less than 15 or if the sum of weights $\omega^{ij}_{i'j'}$ drops down to less than 20% of the original sum, the estimator $\hat{\boldsymbol{\beta}}_{ijk}$ is set to zero. In precipitation post-processing one or both of these criteria are met quite frequently, depending on the occurrence of precipitation. In temperature post-processing one or both of these criteria are met in less than 0.3% of all grid locations.

Example

Fig. 4.13 shows an example of a post-processed LM 2m-temperature forecast. The post-processing procedure is the same than applied to Fig. 4.6 except that the refined sample $\breve{\mathcal{S}}_{ijk}$ (Eq. 4.14) is now used instead of the simple neighbourhood sample $\mathcal{S}_{ijk}$ (Eq. 4.2). Orographic elevation is taken into account as an explanatory variable of the 2m-temperature forecast.

Let us first discuss the estimate of the temperature gradient β_{ijk} (not shown). In large

parts of the domain the value of $\hat{\beta}_{ijk}$ ranges from -0.01 K/m to 0.00 K/m. The values appear realistic, because they are close to the gradient $-0.0065\,K/m$ of the standard atmosphere. It also turns out that the data yield a spatially variable temperature gradient indeed.

Some regions exhibit surprising values of $\hat{\beta}_{ijk}$ that range from $0.00\,K/m$ up to almost $0.04\,K/m$. Typically, such regions are located near the coast. In these regions a slight decrease of orographic elevation coincides with decreasing distance to the coast. Generally, 2m-temperature values systematically decrease with decreasing distance to the coast on a clear summer day. As a result, the linear fit misinterprets the 2m-temperature drop as an effect of changes in orographic elevation. This is a typical example of mistaking a correlation for a causation.

In this case such misinterpretation does not deteriorate the post-processing procedure. First, orographic elevation exhibits only small spatial variability in the coastal regions under consideration so that the correction term $(\boldsymbol{d}_{i'j'k'} - \boldsymbol{d}_{ijk})^T \boldsymbol{\beta}_{ijk}$ (cf. Eq. 4.14) remains small in spite of a large value of $\boldsymbol{\beta}_{ijk}$. Secondly, distance to the coast can be seen as an explanatory variable that needs to be considered anyway. The least squares fit to orographic elevation automatically takes care of it, as the two gradient vectors are mostly parallel in coastal regions.

Now let us turn back to the post-processed forecast in Fig. 4.13. Compared to Fig. 4.6 linear modification retains a considerable amount of spatial variability in the estimated expected value. Regions over the sea are unaffected because linear modification is only applied over land. A very nice effect is the preservation of small-scale 2m-temperature structures in the Alpine region and in regions of low mountain range, because these structures are expected to be deterministically predictable.

Linear modification certainly runs the risk of erroneously retaining features that are not deterministically predictable. This case typically occurs when small-scale 2m-temperature variations are correlated with orographic variations without any causation. Objective verification against observational data provides a more objective and comprehensive evaluation of the refined neighbourhood procedure with linear modification (Chapter 6).

Before concluding the issue of linear modification, a brief note should be made on consistency between several predictands. As mentioned in Section 4.2.5, consistency between several predictands should be preserved. A simple way of achieving consistency is the application of identical neighbourhood shapes and sizes to different predictands. However, linear modification introduces additional difficulties in terms of consistency. This issue remains for future research.

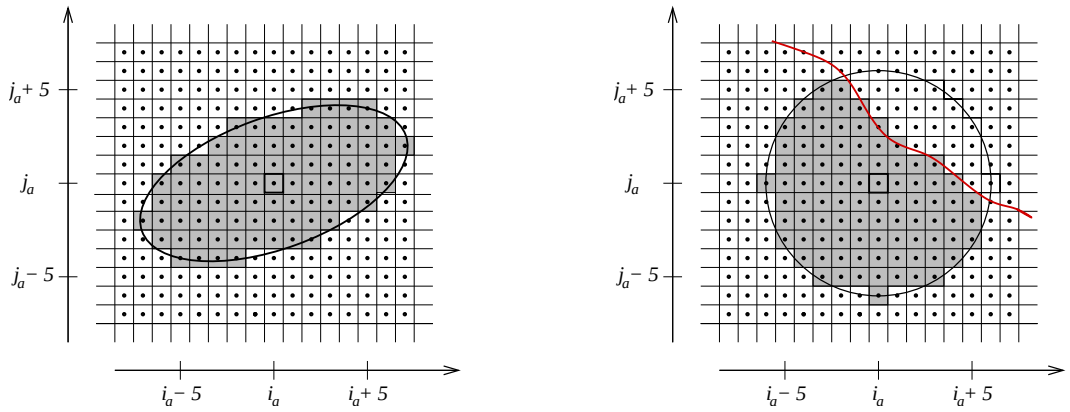

Figure 4.14: *Alteration of neighbourhood shape. Left: Distorted neighbourhood with elliptic shape. The distortion aims at reducing the orographic variability within the neighbourhood by choosing an appropriate eccentricity and orientation of the ellipse. For example, the main axis of the ellipse might align with a valley. Right: The neighbourhood is cut off along coastlines (red line) so as to include land points or sea points only. Boundaries of snow cover are treated similarly.*

4.4.3 Shape of the neighbourhood

Introductory remarks

Linear modification (Section 4.4.2) already alleviates the problem of explanatory variables substantially. However, the modification is restricted to the reduction of the *linear* effect, so it is advisable to apply yet another methodology in addition.

In this section the effect of explanatory variables is additionally alleviated by changing the shape of the neighbourhood $\mathcal{U}_{ijk}$ so as to reduce the variability of the explanatory variables within the neighbourhood. If there is less variability in the explanatory variables, there is also less variability in their effect, and the effect is less disturbing within the sample $\mathcal{S}_{ijk}$.

The shape of the neighbourhood is changed in two ways (cf. Fig. 4.14):

1. The neighbourhood is simply cut off along certain boundaries.
2. The neighbourhood retains the shape of an ellipsoid, but it is distorted and rotated additionally.

In this study, the neighbourhood shape aims to comply with the structure of orographic elevation, snow cover and the land-sea mask. Thus, neighbourhood shape strongly depends on location and rarely on time.

Refining the neighbourhood shape can be seen as a step towards the kriging method, because the refined shape aims at a better conformation with the real covariance structure

of the data. The refined shape implicitly accounts for the fact that the raw model output does not have unit covariance matrix in reality. However, there is still a fundamental difference between the kriging method and the refined neighbourhood shape, because the neighbourhood method exclusively uses external data to draw information about the presumed covariance structure. The smoothing procedure is neither influenced by the predictand itself, nor does it vary with the atmospheric flow.

The refined neighbourhood shape also finds an equivalent in kernel methods, in the sense that kernel functions may be non-isotropic. However, the refined neighbourhood shape is still an extension to the classical kernel method, since it allows the kernel function to vary across the domain.

Neighbourhood cut-off

Neighbourhood cut-off is a method to alleviate the deterministic effect of explanatory variables that can be expressed in a categorical way. For example, surface type can be perceived as a categorical variable: $d = 1$ if there is mainly land surface and $d = 0$ if there is mainly sea surface within a grid box. Another example is the presence of snow cover at the ground.

In terms of 2m-temperature forecasts, the land-sea mask and the presence of snow cover must be considered as explanatory variables. For the moment, let us define the vector of explanatory variables $\boldsymbol{d}_{ijk} = (d^1_{ijk}, d^2_{ijk})^T$ with $d^1_{ijk} \in \{0, 1\}$ expressing land-sea mask and $d^2_{ijk} \in \{0, 1\}$ expressing snow cover at the ground.

If there is a switch from sea to land or if there is a switch from snow-covered ground to snow-free ground within a neighbourhood $\mathcal{U}_{ijk}$, the refined neighbourhood excludes those points $(i'j'k')$ which differ from the central point (i, j, k) in at least one of the two explanatory variables (cf. right portion of Fig. 4.14):

$$\mathcal{U}^{\text{cutoff}}_{ijk} \equiv \mathcal{U}_{ijk} \setminus \left\{ (i', j', k') \middle| \boldsymbol{d}_{i'j'k'} \neq \boldsymbol{d}_{ijk} \right\}. \tag{4.16}$$

The cut-off procedure may induce a problem, as it diminishes the amount of elements within the respective neighbourhood sample $\mathcal{S}$. The reduction can be severe in some cases, for example on a small island on the sea. Reducing the sample size increases the risk of obtaining unreliable statistical estimators from the sample. Especially the quality of quantile estimates is affected. Therefore, post-processing is not applied at points with extremely small sample size. When a neighbourhood $\mathcal{U}^{\text{cutoff}}_{ijk}$ comprises less than 15 points, the estimation at point (i, j, k) is omitted and the estimator is set to a missing value at this point.

Missing values usually occur on less than 1% of the total grid points. Still, operational requirements do not tolerate the existence of missing values. As a very last step of statistical post-processing, all missing values within the post-processed forecast field are

replaced by spatially interpolated values from surrounding valid estimates. The interpolation procedure is bi-linear, isotropic and applies Gaussian weights that decrease with increasing spatial distance from the location under consideration. The Gaussian weighting function is defined analogously to Eq. 4.15 with a spatial smoothing parameter $\varpi_{\text{intpol}} = 2$ instead of γ_{expl}.

Neighbourhood distortion

While the cut-off procedure considers explanatory variables in their categorical form, the procedure of neighbourhood distortion addresses explanatory variables in their continuous form. In contrast to linear modification and neighbourhood cut-off, neighbourhood distortion is able to treat only one explanatory variable at a time. The explanatory variable is chosen to be orographic elevation here, because it is presumed that effects due to orographic elevation enhance deterministic predictability most.

The neighbourhood $\mathcal{U}_{ijk}$ (Eq. 4.1) is determined by spatio-temporal distance to the point (i, j, k). The approach of neighbourhood distortion combines the distance criterion with the criterion of spatial variability of the explanatory variable within the neighbourhood. A predictand $z_{i'j'k'}$ is believed to be more representative of the predictand z_{ijk} if the difference between the corresponding explanatory variables d_{ijk} and $d_{i'j'k'}$ is smaller.

In the simple definition of the neighbourhood $\mathcal{U}_{ijk}$ (Eq. 4.1), the projection of its boundaries on the spatial dimensions forms a sphere with central point (i, j, k) and radius ρ_{xy} (cf. left portion of Fig. 4.4). In the refined version of the neighbourhood this projection forms an ellipse with central point (i, j, k) (cf. left portion of Fig. 4.14). Its eccentricity and the directions of its axes are determined by the quadratic differences $(d_{ijk} - d_{i'j'k'})^2$ of the explanatory variables at the central point (i, j, k) and nearby points (i', j', k').

The methodology is developed along the lines of Hodges (1999). Hodges (1999) constructs a Gaussian kernel function on $\mathbb{R}^2$ which orientates itself according to the spatial density of events in the vicinity of the central point (i, j, k). The kernel function takes on large values in places where the density of events is large and vice versa.

This study replaces the criterion of event density by the criterion $c_{s,s'}$ which is determined by the variability of explanatory variables and by spatial distance. The indices s, s' refer to the 2-dimensional vectors $\boldsymbol{s} \equiv (i, j)^T$ and $\boldsymbol{s'} \equiv (i', j')^T$ which denote spatial locations. The term $c_{s,s'}$ is a product of an orography term $o_{s,s'}$ and a distance weighting term $\omega_{s,s'}$:

$$c_{s,s'} \equiv o_{s,s'} \cdot \omega_{s,s'} \tag{4.17}$$

where

$$o_{s,s'} \equiv \exp\left(- \frac{(d_{ijk} - d_{i'j'k'})^2}{\chi^2} \right) \tag{4.18}$$

and

$$\omega_{s,s'} \equiv \exp\left(- \frac{(\boldsymbol{s} - \boldsymbol{s}')(\boldsymbol{s} - \boldsymbol{s}')^T}{2\varrho^2} \right). \tag{4.19}$$

The term $o_{s,s'}$ has a large value at locations where the values of explanatory variables d_{ijk} and $d_{i'j'k'}$ are close to each other. The weight $\omega_{s,s'}$ depends on the distance between the locations $\boldsymbol{s}$ and $\boldsymbol{s}'$. A larger weight is given to the term $o_{s,s'}$ when the locations $\boldsymbol{s}$ and $\boldsymbol{s}'$ are closer to each other. The exponential form $\exp(-x^2)$ is chosen, because it is a non-negative, symmetric, smooth function with a global maximum at zero. For the sake of computational efficiency, the exponential functions are truncated to zero when their value is very small. The tuning parameter $\chi = 70\,\mathrm{m}$ determines the general degree of influence of the orographic field on the term $o_{s,s'}$ by scaling the orographic differences. The tuning parameter $\varrho \approx \frac{2}{3}\rho_{xy}$ scales the spatial distances within the weighting function $\omega_{s,s'}$.

The term $c_{s,s'}$ is used in the definition of a 2×2 covariance matrix

$$\tilde{R}_{s,s'} \equiv C_{s,s'}^{-1} \sum_{i'=1}^{N_x} \sum_{j'=1}^{N_y} \left[(\boldsymbol{s} - \boldsymbol{s}')(\boldsymbol{s} - \boldsymbol{s}')^T \cdot c_{s,s'} \right] \tag{4.20}$$

with the normalising constant

$$C_{s,s'} \equiv \sum_{i'=1}^{N_x} \sum_{j'=1}^{N_y} c_{s,s'}.$$

As the covariance matrix $\tilde{R}_{s,s'}$ is symmetric, it has two eigenvalues $\tilde{\lambda}_1 \geq \tilde{\lambda}_2$ with the corresponding normalised eigenvectors $\tilde{\boldsymbol{e}}_1$ and $\tilde{\boldsymbol{e}}_2$. The eigenvector $\tilde{\boldsymbol{e}}_1$ points into the direction of smallest local variability of orographic elevation and $\tilde{\boldsymbol{e}}_2$ is perpendicular to $\tilde{\boldsymbol{e}}_1$.

A preliminary definition of a refined neighbourhood is given by:

$$\tilde{\mathcal{U}}_{ijk}^{\mathrm{distort}} \equiv \left\{ (i', j', k') \middle| \left(\boldsymbol{s} - \boldsymbol{s}'\right)^T \tilde{R}_{s,s'}^{-1} \left(\boldsymbol{s} - \boldsymbol{s}'\right) + \left(\frac{k - k'}{\rho_{\updownarrow}} \right)^2 \leq 1 \right\} \tag{4.21}$$

where $\tilde{R}_{s,s'}^{-1}$ denotes the inverse of $\tilde{R}_{s,s'}$ (Eq. 4.20). If the inverse does not exist, neighbourhood distortion is omitted and the simple neighbourhood $\mathcal{U}_{ijk}$ (Eq. 4.1) is applied instead. Such a case is expected to happen in regions of constant orographic elevation, but has not occurred in any of the case studies investigated so far.

The refined neighbourhood $\tilde{\mathcal{U}}_{ijk}^{\mathrm{distort}}$ (Eq. 4.21) rescales the distances between locations $\boldsymbol{s}$ and $\boldsymbol{s}'$ according to the explanatory variables. A projection of the neighbourhood on its spatial dimensions results in an ellipse with the semi-major axis pointing into the direction of eigenvector $\tilde{\boldsymbol{e}}_1$, i. e. the direction of lowest local orographic variability. The eccentricity of the ellipse is determined by the ratio between the two eigenvalues $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$, i. e. the ratio between the orographic variability in the respective directions of the two semi-axes. The time direction is not affected by the distortion.

The definition of the distorted neighbourhood $\tilde{\mathcal{U}}_{ijk}^{\text{distort}}$ (Eq. 4.21) is preliminary, because it does not preserve the volume of the original neighbourhood $\mathcal{U}_{ijk}$ (Eq. 4.1) yet. Volume preservation is achieved by the final definition

$$\mathcal{U}_{ijk}^{\text{distort}} \equiv \left\{ (i', j', k') \middle| \left(\boldsymbol{s} - \boldsymbol{s}'\right)^T R_{s,s'}^{-1} \left(\boldsymbol{s} - \boldsymbol{s}'\right) + \left(\frac{k - k'}{\rho_{\text{tt}}}\right)^2 \leq 1 \right\} \tag{4.22}$$

where the matrix $R_{s,s'}$ is a refined version of matrix $\tilde{R}_{s,s'}$. The matrix $R_{s,s'}$ is determined by its eigenvalues λ_1, λ_2 and by its eigenvectors $\boldsymbol{e}_1 = (e_{11}, e_{12})^T$ and $\boldsymbol{e}_2 = (e_{21}, e_{22})^T$:

$$R_{s,s'} = \begin{pmatrix} e_{11} & e_{21} \\ e_{12} & e_{22} \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}. \tag{4.23}$$

The eigenvectors and eigenvalues of $R_{s,s'}$ are closely related to those of the matrix $\tilde{R}_{s,s'}$. They are defined by the three conditions

$$\begin{array}{ccc} \boldsymbol{e}_1 = \tilde{\boldsymbol{e}}_1 & \text{and} & \boldsymbol{e}_2 = \tilde{\boldsymbol{e}}_2 \\ \lambda_1 \cdot \tilde{\lambda}_2 & = & \tilde{\lambda}_1 \cdot \lambda_2 \\ \lambda_1 \cdot \lambda_2 & = & \rho_{xy}^4. \end{array}$$

The first condition preserves the directions of the semi-axes; the second condition preserves the eccentricity; the third condition restores the original volume of the neighbourhood $\mathcal{U}_{ijk}$.

Example

Fig. 4.15 shows an example of a post-processed LM 2m-temperature forecast with a refined neighbourhood shape. The post-processing procedure is the same than applied to Fig. 4.6 except that the neighbourhood is distorted according to orographic elevation (Eq. 4.22) and cut off along coastlines and boundaries of snow cover (Eq. 4.16).

The effect of neighbourhood cut-off becomes clearly visible in the resulting estimate (Fig. 4.15). Compared to results of simple post-processing (Fig. 4.6) the post-processed field now exhibits temperature discontinuities along coastlines and also along snow boundaries in the Alpine region.

The abrupt changes along coastlines are in perfect accordance with those in the original 2m-temperature forecast (Fig. 4.5). As they are believed to be deterministically predictable, the cut-off procedure improves the estimate of expected value indeed.

Neighbourhood cut-off along snow boundaries produces some cold "islands" within the Alpine region. Such islands are not visible in the original forecast (Fig. 4.5) so that neighbourhood cut-off along snow boundaries apparently results in causing artefacts instead of preserving realistic features. Other case studies, however, show a more favourable effect of neighbourhood cut-off along snow boundaries (not shown). In these cases the original forecast indeed contains temperature discontinuities along snow boundaries

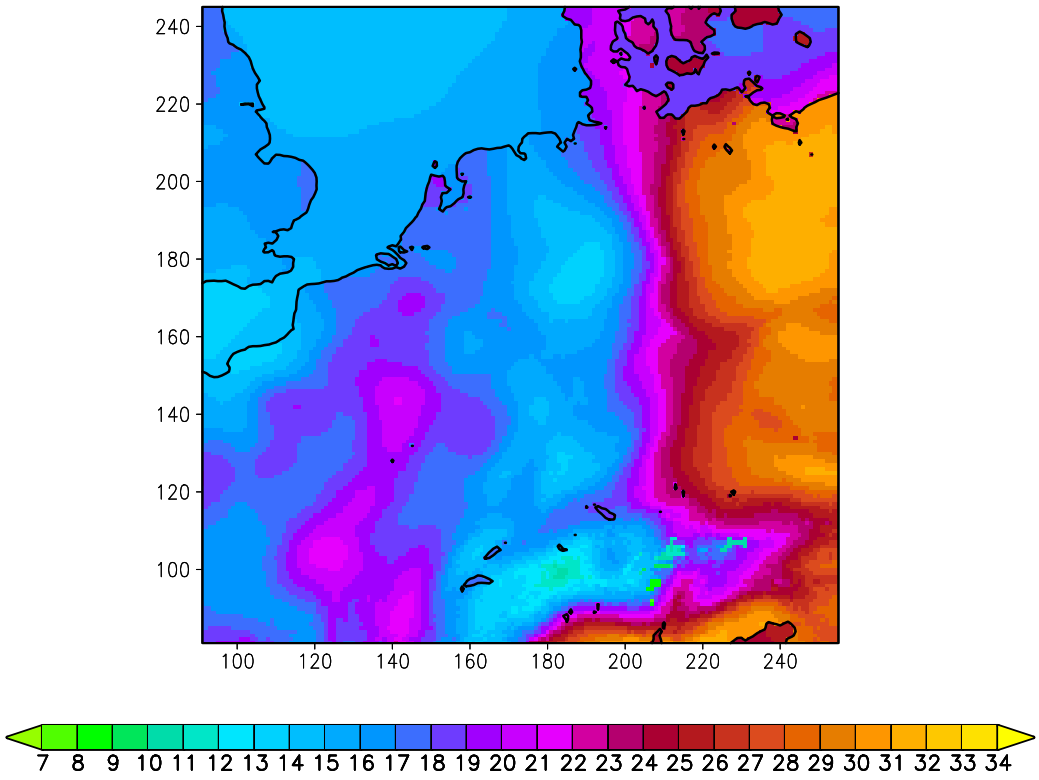

Figure 4.15: *Post-processed version of the 2m-temperature forecast in Fig. 4.5. Unit: Degree Celsius. Estimate of expected value. Post-processing procedure: maximum likelihood estimator (Eq. 4.3) from a neighbourhood that is both distorted according to orographic elevation and cut off along coastlines and boundaries of snow cover (cf. Fig. 4.14).*

that are nicely preserved in the post-processed forecast. In summary, it remains open whether neighbourhood cut-off along snow boundaries is generally beneficial in LM post-processing.

The effect of neighbourhood distortion is extremely small. Compared to simple post-processing (Fig. 4.6) some small-scale temperature variance is additionally recovered. An instance of this becomes slightly visible in the region of Erzgebirge and Böhmerwald where the distorted neighbourhood rudimentarily retains a small part of the orographic 2m-temperature pattern of the original forecast (Fig. 4.5). In other places it is not clear whether the distorted neighbourhood works towards an improved estimate of expected value or not.

All in all, the effect of neighbourhood distortion is small compared to the large effort invested. A possible cause might be an inadequate adjustment of the tunable parameters χ (Eq. 4.18) and ϱ (Eq. 4.19). Further investigations of this issue are needed and remain for a future study.

Table 4.2: *Explanatory variables under consideration and their alleviation in the neighbourhood methodology. Standard version.*

Predictand	Explanatory Variable	Alleviation Method
2m-temperature	land-sea mask	neighbourhood cut-off
	snow cover	neighbourhood cut-off
	orographic elevation	linear modification of DMO
precipitation	none	

4.4.4 Standard version of considering explanatory variables

Aims and motivation

Subsections 4.4.2 and 4.4.3 present several methodologies that alleviate the deterministic effect of explanatory variables. Both the 2m-temperature forecast and the precipitation forecast are examined. In terms of temperature post-processing, the methods include a linear modification according to orographic elevation, neighbourhood distortion according to orographic elevation and neighbourhood cut-off according to the land-sea mask and snow cover. These examples represent only a small fraction of the large number of experiments carried out. Besides orographic elevation, land-sea mask and snow cover, further explanatory variables are considered in addition: orographic slope, distance to the sea and position of the sun. Including the position of the sun is motivated by the diurnal cycle of 2m-temperature. The diurnal cycle can be seen as small-scale variability in time which is deterministically predictable. Linear modification is formulated in a multivariate way which allows a simultaneous consideration of several explanatory variables.

Before post-processing a forecast, one has to specify the explanatory variables and the corresponding method to alleviate their deterministic effect. The post-processing procedure possesses maximum flexibility so that any combination can be set up immediately.

Since the neighbourhood method is envisaged to enter the operational system of DWD, it is desirable to define a standard version of considering explanatory variables that is optimal in some sense. Optimality is not solely attained through best quality of the post-processing forecasts, but through a best trade-off between the complexity of the post-processing procedure and the quality of the resulting forecasts.

Presentation of the standard version

The standard version of considering explanatory variables (Tab. 4.2) is chosen on the basis of subjective and objective evaluation of the post-processed forecasts. The standard version of 2m-temperature post-processing includes linear modification according to orographic elevation only and the neighbourhood is cut off along boundaries of snow

cover and along coast lines. The standard version of precipitation post-processing does not consider any explanatory variables. The reason for this is not a lack of explanatory effects, but the difficulties in dealing with them appropriately.

Variants of this version are taken up again when the results of objective verification are explicitly presented in Chapter 6. The standard version is compared to simpler versions and more complex versions. In this context the question of an optimal neighbourhood size is also raised.

Justification of the standard version

Various different versions of considering explanatory variables are set up and investigated. In its first stage, the investigation consists in the subjective evaluation and in-depth study of several case studies in Europe. In its second stage, a more extensive and objective investigation is carried out in co-operation with DWD. Ulrich Damrath (DWD, Offenbach) and Volker Renner (DWD, Offenbach) selected four summer time periods with a duration of two weeks each and applied different post-processing versions to the respective LM forecasts. They used standard verification measures to evaluate the quality of the resulting estimates of expected value and quantiles. Similar endeavours are also made in this study and are presented in Chapter 6.

Linear modification In-depth studies of several cases indicate the lack of a linear, deterministic association between the explanatory variable and the forecast to be post-processed. For example, the problem occurs for the following pairings: precipitation and orographic elevation, precipitation and orographic slope, 2m-temperature and distance to the coast, 2m-temperature and position of the sun. Several mathematical transformations of explanatory variables and forecast variables are tried out, but do not lead to the desired improvement (not shown).

Only in terms of 2m-temperature and orographic elevation the data roughly exhibit a linear, deterministic relation. Accordingly, results of objective verification indicate a clear improvement when the 2m-temperature forecast undergoes linear modification according to orographic elevation. Due to these results, the standard version of considering explanatory variables (Tab. 4.2) only includes a linear modification with respect to orographic elevation and 2m-temperature.

More sophisticated alleviation methods – possibly going beyond the linear approach – might yields better results. Their development remains for a future study. However, their operational application might be impractical due to a potential increase in computing time.

Neighbourhood cut-off In terms of 2m-temperature post-processing, a subjective assessment of several case studies indicates an improvement due to neighbourhood cut-off along coastlines and boundaries of snow cover. In terms of precipitation post-processing, a neighbourhood cut-off makes less sense, because the precipita-

tion field generally does not regularly exhibit sharp discontinuities at locations that are marked by an explanatory variable. Thus, the choice of the standard version (Tab. 4.2) appears to be justified.

Neighbourhood distortion Recall that neighbourhood distortion according to orographic elevation only leads to tiny improvements in the case study presented (Fig. 4.15). Accordingly, results of objective verification do not reveal a considerable effect of neighbourhood distortion either. This applies both to the temperature and to the precipitation forecast.

Neighbourhood distortion does not contribute to an optimal version, because it requires additional computing time without improving the forecast substantially. For the moment, neighbourhood distortion is not included in the standard version of considering explanatory variables, but further research is recommended.

4.5 Quantile estimation via the neighbourhood

In the previous subsections, neighbourhood estimates of the expected value and the exceedance probability are introduced. This section extends the neighbourhood method to the estimation of quantiles.

4.5.1 Introductory remarks

Definition of the quantile function

Let z be a realisation of the random variable Z which is distributed according to a distribution function F. The quantile function $q(p)$ is defined as

$$q(p) = F^{-1}(p)$$

where $p \in [0, 1]$ is a probability and F^{-1} is the inverse of F. There is a probability of p that the realisation z will be less than the quantile $q(p)$:

$$\Pr\big[Z < q(p)\big] = p \quad \text{for all} \quad p \in [0, 1].$$

If z is the 2m-temperature forecast at a certain location and a certain time, there should be a probability p that the quantile $q(p)$ is greater than the corresponding observation of 2m-temperature, given that the quantile is correctly forecast.

Overview of the estimation procedure

Several approaches to estimating the quantile function are conceivable. This study follows a non-parametric approach, because parametric assumptions would severely con-

Figure 4.16: *Flowchart of quantile estimation. First, the neighbourhood yields a surrogate sample at each grid point and output time of the DMO. Then the elements of the sample are arranged according to an empirical ranking formula or a so-called* plotting position formula. *The formula gives a rough, discrete estimate of the quantile function: the empirical quantile function. Kernel regression transforms the empirical quantile function into a denoised, continuous estimate of the quantile function.*

strict the diversity of resulting estimates and might misinterpret the information contained in the raw data.

This study starts out from the neighbourhood sample and combines two existing methods (cf. Fig. 4.16): (1) application of a ranking formula or a so-called *plotting position formula* to the construction of an *empirical quantile function* and (2) kernel regression of the empirical quantile function by the Gasser-Müller technique. Moon and Lall (1994) propose the combination of ranking formulas and kernel regression for the statistical analyses of hydrological data.

The neighbourhood sample is generated according to the procedure described in the previous subsections. Explanatory variables are considered in the standard way as listed in Tab. 4.2. The plotting position formula sorts the sample elements by their value and assigns probabilities. The resulting empirical quantile function yields unbiased estimates of the unknown quantile function at discrete locations. The kernel regression method interpolates these estimates to the location of interest and reduces their variance error. The reduction of variance error is an essential aspect, because a quantile estimate usually possesses a larger variance error than a corresponding estimate of the expected value.

The estimation procedure takes account of several aspects of forecast uncertainty. While sample generation via the neighbourhood accounts for forecast uncertainty in location and time, processing the sample via kernel quantile regression additionally takes account of forecast uncertainty in precipitation amount or temperature value.

First, the derivation of the empirical quantile function is explained (Subsection 4.5.2) and then the kernel regression procedure is described (Subsection 4.5.3). Subsection 4.5.4 presents some examples of quantile estimates and Subsection 4.5.5 additionally deals with the mixed discrete-continuous distribution of precipitation in quantile estimation.

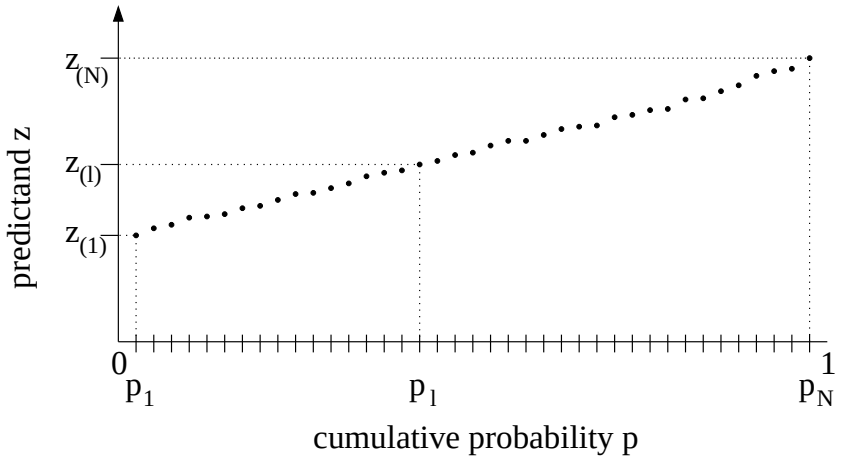

Figure 4.17: *Sketch of an empirical quantile function* $(p_l, z_{(l)})$ $(l = 1, \ldots, N)$. *The sample elements* $z_{(l)}$ *are sorted and arranged according to an empirical ranking formula or a so-called* plotting position formula. *The plotting position formula determines the probability values* p_l $(l = 1, \ldots, N)$.

4.5.2 Estimate of the empirical quantile function

The empirical quantile function

Let $\{z_1, \ldots, z_N\}$ be a set of N independent realisations of the random variable Z and rearrange them in ascending order of magnitude:

$$z_{(1)} \leq \cdots \leq z_{(N)}.$$

Then $z_{(l)}$ is a realisation of the lth *order statistic* $Z_{(l)}$ $(l = 1, \ldots, N)$ (David, 1981). The empirical quantile function q_e associates cumulative probabilities p_l $(l = 1, \ldots, N)$ to the realisations $z_{(l)}$ $(l = 1, \ldots, N)$, as illustrated in Fig. 4.17. The function q_e is determined by an empirical ranking formula or *plotting position formula* that lays down the probability values p_l. Ideally, the plotting position formula should be unbiased so that the data $(p_l, z_{(l)})$ satisfy $\mathrm{E}[z_{(l)}] = \mathrm{E}[q(p_l)]$.

Choice of the plotting position formula

Plotting position formulae have been used for many decades, mainly in the field of hydrology (Cunnane, 1978). They are also a statistical instrument in up-to-date atmospheric research (Folland and Anderson, 2002). A multitude of plotting position formulae have been suggested in the literature (e. g. the Weibull, Beard, Hazen or Adamowski formula). Most of the formulae follow the general form

$$p_l = \frac{l - \alpha}{N + 1 - 2\alpha}$$

Table 4.3: *The chosen value of the plotting position parameter α for different predictands.*

Predictand	Parametric Assumption	Choice of α
hourly 2m-temperature	normal distribution	$\alpha = 0.375$ (Blom, 1958)
max/min 2m-temperature	Gumbel distribution	$\alpha = 0.44$ (Gringorten, 1963)
precipitation amount, conditional on occurrence	gamma distribution (see Folland and Anderson, 2002)	$\alpha = 0.4$

which is also applied here. The parameter α is a constant value mostly ranging between $\alpha = 0$ (Weibull's formula) and $\alpha = 0.5$ (Hazen's formula) (Moon and Lall, 1994).

Choice of the parameter α

The optimal choice of the parameter α depends on the shape of the unknown distribution function F of the random variable Z. Thus, the choice of α is connected to a parametric assumption about the distribution F, but the parametric assumption has much less impact on the resulting estimate than in parametric methodologies. In this study, the choice of α varies with the weather parameter under consideration, because precipitation, hourly 2m-temperature and daily max/min 2m-temperature follow different distributions (Tab. 4.3).

4.5.3 Kernel regression of the empirical quantile function

Introductory remarks

The empirical quantile function $q_e : \{p_1, \ldots, p_N\} \rightarrow \{z_{(1)}, \ldots, z_{(N)}\}$ must be transformed to an estimate of the continuous quantile function $q : [0, 1] \rightarrow \mathbb{R}$. Obviously, some sort of interpolation or curve fitting is necessary, for example by piecewise linear interpolation, least squares fitting or non-parametric regression. Non-parametric regression appears to be the most favourable choice, as the resulting function is differentiable and not confined to a parametric assumption. Further, non-parametric regression contributes to denoising the data $z_{(l)}$.

Discussion of underlying assumptions

In the context of quantile regression, the underlying assumption of non-parametric regression can be phrased as

$$q_e(p_l) = q(p_l) + \varepsilon_l \quad (l = 1, \ldots, N) \tag{4.24}$$

where ε_l is normally distributed with $\mathrm{E}[\varepsilon_l] = 0$ and $\mathrm{VAR}[\varepsilon_l] = \sigma^2 < \infty$. The assumption $\mathrm{E}[\varepsilon_l] = 0$ is justified, as it is in accordance with the assumption of an unbiased plotting position formula in Subsection 4.5.2. The assumption of a normal distribution is disputable though.

The precipitation data follow a mixed discrete-continuous distribution which largely deviates from the normal distribution. For the moment, this fact is ignored and its discussion is postponed to Subsection 4.5.5. Subsection 4.5.5 finds a way for the proper treatment of the mixed discrete-continuous distribution and revises the estimation procedure accordingly.

In terms of 2m-temperature, the data do not follow the normal distribution either, but the deviation is less severe. It is reasonable indeed to assume that the random variable Z is normally distributed, but the assumption of normality really concerns the empirical quantile $q_e(p_l) = z_{(l)}$ which is a realisation of the order statistic $Z_{(l)}$ (Eq. 4.24). Even if the random variable Z itself was normally distributed, the order statistic would follow a different distribution, for example the Gumbel distribution in case of the extremes $Z_{(1)}$ and $Z_{(N)}$ (David, 1981). Moon and Lall (1994) do not mention this inconsistency and this study simply neglects it, too. A possible consequence might be a loss of optimum properties in the kernel quantile estimate, but the effect might be irrelevant compared to the effect of other inconsistencies between the statistical model and the raw data.

Definition of the kernel estimator

After the provisional approval of the regression model above (Eq. 4.24), this study applies the non-parametric kernel regression methodology by Gasser and Müller (1984). The estimator $\hat{q}$ at probability $p_a \in [0, 1]$ is defined as a weighted sum of the realisations $z_{(l)}$ $(l = 1, \ldots, N)$:

$$\hat{q}(p_a) = \sum_{l=1}^{N} z_{(l)} \cdot \omega_l(p_a).$$

As illustrated in Fig. 4.18, the weights $\omega_l(p_a)$ $(l = 1, \ldots, N)$ are given by the definite integrals of a kernel function K:

$$\omega_l(p_a) = \frac{1}{h_p} \int_{\kappa_{l-1}}^{\kappa_l} K\Big(\frac{p_a - s}{h_p}\Big)\, ds \tag{4.25}$$

where κ_l $(l = 0, \ldots, N)$ is an interpolating series of p_l, given by $\kappa_l = (p_l + p_{l+1})/2$, $l = 1, \ldots, (N-1)$, $\kappa_0 = 0$ and $\kappa_N = 1$; h_p is a bandwidth around point p_a, also known as the *smoothing parameter*.

Choice of the kernel function

Usually, a kernel function K is symmetric and satisfies $\int K(t)\, dt = 1$, $\int tK(t)\, dt = 0$ and $\int t^2 K(t)\, dt \neq 0$, where $t = (p_a - s)/h_p$. In this study, the Epanechnikov kernel

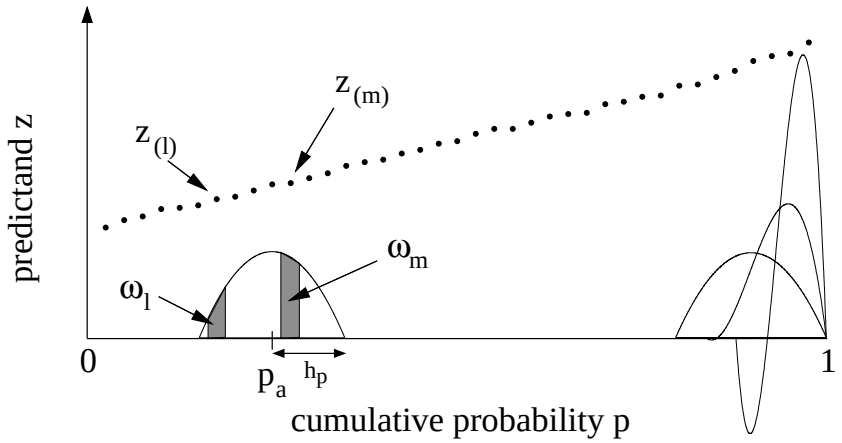

Figure 4.18: *Kernel regression of the empirical quantile function; schematic view. The quantile estimate $\hat{q}$ at probability p_a is the weighted sum of realisations $z_{(l)}$ $(l = 1, \ldots, N)$. The weights $\omega_l(p_a)$ are given by definite integrals of the Epanechnikov kernel K (Eq. 4.25). By definition, the Epanechnikov kernel is symmetric about the origin and has a bandwidth of one (Eq. 4.26). Here, the kernel is shifted to probability p_a and compressed to bandwidth h_p so that a shaded area below $z_{(l)}$ visualizes the value of the respective weight $\omega_l(p_a)$. Curves near the upper end of the quantile domain $[0, 1]$ illustrate the shape of the Müller boundary kernel (Eq. 4.27) in an analogous way. The Müller kernel is applied if $p_a \in [0, h_p[$ or $p_a \in](1 - h_p), 1]$.*

(Epanechnikov, 1969)

$$\begin{array}{rcll} K(t) & = & 0.75(1-t^2) & \text{if} \quad |t| < 1 \\ K(t) & = & 0 & \text{if} \quad |t| \geq 1 \end{array} \tag{4.26}$$

is applied if $h_p \leq p_a \leq (1 - h_p)$ (Fig. 4.18). The Epanechnikov kernel has been shown to have optimum asymptotic properties in terms of the error in the estimated function.

The limitation $0 \leq p \leq 1$ causes a bias in the kernel quantile estimate if $p_a \in [0, h_p[$ or $p_a \in](1 - h_p), 1]$. So as to circumvent this unwanted effect, the corresponding optimum boundary kernel K_r is applied in these cases (cf. Fig. 4.18). The boundary kernel has been derived from Müller (1991):

$$\begin{array}{rcll} K_r(t) & = & 6(1+t)(r-t)\frac{1}{(1+r)^3}\left\{1 + 5\left(\frac{1-r}{1+r}\right)^2 + 10\frac{1-r}{(1+r)^2}t\right\} & \text{if} \quad t \in [-1, r] \\ K_r(t) & = & 0 & \text{if} \quad t \notin [-1, r] \end{array} \tag{4.27}$$

where

$$\left.\begin{array}{rcl} r & = & p_a/h_p \\ t & = & (p_a - s)/h_p \end{array}\right\} \quad \text{if} \quad p_a \in [0, h_p[\quad \text{(left boundary)}$$

and

$$\left.\begin{array}{rcl} r & = & (1 - p_a)/h_p \\ t & = & (s - p_a)/h_p \end{array}\right\} \quad \text{if} \quad p_a \in](1 - h_p), 1] \quad \text{(right boundary).}$$

Note for $r = 1$, the Müller kernel is identical to the Epanechnikov kernel. Whereas the Epanechnikov kernel provides a weight series ω_l that is suitable for interpolating the realisations $z_{(l)}$, the Müller boundary kernel provides extrapolation.

Choice of the smoothing parameter

Compared to the choice of the kernel function, the choice of the smoothing parameter h_p has a much larger impact on the accuracy of the estimated function $\hat{q}(p)$ (Silverman, 1986). A smoothing parameter is globally optimal when it minimizes the average mean integrated squared error (MISE)

$$\mathrm{MISE}(\hat{q}) = \mathrm{E}\Big[\int \big(\hat{q}(p) - q(p)\big)^2 \, dp\Big]$$

over the domain $0 < p < 1$. The MISE can be decomposed into bias and variance components. The smoothing parameter controls the trade-off between bias and variance terms; decreasing the smoothing parameter h_p reduces the bias at the expense of increasing the variance, and vice versa.

In terms of the Epanechnikov kernel and Müller boundary kernel, expressions for the optimum global bandwidth and optimum bandwidth variation near endpoints can be found in Gasser et al. (1991) and Müller (1991), respectively. As the optimum global bandwidth generally depends on the unknown variance σ^2 of the data $z_{(l)} = q_e(p_l)$ and on the roughness of the unknown quantile function q, the derivation of the optimum bandwidth is not straightforward. A multitude of data-driven automatic optimisation techniques exist, e. g. an iterative plug-in method by Gasser et al. (1991) or a variety of cross validation methodologies. There is as yet no universally accepted approach to automatic bandwidth selection.

In LM post-processing, the implementation of a data-driven automatic bandwidth optimisation algorithm is omitted due to its computational effort involved. In addition to the quantile estimation procedure itself, the bandwidth optimisation procedure would have to be called up for each forecast, because the data are expected to yield different optimal bandwidths depending on the weather situation. Thus, the associated computational effort might be substantial and the low-budget character, an essential justification of the post-processing procedure, might be diluted.

Instead, the value of the smoothing parameter is chosen subjectively. Various values of h are tried out and the respective results are inspected visually on the basis of several case studies (not shown). The investigation results in the following choice for all forecasts:

$$h_p = \begin{cases} \frac{1}{16} & : \quad N \geq 40 \\ \frac{5}{2N} & : \quad N < 40 \end{cases} .$$

For $N < 40$ the bandwidth adapts to the sample size N so as to ensure that there are always at least five weights $\omega_l > 0$ if $h_p \leq p_a \leq 1 - h_p$, even if the sample size N

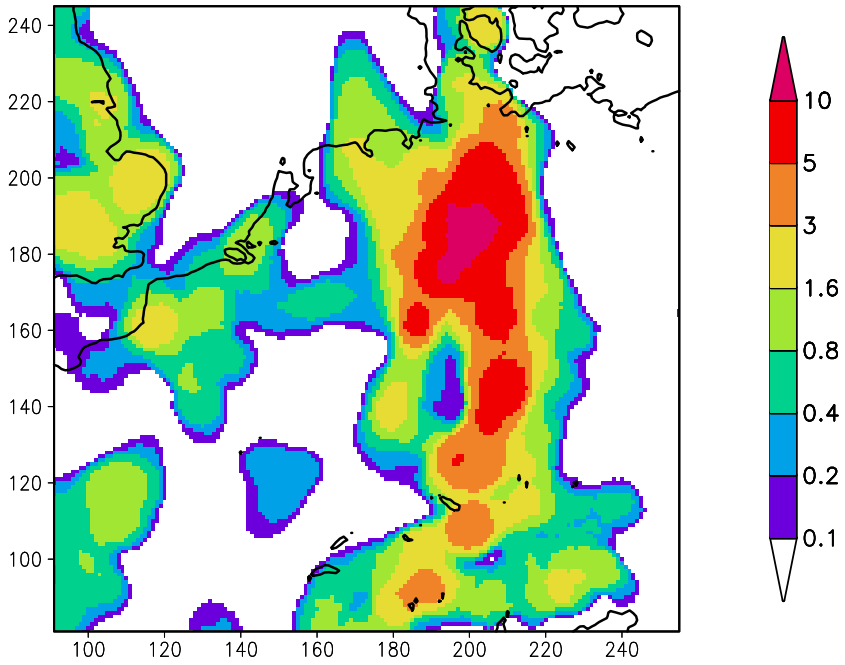

Figure 4.19: *Post-processed version of the precipitation forecast in Fig. 4.7. Unit: mm. Estimate of the 90%-quantile. Post-processing procedure: quantile regression with standard consideration of explanatory variables (Tab. 4.2) and a medium neighbourhood size (Tab. 4.1).*

gets smaller and the spacing $(N + 1 - 2\alpha)^{-1}$ between the discrete probability values p_l $(l = 1, \ldots, N)$ becomes larger. This idea is related to *nearest neighbour smoothing*, a standard method that keeps the number of interpolating data points fixed (Silverman, 1986).

4.5.4 Examples of quantile estimates

Field representation

The field representations in Fig. 4.19 and 4.20 give a first overview of the resulting quantile estimates. Fig. 4.19 depicts the 90%-quantile estimate of precipitation and Fig. 4.20 depicts the 50%-quantile estimate of 2m-temperature. If the quantile estimates are perfect, there will be a probability of 90% that the 90%-quantile estimate at a specific grid point exceeds the corresponding value of the observation. Analogously, there will be a probability of 50% that the 50%-quantile estimate exceeds the corresponding observation. Thus, the values of the 90%-quantiles in Fig. 4.19 exhibit much larger values than the DMO in Fig. 4.7, whereas the values of the 50%-quantiles in Fig. 4.20 have a similar magnitude than the values of the DMO in Fig. 4.5.

Similarly to estimates of the expected value (Fig. 4.6), the fields of estimated quantiles

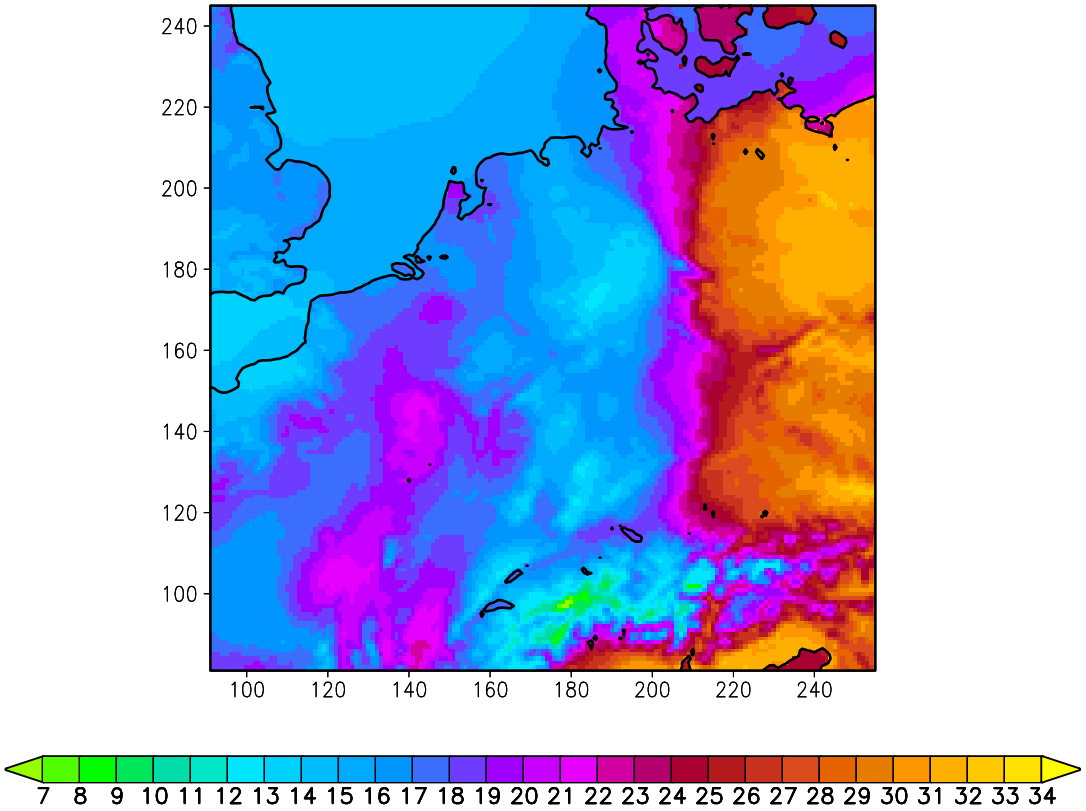

Figure 4.20: *Post-processed version of the 2m-temperature forecast in Fig. 4.5. Unit: Degree Celsius. Estimate of the 50%-quantile. Post-processing procedure: quantile regression with standard consideration of explanatory variables (Tab. 4.2) and a medium neighbourhood size (Tab. 4.1).*

are much smoother than the original LM output. This is a consequence of deriving probabilistic information from the spatio-temporal neighbourhood. One might get the impression that post-processing results in a loss of information, but there is merely a transfer of information. Small-scale spatio-temporal variability of the DMO is transformed into probabilistic information. The probabilistic information in the post-processed forecast becomes visible in a boxplot representation.

Boxplot representation

Fig. 4.21 and 4.22 contain a time series of boxplots at a specific grid point. The boxplots comprise quantiles for five different probabilities, the expected value and the DMO (cf. Fig. 4.23). The 50%-quantile and the expected value are expected to offer a more reliable forecast compared to the DMO. The 90%-quantile is a forecast of risk and the difference between the 90%-quantile and the 10%-quantile is a measure of forecast uncertainty.

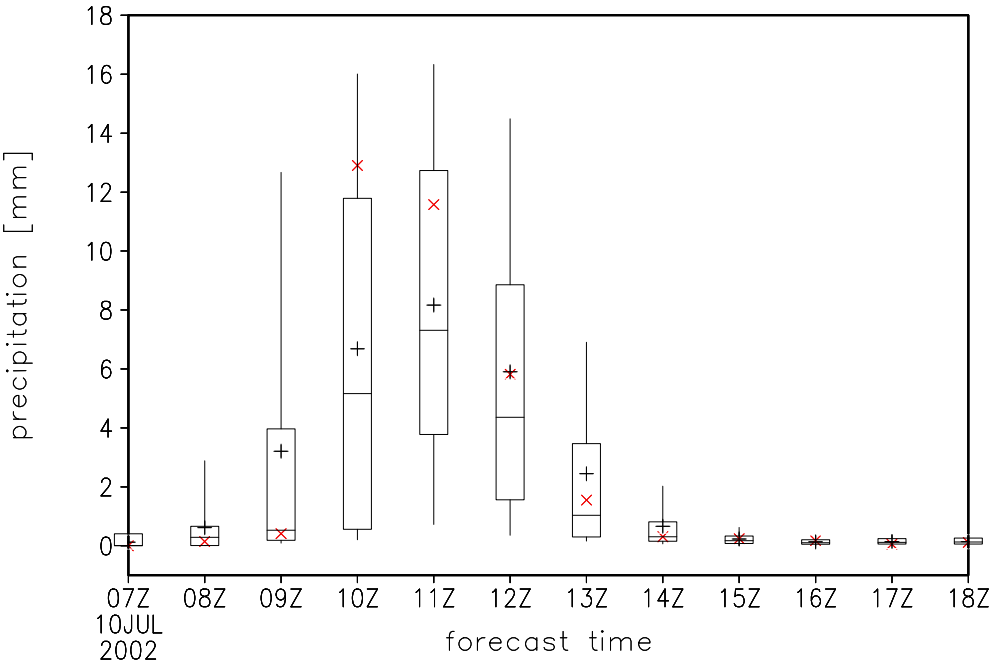

Figure 4.21: *Time series of precipitation boxplots at grid point* (185, 195). *Forecast variable is the 1h-accumulation of the previous hour. Unit: mm. A corresponding field representation of the 90%-quantile for the accumulation time 13–14 UTC is shown in Fig. 4.19. Estimates of five quantiles, estimates of the expected value and the DMO are shown. See Fig. 4.23 for an explanation of the layout.*

As the probabilistic information in the boxplot originates from small-scale spatio-temporal variability in the DMO, there is a clear link between measures of forecast uncertainty in the boxplot and the roughness of the DMO field. Large differences between quantile estimates of different probabilities are found in times and locations where the DMO exhibits large spatio-temporal variability. For example, if a region of fairly low precipitation values contains a marked precipitation peak in the DMO, the 90%-quantiles in the vicinity of the peak shoot up while the 10%-quantiles stay down. The large difference between these two quantiles indicates large forecast uncertainty, as the exact location of the peak is not deterministically predictable. Correspondingly, the difference between different quantiles is much smaller in fairly smooth areas of the DMO, thus indicating low forecast uncertainty.

Note that the boxplot of precipitation (Fig. 4.21) exhibits an estimate of the expected value that is systematically larger than the corresponding estimate of the 50%-quantile. These differences are due to the fact that the probability density function of precipitation is generally not symmetric about its mean, but skewed towards higher values.

As this study also focusses on the needs of the standard user of a forecast, the boxplot representations in Fig. 4.19 and 4.20 deserve even more attention. The boxplots reflect

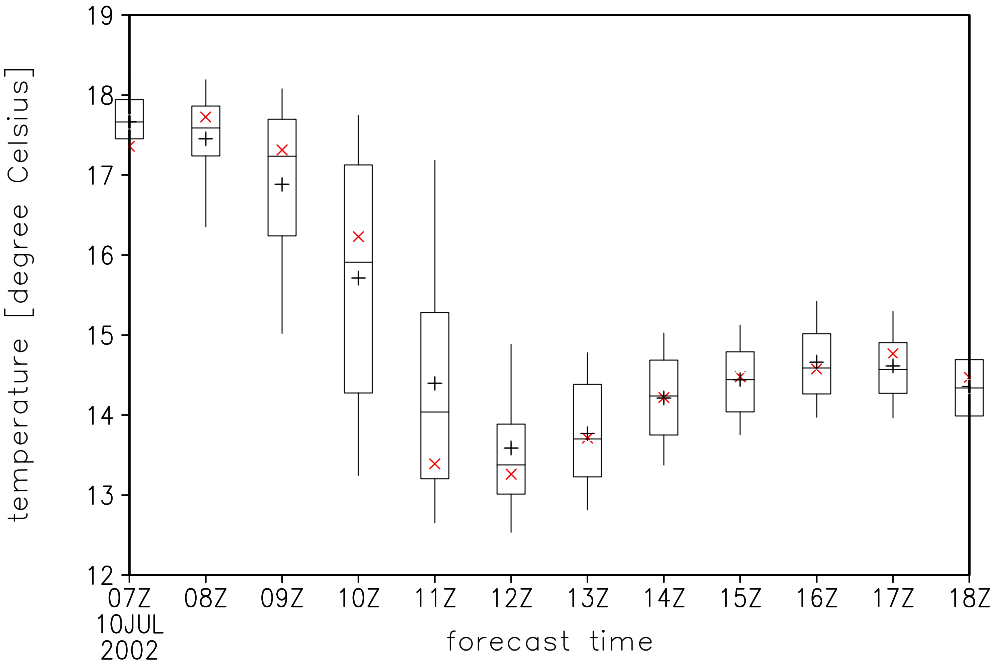

Figure 4.22: *Time series of temperature boxplots at grid point* $(185, 195)$. *Forecast variable is the 2m-temperature at the full hour. Unit: Degree Celsius. A corresponding field representation of the 50%-quantile at time 14 UTC is shown in Fig. 4.20. Estimates of five quantiles, estimates of the expected value and the DMO are shown. See Fig. 4.23 for an explanation of the layout.*

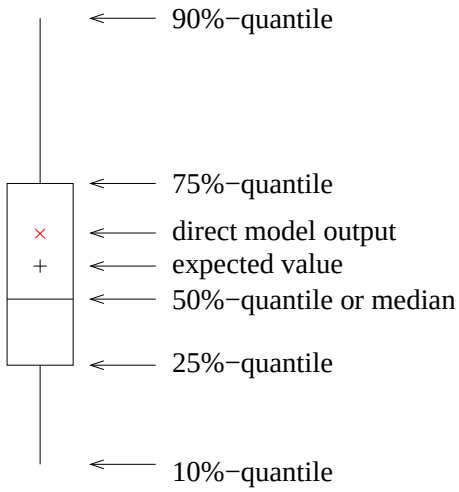

Figure 4.23: *Layout of the boxplots depicted in Fig. 4.21 and 4.22.*

the user's perception of a forecast better than the field representations in Fig. 4.19 and 4.20, because the standard user generally looks into the forecast at a very specific location instead of its geographical distribution. Subsection 6.4.1 resumes this aspect of the post-processed forecast and argues that post-processing enhances the information content of the forecast and thus contributes to forecast goodness.

4.5.5 Considering the mixed discrete-continuous distribution of precipitation

Motivation and notation

The procedure of kernel quantile regression in Subsection 4.5.3 acts on the assumption that the order statistics $\{z_{(1)}, \ldots, z_{(N)}\}$ follow a normal distribution (Eq. 4.24). However, the distribution of precipitation severely conflicts with this assumption so that the kernel estimation procedure needs a revision.

Precipitation values $\{z_1, \ldots, z_N\}$ are realisations of a random variable Z with a probability distribution F that is *mixed discrete-continuous*. The probability π° of *precipitation occurrence* is determined by a *discrete distribution* F° having the two possible outcomes $\{Z|Z > 0\}$ and $\{Z|Z = 0\}$. The probability of exceeding a certain *precipitation amount* is determined by the *continuous distribution* F^+. The distribution F^+ is defined on $\mathbb{R}_0^+$ and gives the probability that a precipitation amount is less than a certain value z_a, given that precipitation occurs.

As the mixed discrete-continuous distribution is a major deviation from the underlying assumption of kernel quantile regression, the estimation of the quantile function $q = F^{-1}$ requires some extra thought when dealing with precipitation. This subsection finds a way how the mixed discrete-continuous distribution can be considered in the estimation procedure of the quantile function $q = F^{-1}$.

A new expression for the quantile function

Let us first seek a new expression for the quantile q at a probability p_a which takes the mixed discrete-continuous distribution into account. If $\pi^\circ = 0$ the random variable Z is simply distributed according to F°, so let us assume that $\pi^\circ > 0$. Then the quantile $q(p_a)$ is not defined for $p_a \in [0, 1 - \pi^\circ[$, so only $p_a \in [1 - \pi^\circ, 1]$ is considered.

Let $z_a > 0$ be an arbitrarily chosen precipitation value with $F(z_a) = p_a$. The probability of $Z \geq z_a$ may be written as the product of a conditional and an unconditional probability:

$$\Pr\big[Z \geq z_a\big] = \Pr\big[Z \geq z_a | Z > 0\big] \cdot \Pr\big[Z > 0\big].$$

This is equivalent to

$$1 - \Pr\big[Z < z_a\big] = \Big(1 - \Pr\big[Z < z_a | Z > 0\big]\Big) \cdot \Pr\big[Z > 0\big]$$

and may be rewritten as

$$1 - p_a = (1 - p_a^+) \cdot \pi^\circ$$

with the additional definition

$$p_a^+ \equiv \Pr\big[Z < z_a | Z > 0\big], \quad p_a^+ \in [0, 1].$$

According to the above-named restrictions $\pi^\circ > 0$ and $p_a \in [1-\pi^\circ, 1]$, the probability p_a^+ can be related to the probabilities p_a and π°:

$$p_a^+ = 1 - \frac{1 - p_a}{\pi^\circ}. \tag{4.28}$$

Let Z^+ be a random variable with distribution F^+ as defined above, then:

$$\Pr\big[Z^+ < z_a\big] = \Pr\big[Z < z_a | Z > 0\big] = p_a^+.$$

With the quantile function $q^+ \equiv (F^+)^{-1}$ it follows that

$$q^+(p_a^+) = z_a = q(p_a) \quad \text{for all} \quad p_a \in [1 - \pi^\circ, 1] \quad \text{and} \quad \pi^\circ \in]0, 1] \quad \text{and} \quad z_a > 0.$$

Estimator of the quantile function

The previous Eq. $q(p_a) = q^+(p_a^+)$ implies that the estimator $\hat{q}(p_a)$ may be defined as

$$\hat{q}(p_a) = \widehat{q^+}(\widehat{p_a^+}) \tag{4.29}$$

with $\widehat{p_a^+} = 1 - (1 - p_a)/\hat{\pi}^\circ$ for $\hat{\pi}^\circ > 0$ and $\hat{\pi}^\circ \geq 1 - p_a$. The probability π° of precipitation occurrence is estimated by the maximum likelihood estimator $\hat{\pi}^\circ = n/N$ where $0 \leq n \leq N$ denotes the number of elements $z_l > 0$ in the set $\{z_1, \ldots, z_N\}$.

The quantile q^+ at probability $\widehat{p_a^+}$ is estimated through the kernel quantile estimation procedure described in the previous section. As the distribution F^+ is conditional on precipitation occurrence, the estimation is based on the subset of the n non-zero precipitation amounts $\big\{z_l | z_l > 0\big\} \subseteq \{z_1, \ldots, z_N\}$. The plotting position parameter α is chosen according to a gamma distribution (see Tab. 4.3).

If $\hat{\pi}^\circ = 0$ or $\hat{\pi}^\circ < (1-p_a)$, the estimator $\hat{q}(p_a)$ is set to zero. If a prohibitively small sample size n does not allow the derivation of $\widehat{q^+}(\widehat{p_a^+})$, the estimator $\hat{q}(p_a)$ remains undefined and remains for spatial interpolation by surrounding estimates that are valid.

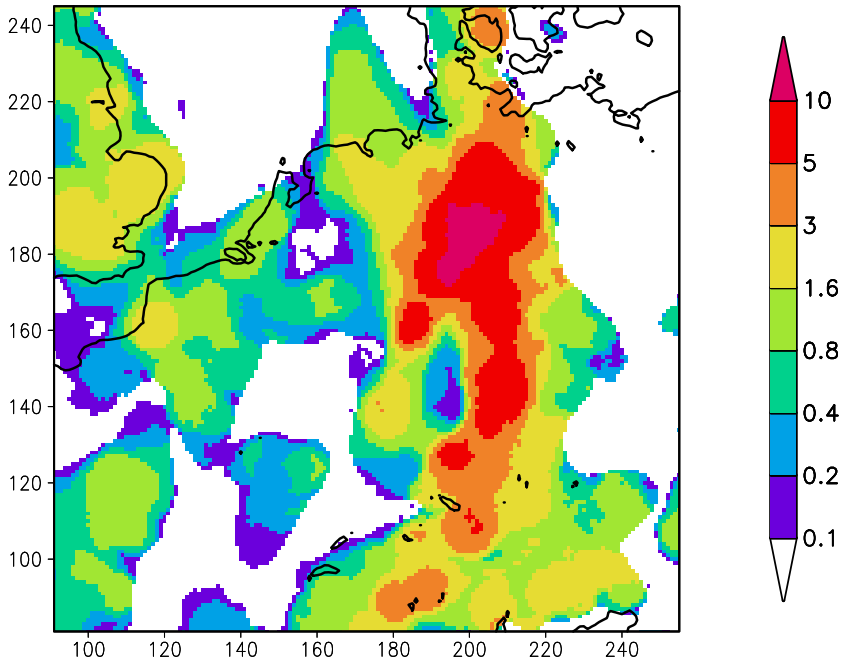

Figure 4.24: *Post-processed version of the precipitation forecast in Fig. 4.7. Unit: mm. Estimate of the 90%-quantile. Post-processing procedure: quantile regression with standard consideration of explanatory variables (Tab. 4.2) and a medium neighbourhood size (Tab. 4.1). As opposed to Fig. 4.19 the mixed discrete-continuous distribution of precipitation is taken into account here.*

Effect of considering the mixed discrete-continuous distribution

When the mixed discrete-continuous distribution is explicitly taken into account quantile estimates according to Eq. 4.29 markedly differ from the conventional quantile estimates of the previous Section 4.5. Fig. 4.24 shows an example of the revised 90%-quantile estimates which can be juxtaposed with the conventional estimates in Fig. 4.19. Several differences can be noticed:

- The field of the revised quantile estimates exhibits discontinuities at the transition from estimates $\hat{q} = 0$ to estimates $\hat{q} > 0$.

- The revised estimates of 90%-quantiles are generally larger than the conventional estimates in regions of $\hat{\pi}^{\circ} \geq (1 - p_a)$.

- The consideration of the mixed discrete-continuous distribution results in a substantial increase in missing values. In the case study presented here, the number of missing values increases from less than 1% to approximately 10%–20%. This effect does not explicitly show in Fig. 4.24, because the missing values are already interpolated by surrounding quantile estimates that are valid.

The discontinuities at the transition from zero estimates to non-zero estimates are due to the explicit distinction between cases of $\hat{\pi}^\circ < (1 - p_a)$ and cases of $\hat{\pi}^\circ \geq (1 - p_a)$. Quantiles in regions of $\pi^\circ < (1-p_a)$ are simply set to zero whereas quantiles in regions of $\hat{\pi}^\circ \geq (1-p_a)$ are estimated by elements of the neighbourhood sample that are exclusively greater than zero.

Revised estimates of 90%-quantiles are generally larger than the conventional estimates in regions of $\hat{\pi}^\circ \geq (1 - p_a)$, because precipitation values of zero are explicitly excluded from the kernel regression procedure. Thus, the resulting quantile estimates are not "diluted" by zero-valued sample elements anymore.

The increased number of missing values arises from the occasional omission of kernel regression. Kernel regression is cancelled when there is a prohibitively small number of non-zero precipitation values that are assigned non-zero kernel weights. This case occurs in regions where the estimated probability $\hat{\pi}^\circ$ exceeds the value $(1 - p_a)$ only slightly. Then the quantile estimator $\widehat{q^+}(\widehat{p_a^+})$ (Eq. 4.29) is not set to zero and the probability $\widehat{p_a^+}$ is close to zero (Eq. 4.28). Probabilities close to zero pose a challenge to quantile estimation. The Müller boundary kernel is applied which has a smaller support than the Epanechnikov kernel (cf. Fig. 4.18). Consequently, an even smaller number of precipitation values is assigned non-zero kernel weights, unless there is an exceptionally large sample of non-zero precipitation values. Thus, it becomes more likely that the kernel regression procedure is cancelled and quantile estimates are set to a missing value.

In summary, the eye-ball inspection of the resulting estimate does not indicate that the revised estimate were superior to the conventional estimate. However, there might be a measurable improvement in the quality of the post-processed forecast, because the revised estimation procedure is theoretically more sound. This is investigated by the verification procedure described in Chapter 6 (not shown). The resulting quantiles estimates of the conventional and the refined version are compared to each other in terms of their reliability, but the differences are marginal.

Although it cannot be demonstrated that the consideration of the mixed-discrete distribution has any influence on the quality of the resulting estimates, the consideration of the mixed-discrete distribution is still included in the standard version of quantile estimation. The corresponding computational cost is negligible, forecast quality does not deteriorate and theoretical arguments still speak in favour of taking the mixed discrete-continuous contribution into account.

5 Wavelet smoothing

5.1 Introductory remarks

In addition to the neighbourhood methodology another post-processing scheme is developed: *wavelet smoothing*. Subsection 5.1.1 outlines the aims and motivations of wavelet smoothing in this study and Subsection 5.1.2 puts them in perspective of existing wavelet applications in meteorology. Subsection 5.1.3 gives an overview of the remaining sections in this chapter.

5.1.1 Overview

Which statistical quantities are estimated?

Wavelet smoothing combines statistical concepts with the mathematical framework of the wavelet transform. Similarly to the neighbourhood methodology, kriging or kernel smoothing, it also aims to derive non-parametric estimates of a large array of parameters on the basis of a single realisation on each parameter.

Unlike the neighbourhood methodology, wavelet smoothing does not provide a tool for the simultaneous estimation of expected value, exceedance probability and quantiles. In this study, wavelet smoothing is exclusively used for the estimation of the *expected value*. As in kernel smoothing, the smoothing procedure is applied within the framework of non-parametric regression of a function.

In a future study, the procedure can easily be extended to the estimation of exceedance probabilities by replacing the data of the raw model output by indicator data (cf. Eq. 4.4) before wavelet smoothing is applied. Then the smoothing procedure would be applied within the framework of non-parametric density estimation or intensity estimation.

Similarly to kriging and kernel smoothing, wavelet smoothing can also be phrased as a weighted sum of the model output values located at different points in space. For this reason, wavelet smoothing cannot provide a direct gateway to the non-parametric estimation of quantiles which is inseparably linked to the operations of ordering and sorting the realisations by their value (see also Subsection 4.1.2).

Note, however, that the procedure of wavelet smoothing markedly differs from the procedures of the kernel, kriging or neighbourhood method. The spatial weights implicitly

result from a selective alteration of the model output in the wavelet domain, as is described below. As compared to the kriging, kernel, and neighbourhood method, this is a completely new approach to the estimation of the expected value in this study so far.

Relation to image processing and the concept of denoising

Wavelet smoothing belongs to the wide class of methods in image processing. Methods of image processing are particularly suitable for the treatment of LM output, because the data is available on a lattice. For comparison, the kriging method does not take advantage of this fact, because it has originally been developed for the treatment of irregularly spaced data.

In the context of wavelet smoothing, some new technical terms are introduced. Estimating the expected value is synonymous with *denoising* the DMO. Wavelet smoothing aims at the decomposition of the raw model output into a *signal* and the *noise*. The signal comprises deterministically predictable features and the noise comprises deterministically unpredictable features. The expected value is estimated by a filtering operation which removes the noise and preserves the signal.

Why using wavelets?

The use of wavelets in meteorology is in some way "natural", as there is often a structural affinity between the mathematical framework they offer and the physical nature of the process under study. We often seek to decompose a function into components using basis functions in the hope that this decomposition will enable us to see something different about the process or enable us to perform certain operations on the process with greater ease. Usually, the choice of the basis set for the decomposition is governed by the property that we wish to be revealed by the decomposition.

In the context of denoising the LM output, the use of wavelets is considered advantageous because of their connection to statistical multi-scale modelling, their ability to represent data in a parsimonious way and their noise-whitening property. When the LM output is transformed to the wavelet domain, the signal may become easily accessible and clearly distinguishable from the noise. Thus, the denoising procedure may become very simple and very flexible.

As opposed to the kernel method, wavelet smoothing can even be *nonlinear*, in the sense that the weights in the spatial averaging procedure are not pre-determined, but directly influenced by the noisy data itself. In this study, wavelet smoothing is set up as a linear filter (Section 5.3) and as a nonlinear filter (Section 5.4). The linear filter is theoretically similar to a kernel-based filter, but computationally much more efficient. The nonlinear wavelet-based filter is not only faster than the kernel-based filter, but also *adapts* to the noisy data in an automatic way. The degree of smoothing is not connected to a pre-defined global bandwidth but locally adapts to the unknown smoothness of

the underlying signal. The nonlinear estimator even has desirable statistical optimality properties if the nonlinear filter is properly tuned and if the noisy data satisfy a certain statistical model.

Pretensions of wavelet smoothing in comparison with the neighbourhood method

Wavelet smoothing does not contain any explicit consideration of explanatory variables in the way the neighbourhood method does (cf. Section 4.4). Thus, wavelet smoothing rather relies upon statistical reasoning than upon physical reasoning. Due to the lack of physical elements in the method, the resulting estimates of wavelet smoothing are not necessarily expected to outperform the resulting estimates of the neighbourhood method. Compared to the design of the neighbourhood methodology, the design of wavelet smoothing lays more emphasis on computational efficiency and flexibility. This study shows whether the underlying statistical concept of wavelet smoothing is also sufficient to yield reasonable results in LM post-processing. If so, wavelet smoothing would be an alternative tool in LM post-processing that requires less computational and conceptual effort than the neighbourhood method.

Although wavelet smoothing does not aim at an immediate improvement of the neighbourhood results, it certainly does aim at an improvement of results obtained from the simple neighbourhood that does without explanatory variables. It has been shown that the simple neighbourhood method is equivalent to kernel smoothing (cf. Subsection 4.3.1). A comparison between kernel smoothing and wavelet smoothing may be informative indeed, because their concepts are closely related. Both methods are based on the statistical model of non-parametric regression (Eq. 4.6) and consist in a filtering operation. Kernel smoothing can be viewed as the conventional method, while wavelet smoothing is an innovative method which has emerged only recently. Therefore, results of simple kernel smoothing play the role of a low benchmark when results of wavelet smoothing are assessed.

5.1.2 Existing wavelet applications

General remarks

The concept of wavelet transforms was formalized in early 1980's and has been extensively applied in a wide area of subjects since then. Wavelet transforms possess a multitude of nice properties and there is a great variety of commonly used wavelet transforms. The choice of a particular type of wavelet depends on the kind of mathematical properties that are desired in the application. Therefore, it is important to be aware which mathematical properties are needed and to understand the particular motivation for the use of wavelets. Appendix A.1 presents a general overview of wavelets and their

applications. In particular, the difference between continuous and discrete wavelets is explained.

Here the focus is on existing wavelet applications in meteorology. Even within the subject of meteorology the motivations of wavelet applications can differ substantially and the motivations for their use are multifaceted. First, a general overview is given and then emphasis is laid on wavelet-based noise reduction of meteorological fields, an area that coincides with the aims of wavelet smoothing in this chapter.

Wavelet applications in meteorology

As wavelet transforms are apt to address different scales while retaining "local" information, they are a popular tool in time series analysis (e. g. Torrence and Compo, 1998) or multi-scale analysis of unsteady and intermittent signals such as rainfall or clouds (e. g. Kumar and Foufoula-Georgiou, 1993a). Wavelets are also used to detect scaling in rainfall and clouds or turbulence (Farge, 1992; Kumar and Foufoula-Georgiou, 1993b,c) and thereby take advantage of the self-similarity of wavelets (Abry et al., 2000).

As there is often a link between the scale of a phenomenon, its statistical properties and the respective physical process (Perica and Foufoula-Georgiou, 1996; Harris and Foufoula-Georgiou, 2001), wavelets also serve as a lens for decomposing meteorological fields according to processes (e. g. Yano et al., 2001).

Another motivation for the use of wavelets is their noise-whitening property (Flandrin, 1992; Masry, 1993) which could be useful in the analysis and approximate synthesis of stochastic processes that are present in meteorology.

Furthermore, wavelets are applied to develop multivariate measures in spatial forecast verification (Briggs and Levine, 1997; Casati et al., 2004), in the analysis of global synoptic data such as atmospheric blocking (Fournier, 2000), and for data compression of global atmospheric data on the sphere (Göttelmann, 1999). In these studies it is the connection to the mathematical concept of *multiresolution analysis* (MRA) which makes the use of wavelets attractive, together with the fact that the wavelet transform of many meteorological phenomena is a sparse signal. These mathematical properties are also essential in wavelet-based post-processing of the LM output.

Wavelet-based noise reduction of meteorological forecast fields

Wavelet-based noise reduction of spatial fields belongs to the subject "Wavelets and Statistics" which includes the use of wavelets in non-parametric regression and density estimation. The subject emerged in the early 1990's, has gained increasing attention and has undergone further development since then. Ogden (1997), Härdle et al. (1998) and Vidakovic et al. (2000) are good references to statistical applications of wavelet transforms.

Wavelet-based noise reduction of meteorological forecast fields has been demonstrated in the context of field forecast verification. Briggs and Levine (1997) successfully apply the technique to synthetic forecast data. The data are generated by adding idealised noise to a meteorological analysis field. The design of the synthetic forecast data guarantees that the statistical properties of the data are in accordance with the underlying statistical model of the noise reduction technique.

Still, it remains open whether "real" forecast data yield similar results. It is even unclear to which extent "real" forecast data are statistically similar to their synthetic counterpart. LM post-processing presents a first attempt to apply wavelet-based denoising techniques to "real" forecast data and to derive a novel forecast product by doing so.

5.1.3 Organisation of the chapter

Section 5.2 outlines the general procedure of wavelet smoothing. Then Section 5.3 presents a linear wavelet smoothing procedure which is scale-dependent. Section 5.3 also points to a lack of robustness in the resulting estimate and cures the problem by developing the *translation-invariant* wavelet smoothing procedure. Section 5.4 presents a nonlinear wavelet smoothing procedure: *wavelet thresholding*. Section 5.5 introduces a refinement to wavelet thresholding: the consideration of another forecast variable which provides additional information about the signal-to-noise ratio in the field to be post-processed. Section 5.6 presents a summary and an outlook.

Resulting estimates of the expected value are presented and subjectively assessed throughout this chapter. An objective assessment of their forecast goodness by comparison with observational data follows in Chapter 6.

Technical and mathematical details of the wavelet transform are transferred to Appendix A and Appendix B whenever possible. The interested reader is referred to these appendices if she desires more background information.

5.2 General methodology

5.2.1 Overview

The noise filtering procedure consists of three steps (Fig. 5.1): (1) transformation of the forecast field into wavelet domain, (2) curtail those wavelet coefficients which are believed to be primarily associated with noise and retain the others, (3) backward transformation of the data to the original spatial grid.

As opposed to the spatio-temporal neighbourhood methodology of the previous chapter, wavelet smoothing is confined to two dimensions, i. e. to purely spatial forecast fields. From a physical point of view, an additional inclusion of the time dimension would be

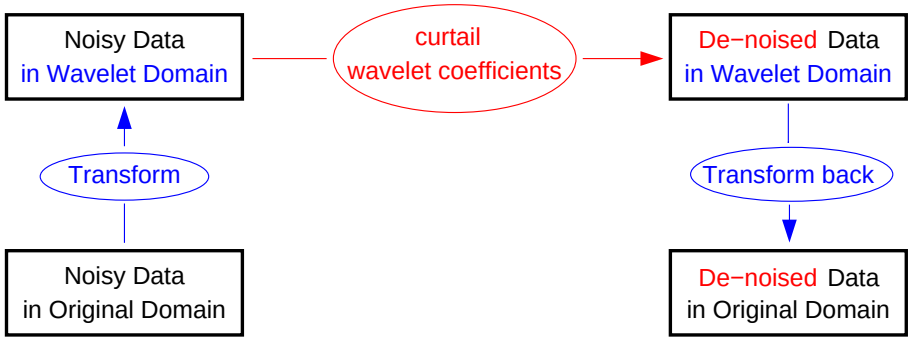

Figure 5.1: *Procedure of wavelet-based post-processing.*

desirable. Theoretically, this would be feasible by extending the wavelet method to a 3D wavelet transform. Due to technical reasons, however, the time dimension is still omitted in this study, because a 3D wavelet transform would require a much finer grid of output times than currently available. The 3D version of wavelet smoothing would not be applicable to the operational environment of DWD, since the operational output interval of $\Delta t = 1$ hour is far too large.

5.2.2 Wavelet transform

The wavelet transform algorithm

The wavelet transform algorithm is chosen to be the two-dimensional lifting scheme. Appendix A.4 describes the algorithm in detail. Thomas Gerstner (Institute for Numerical Simulation, University of Bonn) has kindly provided the corresponding software. The wavelet transform of the lifting scheme possesses several nice properties which are further explained in Appendix A. In this study, the following are beneficial:

- The wavelet transform is bi-orthogonal, i. e. the transform is directly linked to the mathematical concept of multiresolution analysis (MRA).
- The wavelet transform is moment-preserving, i. e. the first moment of the function is preserved in every *approximation space* of the MRA (cf. Appendix A.2).
- The corresponding wavelets are defined on a lattice, so discretisation is unnecessary.
- The transformation algorithm is highly efficient.

The connection to MRA is indispensable in this study, because the MRA framework is essential to the statistical wavelet regression concept (Ogden, 1997; Härdle et al., 1998; Vidakovic, 1999). MRA holds the key to a scale-dependent field decomposition, to a parsimonious data representation and to approximation in Besov spaces (Härdle et al., 1998). A comprehensive definition of MRA can be found in Appendix A.2.

Bi-orthogonality is a slightly weaker condition than orthogonality (cf. Appendix A). Orthogonality is set aside in this study, because the only real and compactly supported symmetric or antisymmetric wavelet that fits the orthogonal MRA is the Haar wavelet. The Haar wavelet is abandoned in favour of a smoother function, because the square-edged shape of the Haar wavelet might cause artefacts in the post-processed forecast field.

Moment preservation (Härdle et al., 1998; Jensen and la Cour-Harbo, 2001) is especially desirable, as this study deals with physical fields. When a meteorological forecast field is post-processed by linear wavelet smoothing, the area average of the forecast variable is preserved.

The fact that the lifting scheme is a highly efficient algorithm further maintains the low-budget character of LM post-processing and lays the technical foundations for possible operational use of the method.

The lifting scheme requires that the side length of the model grid consists of a number of points that is a power of 2 plus 1 (Appendix A.4), i. e. $N_x = N_y = 2^n + 1$, $n \in \mathbb{N}$. In general, this is not satisfied by the spatial domain of the LM. Adapting the wavelet transform algorithm to arbitrary-length data sets is not trivial. Another option is to manipulate the data set so that the size of the data set is increased to the next larger power of two (Ogden, 1997). However, these attempts introduce problems such as a dilution of the signal or the introduction of an artificial auto-correlation between wavelet coefficients. For these reasons a very simple solution is selected: Only a fraction of the entire LM domain is post-processed which has the desired size of $(2^n + 1) \times (2^n + 1)$ grid points.

Illustration of the wavelet transform

The following example illustrates the wavelet transforms of a LM precipitation forecast and a LM 2m-temperature forecast. First, a section of $(2^8 + 1) \times (2^8 + 1)$ $= 257 \times 257$ grid points in size is selected from the entire LM domain (Fig. 5.2 and 5.4). The entire LM domain comprises 325×325 grid points.

Then the respective wavelet transforms are derived. The absolute values of the wavelet coefficients are depicted in Fig. 5.3 and 5.5. Fig. 5.6 illustrates the layout of their presentation. Each square represents one of the three directional subspaces of the detail space $\boldsymbol{W}_l$ and contains wavelet coefficients that pertain to a specific level l and to a specific direction v, h or d. Within a square, the coefficients are arranged in order of their two location indices m_1, m_2. In the following paragraphs, the terms *level*, *direction* and *location* are explained very briefly. More comprehensive explanations can be found in Appendix A.

The level $l \in \{-n, \ldots, -1\}$ relates the wavelet coefficient to a spatial scale. The finest scale is $l = -1$ and the coarsest scale is $l = -n$. The coarser the level, the smaller the

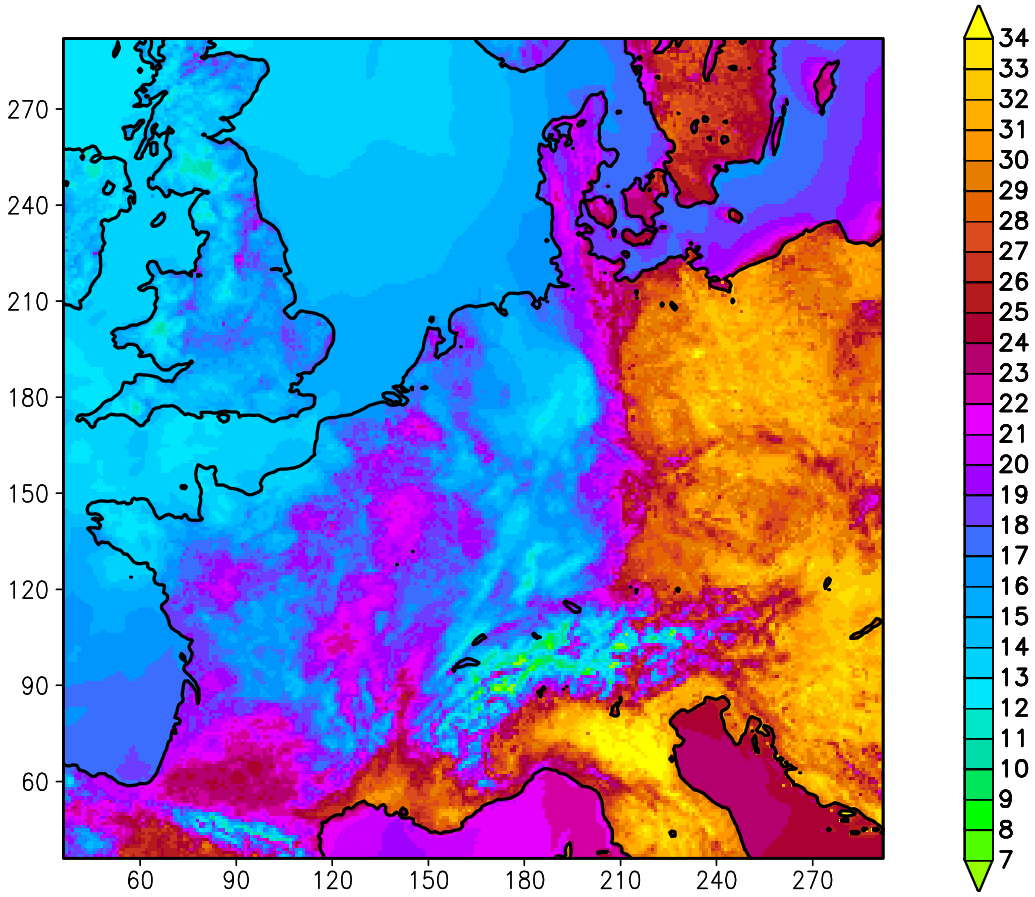

Figure 5.2: *Original LM forecast of 2m-temperature on July 10, 2002, 14 UTC. Unit: Degree Celsius. A section of* $(2^8 + 1) \times (2^8 + 1)$ *grid points is selected from the LM domain.*

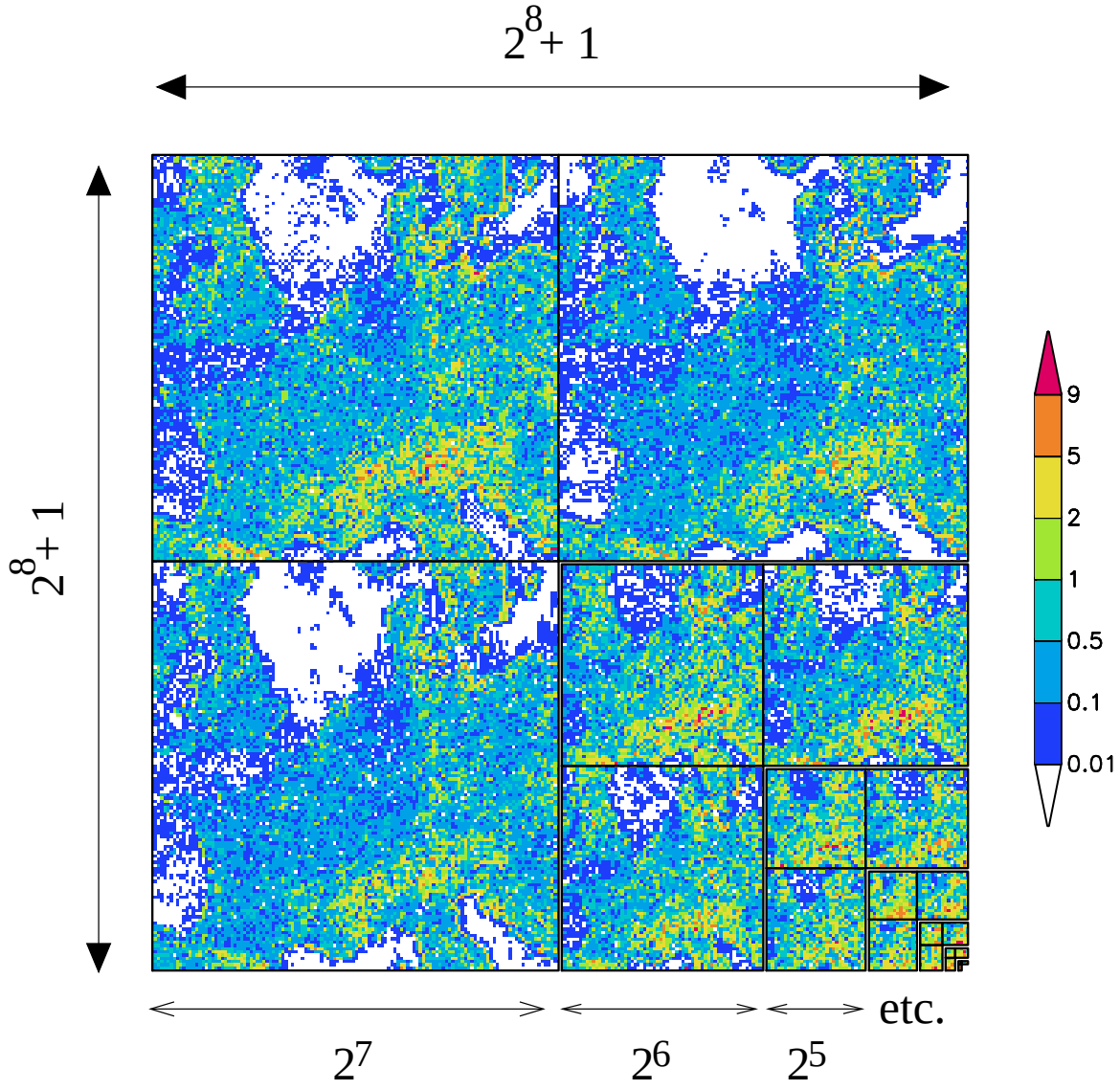

Figure 5.3: *Wavelet transform of Fig. 5.2. Absolute value of the wavelet coefficients. Unit: Kelvin.*

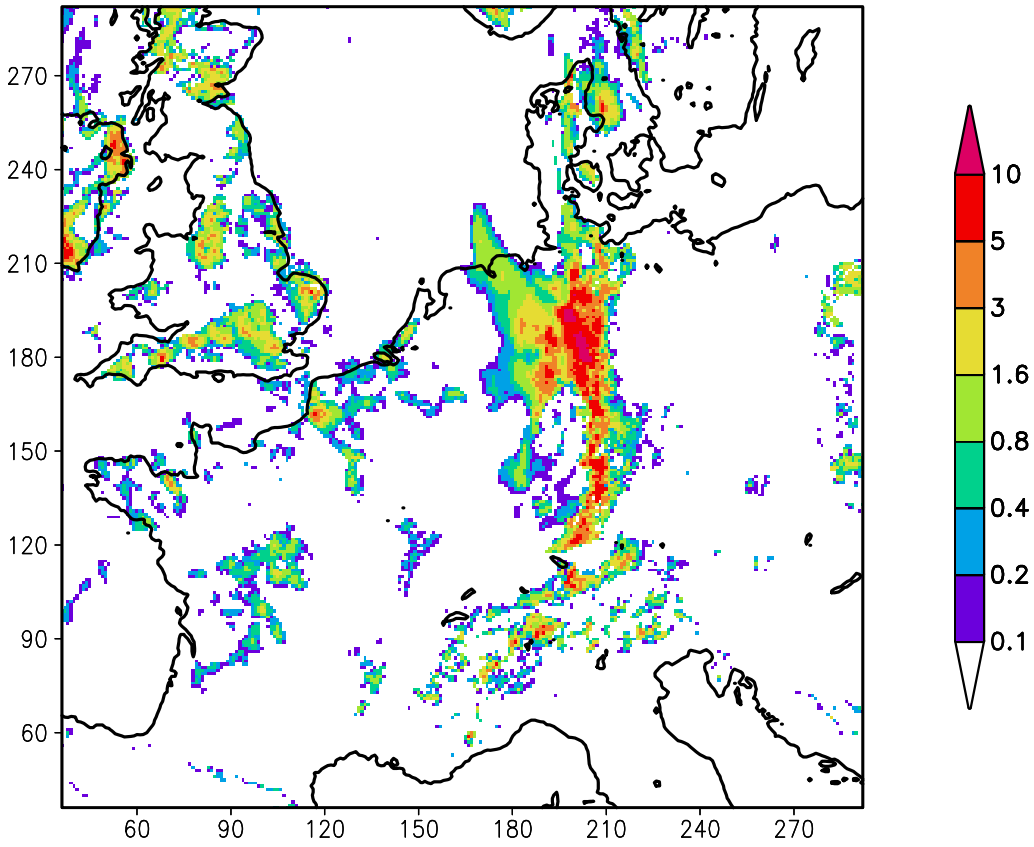

Figure 5.4: *Original LM forecast of 1h-accumulations of precipitation on July 10, 2002, 13–14 UTC. Unit: mm. A section of* $(2^8+1)\times(2^8+1)$ *grid points is selected from the LM domain.*

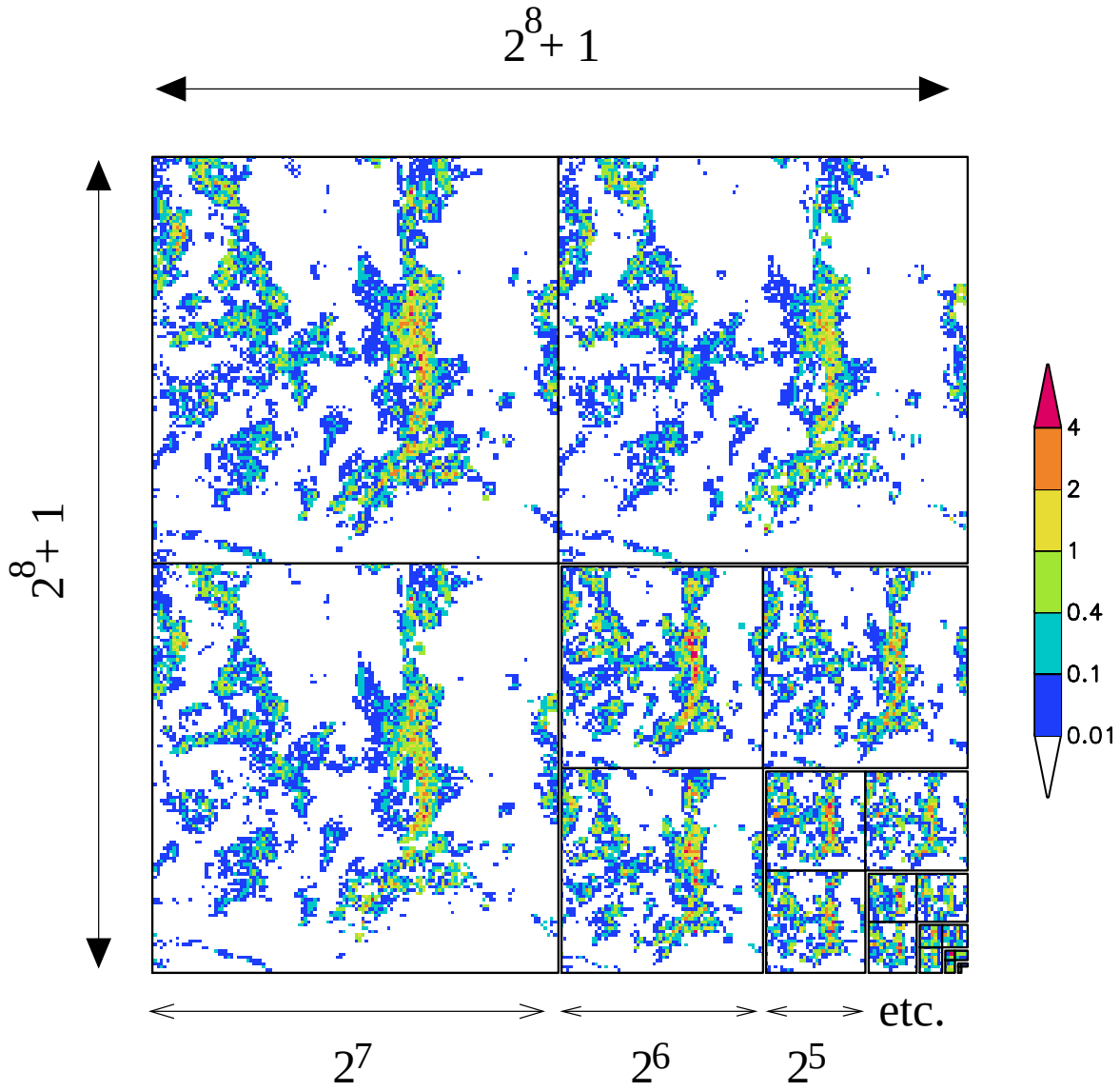

Figure 5.5: *Wavelet transform of Fig. 5.4. Absolute value of the wavelet coefficients. Unit: mm.*

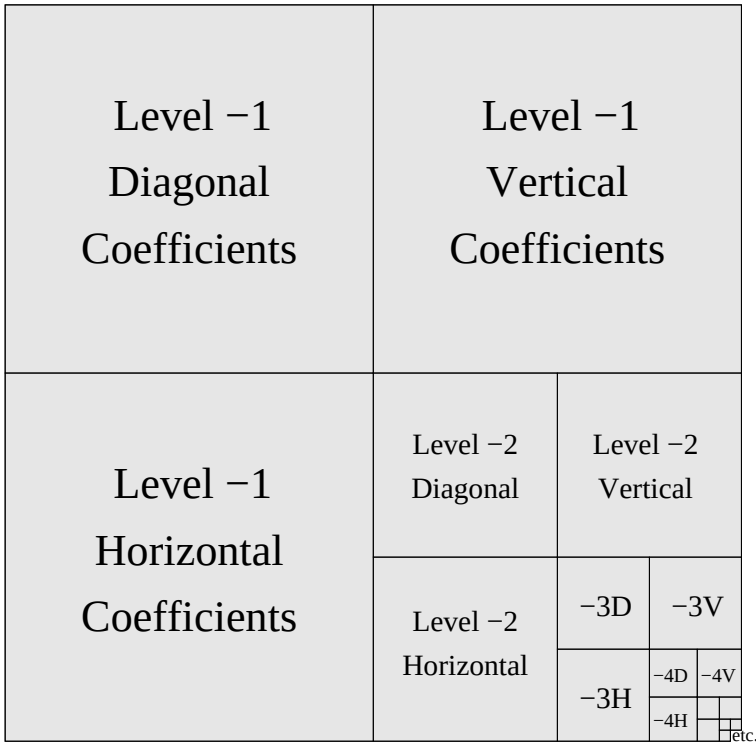

Figure 5.6: *Layout of wavelet coefficients depicted in Fig. 5.3 and 5.5. The horizontal, vertical and diagonal directions are perpendicular to the direction of the gravitational force. See text for details.*

number of different locations (cf. Appendix A.4). In Fig. 5.3 and 5.5 the coarsest levels are omitted.

Each level contains three directional subspaces: the horizontal (h), vertical (v) and diagonal (d) direction. The terms "horizontal", "vertical" and "diagonal" are borrowed from the language of two-dimensional image processing. When the wavelet transformation is applied to meteorological fields in this study, the horizontal, vertical and diagonal directions are perpendicular to the direction of the gravitational force. The vertical direction is parallel to the northward direction and the horizontal direction is parallel to the eastward direction. The explicit consideration of the diagonal directions originates from the fact that the wavelet transform is restricted to only one level index l although the transform relates to a function that has two dimensions in space (cf. Appendix A.3).

Wavelet coefficients of the same level and direction belong to different locations. When they are arranged in order of their two location indices m_1 and m_2, they roughly reflect the structure of the original function in space. For example, this becomes visible in the various "tiles" of Fig. 5.3 and 5.5.

In this section, Fig. 5.3 and 5.5 are simply used to illustrate the technical issues of the wavelet transform. Further discussion of Fig. 5.3 and 5.5 follows in Section 5.4 which reveals interesting characteristics in the wavelet transform of the LM forecasts.

5.2.3 Wavelet curtailing

The key element of the post-processing procedure is the modification of the wavelet coefficients. This step is supposed to separate deterministically predictable features from random forecast error. Coefficients that are suspected to have a high signal-to-noise ratio are retained and coefficients with a presumably low signal-to-noise ratio are curtailed in some way.

Criteria are sought that enable to estimate the approximate signal-to-noise ratio of a wavelet coefficient. Several criteria are conceivable:

- level l (= associated spatial scale) of the wavelet coefficient
- location m_1, m_2 of the wavelet coefficient
- direction h, d, v of the wavelet coefficient
- absolute value of the wavelet coefficient (*thresholding*)
- properties of an additional field
- combinations of the criteria above

The level criterion implies that the signal-to-noise ratio depends on spatial scale and allows for scale-dependent smoothing. The location criterion assumes that the signal-to-noise ratio varies with spatial location. For example, the location criterion can be used to explicitly prevent spatial smoothing of 2m-temperature along coastlines. The direction criterion allows for anisotropic smoothing, for example in line with the spatial orientation of orographic structures. The absolute value criterion allows for nonlinear smoothing and is motivated in Section 5.4. The additional field criterion looks in the wavelet transform of an additional field in order to get information on the signal-to-noise ratio of the field to be post-processed. This approach is promising if there are interrelations between the deterministic predictability of one variable and field characteristics of an additional variable.

As to give a first insight into the technique of wavelet smoothing, the level criterion is presented in Section 5.3. The absolute value criterion and the additional field criterion are investigated and combined in the following sections. The application of other criteria remains for future research.

5.3 Level criterion

5.3.1 Curtailing procedure

Motivation

Using the level as a criterion in curtailing wavelet coefficients is motivated by the scale-dependency of predictability that is discussed in the context of the neighbourhood methodology (Section 4.2.3).

Methodology

In order to remove the statistical information and retain the deterministic information of a model forecast, it is reasonable to set the wavelet coefficients of the finest levels to zero. The rule of modifying the wavelet coefficients reads as follows:

$$\gamma^{\text{mod}}_{l,m_1,m_2} = \begin{cases} \gamma_{l,m_1,m_2} & : \quad l \geq l_{\text{curtail}} \\ 0 & : \quad l < l_{\text{curtail}} \end{cases} \tag{5.1}$$

with γ_{l,m_1,m_2} denoting the wavelet coefficient of the forecast field on level $l \in \{-n, \ldots, -1\}$ and at location (m_1, m_2) and $\gamma^{\text{mod}}_{l,m_1,m_2}$ denoting the modified wavelet coefficient. The level $l_{\text{curtail}} \in \{-n+1, \ldots, -1\}$ is chosen *a priori*. The level $l_{\text{curtail}} - 1$ is the finest level where the corresponding wavelet coefficients remain untouched.

Linearity of the method and relation to kernel smoothing

Wavelet regression according to the level criterion is a *linear* filter, in the sense that the estimator of the expected value is a linear function of the original noisy forecast field. The estimate of the expected value is a projection of the original forecast field on the subspace $\boldsymbol{V}_{l_{\text{curtail}}}$ (cf. Appendix A.2). The wavelet functions Ψ_{l,m_1,m_2} and the scaling functions Φ_{l,m_1,m_2} (cf. Appendix A.2) can be regarded as certain scaled "kernels", and their scale is defined by the value $l \in [l_{\text{curtail}}, 0]$ (Vidakovic, 1999).

5.3.2 Level criterion in action

Wavelet smoothing according to the level criterion is applied to the original LM forecast of 2m-temperature and precipitation. The forecasts are transformed to the wavelet domain, all coefficients of the finest two levels are set to zero and then the spatial fields are reconstructed. The choice of $l_{\text{curtail}} = -2$ merely serves as a first example and might not be optimal yet. The procedure is expected to coarsen the direct model output to an effective grid box size of approximately $4\Delta x$. Fig. 5.7 and 5.8 present the respective results which are discussed in the following paragraphs.

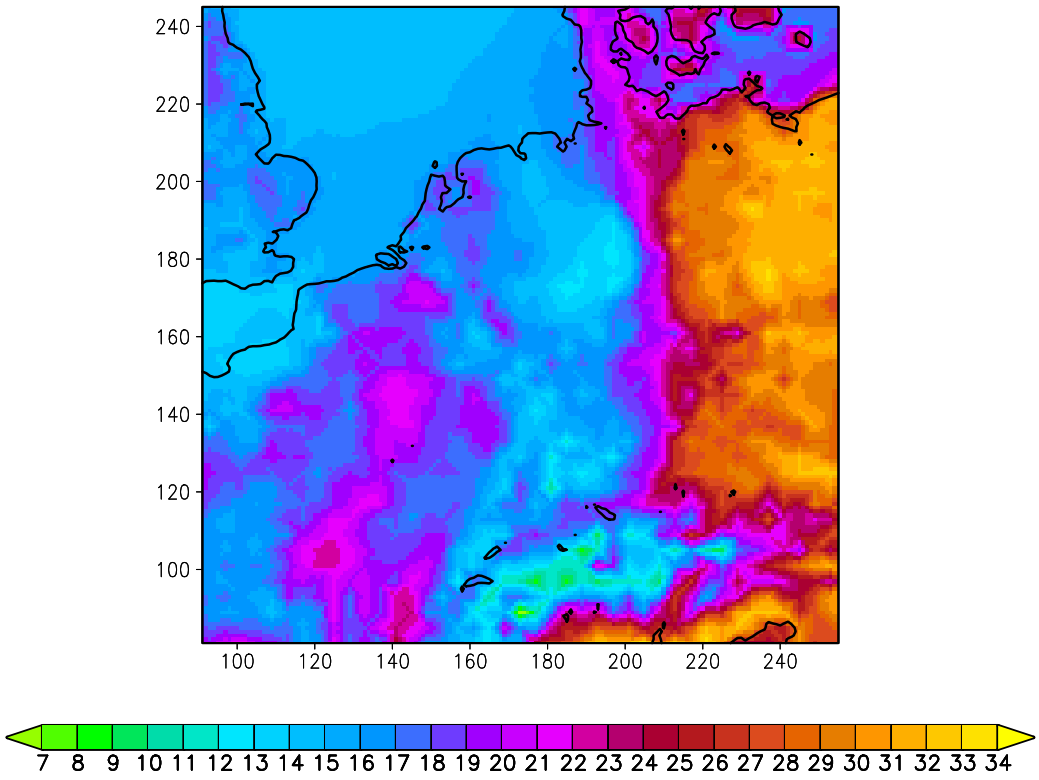

Figure 5.7: *Post-processed version of the 2m-temperature forecast in Fig. 5.2. Unit: Degree Celsius. Reconstruction after curtailing the two finest levels* ($l_{\text{curtail}} = -2$). *The full domain of the post-processed forecast comprises* 257×257 *grid points; here a section of* 165×165 *grid points is shown.*

Spatial smoothing

A comparison between the original forecasts (Fig. 5.2 and 5.4) and the post-processed forecasts (Fig. 5.7 and 5.8) shows that the main effect of the linear wavelet filter consists in spatial smoothing. The coarser the level l_{curtail}, the larger the spatial scales which are smoothed.

This is in accordance with the notion that small scales are associated with processes that possess a smaller amount of deterministic predictability. Therefore, the estimate of the expected value should not exhibit any small-scale details indeed.

Artefacts

In addition to the smoothing effect, the post-processed fields appear ragged in some places. Here and there narrow streaks intersperse with the smooth field. The streaks are strictly horizontal, vertical or diagonal. Subsection 5.3.3 shows that they arise from the lifting scheme and develops a method for their elimination.

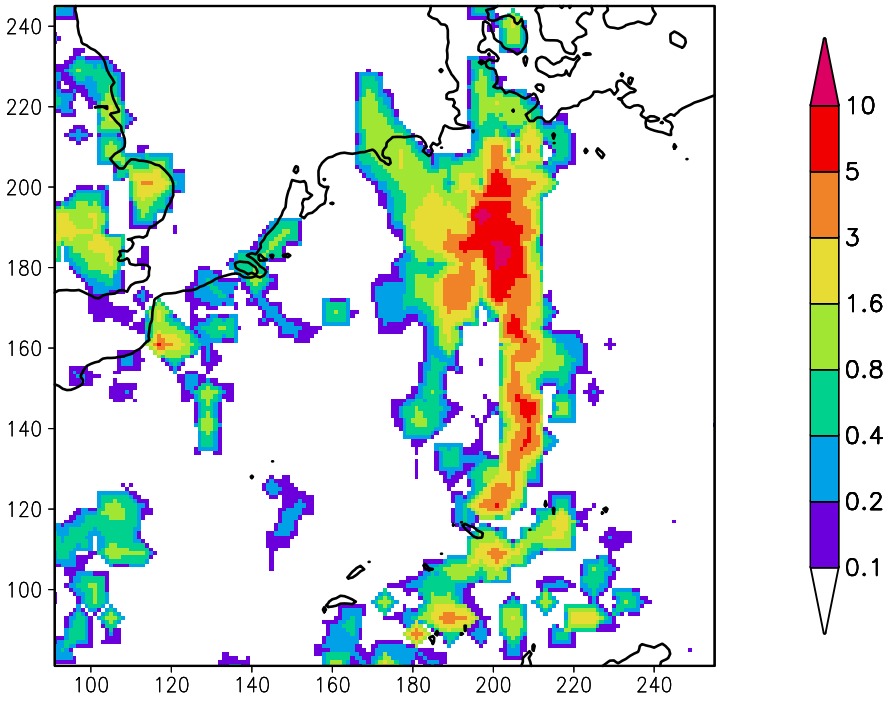

Figure 5.8: *Post-processed version of the precipitation forecast in Fig. 5.4. Unit: mm. Reconstruction after curtailing the two finest levels* ($l_{\text{curtail}} = -2$). *The full domain of the post-processed forecast comprises* 257×257 *grid points; here a section of* 165×165 *grid points is shown.*

Explanatory variables

Explanatory variables (cf. Section 4.4) are not taken into account in wavelet smoothing. This becomes particularly obvious in deficiencies of the post-processed 2m-temperature forecast (Fig. 5.7). Sharp edges in the temperature field that are associated with the land-sea mask or complex orography are eliminated in the estimated expected value even though they are deterministically predictable and are therefore part of the signal. This issue is followed up by *wavelet thresholding* in Section 5.4.

Preservation of the area mean

The moment-preserving property of the lifting scheme ensures that linear wavelet smoothing does not change the area mean of the forecast field. Due to the conservation principles of physical quantities the preservation of the area mean appears to be very desirable in this study.

When taking a closer look at the post-processed field of precipitation, it becomes visible that a few values are slightly below zero (not shown in Fig. 5.8). This is a consequence of the negative values in the wavelet function (cf. Appendix A). The example shows that

the moment-preserving property is only of mathematical nature and does not necessarily contribute to a compliance with physical conservation principles.

As negative precipitation values might be disturbing to the user of the forecast, all negative values of the post-processed forecast are set to zero. Consequently, the area mean of the post-processed forecast is slightly larger compared to the original forecast. This fact is accepted here, because the difference amounts to only 0.7% of the area mean in the case study shown.

Is linear wavelet smoothing superior to kernel smoothing?

Subsection 5.1.1 already indicates that results of wavelet smoothing should not be compared to results of the standard neighbourhood method, but to results of the simple neighbourhood method without explanatory variables or to results of kernel smoothing. Subsection 4.3.2 shows that the simple neighbourhood method is equivalent to the use of the spherical kernel K_{ball} (Eq. 4.10) in kernel smoothing.

Since the results of kernel smoothing are supposed to play the role of a low benchmark in this study, an even simpler kernel function is used here, namely the two-dimensional rectangular kernel K_{rect}:

$$K_{\text{rect}}(u,v,w) = \begin{cases} 1 & : \quad |u| \leq 1 \wedge |v| \leq 1 \wedge |w| = 0 \\ 0 & : \quad |u| > 1 \vee |v| > 1 \vee |w| > 0 \end{cases} \quad \text{with} \quad u,v,w \in \mathbb{R}. \tag{5.2}$$

In order to reach best correspondence with the procedure of wavelet smoothing, the rectangular kernel is confined to the two spatial dimensions and omits the temporal dimension. The bandwidths take the values $h_x = h_y = 2$ so that the smoothing intensity of the kernel method is similar to that of the linear wavelet method when the two finest wavelet levels are curtailed ($l_{\text{curtail}} = -2$).

A subjective inspection of the post-processed precipitation forecasts does not reveal any decisive superiority of linear wavelet smoothing as compared to kernel smoothing (cf. Fig. 5.9 and 5.8). On the contrary, linear wavelet smoothing introduces several disadvantages: the size of the LM domain is required to have a certain size, the degree of smoothing is confined to a few categories, the post-processed field exhibits artefacts, some values of precipitation become negative and the inclusion of the time dimension is hampered by technical obstacles.

However, these shortcomings are more than balanced by the massive gain of computational efficiency that comes with wavelet smoothing. Especially when the degree of smoothing is large, even the simple neighbourhood method consumes an immense amount of computational effort so that the desired low-budget character of LM post-processing is lost. In this respect, the computationally elegant wavelet methodology comes to the rescue.

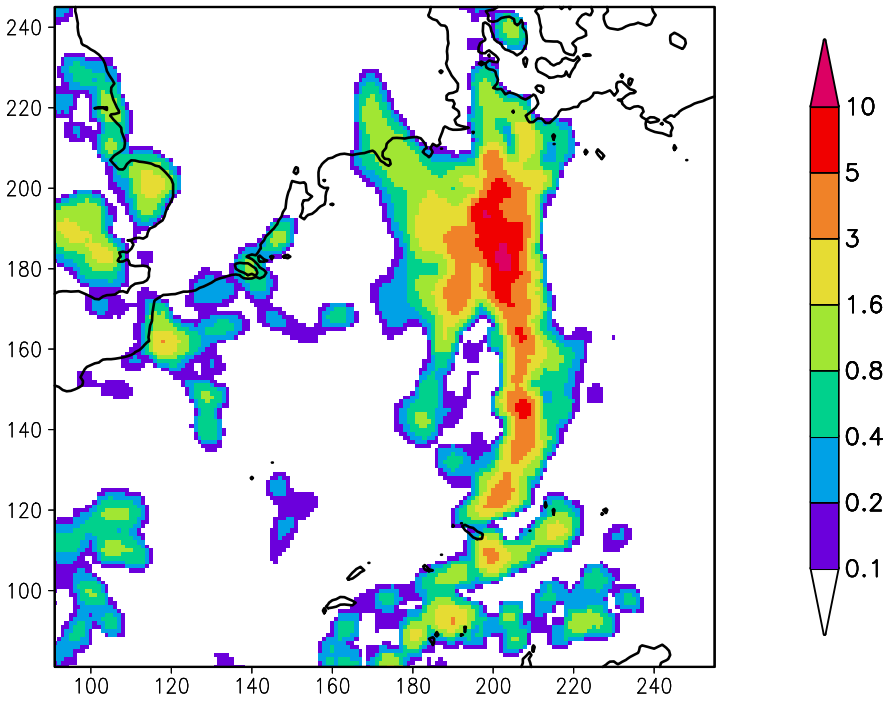

Figure 5.9: *Example of kernel smoothing with the rectangular kernel function* K_{rect} *and bandwidths* $h_x = h_y = 2$ *(Eq. 5.2). Unit: mm. The original data come from the LM precipitation forecast in Fig. 4.7.*

5.3.3 Translation-invariant wavelet smoothing

The previous subsection points to artefacts in the post-processed 2m-temperature and precipitation fields (cf. Fig. 5.7 and 5.8). They originate from the particular wavelet transform in this study, the lifting scheme. This subsection introduces the so-called *translation-invariant* wavelet transform which succeeds in eliminating the artefacts from the post-processed fields. If the translation-invariant transform is applied, level curtailing becomes equivalent to kernel smoothing.

Illustration of the problem

Linear wavelet smoothing does not only produce artefacts; the post-processed value at a specific location also depends on its position within the domain. If the post-processing procedure is repeated for a domain that is shifted by a few grid points, the same geographical location can exhibit a different estimate of the expected value. An illustration of this behaviour is presented in Appendix B.1.

The effect of shifting is due to the dependence of the wavelet transform on details of the lattice. The lattice is determined by the choice of its origin, i. e. the geographical location of the south-western corner of the domain. Further explanations are given in

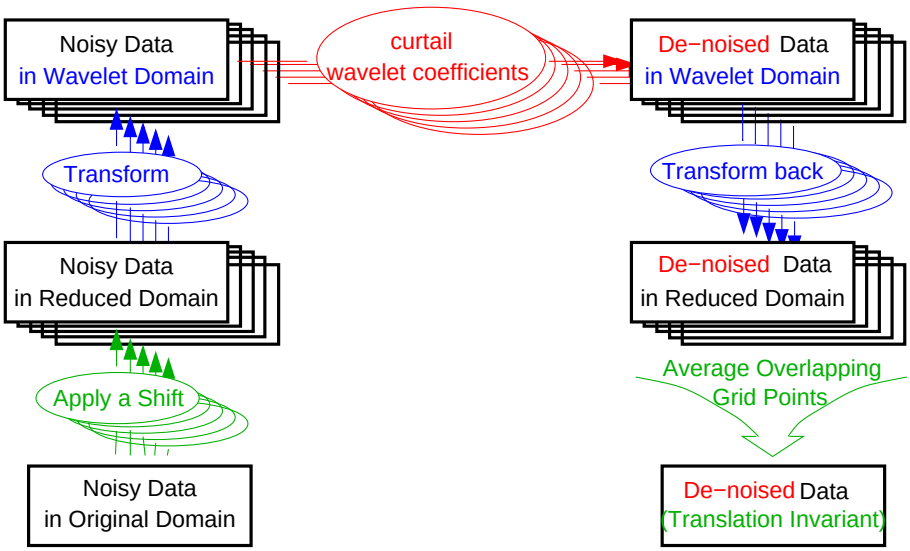

Figure 5.10: *Translation-invariant wavelet denoising by averaging shifts. Application of a range of shifts, reconstruction of the respective fields after curtailing some wavelet coefficients and averaging the several results so obtained.*

Appendix B.1.

As the particular choice of origin is made in an arbitrary way, the resulting estimate is arbitrary to a certain extent, too. In this sense, the estimator lacks robustness. In order to attain an improved estimate and to avoid artefacts in the field, the effect of the origin needs to be eliminated.

Methodology

The problem can be solved by applying a range of shifts, and average over the several results so obtained (Fig. 5.10). The approach of averaging shifts is also called *translation-invariant* wavelet denoising (cf. Coifman and Donoho, 1995).

Appendix B.2 elaborates the details of the method. A key question is the optimal range of shifts. A large number of shifts results in a substantial increase in computational effort and in a smaller number of grid locations where the estimator is derived. This study succeeds in determining a limited range of shifts by discovering redundancies in the effect of different shifts (cf. Appendix B.2). Therefore, the computational effort is kept within the bounds of feasibility.

Before delving into LM applications of the averaging shifts approach, an alternative approach to translation-invariant wavelet smoothing should be mentioned: the so-called *stationary wavelet transform* (Nason and Silverman, 1995). The stationary wavelet transform can be constructed from any ordinary wavelet transform and directly enables translation-invariant wavelet denoising without repeating the denoising procedure sev-

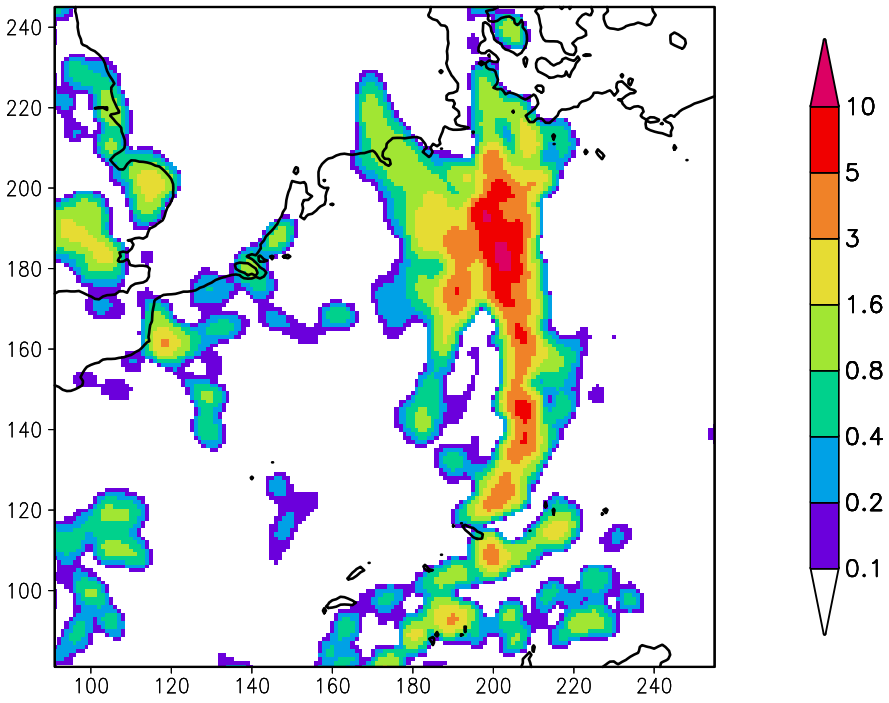

Figure 5.11: *Post-processed version of the precipitation forecast in Fig. 5.4. Unit: mm. Same post-processing procedure than applied in Fig. 5.8, but with additional shift averaging (cf. Fig. 5.10). The full domain of the post-processed forecast comprises* 250×254 *grid points; here a section of* 165×165 *grid points is shown.*

eral times. The stationary wavelet transform is equivalent to averaging different shifts of the ordinary wavelet transform, but it is more elegant and computationally more efficient (Nason and Silverman, 1995). However, the implementation of the stationary wavelet transform is conceptually more complicated than the brute force approach of averaging shifts, so LM post-processing starts out with the latter. Nevertheless, the stationary wavelet transform still appears to be an attractive concept in LM post-processing and its implementation remains for future research.

Effect of shift averaging

The examples in Fig. 5.11 and 5.12 demonstrate that shift averaging does not only result in improved robustness but also in improved smoothness when compared with results of single shift wavelet smoothing (Fig. 5.8 and Fig. 5.7). The artefacts disappear with shift averaging.

These subjective impressions are confirmed by objective verification of a larger set of forecasts (Section 6.6.4). As both subjective and objective assessment indicate a beneficial effect of the translation-invariant wavelet transform, averaging shifts is included in

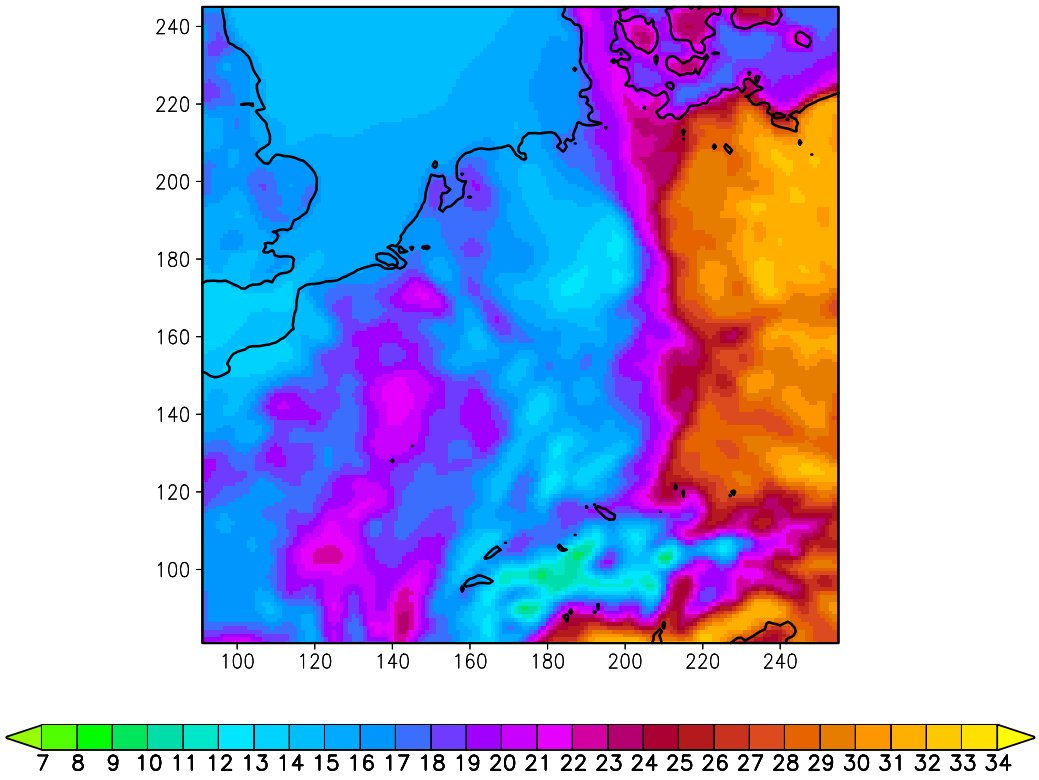

Figure 5.12: *Post-processed version of the 2m-temperature forecast in Fig. 5.2. Unit: Degree Celsius. Same post-processing procedure than applied in Fig. 5.7, but with additional shift averaging (cf. Fig. 5.10). The full domain of the post-processed forecast comprises* 250×254 *grid points; here a section of* 165×165 *grid points is shown.*

all applications of the wavelet transform that follow in this study.

Equivalence between level curtailing and kernel smoothing

The results of translation-invariant wavelet level curtailing resemble those of kernel smoothing (Fig. 5.9) very much. In the following paragraphs, it is shown that translation-invariant wavelet smoothing according to the level criterion (Eq. 5.1) is even equivalent to the kernel method. The kernel method is assumed to leave kernel function and bandwidth unchanged on the whole domain.

Recall that linear wavelet smoothing can be rephrased as a weighted average of the original forecast. When applying the level criterion in combination with shift averaging, the spatial weights only depend on the location relative to the point where the estimate is derived. Fig. 5.13 and 5.14 depict the respective weights for two different choices of the curtailing level l_{curtail}. The different weighting functions only differ by a dilation factor.

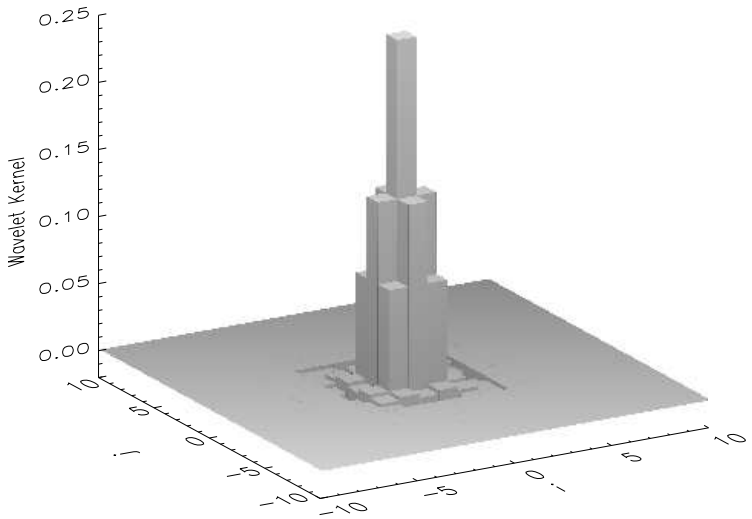

Figure 5.13: *The "wavelet kernel". Spatial weighting function that results from the wavelet smoothing procedure if the level criterion is applied in combination with shift averaging. The weighting function corresponds to a curtailing level of* $l_{\mathrm{curtail}} = -1$.

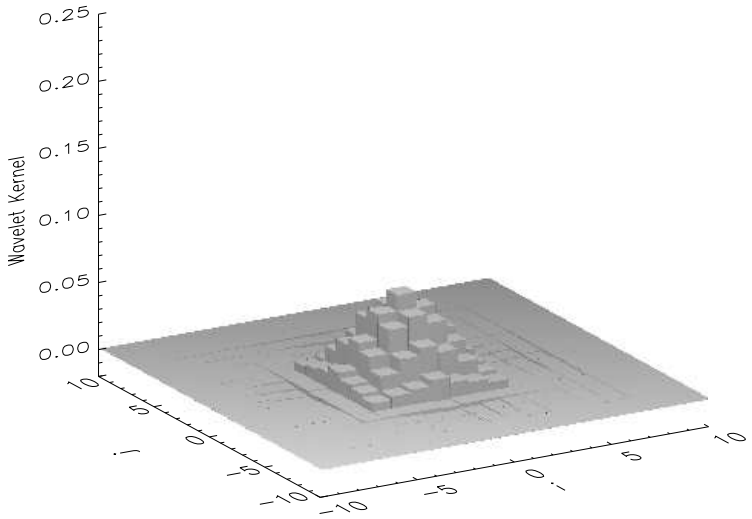

Figure 5.14: *The "wavelet kernel". Same as Fig. 5.13, but for a curtailing level of* $l_{\mathrm{curtail}} = -2$.

The weighting function can even be viewed as a kernel function, because it indeed satisfies the requirements of a discrete, two-dimensional kernel $K(i, j)$ (Rajagopalan and Lall, 1995); the function is symmetric, the weights are centred around zero and sum up to unity. Note that the "wavelet kernel" exhibits some negative values. This is unusual in a kernel function, but does not contradict the general requirements. Kernel bandwidth is fully determined by the level l_{curtail}. In contrast to kernel smoothing, the bandwidth is confined to a few categories.

Differences between the kernel smoothed field (Fig. 5.9) and the wavelet smoothed field (Fig. 5.11) are simply due to the impact of kernel shape and slight differences in kernel bandwidth. The most obvious difference between the kernel smoothed field and the wavelet smoothed field consists in an improved preservation of relative maxima and minima. A detailed interpretation of the effect due to different kernel shape is probably not worthwhile at this stage, because bandwidth selection and the selection of l_{curtail} have not been optimised yet. In comparison to kernel bandwidth, kernel shape only plays a minor role in terms of the root mean squared error of the estimate (Silverman, 1986). Linear wavelet smoothing is expected to achieve a similar accuracy in the estimate than kernel smoothing if their smoothing degree is equivalent.

Is there a relation between shift variability and forecast uncertainty?

As different smoothing results due to different shifts are treated as equally valid, they may be able to serve as a means to synthesize noise and derive an estimate of forecast uncertainty ε_{ij} in addition to the estimate of expected value μ_{ij} (Eq. 4.6). Appendix B.3 briefly discusses whether the variability within the set of smoothed forecasts is estimating something. It is argued that there is no equivalence to forecast uncertainty. Further research is needed until a statistically sound interpretation of the inflated variability field might become possible.

5.4 Absolute value criterion: Wavelet thresholding

Up to this point, this work has dealt with *linear* smoothing methodologies only. Wavelet smoothing according to the level criterion is even equivalent to kernel smoothing. This section moves on to another item of the list of curtailing criteria (cf. Section 5.2.3): the absolute level criterion, also known as *wavelet thresholding*. Wavelet thresholding clearly goes beyond kernel smoothing, as it is a *nonlinear* wavelet filter. The corresponding spatial weights are no longer pre-determined, but adapt to local properties of the underlying data.

5.4.1 Introductory remarks

Potentials of a nonlinear approach

Linear filters can only be optimised with regard to one particular degree of smoothness. It has been pointed out by many researchers that linear filters are not efficient when signals have considerable inhomogeneity such as varying degrees of smoothness (see introductory section of Donoho and Johnstone, 1998). Varying degrees of smoothness are certainly typical of spatial precipitation fields and 2m-temperature fields so that the application of nonlinear filters appears promising in LM post-processing.

In meteorology there is a structural affinity between the physical origin of the noisy fields under study and the mathematical framework a nonlinear filter offers. Noise in meteorological fields arises from a multitude of different stochastic processes. Whereas a linear approach to noise filtering implicitly assumes that one stochastic process explains the full range of noise in meteorological fields, a nonlinear approach offers the flexibility to account for multiple sources of noise.

Wavelet thresholding

The simplest nonlinear wavelet filter is *thresholding*. Wavelet thresholding was developed in the early 1990's (Donoho and Johnstone, 1994; Donoho et al., 1995) and has been used in a wide range of applications including meteorology (Briggs and Levine, 1997; Katul and Vidakovic, 1998; Yano et al., 2001). It uses the absolute value of the wavelet coefficient as a criterion in the curtailing step. This is motivated by the fact that the measures of smoothness of a function often depend on the magnitudes of its wavelet coefficients. Further explanations about the underlying theory are given in Subsection 5.4.3.

5.4.2 Curtailing procedure

Universal threshold

The two most common thresholding policies are *hard* and *soft* thresholding (Donoho and Johnstone, 1994). Hard thresholding retains wavelet coefficients that possess an absolute value greater than or equal to a universal threshold ν and sets others to zero:

$$\gamma_{l,m_1,m_2}^{\mathrm{mod}} = \begin{cases} \gamma_{l,m_1,m_2} & : \quad |\gamma_{l,m_1,m_2}| \geq \nu \\ 0 & : \quad |\gamma_{l,m_1,m_2}| < \nu \end{cases} . \tag{5.3}$$

Soft thresholding additionally reduces the absolute values of the retained coefficients by the value of ν.

Level-dependent thresholds

The procedure of wavelet thresholding is extended from using a universal threshold ν to using level-dependent thresholds ν_l. Statistical arguments maintain that the largest scale's wavelet coefficients should be left un-thresholded regardless of their size (Donoho and Johnstone, 1994). Physical arguments state that signal characteristics differ at different levels, because coarse levels show a larger proportion of important signal features (cf. Section 4.2.3). Even if noise is the same on each level, smaller thresholds on coarse levels should be preferred.

5.4.3 Why thresholding works well in theory

Intuitive arguments

Wavelet thresholding is a successful denoising procedure if small wavelet coefficients are mainly associated with noise and large wavelet coefficients are mainly associated with the signal. This assumption is justified if all of the following is true:

- The noisy data satisfies the two-dimensional statistical regression model (Eq. 4.6), i. e. the noise is Gaussian, stationary, white and zero-mean in the grid point domain.
- The chosen wavelet transform is connected to MRA and possesses the noise-whitening properties, i. e. the wavelet transform is orthogonal.
- The signal is intermittent.
- The noise does not dominate the signal.

With these prerequisites the wavelet transform of the noisy data satisfies the following conditions:

- Noise is stationary and white, i. e. it is spread out evenly over all wavelet coefficients.
- The signal is a *sparse signal*, i. e. the signal is represented by only few wavelet coefficients greater than zero.
- The noise level is low enough, so that wavelet coefficients are still distinguishable in the noisy field.

These characteristics of the wavelet transform allow to distinguish the signal from noise by merely looking at the magnitude of the wavelet coefficients.

Statistical optimality properties

In addition to the intuitive arguments above, several statistical optimality properties have been proven for the wavelet thresholding approach (Donoho and Johnstone, 1994; Donoho et al., 1995; Härdle et al., 1998). If the conditions above are satisfied and if the threshold ν is optimally chosen, the estimator attains *smoothness* and *adaptation*.

Smoothness states that there is a high probability that the estimator is as smooth as the unknown function to be estimated (Vidakovic, 1999). Adaptation refers to the so-called *minimax risk* of the estimator. The risk is the expected distance between the estimator and the unknown function, where the distance is determined by a norm which can be quite general. The minimax risk minimizes the worst case risk of the estimator within the function space. Adaptation states that the estimator achieves almost minimax risk over one of a wide range of smoothness classes, including the classes in which linear estimators do not achieve the minimax rate (Vidakovic, 1999).

These vaguely described properties of smoothness and adaptation can be formalised with the help of Besov spaces and are presented in mathematical detail by Härdle et al. (1998). Adaptation and smoothness are often illustrated by the notion of an "oracle" which provides best information on the locally optimal bandwidth in case of spatially variable bandwidth kernel estimation.

As this study uses the average shift approach (cf. Fig. 5.10), it should be mentioned that these optimality properties have only been proven for wavelet thresholding with a single shift. Nason and Silverman (1995) presume that the averaging shifts approach retains the adaptivity of ordinary wavelet thresholding and improves its performance on smooth parts of the function of interest.

5.4.4 Possible problems in practice

The above-named optimality properties might not be attained (1) if the wavelet transform is far from orthogonal or (2) if the noisy data does not even crudely satisfy the necessary conditions. In the following paragraphs, the necessary prerequisites of wavelet thresholding are discussed in view of LM post-processing. Furthermore, the wavelet coefficients of LM forecasts undergo a visual inspection with regard to their statistical properties.

Wavelet transform

Problems may arise in wavelet thresholding because the lifting scheme

- is only bi-orthogonal instead of orthogonal,
- is a *second generation* wavelet transform.

As the wavelet transform is not strictly orthogonal, it is no longer guaranteed that the noise is stationary in the wavelet transform even if it is stationary in the grid domain. In this case thresholding becomes difficult, because the amount of noise is different for each wavelet coefficient so that it is hard to remove it decently by a universal threshold. A possible remedy is to normalise the wavelet coefficients before thresholding (Jansen, 2001). The normalisation requires knowledge of the noise covariance structure in the wavelet domain. If the noise is stationary in the grid domain, the non-orthogonal transformation algorithm itself provides all necessary information. Implementing this proposal into LM post-processing remains for a future study.

Further problems may arise because the lifting scheme is a second generation wavelet transform, i. e. the transform is purely spatial and does without Fourier techniques (see Appendix A.4 for details). Jansen (2001) calls attention to problems with the condition number of the transform. A badly conditioned transform makes it difficult, if not impossible, to predict the effect of a threshold. However, it is not clear whether this issue is relevant in this study.

Conditions of the data

Ideally, the noisy data should satisfy the regression model (Eq. 4.6) and should possess an intermittent, sufficiently large signal. However, random error in original LM forecasts is suspected to exhibit the following deviations from these conditions (cf. Section 4.2.3):

- heteroscedastic noise,
- colored noise,
- non-Gaussian noise (precipitation only),
- weak signal-to-noise ratio (precipitation only).

In this study the wavelet thresholding procedure simply neglects these data deficiencies and a proper treatment remains for future research. The data deficiencies, their expected effect on wavelet thresholding and possible remedies are discussed in the following paragraphs.

The noise is heteroscedastic, because forecast uncertainty is generally not uniform in space (cf. Section 4.2.3). As a consequence, the expected noise amplitude is also different for each wavelet coefficient, and it may be impossible to remove noise by a universal threshold. The application of individual thresholds could be an alternative, but new problems may arise when trying to estimate the individual thresholds.

The noise is colored, since spatial coherence of forecast uncertainty is omnipresent in meteorological fields. As a consequence, the noise is colored in the wavelet domain, too, and thresholding becomes difficult. If the spatial covariance matrix of the noisy field was known up to constant factor, the problem can be solved by normalisation of the wavelet coefficients (Jansen, 2001). However, it is difficult to attain information about

the spatial covariance matrix without an ensemble at hand. Then the covariance matrix may be derived via a model, in a similar way as common techniques of meteorological data assimilation.

Deviations from the Gaussian distribution and a weak signal-to-noise ratio occur in the fields of precipitation. Both facts increase the risk that wavelet thresholding may fail to suppress the noise. The weak signal-to-noise ratio originates from the relatively small degree of deterministic predictability. The non-Gaussian distribution causes substantial problems if the distribution is heavy-tailed. As wavelet based methods are sensitive to outliers, the associated wavelet coefficients have a large magnitude and are wrongly classified as part of the signal.

Visual inspection of the LM wavelet coefficients

If the lifting scheme and the original LM forecasts satisfied the prerequisites of wavelet thresholding, the wavelet coefficients of LM forecasts would form the following picture: The noise is spread out evenly over all wavelet coefficients and the noise amplitude would be the same for each location and each level. Only few coefficients would represent the signal and would clearly protrude the noise amplitude.

When applying the lifting scheme to real LM forecast data, a somewhat different picture is generated (Fig. 5.3 and Fig. 5.5). In case of 2m-temperature one can distinguish regions with complex orography, smooth orography and the sea. In case of precipitation one can distinguish regions of convective rain, stratiform rain and no rain.

The differences between those regions could either originate from non-stationary noise or from variable signal characteristics. If only the signal were responsible then the prerequisites of wavelet thresholding would not be violated. However, this possibility is unrealistic, because it implies that the noise amplitude does not vary with location and therefore must comply with the lowest noise level that is apparent in Fig. 5.3 and 5.5. In case of 2m-temperature it is the noise amplitude over the sea that would control the standard and in case of precipitation it is the amplitude in non-precipitating regions. However, both appear far too low as a representative of the entire region. Therefore, it must be concluded that the wavelet transforms of LM forecast data contain non-stationary noise indeed. It remains to be seen whether this is an obstacle in wavelet thresholding.

5.4.5 Choice of the threshold

Existing methods

A central question in wavelet thresholding is the choice of the threshold value. The choice is either *a priori* or data driven. Numerous data driven approaches exist in

the literature, for example the universal threshold and the minimax threshold (Donoho and Johnstone, 1994), the SURE (Stein's unbiased risk estimate) threshold (Donoho et al., 1995), generalised cross validation procedures with regard to minimisation of mean squared error (Nason, 1995), application of multiple hypothesis tests with regard to the false discovery rate (Abramovich and Benjamini, 1995; Benjamini and Hochberg, 1995) and Bayesian approaches (Abramovich et al., 1998; Vidakovic, 1998).

In contrast to kernel bandwidth selection methods, wavelet threshold selection methods focus on rather different criteria than just balancing bias and variance. In fact, there is a great variety of different criteria: minimising the risk of erroneously curtailing the signal, minimising the risk of erroneously keeping a noisy coefficient or achieving a certain degree of smoothness in the estimate, to name only few. Consequently, there can be huge differences between the results of different threshold selection methods.

Approach in LM post-processing

In general, data driven methods should be preferred to an *a priori* approach, because they tend to be more objective. In case of threshold selection, however, the mere choice of a certain data driven method is already of subjective nature, because it is governed by the *a priori* vision of the desired result. Therefore, it is justifiable to set data driven methods aside for the moment and to start with a visual inspection of the data first.

Furthermore, an ordinary application of standard data driven methods is expected to fail in LM post-processing, because they rely on the assumption that the noisy data follows a uniform statistical model. This assumption is not satisfied by LM forecast data (cf. Subsection 5.4.4) so that existing procedures would have to be modified before they are applied to LM forecast data.

The very first attempt at choosing a threshold is based on a visual inspection of the wavelet coefficients from noisy LM forecast data (Fig. 5.5 and Fig. 5.3). A threshold value is sought that classifies the wavelet coefficients into signal and noise. The selection is roughly led by the notion that the signal must pertain to only few coefficients that clearly protrude the noise amplitude. Thus, the chosen threshold must be slightly smaller than the values of the largest wavelet coefficients.

Obviously, the selection is hampered by the fact that LM forecast noise is heteroscedastic (cf. Section 5.4.4), but a very crude guess is still made: 1.8 K for the 2m-temperature forecast and 1 mm for the precipitation forecast. In correspondence with the arguments of Section 5.3 coefficients of the coarsest levels remain un-thresholded so that the threshold is level-dependent (cf. Tab. 5.1).

In terms of 2m-temperature, the threshold value of 1.8 Kelvin is geared to land regions with smooth orography (cf. Fig. 5.3). In terms of precipitation, it is difficult to identify single coefficients that protrude the others. Therefore, the threshold value of 1 mm is just an arbitrary guess which roughly corresponds to the magnitude of wavelet coefficients

Table 5.1: *Level-dependent thresholds for wavelet thresholding.*

level l	-1	-2	-3	-4	-5	-6	-7	-8
2m-temperature: threshold ν_l [Kelvin]	1.8	1.8	1.8	1.8	0	0	0	0
precipitation (version A): threshold ν_l [mm]	1	1	1	1	0	0	0	0
precipitation (version B): threshold ν_l [mm]	∞	1	1	0	0	0	0	0

in regions of precipitation occurrence (cf. Fig. 5.5). There are two versions of threshold values (cf. Tab. 5.1). Version A is a first guess according to the considerations above; Version B is motivated in the next subsection.

5.4.6 Thresholding in action: 2m-temperature forecast

Wavelet thresholding is now applied to the 2m-temperature forecast field in Fig. 5.2. A visual inspection of the result in Fig. 5.15 gives a first impression whether wavelet thresholding is able to preserve features that are deterministically predictable and remove the others. The threshold is chosen according to Tab. 5.1. The averaging shifts approach is also applied so that the wavelet transform is translation-invariant (cf. Subsection 5.3.3).

Data-driven variations in the amount of spatial smoothing

The main effect of wavelet thresholding is spatial smoothing. This is in accordance with the notion that small scales are deterministically less predictable than large scales. In contrast to level curtailing (Fig. 5.12) the amount of smoothing varies much with location though. Whereas medium gradients and small temperature gradients are smoothed, steep temperature gradients are mostly retained. This becomes especially obvious in regions of complex orography, along the cold front, along the coastlines of Denmark or at the cities of Berlin and Prague.

The preservation of steep gradients can be explained by the way wavelet thresholding operates. Wavelet thresholding retains wavelet coefficients with large magnitudes. Large magnitudes indicate large spatial variability, because a wavelet coefficient is the deviation from a piecewise linear interpolation of the spatial field (cf. Appendix A).

Principally, variations in the amount of spatial smoothing can also be achieved by kernel smoothing with a spatially variable bandwidth. When the underlying signal is assumed to be rough, the amount of smoothing is small and vice versa. The concepts of wavelet thresholding and kernel smoothing are still entirely different; kernel smoothing uses a

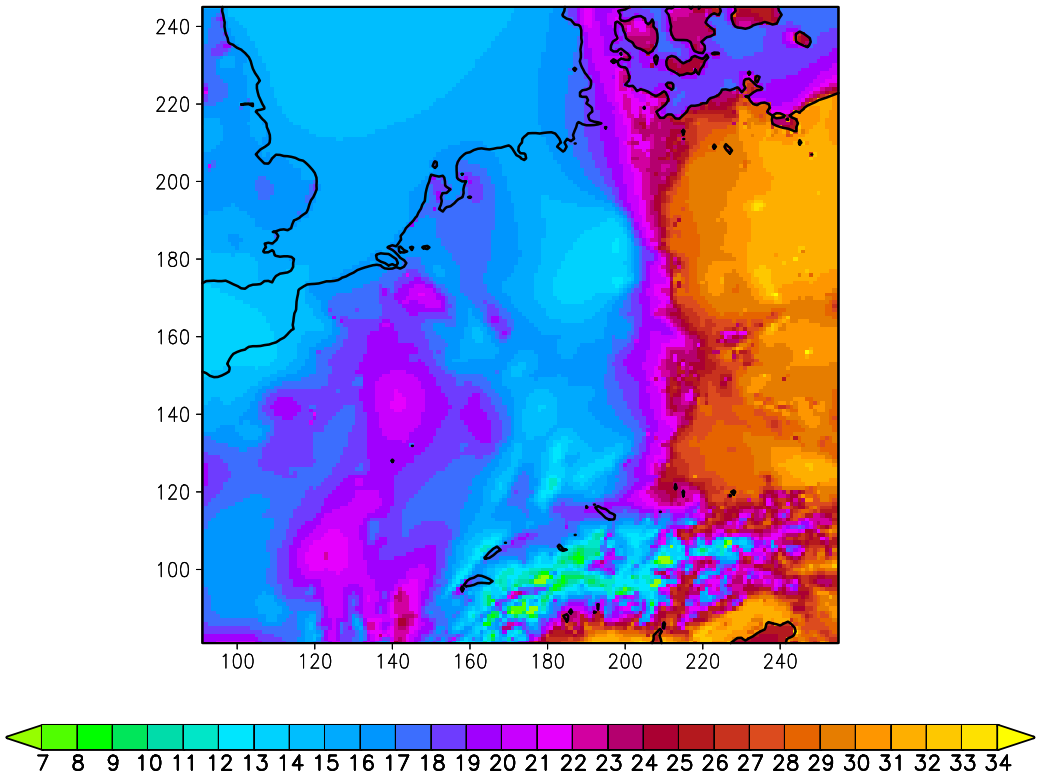

Figure 5.15: *Post-processed version of the 2m-temperature forecast in Fig. 5.2. Unit: Degree Celsius. Reconstruction after hard thresholding (Eq. 5.3) with level-dependent thresholds (Tab. 5.1). Shift averaging is also applied (cf. Fig. 5.10). The full domain of the post-processed forecast comprises* 226×242 *grid points; here a section of* 165×165 *grid points is shown.*

pre-determined bandwidth, whereas wavelet thresholding automatically adapts to local properties of the data. It is not prior knowledge, but the data itself that indicates the local roughness of the signal and controls the amount of smoothing.

Note, however, that it is unknown at this stage whether wavelet thresholding really attains adaptivity in a mathematical sense (cf. Subsection 5.4.3).

Successful preservation of the signal

Recall that wavelet smoothing aims at the preservation of the signal, i. e. the preservation of deterministically predictable features. In comparison to level curtailing (Fig. 5.12) the thresholding procedure is very good at preserving the signal.

Fig. 5.15 shows that wavelet thresholding generally preserves abrupt jumps within the 2m-temperature field. They are often associated with coastlines, orography, cities or changes in snow cover. All of these belong to the group of explanatory variables (cf. Sec-

tion 4.4) which induce deterministic predictability in temperature. Thus, wavelet thresholding preserves exactly those patterns that represent the signal.

While the neighbourhood methodology needs a lot of additional effort to retain this part of the signal, wavelet thresholding achieves the same "for free". The simple statistical model of wavelet thresholding achieves the same result than the explicit implementation of physical interrelations in the neighbourhood.

Successful removal of noise

Wavelet thresholding does not only succeed in signal preservation, but also in noise removal. A successful removal of noise eliminates all field elements that are deterministically unpredictable. For example, small-scale variations of 2m-temperature that are not associated with explanatory variables are suspected to represent the noise.

Fig. 5.15 shows indeed that wavelet thresholding has a spatial smoothing effect in those regions that are free of explanatory variables. Compared to level curtailing (cf. Fig. 5.12) the amount of smoothing is even more intense in these regions. The amount of smoothing is largest just behind the cold front where deterministic predictability is expected to be especially low.

Erroneous preservation of noise

We have seen that wavelet thresholding is surprisingly often correct in identifying the signal. However, sometimes the procedure erroneously classifies noisy coefficients as part of the signal, too, and wrongly preserves features that are deterministically unpredictable.

Such a case becomes obvious when looking at the cold front in Fig. 5.15. As the cold front is marked by steep temperature gradients, wavelet thresholding interprets the front as part of the signal, i. e. it assumes that the front is deterministically predictable in every detail. This assumption is not valid, with regard to the exact shape of the frontal line and with regard to the exact location of the front within the domain.

First, wavelet thresholding leaves the frontal line very ragged. As such details within the front are not deterministically predictable, a properly post-processed version of the frontal line should rather resemble a smooth curve instead. The simple thresholding procedure is not able to deal with discontinuities along a smooth curve, but Do et al. (2001) have recently developed a directional smoothing technique for the filtering of smooth curves. This algorithm may be able to cope with this challenge so that it is recommended for implementation into a future version of the LM post-processing procedure.

Secondly, the exact location of the front is associated with forecast uncertainty, because the front velocity is difficult to predict. In this case the proper post-processing technique

depends on the intended use of the post-processed forecast. If the post-processed forecast is strictly used as an estimate of the expected value, the field should indeed contain smoother temperature gradients in the frontal region than the original field. If the post-processed forecast is merely intended as a visual forecast guidance, it is important to keep patterns visible that are meteorologically relevant. In this case it would be suggested to retain the frontal gradients so that the front is not disguised by spatial smoothing.

Erroneous removal of the signal

Another potential weakness of wavelet thresholding is the erroneous removal of a coefficient that possesses a large signal-to-noise ratio. This case would result in spatial smoothing of deterministically predictable features in the field reconstruction of Fig. 5.15.

If at all, this case might occur in regions of the sea, since wavelet thresholding entirely removes the respective coefficients of the four finest levels there. However, excessive smoothing cannot introduce large errors over the sea, because the original 2m-temperature forecast is already very smooth.

Summary and outlook

Wavelet thresholding does a very good job in preserving elements that are deterministically predictable and in retaining the others. The criteria of wavelet thresholding appear to match our physical notion of deterministic predictability of temperature. This is due to the fact that abrupt jumps in the temperature field generally coincide with the presence of explanatory variables.

The results of wavelet thresholding appear in an even better light when considering that wavelet thresholding is such a simple and efficient methodology. In contrast to the neighbourhood methodology an explicit consideration of explanatory variables is not implemented.

Demonstrated shortcomings within the frontal area are tackled in Section 5.5. Above all, more work needs to be done on optimal threshold selection. In the long run, an automatic method is needed which is more efficient and objective than visual inspection of the data.

5.4.7 Thresholding in action: Precipitation forecast

Fig. 5.16 shows the result of applying wavelet thresholding to a LM precipitation forecast (cf. Fig. 5.4). The threshold is chosen according to version A in Tab. 5.1. Shift averaging is applied so that the wavelet transform is translation-invariant (cf. Subsection 5.3.3). Again, the post-processed forecast is examined with regard to data-driven variations in

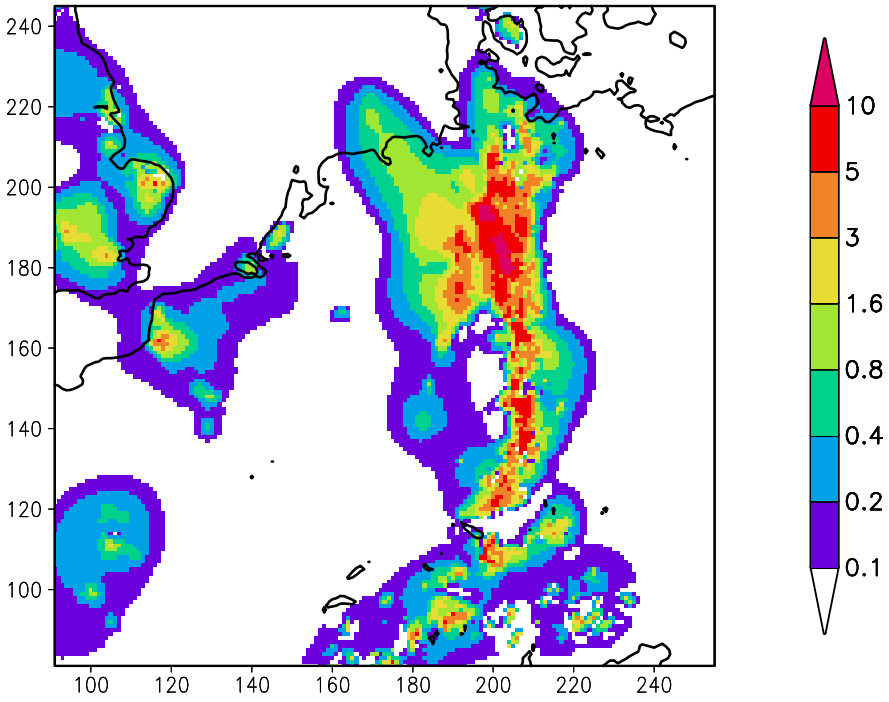

Figure 5.16: *Post-processed version of the precipitation forecast in Fig. 5.4. Unit: mm. Reconstruction after hard thresholding (Eq. 5.3) with level-dependent thresholds (Tab. 5.1, version A). Shift averaging is also applied (cf. Fig. 5.10). The full domain of the post-processed forecast comprises* 226×242 *grid points; here a section of* 165×165 *grid points is shown.*

the amount of spatial smoothing, the successful preservation of deterministically predictable features and removal of others.

Data-driven variations in the amount of spatial smoothing

The main effect of wavelet thresholding is spatial smoothing of the original forecast. As opposed to level curtailing (Fig. 5.11) the amount of smoothing varies with location. The transition between precipitation occurrence and no precipitation occurrence is diluted very much. The interior of precipitating regions almost remain untouched, especially within the rainband along the cold front.

As explained in the context of 2m-temperature, the amount of smoothing is governed by the roughness of the original precipitation field (cf. Fig. 5.4). Abrupt jumps are preserved and smoothing is applied to regions that already possess a certain degree of smoothness.

Are deterministically predictable features preserved and others removed?

Whereas the smoothing strategy of wavelet thresholding is absolutely suitable for 2m-temperature fields, it is totally inconsistent when it comes to precipitation fields.

Large gradients in the original precipitation field are often associated with convective, small-scale, intense precipitation and small gradients are generally associated with stratiform, large-scale, slight precipitation. The former is expected to possess much less deterministic predictability than the latter. Following this argumentation, the amount of smoothing should ideally be larger at locations that exhibit large gradients in the original forecast and smaller at locations that exhibit small gradients. This is exactly the opposite of the effect wavelet thresholding achieves.

The poor result of wavelet thresholding is probably due to the fact that the noise is heavy-tailed and dominates the signal. Furthermore, the noise is heteroscedastic and there appears to be a coherence between noise amplitude and signal amplitude. All these properties contradict the plan to distinguish signal and noise via the magnitude of a wavelet coefficient.

The only positive aspect of wavelet thresholding is its occasional ability to preserve orographically induced variability of precipitation. Orographically induced precipitation patterns are generally deterministically predictable and need to be preserved. As orographically induced precipitation is sometimes associated with abrupt jumps within the spatial field, wavelet thresholding preserves this part of the signal indeed.

Mixture of level curtailing and wavelet thresholding

As wavelet thresholding of precipitation fields does not lead to the desired results, another attempt is made on precipitation post-processing. The new version is a mixture of level curtailing and wavelet thresholding (see version B in Tab. 5.1). Coefficients of the finest level are entirely discarded, because the smallest scales of precipitation are generally not expected to possess deterministic predictability. Coefficients of levels coarser than $l = -3$ are entirely retained so as to limit the excessive smoothing at the transition between precipitation occurrence and no precipitation occurrence. The remaining levels are exposed to thresholding again.

The mixture of level curtailing and wavelet thresholding leads to much better result than thresholding alone (Fig. 5.17). Convective precipitation is less noisy and the patterns of stratiform precipitation are preserved in a better way.

However, it is not clear whether the result is superior to that of simple level curtailing (cf. Fig. 5.11). Wavelet thresholding produces a more detailed pattern in regions of complex orography and in regions of convective precipitation. The former may be part of the signal, but the latter may be part of the noise. Furthermore, wavelet thresholding produces greater precipitation maxima and a more frequent occurrence of small precip-

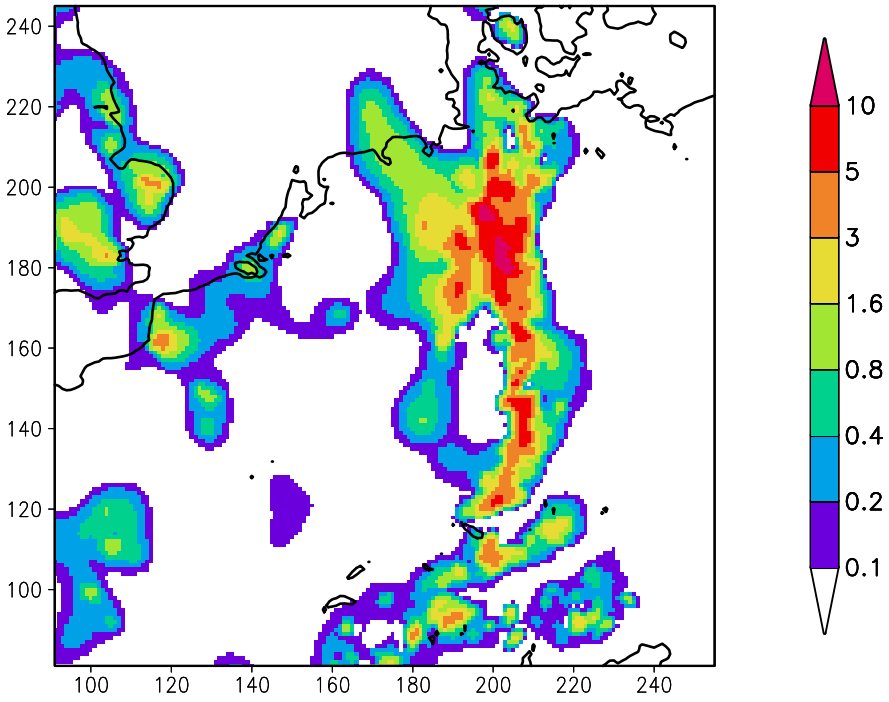

Figure 5.17: *Same as Fig. 5.16 except that version B of the thresholds is used instead of version A (cf. Tab. 5.1).*

itation amounts. As the original LM forecasts already suffer from overforecasting small amounts (cf. Chapter 6), the more frequent occurrence of small precipitation amounts may be unfavourable.

Summary and outlook

In terms of precipitation, wavelet thresholding is not able to improve results of simple level curtailing. This is due to the fact that the statistical properties of precipitation data are not in line with the underlying assumptions of wavelet thresholding. Wavelet thresholding of precipitation is not as successful as wavelet thresholding of 2m-temperature, because quantitative precipitation forecasts lack an eminent link between large gradients in space and deterministic predictability. Further attempts on wavelet thresholding of quantitative precipitation fields do not appear promising.

However, it would be interesting to investigate the use of wavelet thresholding in the estimation of exceedance probabilities of precipitation (cf. Subsection 5.6.2). Post-processing would then start from fields of precipitation occurrence (e. g. Fig. 4.8) instead from fields of quantitative precipitation (e. g. Fig. 4.7). Wavelet thresholding may be able to automatically recognise and preserve characteristic shapes within the field of precipitation occurrence, for example rainbands.

5.5 Additional field criterion

5.5.1 Introductory remarks

Motivation

In the previous section we have seen that the correct differentiation between signal and noise does not always become obvious from the LM forecast field alone. For example, orographically induced precipitation is difficult to discriminate from convective precipitation. Deterministically unpredictable temperature patterns along the front resemble those that are associated with explanatory variables.

Sometimes there are interrelations between the deterministic predictability of one variable and field characteristics of an additional variable. The additional variable could either be another forecast variable or a constant variable such as the orographic height. For example, deterministically predictable precipitation patterns are often associated with orographic variability. Furthermore, ensemble experiments (Chapter 3) indicate an interrelation between predicted convective precipitation and the ensemble spread of 2m-temperature. Thus, noisy temperature discontinuities along the front may be associated with the predicted field of convective precipitation.

Wavelet smoothing may take advantage of such interrelations between several variables. In order to account for the interrelation between two different variables, another wavelet curtailing criterion is introduced: the *additional field criterion* (cf. Subsection 5.2.3).

Related concepts

The additional field criterion can be viewed as an attempt to enrich the statistical procedure by physical information. The idea is not new in non-parametric estimation; the kriging method (Section 4.3) pursues similar aims in the framework of cokriging and multivariate kriging.

In the context of wavelet thresholding in meteorology, Yano et al. (2001) elaborate a concept of using additional fields, but with a slightly different intention. Instead of filtering a deterministically predictable pattern, they decompose a meteorological field according to its governing processes.

Overview

In this study, the precipitation forecast plays the role of the additional field in the post-processing procedure of the 2m-temperature forecast. It is suspected that the deterministic predictability of small-scale temperature decreases if convective precipitation occurs simultaneously. This interrelation is incorporated into the procedure of wavelet

curtailing (Subsection 5.5.2). An example of the post-processed 2m-temperature forecast is presented in Subsection 5.5.3.

5.5.2 Curtailing procedure

The additional field criterion is used in combination with the criterion of absolute value (cf. Section 5.4). An extension of the wavelet thresholding procedure (Eq. 5.3) is developed. The level-dependent thresholds ν_l (Tab. 5.1) are upgraded to thresholds ν_{l,m_1,m_2} that depend on level and location. The thresholds are derived via the wavelet coefficients of the additional field. The additional field must be available on the same lattice than the forecast field itself so that the respective wavelet coefficients are compatible.

The additional field criterion is related to the criterion of location (cf. Subsection 5.2.3), because it causes the post-processing procedure to be location-dependent. However, the role of different locations m_1, m_2 is governed by the additional field. The relation between the additional field criterion and the criterion of location becomes especially close when the additional field is static.

In the following example, the 2m-temperature field is post-processed and the precipitation field plays the role of the additional field in the curtailing procedure. The predictability of small-scale temperature is assumed to depend on the occurrence of precipitation. An indicator of local precipitation occurrence is small-scale variability within the precipitation field, so that the corresponding wavelet coefficient γ^z_{l,m_1,m_2} of the finest level $l = -1$ deviates from zero: $|\gamma^z_{-1,m_1,m_2}| \geq 0.001$. In order to affect the small-scale temperature field, this condition is used as a criterion to modify the threshold value of the finest level:

$$\tilde{\nu}_{-1,m_1,m_2} = \begin{cases} \nu_{-1} & : \quad |\gamma^z_{-1,m_1,m_2}| < 0.001 \\ \infty & : \quad |\gamma^z_{-1,m_1,m_2}| > 0.001 \end{cases} . \tag{5.4}$$

As a result, the finest level of the 2m-temperature forecast is entirely eliminated at locations with precipitation occurrence.

5.5.3 Additional field criterion in action

The additional field criterion is now applied to the 2m-temperature forecast field in Fig. 5.2. Wavelet thresholding is applied according to Eq. 5.3, the threshold value is chosen according to Tab. 5.1 and modified according to the additional field criterion in Eq. 5.4. The precipitation forecast in Fig. 5.4 serves as the additional field.

Shift averaging is also applied (cf. Fig. 5.10), both to the temperature field and to the additional field of precipitation. The temperature field undergoes the wavelet thresholding procedure of Eq. 5.4 for a range of shifts. When the temperature field and the precipitation field are combined in the thresholding procedure, identical shifts are chosen. Each

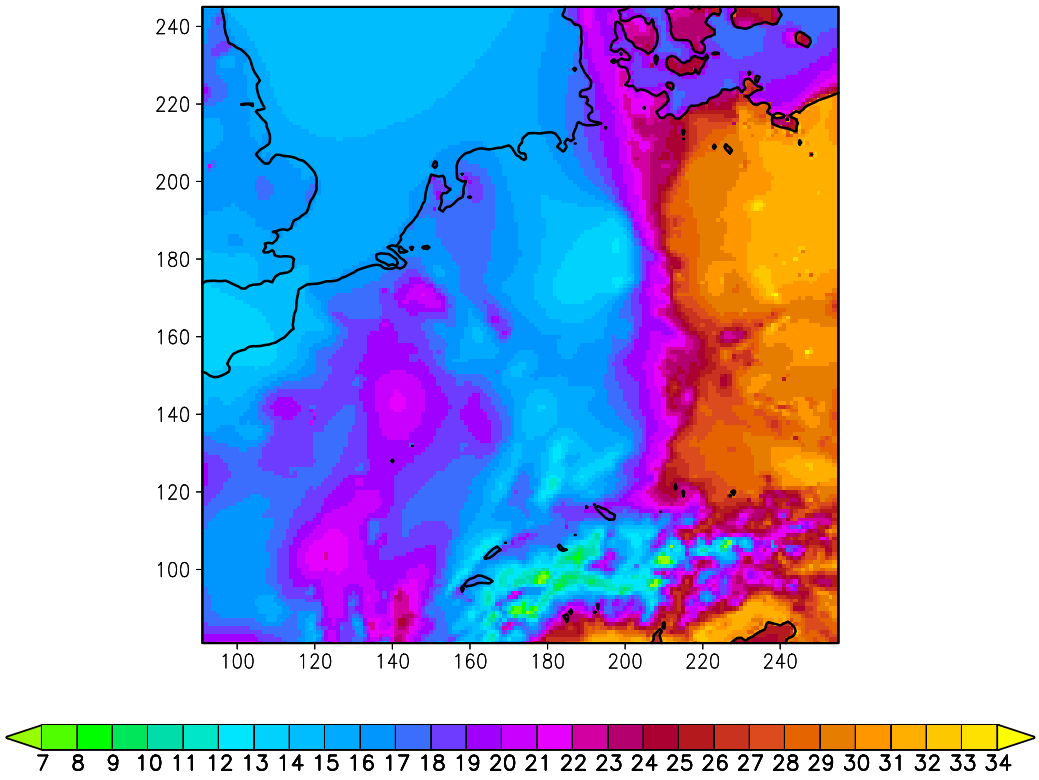

Figure 5.18: *Post-processed version of the 2m-temperature forecast in Fig. 5.2. Unit: Degree Celsius. Reconstruction after thresholding with the additional field criterion. The level-dependent thresholds are selected according to Tab. 5.1 and location dependency is introduced via the additional field (Eq. 5.4). Shift averaging is also applied (cf. Fig. 5.10). The full domain of the post-processed forecast comprises* 226×242 *grid points; here a section of* 165×165 *grid points is shown.*

result is transformed back to grid point space and then the smoothed temperature fields are averaged.

Fig. 5.18 shows the resulting post-processed forecast. Differences between the post-processed forecasts in Fig. 5.15 and 5.18 are exclusively due to the modification of the threshold value according to the precipitation field. As expected, the additional field criterion affects regions with predicted occurrence of precipitation and increases the spatial smoothness of the forecast. The temperature gradients within the front and within the Alpine region are decreased. In the frontal region this effect is desirable, as frontal temperature gradients are considered to be noisy indeed. In the Alpine region the benefit of the effect is questionable, as temperature gradients that are associated with complex orography may well be part of the signal, especially if the small-scale precipitation forecast is induced by orography as well.

Consequently, an even more sophisticated version of wavelet thresholding may benefit

from introducing orographic height as another additional field. The inclusion of orography might not only be useful in terms of 2m-temperature, but also in terms of separating convective precipitation from orographically induced precipitation.

Note, however, that the additional field criterion increases the computational effort of wavelet thresholding, because it requires the wavelet transform of the additional field. If the additional field is constant with time, the increase is only marginal, because the wavelet transform is derived only once for all times. If the additional field is time-dependent, however, the associated gain in forecast quality should always be rated in connection with the increase in computational cost.

Up to this point, it can be concluded that the additional field criterion apparently has the potential to improve signal detection, noise detection and consistency between several forecast variables in LM post-processing. Further investigations are recommended, but remain for future research.

5.6 Summary and outlook

5.6.1 Summary

New paths trodden

In addition to the neighbourhood, kriging and kernel smoothing, wavelet smoothing is another attempt to derive non-parametric estimates of a large array of parameters on the basis of a single realisation on each parameter. In this study, the wavelet method is used to estimate the expected value, but may be easily extended to an additional estimation of the exceedance probability.

The statistical concepts of wavelet smoothing and kernel smoothing are the same, but the wavelet method extends the scope of the kernel method considerably, because the wavelet-based filter is able to *adapt* to the noisy data in an automatic way. This study is the first to apply a wavelet-based denoising techniques to "real" forecast data and to derive a novel forecast product by doing so.

The wavelet transform algorithm is chosen to be the lifting scheme, a highly efficient, bi-orthogonal, second generation wavelet transform. Wavelet curtailing is governed by certain criteria chosen *a priori*. Spatial scale, location and the absolute value of the coefficient are explored as curtailing criteria. The results of wavelet smoothing are discussed on the basis of a case study and its visual inspection.

Lessons learned

Wavelet smoothing proves to be much more efficient than kernel smoothing. This achievement is very desirable, because computational efficiency is generally one of the decisive advantages of LM post-processing compared to ensemble techniques.

The use of a translation-invariant wavelet transform is indispensable in order to prevent disturbing artefacts in the post-processed forecast. In this work, a brute force approach to translation-invariant wavelet smoothing is explored: averaging shifts. Averaging shifts has the side effect of an increased computational expense. This is accepted, because the averaging shifts approach successfully removes the artefacts.

Among the criteria used in wavelet curtailing, wavelet thresholding is most interesting. In terms of 2m-temperature post-processing the method scores a resounding success with regard to the effortless preservation of small-scale structures associated with explanatory variables. In terms of precipitation post-processing, a clear superiority to the results of kernel smoothing has not become apparent. This demonstrates that the result of wavelet thresholding largely depends on the underlying data. Certain statistical properties in the noisy data are indispensable for the success of wavelet thresholding.

Note that all conclusions drawn so far are based on the visual inspection of one case study. Such an investigation serves as a first indicator of the potentials of wavelet smoothing. Further affirmation follows in Chapter 6 which contains the objective evaluation of a larger set of case studies.

5.6.2 Outlook

Suggestions for possible refinements

As the computational efficiency of wavelet smoothing is still hampered by shift averaging, additional attempts should be made to decrease the computational cost of the translation-invariant wavelet transform. For example, it could be investigated whether the application of only few shifts might yield a result that is already very close to the optimum.

More thought should be given to threshold selection in wavelet thresholding. Visual inspection is no longer an option when a larger set of forecasts is post-processed. Instead, one of the existing data driven threshold selection methodologies should be adapted to the problem of LM post-processing so that the method is able to cope with the special properties of LM forecast data. In particular, the method may have to account for location-related differences in noise amplitude.

Furthermore, the additional field criterion could be refined. Especially the consideration of orography appears promising in terms of improved discrimination between signal and noise.

The two-dimensional wavelet transform could be extended to three dimensions so as to enable smoothing in time, too. In this case the option of asymmetric wavelet smoothing would gain in importance so that the direction criterion would have to be applied additionally. Note, however, that the inclusion of the time direction would also result in an increased computational effort and would require a smaller increment Δt between consecutive output times.

Possible applications in related contexts

Wavelet smoothing may not only be used for the non-parametric estimation of the expected value, but also for the estimation of the exceedance probability. As explained in Section 4.3 the exceedance probability can be expressed by the intensity function of a point process. Non-parametric estimation of the intensity function is closely related to the problem of non-parametric density estimation. The use of wavelet transforms in non-parametric density estimation has been well tried and tested (Ogden, 1997; Härdle et al., 1998; Vidakovic et al., 2000). The procedure is similar to the use of wavelet transforms in non-parametric regression which is already demonstrated in this study. Thus, the application of wavelet transforms in the estimation of exceedance probabilities only requires another tiny step.

Furthermore, wavelet transforms can be very useful in connection with ensembles, both as an analysing tool of experimental ensembles and as a post-processing tool of operational ensembles. Wavelet analysis of LM ensemble experiments may reveal general information on signal and noise characteristics that are present in an operational LM forecast. This information can be very valuable in the development stage of LM post-processing methodologies such as those investigated in this study.

Interpretation of an operational ensemble forecast may also benefit from wavelet post-processing. For example, the ensemble mean can be derived by a combination of wavelet thresholding and ensemble averaging. This combination is explicitly discussed in the context of image processing, namely with the aim to denoise multiple noisy image copies (Chang et al., 2000). Further investigations in this area are strongly suggested, especially since the idea of combining statistical and dynamical ensembles has lately become an issue in meteorological research (Wilks, 2002; Roulston and Smith, 2003).

6 Objective verification

Chapters 3–5 present the motivations, methodologies and results of the LM ensemble, neighbourhood post-processing and wavelet smoothing. The respective results are assessed by subjective eye-ball inspection of a case study.

This chapter supplements these investigations by a statistical evaluation of a larger data base of forecast data. The forecasts are compared to corresponding observational data. Forecast assessment is condensed into objective verification measures of forecast goodness which allow for a direct comparison between results of different forecast systems.

6.1 Introductory remarks

6.1.1 Aims

The aim of objective verification in this chapter is to

- assess forecast goodness of the various forecast systems in this study,
- decide on the optimal configuration of the neighbourhood method,
- assess the user's gain when applying a forecast system to her individual needs,
- identify the forecast system with the optimal ratio of forecast goodness and computational cost.

6.1.2 Organisation of the chapter

The chapter is organised as follows. The verification strategy is outlined in Section 6.2. Sections 6.3–6.5 deal with the evaluation of neighbourhood post-processing. Section 6.3 contains a comprehensive forecast quality assessment of the standard neighbourhood post-processing procedure. Estimates of expected value, quantiles and exceedance probability are examined. Other important aspects of forecast goodness, such as the user's gain in applying the forecast, are investigated in Section 6.4. Section 6.5 examines various configurations of the neighbourhood method and looks for an optimal neighbourhood size and an optimal consideration of explanatory variables. Section 6.6 assesses forecast goodness of wavelet smoothing. Section 6.7 assesses the experimental LM ensemble of

Chapter 3 and compares it to the post-processing systems developed in this study. In order to avoid repetitions, intermediate recapitulations of the respective sections are combined and postponed to Section 6.8 which contains a summary of the entire chapter.

6.2 Verification strategy

6.2.1 Verification data

Observational data

Observed precipitation data from rain gauges are collected for the time period July 10–24 2002. The rain gauges are located within the area of Germany and its close surroundings (47.0°N–55.5°N; 5.5°E–15.5°E). Precipitation measurements are obtained from approximately 200–300 stations per observation time so that the ratio of the number of reporting stations to the number of model grid boxes is approximately 1/50. The observational data set comprises measurements of 1h-accumulations, 6h-accumulations, 12h-accumulations or 24h-accumulations of precipitation. Some additional values of 6h-accumulations are derived by subtracting measurements of 6h-accumulations from measurements of 12h-accumulations.

Temperature data in two meters height are collected for the same time period. The data comprises hourly measurements as well as the daily minimum and daily maximum. The measurements also originate from SYNOP stations and are obtained from approximately 310–330 stations per observation time. This study focusses on the hourly measures of 2m-temperature.

At DWD an additional verification study exists in parallel which uses a more extensive data set of observations of the warm season (Tab. 6.1). The DWD study and this study have been carried out independently. Verification at DWD exclusively looks at the neighbourhood method and assesses the quality of the expected value and quantiles. Sensitivity studies concerning neighbourhood size and explanatory variables are carried out as well. The results of DWD agree with the findings of this study whenever a direct comparison is possible. Thus, the findings of this study are indirectly supported by a larger data base.

Forecast data

The verification procedure is applied to LM short-range simulations (version 2.19) that are started at 00 UTC every day and run for 48 hours. The model configuration is identical to the operational setting except for the lateral boundary conditions which are provided by LM analyses instead from predictions by the global model GME.

Table 6.1: *Verification periods considered in this study and by a parallel study at DWD.*

Verification Period	Predictand	Area	compiled by
Jul 10–24, 2002	precipitation (1h, 6h, 12h, 24h) 2m-temperature max and min temp.	Germany and close surroundings	Susanne Theis
Sep 01–15, 2001	precipitation (1h, 6h, 12h, 24h) 2m-temperature	Germany	V. Renner (DWD) U. Damrath (DWD)
May 01–15, 2001	precipitation (1h) 2m-temperature	Germany	V. Renner (DWD) U. Damrath (DWD)
Aug 01–15, 2001	precipitation (1h) 2m-temperature	Germany	V. Renner (DWD) U. Damrath (DWD)
Aug 01–15, 2002	max and min temp.	Germany	V. Renner (DWD) U. Damrath (DWD)

The choice of LM analyses was originally motivated by the desire to reduce the large-scale error induced by imperfect lateral boundary conditions. The existence of large-scale errors might blur the benefits of the experimental ensemble forecast and the post-processed forecast, because both methods are especially designed to address small-scale uncertainties of the forecast. If the observations and the forecast field already differ on the large scale, small-scale features might become irrelevant in forecast verification so that achievements of the experimental ensemble or the post-processing procedure may become invisible. For the sake of comparison, forecast verification is repeated for simulations driven by GME boundary conditions (not shown) and it turns out that the choice of LM analyses has only a minor impact on verification results. This implies that the findings of this chapter also apply to the operational setup with GME boundary conditions.

The short-range simulations comprise 48 hours each, but only forecast hours 7–30 are used in the verification procedure of this chapter. This time range is considered to yield optimal forecasts, because the spin-up time is excluded and forecast lead time is still very short. Furthermore, the forecast data sets do not overlap, if they originate from simulations that are started on consecutive days. For the sake of comparison, forecast verification is also carried out for the full time range of the 48h-forecasts (not shown). The results are very similar to those presented in this chapter.

Objective verification of this study investigates the raw LM output, the LM ensemble forecast and the LM post-processed forecast (Tab. 6.2). In the experimental ensemble, the estimates of expected value and exceedance probability are derived by assigning equal weights to each of the 10 ensemble members.

Various configurations of the neighbourhood procedure and wavelet smoothing are

Table 6.2: *Matrix of forecast methods and their respective forecast products in this study. The entry* RR *denotes that the corresponding forecast system derives a respective estimate for the precipitation forecast; the entry* T *denotes that the corresponding forecast system derives a respective estimate for the forecast of 2m-temperature. Neighbourhood PP denotes* neighbourhood post-processing.

Forecast Procedure	Single Realisation	Expect. Value	Quantile	Exceed. Prob.
single LM forecast	RR, T			
LM ensemble forecast		RR, T		RR
neighbourhood PP		RR, T	RR, T	RR
wavelet smoothing		RR, T		
observation	RR, T			

tested. A direct comparison between the whole set of forecast methodologies is not always possible, because the different methodologies estimate different sets of statistical quantities. For example, quantile estimates are only derived by the neighbourhood methodology (Tab. 6.2).

Matching pairs of forecast and observation

Any forecast verification method involves comparisons between matched pairs of forecasts and the observations to which they pertain. In this study, for each observation the nearest grid point of the model output is selected for comparison. In case of temperature data the raw model output and estimates of expected values are adjusted to the orographic height of the observation site using a standard temperature gradient of -0.0065 K/m.

Note that this approach is different from the linear modification of the model output in the neighbourhood procedure (cf. Section 4.4.2). In the neighbourhood procedure the temperature gradient is derived by linear regression of the forecast data itself (cf. Fig. 4.12) and varies with location and time.

6.2.2 Forecast goodness and corresponding measures

Forecast goodness is not only determined by forecast quality, but also by consistency and value (Murphy, 1993). Forecast goodness is therefore defined by the following three determinants:

Forecast quality Forecast quality relates to the degree of correspondence between forecasts and observations.

Table 6.3: *Matrix of forecast products and their verification measures in this study. RR denotes precipitation, T denotes 2m-temperature, BS denotes* Brier score *and its decomposition, ROC denotes* relative operating characteristics*, rmse denotes* root mean squared error*. For definitions of the measures see Appendix C.*

Predictand	Single Realisation and Expected Value	Quantiles	Exceed. Probability
RR	2×2 contingency table BS, economic value	reliability economic value	BS and its decomposition ROC curve reliability diagram economic value
T	*rmse*, correlation, etc.	reliability	

Forecast consistency Forecast consistency relates to the correspondence between the forecast and the forecaster's best judgment.

Forecast value Forecast value relates to the expenses incurred by a user who employs the precipitation forecast as an input into her decision making process. Measures of forecast value particularly stress the user's perspective.

Each of these forecast characteristics therefore needs to be considered when evaluating forecast goodness.

There is a multitude of different verification measures and each measure reflects a different aspect of forecast goodness. Forecast quality is expressed by the joint distribution of forecasts and observations. A single forecast measure reflects only a certain aspect of the joint distribution. Therefore, a variety of measures is consulted in order to give a general judgment of forecast quality. Consistency is difficult to measure, so this study presents a qualitative discussion of consistency instead of a quantitative assessment. Forecast value is investigated in the context of a prototypical decision making situation, the cost-loss scenario (Appendix C.4). The corresponding verification measure of *relative value* (Appendix C.4) quantitatively assesses the economic value of the forecast and relates it to the economic value of a perfect forecast and of a benchmark forecast.

Furthermore, probability forecasts and categorical forecasts require different measures of forecast quality (cf. Appendix C). As visible from Tab. 6.2, the forecast data of this study are a mixture of probability forecasts and categorical forecasts. Quantiles and exceedance probabilities refer to probabilities while the expected value and the raw model output consist of a flat statement that exactly one of a set of possible events will occur. Their corresponding measures of forecast quality are listed in Tab. 6.3.

Even though the expected value is a categorical forecast in terms of forecast verification, it is essentially of probabilistic nature when compared to the direct model output. The definition of the expected value still refers to the probability density function of the

forecast. The raw model output, on the contrary, is a single realisation and therefore not probabilistic.

6.2.3 Choosing a benchmark

In order to evaluate the experimental ensemble forecast and the post-processed forecast it is useful to relate its measures of forecast quality to those of a reference forecast. In this study the following references are applied:

- The raw LM output,
- a "climatological" forecast,
- various versions of the post-processing procedures,
- simple kernel smoothing with the rectangular kernel (Eq. 5.2),
- the *inflated* LM output which is introduced in Section 6.4.2,
- the perfect forecast.

The most prominent reference forecast is the raw LM output of a single simulation. It serves as a low benchmark, because the LM ensemble and the post-processing methods aim at outperforming this reference.

Another low benchmark is the climatological forecast, i. e. a simple forecast that consists in the average of a historical data set. In this study, historical data sets are not available so that a substitute is derived from the observational data of the verification period. As the period only comprises 15 days, the substitute is far more specific than a typical climatology. Nevertheless, the term "climatological" refers to the 15-day-mean in this study, because there is currently no alternative. As a consequence, this study clearly overestimates the forecast quality of the climatological forecast and corresponding conclusions must be handled with caution.

Other low benchmarks are derived through simplified variants of the post-processing methods, simple kernel smoothing and inflation of the forecast (cf. Section 6.4.2).

High benchmarks include the perfect forecast and refined versions of the post-processing procedures. The perfect forecast enables the interpretation of the degree of forecast improvement between a low benchmark and the forecast system under consideration.

6.3 Quality of the standard neighbourhood results

This section contains a comprehensive forecast quality assessment of the standard neighbourhood post-processing procedure. The standard version uses medium neighbourhood size (cf. Tab. 4.1) and considers explanatory variables in the way as listed in Tab. 4.2.

The central question of this section is whether neighbourhood post-processing improves the original LM forecast in terms of forecast *quality*, i. e. the correspondence between forecasts and observations. Estimates of expected value (Sections 6.3.1 and 6.3.2), quantiles (Section 6.3.3) and exceedance probability (Section 6.3.4) are examined. Other aspects of forecast goodness, such as the user's gain in applying the forecast, are investigated in Section 6.4.

6.3.1 Expected value of precipitation

Overview

The expected value of precipitation is a categorical forecast of a continuous predictand (cf. Appendix C). An exceedance threshold is chosen which converts the continuous precipitation forecast into a dichotomous forecast. Then several verification measures are derived from the respective contingency table (Tab. C.1): the probability of detection (POD), the false alarm rate (FAR), the frequency bias (FBI), the equitable threat score (ETS) and the logarithm of the odds ratio (LOR). For definitions of these measures see Appendix C.2.

Several thresholds are chosen and the procedure is repeated for each of them. This approach enables a separate forecast assessment of light precipitation events and heavy precipitation events. A measures-oriented investigation of extreme precipitation is omitted here, because the observational data set contains only very few extreme events and the expected value does not capture them. It can already be concluded that the expected value does not properly represent extreme events.

Fig. 6.1 and 6.2 show the resulting verification measures as a function of the exceedance thresholds. Fig. 6.1 refers to 1h-accumulations and Fig. 6.2 refers to 24h-accumulations. The respective results of 6h-accumulations and 12h-accumulations are not shown here, because they lead to the same conclusions. Results are shown for the raw model output and neighbourhood estimates of the expected value.

Frequency bias

The FBI is the ratio of two unconditional frequencies: the frequency of forecast events and the frequency of observed events. Ideally, a forecast should exhibit a frequency bias of 1.

Fig. 6.1 and 6.2 show that the raw model output suffers from overforecasting light precipitation and from underforecasting heavy precipitation. Neighbourhood post-processing intensifies this unwanted effect for almost all precipitation thresholds and accumulations because the post-processing procedure effectively smooths the model output.

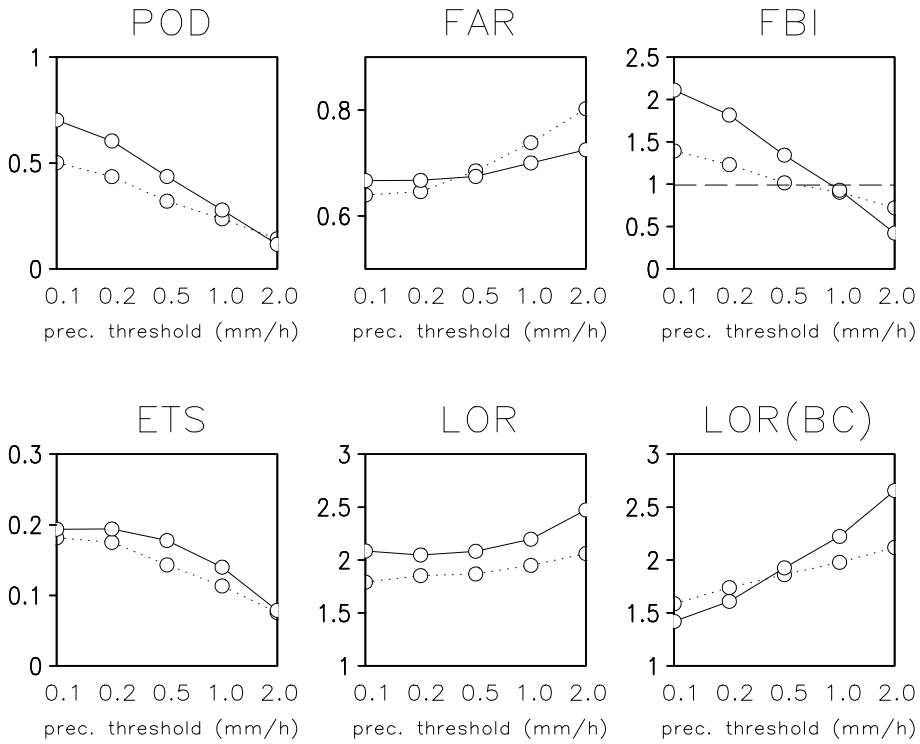

Figure 6.1: *Measures of forecast quality from the contingency table. POD: Probability of detection. FAR: False alarm rate. FBI: Frequency bias. The long dashed line denotes a perfect FBI. ETS: Equitable threat score. LOR: Log odds ratio. LOR(BC): Log odds ratio with bias correction. Predictand: 1h-accumulations of precipitation, original forecast (dotted line) and post-processed forecast (solid line). The standard version of the neighbourhood procedure is applied (cf. Tab. 4.2). The abscissa lists various precipitation thresholds.*

Mullen and Buizza (2000) note that this is also typical of most ensemble prediction systems (e. g. Eckel and Walters, 1998; Hamill and Colucci, 1998). Thus, the deterioration in FBI also occurs in more sophisticated methods of expected value estimation.

Probability of detection

Except for heavy precipitation events the POD of the expected value is larger than the POD of the direct model output. Note that the increase in POD is at least partly a consequence of increased overforecasting. The POD benefits from overforecasting, because it is a measure of correct hits and not a measure of false alarms, misses or correct rejections.

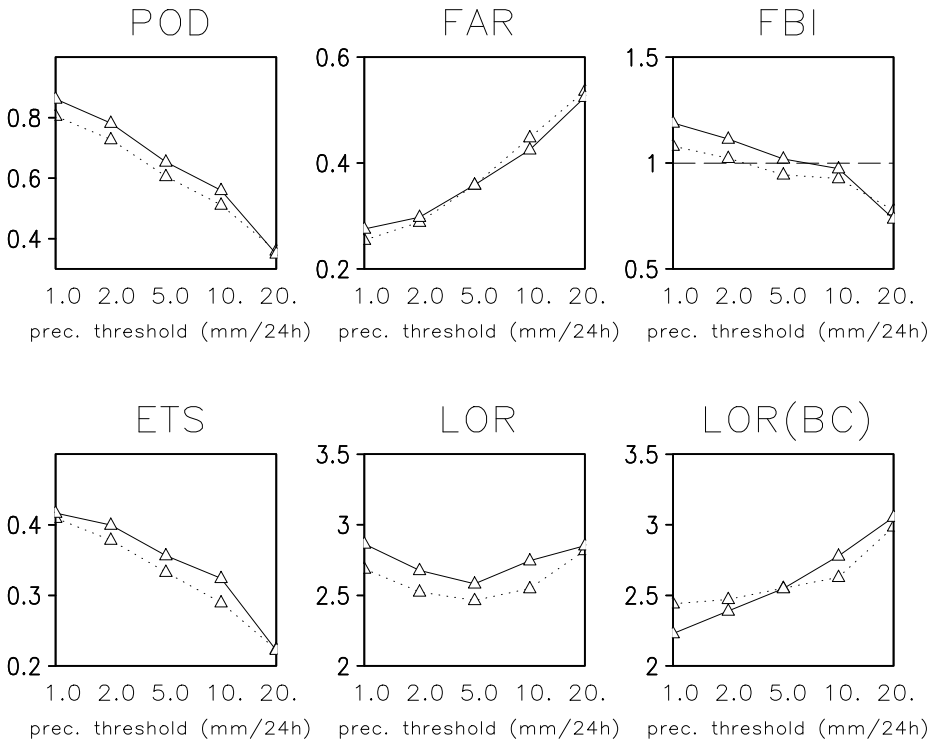

Figure 6.2: *Same as Fig. 6.1, but for 24h-accumulations of precipitation.*

Equitable threat score

Similarly to the POD, the ETS refers to correct hits, but additionally accounts for forecasts that verify by chance. Thus, it implicitly compensates for a simultaneous change in FBI. Hit rates are penalised by a high FBI and vice versa. This becomes visible in Fig. 6.1 and 6.2: the ETS exhibits slightly different results than the POD whenever the FBI deviates very much from 1. All in all, the estimated expected value exhibits a better ETS than the raw model output.

A striking feature is the extraordinarily high ETS of 24h-accumulations. Such high values are unprecedented in the results of other verification periods investigated by DWD (cf. Tab. 6.1). Similar results are attained when the lateral boundary conditions are provided by GME forecasts instead of LM analyses (not shown). Therefore, these high scores do not originate from the application of LM analyses.

False alarm rate

The FAR complements the POD and the ETS as it refers to false alarms instead of hits. Neighbourhood post-processing increases the FAR of light precipitation and decreases the FAR of heavy precipitation.

As long as these changes in FAR are caused by changes in the FBI, they do not truly indicate a change in forecast quality yet. Only a simultaneous decrease of FAR and increase of POD would indicate a true enhancement of forecast quality.

A simultaneous decrease of FAR and increase of POD is visible indeed for the 10 mm-thresholds of 24h-accumulations (Fig. 6.2), for the 10 mm-threshold of 12h-accumulations (not shown), for the 5 mm-threshold of 6h-accumulations (not shown) and for the 0.5 mm- and 1.0 mm-threshold of 1h-accumulations (Fig. 6.1). In other words, the estimated expected value improves forecast quality of the raw model output in the range of moderate to large precipitation amounts.

Log odds ratio

The LOR relates the odds of a hit and the odds of a false alarm. It deserves special attention for several reasons. It combines measures of hits and false alarms, it implicitly accounts for the unconditional frequency of an event, and it allows for a simple test of statistical significance.

Fig. 6.1 and 6.2 show that the LOR of the estimated expected value outperforms the LOR of the raw model output. However, the increase in LOR may not indicate a true enhancement of forecast quality, because the LOR benefits from a large FBI. In order to separate the effect of the FBI, a bias-corrected LOR (LOR(BC)) is calculated in addition (cf. Appendix C.2).

Fig. 6.1 and 6.2 show that the LOR(BC) of the estimated expected value is larger than the LOR(BC) of the raw model output for exactly those precipitation amounts that attain an improvement in both FAR and POD. Thus, the LOR(BC) nicely summarises and reconfirms the findings above.

The differences between the log odds ratios (LOR) of the post-processed output and the raw model output are tested for statistical significance. In case of 1h-accumulations, the differences are statistically significant to the 95% confidence level. Other accumulations do not attain statistical significance. This is probably due to the fact that the respective observational data sets are substantially smaller than the set of 1h-accumulations. However, the 95% confidence level of 1h-accumulations may be overestimated, because precipitation data from different locations and timings may not satisfy the requirement of independence.

Synthesis of the results

Compared to the raw model output, the expected values achieve more hits and more false alarms for light precipitation and achieve less hits and less false alarms for heavy precipitation. Such a forecast can be beneficial to some users and disadvantageous to

Table 6.4: *Forecast quality measures for the estimate of the expected value. Predictand: 2m-temperature. Comparison between the post-processed forecast and the direct model output of the LM. The standard neighbourhood procedure is applied (cf. Tab. 4.2).*

Measure	Single LM Forecast	Neighbourhood PP
rmse	1.92 K	1.81 K
mae	1.40 K	1.33 K
bias	-0.01 K	-0.05 K
stdv_fc	4.04 K	3.86 K
corr	0.89	0.90
skill_DMO	0.000	0.054

others, depending on their individual ability of tolerating high false alarm rates or low hit rates.

For most precipitation thresholds it cannot be concluded that the expected value is a better precipitation forecast compared to the raw model forecast. Only for a small range of moderate to heavy precipitation a general improvement can be noted indeed.

Note that these results need to be affirmed by a larger data base, because some of them do not attain statistical significance. As the parallel study of DWD (Tab. 6.1) yields similar results, it is likely that statistical significance can be reached if the data base of the verification procedure is extended.

6.3.2 Expected value of 2m-temperature

The expected value of 2m-temperature is a categorical forecast of a continuous predictand and is evaluated by calculating the following verification measures: the root mean squared error (*rmse*), the mean absolute error (*mae*), the bias (*bias*), the standard deviation of the model data itself (*stdv_fc*), the correlation (*corr*) between forecasts and observations and the *skill_DMO* which uses the direct model output (DMO) of the single unperturbed LM simulation as a reference forecast. For definitions and explanations of these measures see Appendix C.2.

The verification measures are derived for the direct model output and the post-processed forecast. They are listed in Tab. 6.4. The post-processed forecast exhibits a reduction in *rmse*, a reduction in forecast variability *stdv_fc*, an increase in forecast *bias* and a slight increase in correlation *corr*. The DWD verification studies (Tab. 6.1) produce quite similar results in terms of the post-processing effect. However, the *rmse* of the raw model output substantially varies with the verification period. Its value ranges from 1.37 K to 2.25 K.

Neighbourhood post-processing improves the quality of 2m-temperature forecasts

slightly. The reduction in root mean squared error *rmse* is exclusively attained through a reduction in random error. A reduction in random error is especially valuable, because random error cannot be tackled by re-calibrating the forecast. Moreover, Tab. 6.4 shows that the reduction in forecast variability is attained without any loss of correlation between forecast and observation so that it is not only achieved through a reduction in forecast variance $(stdv\_fc)^2$ (see Taylor, 2001). The improvement in correlation is very small though. The physical relevance and the statistical significance of this increase should be tested first before drawing any further conclusions.

Note that there is a reduction in forecast variability (*stdv_fc*). At first glance, this seems unfavourable, since the variability of the direct model output is already lower than the variability of the observations ($stdv\_obs = 4.11\ K$) and a further decrease of model variability might seem equivalent to a less realistic forecast. However, reducing forecast variability may be a step into the right direction since it acknowledges limits of deterministic predictability. The raw model output may contain small-scale details that are not deterministically predictable. Spatial smoothing may account for the presence of noise and may yield a more reliable forecast.

Although the theory of neighbourhood post-processing assumes an unbiased estimate (cf. Subsection 4.3.3), the neighbourhood estimate of the expected value has a larger bias than the raw model output. This contradiction is solved by the insight that the simple version of neighbourhood post-processing is equivalent to kernel smoothing (Subsection 4.3.2). The statistical model of kernel smoothing clearly shows that a reduction in random error must be paid by an increase in bias (cf. Subsection 4.3.3).

6.3.3 Quantiles of precipitation and 2m-temperature

In contrast to verification measures of categorical forecasts and forecasts of exceedance probability, the literature offers only few verification measures that refer to quantiles (cf. Appendix C.3.2). Due to this lack of standard measures, verification of quantiles is restricted to the assessment of forecast reliability here (cf. Appendix C.3.1). An assessment of forecast resolution (cf. Appendix C.3.1) is omitted, even though resolution is also a very essential property of any probability forecast, because it cannot be improved by forecast calibration.

The reliability diagram (Fig. 6.3) shows the observed frequency of the event [obs $<$ quantile]. So as to account for the mixed discrete-continuous probability distribution of precipitation, only quantiles greater than 0.01 mm enter the reliability diagrams of precipitation (see Krzysztofowicz and Sigrest, 1999). Reliability is not only juxtaposed with the original forecast (closed circle), but also with the perfect forecast (diagonal).

This study and the parallel study at DWD reveal that

- the quantile forecast is superior to the raw model output,

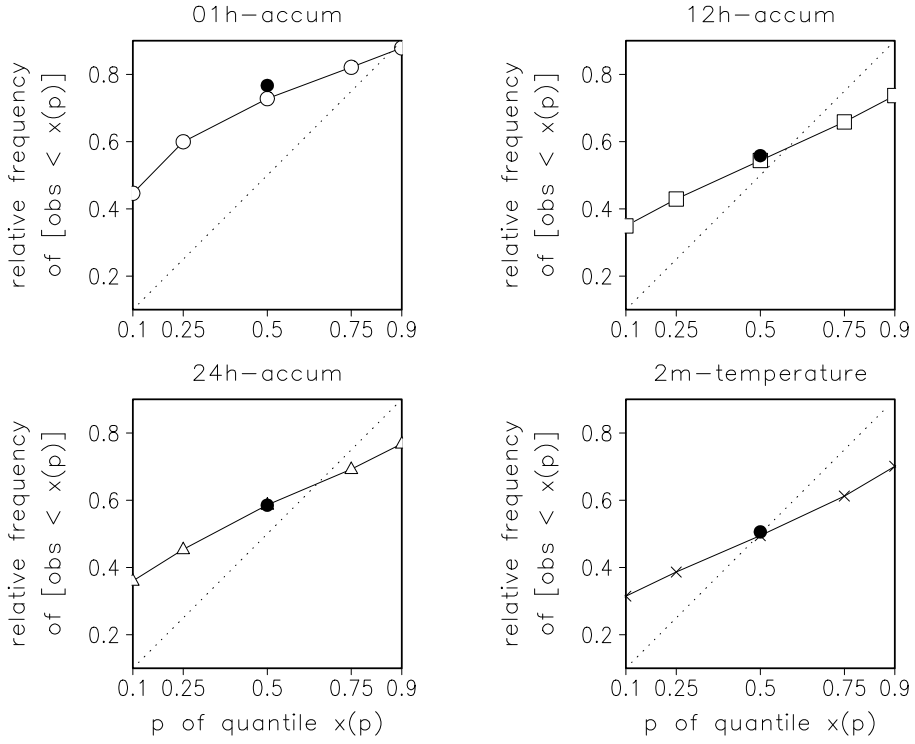

Figure 6.3: *Reliability diagrams for quantiles of precipitation and 2m-temperature. The 10%-, 25%-, 50%-, 75%- and 90%-quantiles are evaluated. The standard version of the neighbourhood procedure is applied in the derivation of quantiles (cf. Tab. 4.2). Precipitation quantiles enter the reliability calculation only if they are greater than 0.01 mm. The dotted line denotes perfect reliability. The closed circle shows the unconditional frequency of [obs < DMO].*

- the quantile forecast is not perfectly reliable,
- the quantile forecast is generally underdispersive,
- the quantile forecast does not always capture variations in forecast uncertainty,
- there is overforecasting in 1h-accumulations.

Among these results, underdispersiveness is probably most striking in the reliability diagram. Compared to a perfectly reliable forecast the reliability curve of the quantiles is tilted clockwise, i. e. the 90%-quantile is usually too low and the 10%-quantile is too high. This means that forecast uncertainty is generally underestimated. Underdispersiveness may result from the fact that neighbourhood post-processing exclusively focusses on small-scale forecast uncertainty whereas the comparison to observational data automatically accounts for the overall forecast error.

Simple inflation of the probability forecast is not able to cure the problem entirely, because it turns out that the reliability of a certain quantile estimate also depends on its

value. For example, observations exceed a fairly high estimate of the 90%-quantile less often than a low estimate (not shown). This indicates that the neighbourhood procedure does not always capture variations in forecast uncertainty.

The 1h-accumulations show slightly different results than other accumulations and 2m-temperature; the observed frequency is additionally shifted towards higher values. This reveals that overforecasting is an ubiquitous feature of the post-processed forecast. This feature is not introduced by neighbourhood post-processing, but originates from the frequency bias of the raw model output. The frequency bias of the raw model output becomes obvious by looking at the unconditional frequency of [obs $<$ DMO] (closed circle, Fig. 6.3). It also becomes visible in the FBI of Fig. 6.1 and is most eminent for a precipitation threshold of 0.01 mm (not shown in Fig. 6.1).

Although the reliability of the post-processed quantile forecast is far from perfect, it is still superior to the raw model output. As the raw model output does not deliver any estimate of quantiles, the reliability curve of the direct model output can be imagined as a horizontal line intersecting the closed circle in Fig. 6.3. Thus, the reliability of neighbourhood temperature quantiles is approximately half-way between the reliability of the raw model output and of a perfect forecast. This is also true for quantiles of 12h-accumulations of precipitation, but is not achieved by all quantiles of 1h-accumulations and 24h-accumulations because of their frequency bias.

Before concluding this subsection it should be noted that operational ensemble prediction systems do not attain perfect reliability either (Hamill and Colucci, 1998; Eckel and Walters, 1998). The ensemble forecast is often recalibrated using historical model data sets.

Richardson (2001b) shows that limited sample size is one explanation of imperfect reliability. Even a perfect ensemble with finite ensemble size would exhibit shortcomings in reliability, depending on the frequency and predictability of the event under consideration. The effect of limited sample size may also play a role in neighbourhood post-processing, because the neighbourhood only derives a finite sample. This study investigates the quantitative influence of limited sample size on the quantile estimate (not shown). The investigation offers an explanation why the observations exceed a fairly high value of the estimated 90%-quantile more often than a fairly low value. However, it turns out that limited sample size explains the lack of reliability only to a very small extent. Therefore, limited sample size appears to be an irrelevant cause.

6.3.4 Exceedance probability of precipitation

Forecast skill compared to the climatological forecast

Accuracy and skill encompass a relevant portion of forecast quality aspects. The accuracy of probability forecasts is measured by the Brier score (cf. Appendix C.3). The Brier score is calculated for several exceedance probabilities of precipitation, each using

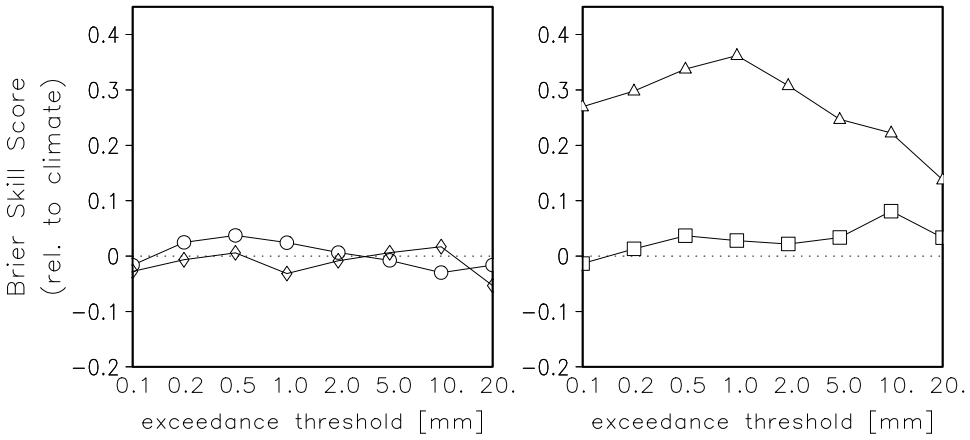

Figure 6.4: *Brier skill score of exceedance probabilities of precipitation. The skill score relates to a climatological forecast as a reference. The climatological reference forecast is based on the observational data of the verification period only. The standard version of the neighbourhood procedure is applied in the derivation of exceedance probabilities (cf. Tab. 4.2). The abscissa lists various exceedance thresholds. Several accumulations are shown: 1h-accumulations (circle), 6h-accumulations (diamond), 12h-accumulations (square), 24h-accumulations (triangle). The dotted line denotes zero skill.*

a different exceedance threshold. Low exceedance thresholds represent the ability to predict light precipitation and high exceedance thresholds represent the ability to predict heavy precipitation. Skill of the exceedance probability is measured by the Brier skill score which relates the Brier score of the forecast to the Brier score of a reference forecast (cf. Appendix C.3).

Fig. 6.4 depicts the Brier skill score with the climatological forecast as a reference forecast. The Brier skill score is shown as a function of the exceedance thresholds. When the climatological forecast is used as a reference, the estimated exceedance probabilities hardly attain any positive skill. Only exceedance probabilities of 24h-accumulations clearly outperform the climatology in terms of Brier score.

Using the climatological forecast as a reference yields misleading results. Recall that this study derives the climatological forecast from the observational data of the verification period itself which is far too short for an appropriate representation of climate. The quality of the climatological forecast is overestimated and it does not really play the role of a low benchmark. Therefore it makes the quality of the post-processed forecast appear in a bad light which is not justified.

Especially the reliability of the climatological forecast is overestimated. Due to this fact, the estimated exceedance probabilities are mainly punished for their imperfect reliability which does not even originate from the post-processing procedure, but is mainly caused by the raw model output. Thus, the Brier skill score of Fig. 6.4 merely visualises the

deficiencies of the direct model output as compared to the observational mean of the verification period.

The good Brier score of 24h-accumulations is probably due to good forecast quality of the direct model output itself. The exceptionally good forecast quality of 24h-accumulations has already become obvious in the verification measures of the direct model output, for example in the ETS (cf. Fig. 6.2). As other verification periods at DWD do not exhibit such good results of ETS, the good Brier skill scores of 24h-accumulations are exclusively associated with the particular verification period here.

As the Brier skill score is a very common forecast measure of probabilistic forecasts, it has been calculated for many other forecast systems such as model output statistics or ensemble prediction. Although the comparison between two completely different forecast systems must be handled with much caution, some other values of the Brier skill score are listed here. They refer to short-range forecasts of exceedance probabilities of precipitation and use a climatological forecast as a reference. Note that the scores are not fully compatible, because they result from different verification strategies, different verification areas and time periods and also from different definitions of the climatological forecast. Mentioning the scores of other forecast systems only aims at a very rough juxtaposition.

Current high-resolution short-range ensemble prediction systems do not necessarily achieve a better Brier skill score than the post-processed forecast in this study. Hamill and Colucci (1998) report Brier skill scores of less than 0.1 for their Eta-Regional Spectral Model short-range ensemble forecasts of 12h-accumulations. Bright and Mullen (2002) investigate numerous versions of MM5 short-range ensemble precipitation forecasts. Their Brier skill scores mostly do not exceed a value of 0.2 for an exceedance threshold of 0.25mm/24h and are thus even lower than the score achieved in this study (cf. Fig. 6.4).

Applequist et al. (2002) evaluate current model output statistics methodologies that forecast exceedance probabilities of 24h-accumulations in the short-range. These techniques achieve similar Brier skill scores or slightly better Brier skill scores than the neighbourhood method. This is probably due to their good calibration which originates from their use of historical error statistics. Note that historical error statistics are not used in the post-processing methods of this study, because they are currently not available. Thus, the methods are still open for an additional application of historical error statistics. Their application may lead to a substantial increase of the Brier skill score, maybe even beyond the skill of model output statistics.

Forecast skill compared to the raw model output

A more appropriate reference forecast is the raw model output, because the main aim of the post-processing procedures consists in the improvement of this forecast. Fig. 6.5

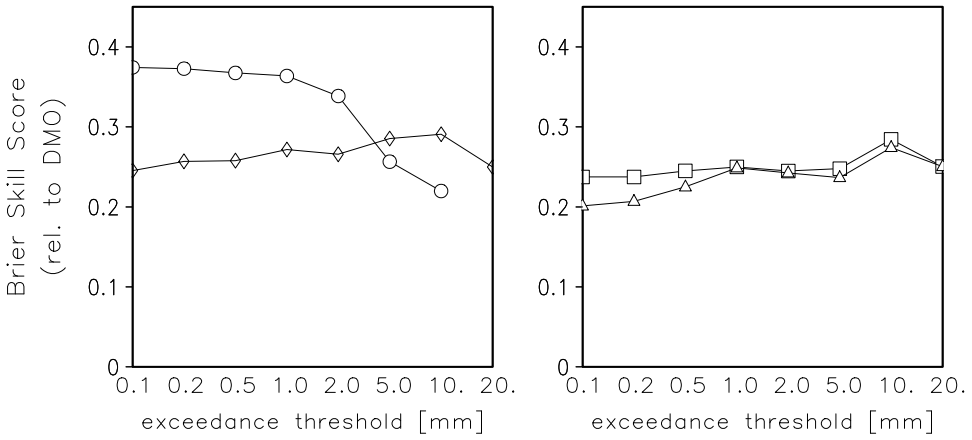

Figure 6.5: *Same as Fig. 6.4, but with the direct model output as a reference. The Brier skill score of the 20mm-threshold of 1h-accumulations is not shown, because the sample is very small.*

shows that this aim is reached indeed, as all of the estimated exceedance probabilities possess positive skill.

When using the raw model output as a reference, the estimated exceedance probability is not punished for shortcomings that are already contained in the raw data. Thus, the lack of reliability in the exceedance probabilities is rated less and resolution becomes more relevant. Resolution is an even more essential aspect of forecast quality as it cannot be improved by calibration.

Fig. 6.5 shows that shorter accumulations of precipitation mostly attain a slightly better skill than longer accumulations. This might be due to the fact that the post-processing procedure is designed to alleviate random error and runs the risk of deteriorating systematic error (cf. Subsection 4.3.3). Furthermore, shorter accumulations are associated with smaller spatio-temporal scales and are therefore expected to possess a smaller amount of deterministic predictability and a larger ratio of random error and systematic error. In this case neighbourhood post-processing is expected to attain a larger improvement of the raw model output and the Brier skill score is expected to be better indeed.

Decomposition of the Brier score: Forecast reliability, resolution and uncertainty

After having looked at the overall accuracy of the forecast, the distribution-oriented approach of forecast verification is followed up now. The Brier score is decomposed into resolution, reliability and uncertainty which all refer to the conditional distribution of the observations given the forecasts (cf. Appendix C.3.1). Whereas resolution pertains to the differences between the conditional average frequencies of the observations for different values of the forecast, reliability compares the conditional average frequencies

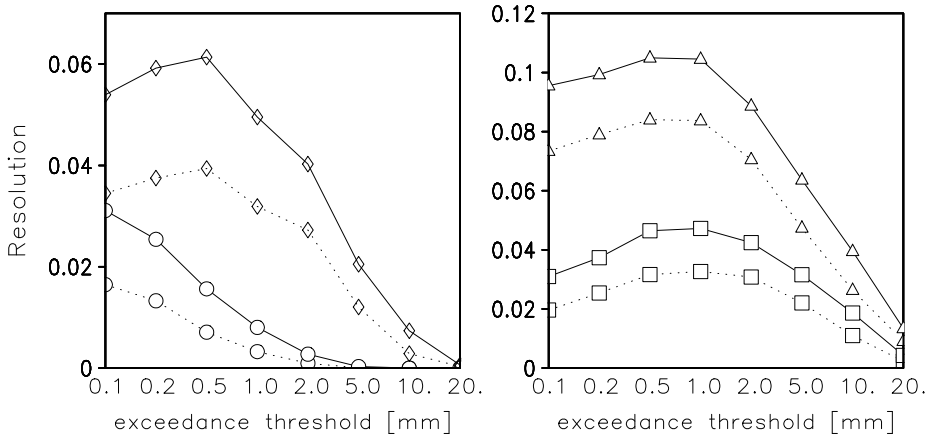

Figure 6.6: *Decomposition of the Brier score into resolution, reliability and uncertainty: resolution of precipitation exceedance probability as a function of exceedance threshold. Resolution of the deterministic direct model output (dotted lines) and of the post-processed probability forecast (solid lines). The standard version of the neighbourhood procedure is applied in the derivation of exceedance probabilities (cf. Tab. 4.2). Resolutions of several accumulations are shown: 1h-accumulations (circle), 6h-accumulations (diamond), 12h-accumulations (square), 24h-accumulations (triangle).*

of the observations with the forecast values themselves. Uncertainty depends only on the variability of the observations and is not influenced by the forecast system.

The decomposition is derived for the estimated exceedance probabilities and for the raw model output. Fig. 6.6, 6.7 and 6.8 depict the results as a function of the precipitation threshold. In the language of the Brier score, high values of resolution and low values of reliability contribute to good forecast quality. Fig. 6.6 and 6.7 show that neighbourhood post-processing improves both reliability and resolution. Thus, using a spatio-temporal neighbourhood for the generation of probabilistic forecasts appears to be a useful approach.

Forecast uncertainty (Fig. 6.8) is shown as well, although it is independent of the forecast system. The accuracy of weather forecasts is not only influenced by the forecasting system, but also by the predictability of the weather (Göber et al., 2004). In order to allow for a meaningful interpretation of resolution and reliability, their values must be related to the respective values of uncertainty. For example, the decrease in reliability and resolution towards higher exceedance thresholds (Fig. 6.6 and 6.7) is mainly due to a decrease in forecast uncertainty (Fig. 6.8) and cannot be accredited to the forecast system itself.

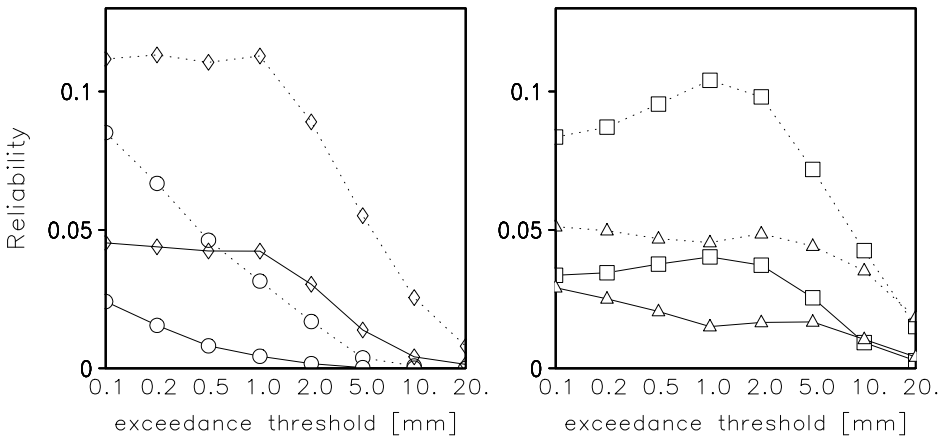

Figure 6.7: *Decomposition of the Brier score into resolution, reliability and uncertainty: reliability of precipitation exceedance probability as a function of exceedance threshold. Reliability of the deterministic direct model output (dotted lines) and of the post-processed probability forecast (solid lines). The standard version of the neighbourhood procedure is applied in the derivation of exceedance probabilities (cf. Tab. 4.2). Reliabilities of several accumulations are shown: 1h-accumulations (circle), 6h-accumulations (diamond), 12h-accumulations (square), 24h-accumulations (triangle).*

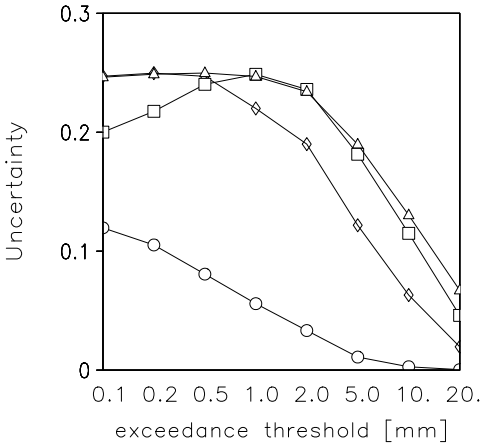

Figure 6.8: *Decomposition of the Brier score into resolution, reliability and uncertainty: uncertainty of precipitation exceedance probability as a function of exceedance threshold. The uncertainty depends on observational data only and is independent of the forecast system. Uncertainties of several accumulations are shown: 1h-accumulations (circle), 6h-accumulations (diamond), 12h-accumulations (square), 24h-accumulations (triangle).*

Which degree of forecast sharpness is beneficial?

Improvements in reliability and resolution are presumably consequences of reducing random error and of creating alternatives to the forecast values "yes" and "no" of the raw model output. However, moving away from the categorical "yes/no" forecast comes with a price tag: a reduction in *forecast sharpness* (Wilks, 1995).

Good sharpness is achieved when the forecasts deviate much from the climatological frequency. Thus, a categorical "yes/no" forecast is expected to attain better sharpness than a probabilistic forecast with intermediate values between 0 and 1.

In cases of limited predictability, however, a reduction in forecast sharpness may still be useful, because then the categorical "yes/no" forecast is not able to attain maximum resolution and reliability. In other words, the categorical "yes/no" forecast has difficulties to discern subsample forecast periods with different relative frequencies of the event and to forecast these frequencies correctly. Cases of limited predictability require a reduction of forecast sharpness in order to retain resolution and reliability. A good forecast maximises forecast sharpness without a loss in resolution and reliability.

The reliability diagram: Forecast reliability, resolution and sharpness

The reliability or attributes diagram (e. g. Wilks, 1995) interprets the Brier score and its decomposition in a geometric way (Fig. 6.9). Contributions from the eleven forecast categories of exceedance probability are shown separately. Additionally, the frequency of use of the forecast categories is depicted in the right boxes of Fig. 6.9. They indicate the relative amount of data that contribute to the eleven forecast categories of the reliability diagram.

The reliability diagrams in Fig. 6.9 indicate that the exceedance probabilities of 1h-accumulations

- possess good forecast resolution,
- are underdispersive,
- suffer from the frequency bias of the raw model output,
- possess little sharpness especially for heavy precipitation.

Good forecast resolution is an essential property of a probability forecast, because resolution cannot be improved by an additional calibration procedure. Good resolution becomes visible, because there is a monotonic relation between forecast probabilities and conditional observed frequencies. Deviations from the monotonic relation only occur at high forecast probabilities of high precipitation amounts. They are most probably due to effects of a small verification data set. Due to a lack of an appropriate verification measure, forecast resolution is not investigated in terms of quantiles (cf. Subsection 6.3.3).

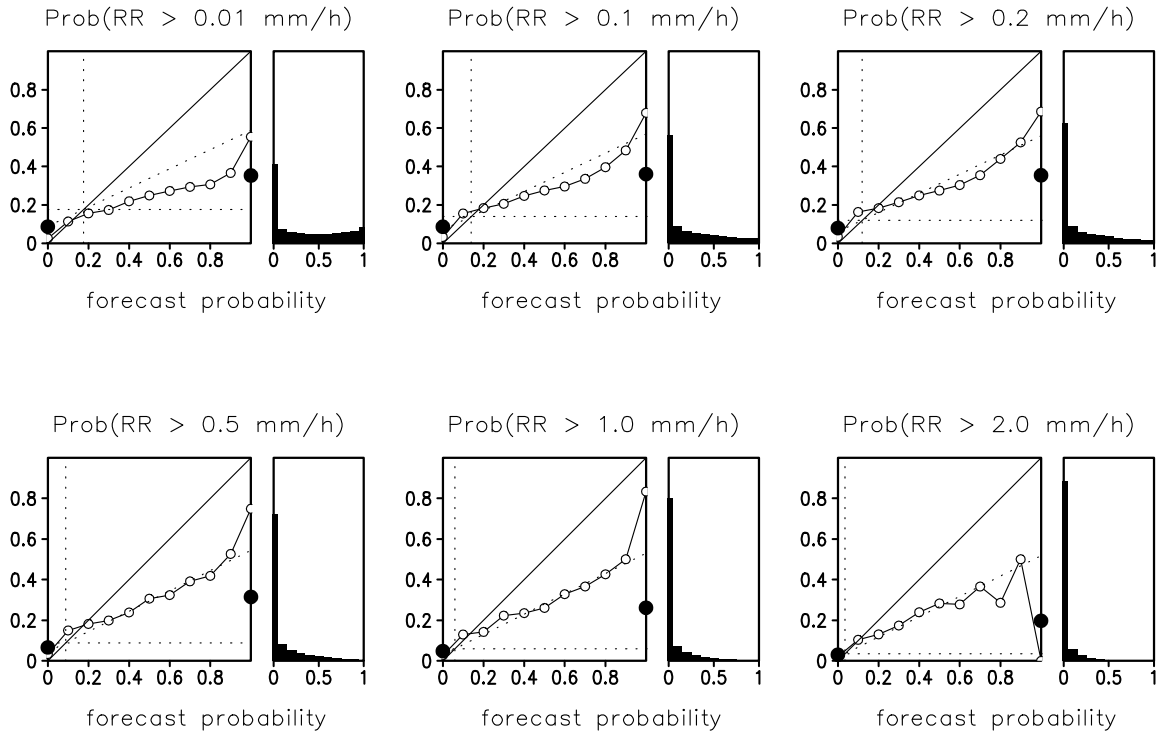

Figure 6.9: *Reliability diagrams for various precipitation events. The events are defined by the exceedance thresholds which are indicated by the respective titles of the diagrams. The predictand is the probability of such an event. The probability is split up into 11 forecast categories. The open circles in the left boxes show the observed relative frequency of precipitation occurrence as a function of the probability that is forecast. The histograms in the right boxes indicate frequency of use of the 11 forecast categories. They are independent of the observations and indicate the relative number of forecasts that contribute to one of the open circles in the left diagrams. The diagonal in the right boxes denotes perfect reliability. The dotted horizontal line denotes zero resolution which is plotted at the level of the sample climatological frequency. The dotted line between the "perfect reliability" and "no resolution" line is the "no Brier skill" line (referring to a climatological reference forecast). Points falling into the region between the vertical line and the "no Brier skill" line contribute positively to forecast skill (referring to a climatological reference forecast). The closed circles indicate the reliability of the original deterministic LM forecast.*

As the quantile estimates originate from the same neighbourhood sample than the estimates of exceedance probability, it can be assumed that the quantile estimates possess good forecast resolution as well.

Underdispersiveness and the frequency bias are also detected in the context of quantiles and discussed in Subsection 6.3.3. They lead to poor reliability and small skill. Underdispersiveness becomes visible, because the reliability curve is tilted clockwise compared to the diagonal of perfect reliability. The frequency bias becomes obvious, because the observed relative frequency of precipitation occurrence is always lower than the forecast probability. Note that the bias does not originate from the post-processing procedure, but from the raw model output, indicated by the two closed circles in the diagram.

The sharpness of the forecast can be inferred from the right boxes of the reliability diagrams (Fig. 6.9). The histograms indicate frequency of use of the forecast categories and are independent of the observations. Good sharpness is achieved when the histogram deviates much from a single column at the climatological frequency of the event. Fig. 6.9 shows that the probability forecasts exhibits especially poor sharpness for those events that are defined by large exceedance thresholds, for example RR $> 1.0\,\mathrm{mm/h}$ and RR $> 2.0\,\mathrm{mm/h}$. On the one hand, this might seem as a disadvantage. On the other hand, the deterministic predictability of heavy precipitation may be so small that the forecasts must be kept close to the sample climatology in order to retain reliability and resolution of the forecast.

Reliability diagrams of 6h-, 12h-, and 24h-accumulations (not shown) give slightly different results than the reliability diagrams of 1h-accumulations. Resolution becomes more difficult to estimate, overforecasting is less prevalent and the forecast is still underdispersive. The estimation of resolution becomes difficult, because the respective verification data sets contain less observations so that sampling effects become omnipresent in the reliability diagram. This is especially a problem in case of 24h-accumulations. Overforecasting is reduced, because the raw model output is more reliable for 6h-, 12h-, and 24h-accumulations. In terms of 24h-accumulations, forecast sharpness is also increased, but not punished by a simultaneous reduction in Brier skill score. The increased sharpness may be due to the fact that 24h-accumulations possess better predictability than shorter accumulations.

The ROC curve: Discriminating between occurrences and non-occurrences

The relative operating characteristics (ROC) curve (cf. Appendix C) is qualitatively similar to resolution in the sense that it assesses the ability of the forecast system to discriminate between occurrences and non-occurrences. The ROC curve is generated both for the raw model output and for the exceedance probabilities derived from the neighbourhood. The probability forecast is classified into eleven evenly spaced categories of forecast probabilities. The ROC curve of the deterministic forecast system is represented by a single point, complemented by the points (0,0) and (1,1) for “never” and “always

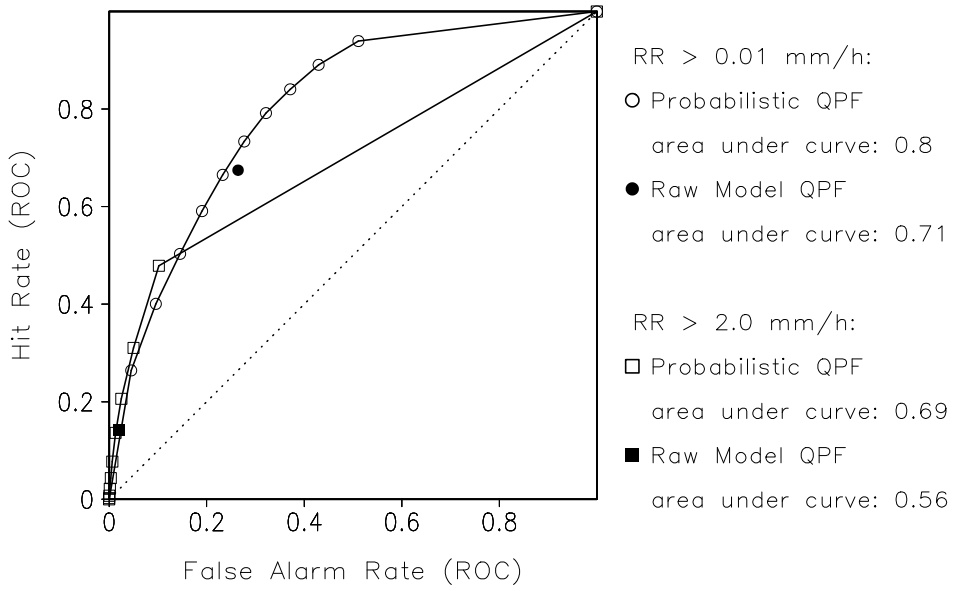

Figure 6.10: *Relative operating characteristics (ROC) for 1h-accumulations of precipitation. A ROC curve lying above the diagonal line indicates a skillful forecast system. The curves are shown for two different thresholds: 0.01 mm/h (circle) and 2.0 mm/h (square). Probabilities of exceedance (open symbols) and the direct model output (closed symbols). The standard version of the neighbourhood procedure is applied in the derivation of exceedance probabilities (cf. Tab. 4.2).*

forecast". ROC curves are obtained for 1h-accumulations of precipitation. Results are shown for a low and a high precipitation threshold, respectively (Fig. 6.10). For the exceedance threshold of 2.0 mm/h, many points of the ROC curve are indistinguishable from each other, because they cluster at the lower left of the diagram.

When compared to the raw model output, the exceedance probabilities of all exceedance thresholds are skillful in terms of the ROC curve. The lower the exceedance threshold, the better the results become. An exception is the threshold of 0.01 mm, as these results are slightly worse than results of the 0.1 mm threshold (not shown).

The superiority of the probabilistic forecast becomes obvious when the area under the ROC curve is calculated (cf. Fig. 6.10). The area is obtained by using the trapezoid rule. Due to the different number of points the areas for the categorical forecast suffer in comparison to areas calculated from the probability forecast. However, this reflects the additional ability to discriminate between events and nonevents associated with having multiple decision criteria (Wandishin et al., 2001).

The ROC curves of the exceedance probabilities are not perfect yet. Existing deficiencies partly result from the frequency bias of the raw model output again. ROC curves of light precipitation suffer from high false alarm rates and ROC curves of heavy precipitation

suffer from low hit rates. This is due to the fact that the raw model output suffers from overforecasting light precipitation and from underforecasting heavy precipitation (see FBI in Fig. 6.1).

Is there a trade-off between random error and systematic error?

Neighbourhood smoothing is expected to reduce random error at the expense of an increased systematic error (Subsection 4.3.3). Resolution does not penalise forecast bias, so this aspect is not reflected by resolution, but might show in reliability. However, reliability does not deteriorate either (Fig. 6.7).

It is unclear why the increase in systematic error does not become visible in forecast reliability although it is expected theoretically. Several explanations are possible. First, random error might be such a dominating source of overall forecast error in the raw model output that a reduction in random error outweighs any increase in systematic error. Secondly, an additional introduction of systematic error might incidentally compensate for an existing bias in the raw model output. Thirdly, neighbourhood post-processing generally brings the forecast closer to a climatological forecast which yields improved reliability at the expense of declined sharpness.

6.4 Forecast consistency and forecast value

As pointed out in Subsection 6.2.2, forecast goodness consists in forecast quality, forecast consistency and forecast value. After having discussed forecast quality in the previous section, this section discusses forecast consistency (Subsection 6.4.1) and forecast value (Subsection 6.4.2) of the standard neighbourhood results. As consistency is difficult to measure, this study presents a qualitative discussion instead of a quantitative assessment. The assessment of forecast value is based on the cost-loss scenario and applies the quantitative measure of *relative value*.

6.4.1 Forecast consistency of the neighbourhood results

An essential aspect of consistency is the inclusion of forecast uncertainty into the forecast (Murphy, 1993). If a description of the inherent uncertainty is missing, the forecast does not correspond to the forecaster's best judgment. The approach wrongly suggests that the forecaster is 100% sure of the precipitation amount at a specific time and a specific location.

As the LM is a deterministic model, the raw model output does not contain any estimate of forecast uncertainty or risk and therefore lacks consistency. The time series of the DMO in Fig. 4.21 and 4.22 reflect the way a user without meteorological background

Table 6.5: *Contingency table for the cost-loss scenario. A user suffers a loss L if an event occurs and the user is not protected. If the user takes preventive action to guard against this potential loss, the user will incur a cost C.*

		Observed	
		Yes	No
Forecast	Yes	C	C
Forecast	No	L	-

typically interprets the model forecast, i. e. she mistakes grid box size with forecast resolution, isolates the grid box nearest her specific location of interest and takes the forecast at face value.

Post-processing clearly enriches the information content of the meteogram. For example, adding probabilistic information via the quantile estimates (Fig. 4.21 and 4.22) informs the user of forecast uncertainty. Although it is not known whether the given estimate of forecast uncertainty reflects a hypothetical forecaster's judgment correctly, the mere existence of a probabilistic forecast formulation is certainly a step towards the right direction.

6.4.2 Forecast value of the neighbourhood results

Introduction to the concept of forecast value

The concept of forecast value presupposes that the user employs the precipitation forecast as an input into her decision making process. Forecast value then refers to the economic saving a user gains by using the forecast. Forecast value does not only depend on forecast quality, but also on the particular decision making situation of the individual user.

In this study, the prototypical decision making situation is described by the cost-loss scenario (Tab. 6.5) which is already proposed by Thompson and Brier (1955). The cost-loss scenario is based on the idea that a user suffers a loss L if an event occurs and the user is not protected. If the user takes preventive action to guard against this potential loss, the user will incur a cost C.

Forecast value is assessed by the verification measure *relative value* (Richardson, 2000; Zhu et al., 2002; Mylne, 2002). Its definition can be found in Appendix C.4. Relative value follows the form of a skill score, i. e. it is related to the perfect forecast and to a reference forecast. In this study, the reference forecast is the climatological forecast from the verification period. It can be shown that relative value depends on the values a, b, c, d from the contingency table (Tab. C.1) and on the cost-loss ratio C/L (Richardson, 2000). The individual values of C and L do not need separate consideration. Different users and decision making situations are sufficiently represented by the range of values $C/L \in [0, 1]$.

If $C > L$ the cost-loss scenario becomes trivial, because then the user should never take protective action, irrespective of the forecast.

Relative value is estimated for the expected value, the exceedance probabilities and the 90%-quantile of the standard neighbourhood results. The relative value of the neighbourhood estimates is compared to the relative value of low benchmark foreasts: the direct model output and a so-called *inflated* version of the direct model output.

It should be mentioned that relative value has originally been designed for the assessment of categorical and probability forecasts. This study is the first to derive relative value of a quantile forecast in addition. This requires a minor reformulation of the decision making situation (cf. Appendix C.4).

Characteristics of the *ex ante* approach

Before delving into the assessment of forecast value, some remarks should be made about the specific approach in this study. If the forecast is expressed in terms of probabilities, the user has to define a threshold probability p_t in order to convert the probability forecast into a categorical guidance in her decision making process (App. C.4). If the threshold probability p_t is exceeded, the user takes protective action and vice versa. In this study the *ex ante* approach (Murphy, 1993) is chosen, i. e. the optimal probability threshold p_t for a particular cost/loss ratio is directly identified by evaluating the relative value associated with the use of different decision criteria and taking the maximum (cf. Richardson, 2000; Mylne, 2002).

In case of the *ex ante* approach, the relative value is automatically non-negative for all cost-loss ratios if the climatological forecast is chosen as a reference forecast. When the relative value is zero, the decision criterion of the respective user is "never protect" (if C/L is high) or "always protect" (if C/L is low), irrespective of the forecast. Only when the forecast system yields a positive relative value, the user really bases her decision on the forecast.

Note that the *ex ante* approach implicitly involves a recalibration of the forecasts on the basis of the observations (Murphy, 1993). The recalibration is encoded in choosing the optimal probability threshold p_t by maximizing relative value, a procedure that is not feasible in an operational forecast setting. The results of the study point out a *potential* value of the exceedance probability, assuming that the forecast can be perfectly calibrated.

Expected value and exceedance probability

Fig. 6.11 shows the relative value of the expected value and the exceedance probability. For comparison, the relative value of the direct model output is presented as a low benchmark. Value of the climatological forecast and value of the perfect forecast are

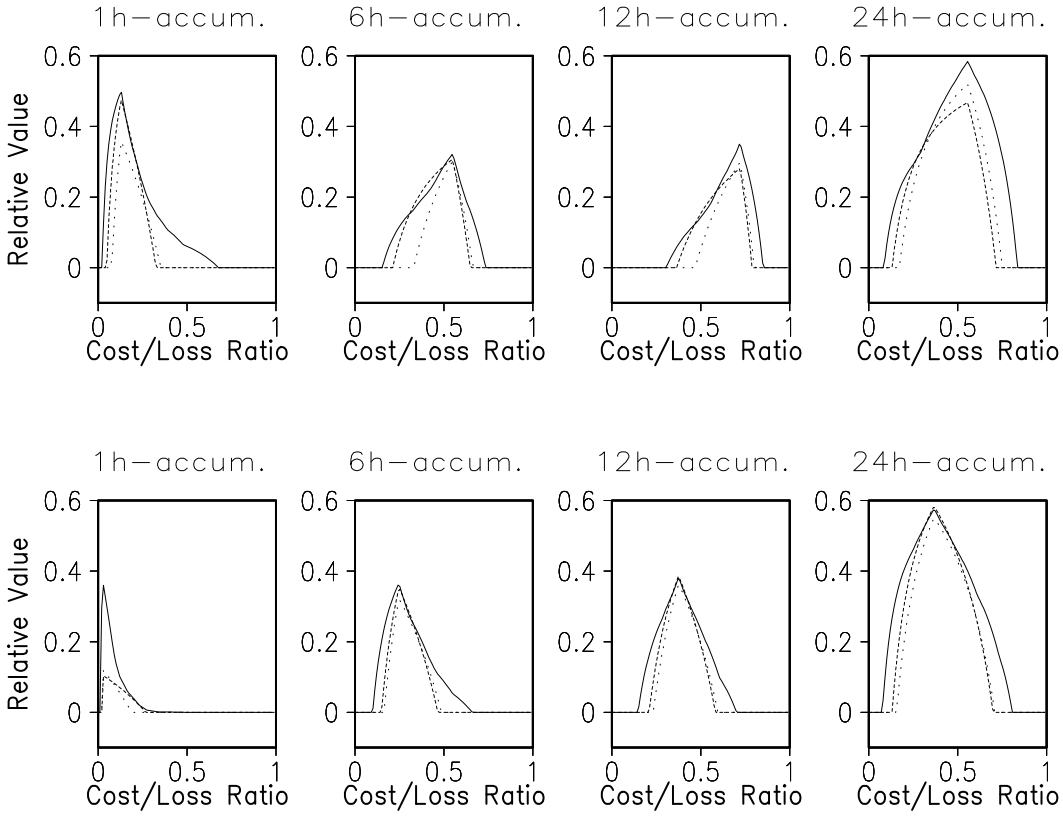

Figure 6.11: *Relative value of exceedance probabilities as a function of the user's cost-loss ratio. Upper row: Exceedance threshold of 0.1 mm. Lower row: Exceedance threshold of 2.0 mm. Results are shown for the probability forecast (solid line), the expected value (dashed line) and the original forecast (dotted line). The standard version of the neighbourhood procedure is applied in the derivation of the probabilities and expected values (cf. Tab. 4.2).*

already contained in the verification measure itself (cf. Appendix C.4). As outlined in Subsection 6.2.3, this study uses a substitute of the climatological forecast which is derived from the 15-day-mean of the verification period itself.

Relative value is estimated as a function of the user's individual cost-loss ratio. Again, forecast variables under consideration are precipitation accumulations over time periods of 1 hour, 6 hours, 12 hours and 24 hours. The estimates of relative value are shown for two exceedance thresholds: 0.1 mm and 2.0 mm.

Fig. 6.11 shows that the exceedance probability outperforms the direct model output and the expected value. This is the case for most cost-loss ratios, accumulation times and exceedance thresholds. Furthermore, the exceedance probability also widens the range of cost-loss ratios for which the forecasts exhibit positive relative value, indicating that a larger group of users can benefit from the exceedance probability as compared to the direct model output and the expected value.

The maximum of relative value is always attained by users whose cost-loss ratio equals the climatological frequency of the event, as expected theoretically (Richardson, 2000). Note that the climatological frequency of the event is lower for 24h-accumulations than for 12h-accumulations. This is not necessarily a contradiction, because a substantial part of the 24h-accumulations are obtained from stations that do not report 12h-accumulations.

Differences between expected value and direct model output are not as clear as the superiority of the exceedance probability, but are in accordance with the result of Subsection 6.3.1. For small cost-loss ratios the expected value often possesses a slightly larger relative value than the direct model output. This corresponds with the fact that the expected value of low precipitation thresholds also exhibits a higher frequency bias, a higher probability of detection and a higher false alarm rate than the direct model output (Fig. 6.1 and 6.2). Such a forecast is more beneficial to users with a small cost-loss ratio, because they can tolerate high false alarm rates better than low hit rates.

As the *ex ante* approach implicitly assumes that forecast calibration is perfect, the superiority of the exceedance probability can be explained by the findings of Murphy (1977). He demonstrates that the value of a probabilistic forecast is always greater than or equal to the value of a categorical forecast, if the probabilistic forecast of concern is perfectly calibrated. This insight does not belittle the results of the forecast-value study here, but emphasises the need for probabilistic forecast formulation even more strongly.

In terms of forecast value, the main advantage of the exceedance probability is its variety of forecasts for different users and their different needs. In case of categorical forecasts, the forecaster adopts the role of a decision maker and "tailors" the probabilistic forecast to a particular user without considering her individual needs. When issuing warnings, this often results in high false alarm rates, since many forecasters tend to err on the safe side. The advantage of a probabilistic forecast is the possibility to clearly separate the roles of the forecaster and the decision maker.

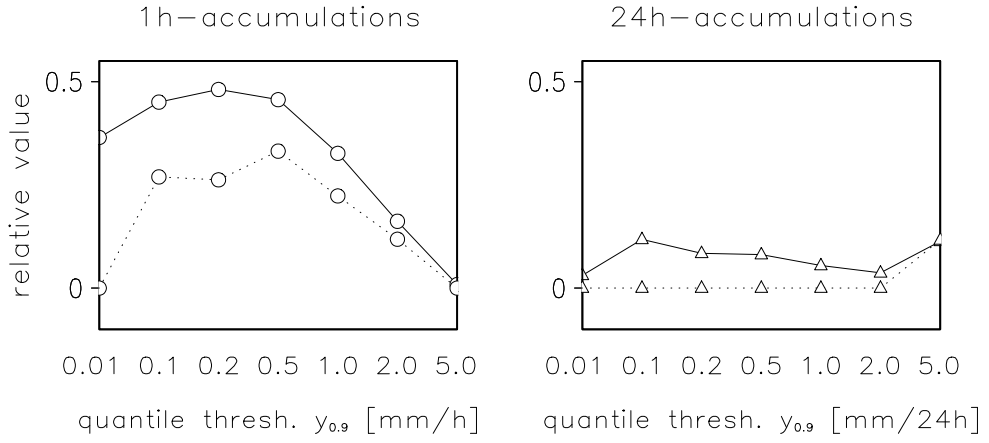

Figure 6.12: *Relative value of 90%-quantiles as a function of quantile threshold $y_{0.9}$. The quantile threshold denotes the lowest precipitation amount that corresponds to a user's cost-loss ratio of 10%. It is assumed that the cost-loss ratio is a strictly monotonically decreasing function of precipitation amount. Results are shown for quantiles derived by the neighbourhood post-processing procedure (solid line) and for quantiles derived by simple inflation of the direct model output (dotted line) (cf. Tab. 6.6).*

90%-quantile

Additionally, forecast value of the 90%-quantile (cf. Appendix C.4) is addressed. The 90%-quantile is an appropriate forecast for those users with a cost loss ratio of 10% or less. Therefore, the 90%-quantile is useful for issuing warnings, because warnings mostly refer to events that cause an immense loss compared to a fairly low cost of preventive action.

In order to assess forecast value of a quantile, the decision making problem must be reformulated. In contrast to the previous section the cost-loss ratio possesses a fixed value of 10% now. Different users are represented by so-called quantile thresholds $y_{0.9}$. The quantile threshold is defined as the precipitation amount for which the user's cost-loss ratio is less than or equal to 10%. A monotonic relation between precipitation amount and cost-loss ratio is assumed. Only if the forecast of the 90%-quantile exceeds the lowest precipitation amount which corresponds to a cost-loss ratio of 10%, the user takes preventive action.

Fig. 6.12 shows the relative value of the 90%-quantile as a function of quantile thresholds that represent the decision making situation of different users. For 1h-accumulations and 24h-accumulations the neighbourhood estimate of the 90%-quantile possesses positive relative value (solid line in Fig. 6.12) whereas relative value is zero for almost all quantile thresholds of 6h-accumulations and 12h-accumulations (not shown). In the following paragraphs, the value of 1h- and 24h-accumulations is discussed further. An explanation is given why they possess larger value than 6h-accumulations and 12h-accumulations.

Highest values are achieved for low quantile thresholds of 1h-accumulations. Low quantile thresholds imply that a cost-loss ratio of $\leq 10\%$ is already attained at low precipitation amounts, i. e. the corresponding user suffers a relatively high loss even in case of light precipitation.

The relatively high value of 1h-accumulations is probably due to good reliability (Fig. 6.3). However, the good reliability of 90%-quantiles originates from the coincidence of two shortcomings: strong overforecasting of 1h-accumulations and underdispersiveness (cf. Subsection 6.3.3). For low quantile thresholds, their negative effects cancel each other out. When the quantile threshold is increased, overforecasting disappears (cf. FBI in Fig. 6.1) and underdispersiveness remains so that relative value deteriorates (cf. solid line in Fig. 6.12). It is likely that the good forecast value of low quantile thresholds will be reduced if a new version of the LM is introduced that alleviates the problem of overforecasting. As overforecasting is less prevalent for 6h-accumulations and 12h-accumulations, the respective 90%-quantiles do not achieve relative values greater than zero.

Another low benchmark for the 90%-quantile: The *inflated* model output

As the verification procedure looks at relative value, the 90%-quantile is implicitly compared to the 15-day-mean of the verification period and to the perfect forecast. It would be desirable to compare the 90%-quantile to another benchmark. A comparison to the DMO would be interesting, because it might conceal shortcomings that originate from the LM itself and reveal the benefits of the post-processing procedure more clearly. However, a direct comparison between the DMO and the 90%-quantile would be unfair to the DMO. By prescribing a fixed cost-loss ratio of 10%, the DMO would be silently assumed to be a forecast of risk. This is obviously not the case. In order to convert the DMO into a very simple forecast of risk, the DMO is *inflated*.

Inflation means that the forecasts from the DMO are simply replaced by somewhat larger values. For example, the inflation rule might consist in adding a constant value to the forecast or in multiplying the forecast by a constant factor. In this study, a slightly different procedure is applied so as to roughly preserve the unconditional distribution of precipitation amounts. The precipitation values are classified into several categories of precipitation amount and the inflation rule moves the forecast value up by one category (Tab. 6.6). In contrast to the neighbourhood estimates, the inflated forecast is inferred from the precipitation amount alone instead of using a spatio-temporal neighbourhood.

Details of the inflation rule (Tab. 6.6) are somewhat arbitrary. All 90%-quantiles are assumed to exceed 0.01 mm; if the direct model output exceeds 0.01 mm, the 90%-quantile is assumed to exceed 0.1 mm; if the direct model output exceeds 0.1 mm, the 90%-quantile is assumed to exceed 0.2 mm etc. Larger intervals are chosen for higher precipitation amounts so as to roughly account for the quasi-logarithmic distribution of precipitation amounts.

Table 6.6: *Inflation of the direct model output (DMO) so as to infer a simple 90%-quantile estimate from the precipitation amount alone instead of using a spatio-temporal neighbourhood.*

DMO	→	estimated 90%-quantile
all	→	$> 0.01\,\mathrm{mm}$
$> 0.01\,\mathrm{mm}$	→	$> 0.1\,\mathrm{mm}$
$> 0.1\,\mathrm{mm}$	→	$> 0.2\,\mathrm{mm}$
$> 0.2\,\mathrm{mm}$	→	$> 0.5\,\mathrm{mm}$
$> 0.5\,\mathrm{mm}$	→	$> 1.0\,\mathrm{mm}$
$> 1.0\,\mathrm{mm}$	→	$> 2.0\,\mathrm{mm}$
$> 2.0\,\mathrm{mm}$	→	$> 5.0\,\mathrm{mm}$

More systematic approaches to deriving an alternative 90%-quantile estimate would become possible if historical error statistics were available for the LM. Then the 90%-quantile could be inferred through quantile regression or recalibration. Quantile regression involves a minimisation of the weighted mean absolute error (Koenker and Hallock, 2001; Bremnes, 2004) and recalibration involves a maximisation of reliability. These approaches are not followed up here, because the continuous refinement of the LM does not allow for the establishment of error statistics that are consistent throughout a long time period.

Relative value of the inflated DMO is derived in the same way than the relative value of the 90%-quantiles in the previous subsection. The result is shown in Fig. 6.12. As expected, the inflated forecast achieves a larger relative value than the DMO itself (not shown). In comparison to the neighbourhood estimate, the inflated DMO is still inferior (Fig. 6.12).

This result gives the impression as if the local spatio-temporal neighbourhood of a forecast revealed more valuable information about forecast uncertainty than the precipitation amount at an isolated grid point. Recall, however, that the neighbourhood estimates of the 90%-quantile (1h-accumulations) benefit from the accidental combination of two shortcomings: overforecasting and underdispersiveness. If the inflated forecast should be affected by only one of these shortcomings, objective verification will make it appear in a worse light than the neighbourhood estimate. Therefore, partial improvements of the post-processing methodology may not become obvious in the 90%-quantile of 1h-accumulations. Caution must be exercised in the interpretation of Fig. 6.12.

6.5 Optimal configuration of the neighbourhood method

The neighbourhood method includes several free parameters, such as neighbourhood size and the way of considering explanatory variables. A standard version of the neighbourhood method is defined in Chapter 4. The standard version uses a *medium* neighbourhood size (cf. Tab. 4.1) and considers explanatory variables in the way as listed in Tab. 4.2. Chapter 4 justifies the particular choice of the standard version by subjective assessment of case studies and by theoretical arguments. This section complements these investigations by a sensitivity study which is based on a larger data base and on objective verification measures. A configuration is sought that maximises forecast quality while keeping the computational cost low.

6.5.1 Neighbourhood size

Introductory remarks

In terms of neighbourhood size, three configurations "small", "medium" and "large" are defined (cf. Tab. 4.1). Theoretical implications of neighbourhood size are discussed in Chapter 4. Some indication of the order of magnitude can be inferred from existing studies (Subsection 4.2.4). Chapter 4 and Section 6.3 focus on results from the medium size. This subsection tackles the following questions:

- How large is the effect of neighbourhood size on forecast quality?
- Does neighbourhood size really result in a trade-off between bias and variance error?
- Is there an optimum neighbourhood size?

As outlined in Subsection 4.2.4), neighbourhood size is tantamount to the uncertainty in location and time we expect of a precipitation amount or temperature value. When neighbourhood size is increased, a grid box forecast is taken to be representative of a larger area, the degree of expected smoothness in the field of the resulting estimate is increased and the location and time of an event is assumed to be deterministically less predictable.

The link between simple neighbourhood post-processing and kernel smoothing (Subsection 4.3.2) offers a more detailed explanation of the effect of neighbourhood size on the resulting estimates. Neighbourhood size is equivalent to the bandwidth in kernel regression and kernel intensity estimation problems. In kernel smoothing the mean squared error of the estimate can be expressed as a function of kernel bandwidth, showing that kernel bandwidth plays a key role in the error of the resulting estimate as it determines the balance in the trade-off between bias and variance error (Subsection 4.3.3). Larger

Table 6.7: *Forecast quality measures of the expected value of 2m-temperature. Comparison of different neighbourhood sizes (Tab. 4.1) in the post-processing procedure. Standard neighbourhood post-processing is applied (cf. Tab. 4.2).*

Measure	Small	Medium	Large
rmse	1.82 K	1.81 K	1.81 K
mae	1.35 K	1.33 K	1.34 K
bias	-0.03 K	-0.05 K	-0.05 K
stdv_fc	3.92 K	3.86 K	3.76 K
corr	0.90	0.90	0.90
skill_DMO	0.048	0.054	0.057

bandwidths lead to a larger bias; the estimate is influenced by values from a larger region in which the shape of the unknown function may vary substantially so that the realisations from the region are not representative of each other. Smaller bandwidths lead to a larger variance error; the estimate is based on comparatively few realisations.

In this section it is shown whether the error concept above actually becomes visible in the verification results. If so, it should be possible to determine an optimal neighbourhood size. However, it is also conceivable that the error concept might not be applicable to LM post-processing, because the fundamental assumptions in kernel smoothing might be too simplistic when applied to LM post-processing.

Expected value of 2m-temperature

Estimates of the expected value of 2m-temperature are derived from the small, medium and large neighbourhood, respectively. The resulting estimates are compared and evaluated by standard measures of forecast quality (Tab. 6.7). A variation of neighbourhood size amplifies or diminishes some of the post-processing effects described in the comparison between the post-processed forecast and the direct model output (cf. Tab. 6.4). Increasing the neighbourhood size generally leads to an increase in bias (*bias*), a slight reduction in root mean squared error (*rmse*) and a strong reduction in the estimate's variability (*stdv_fc*).

The expected trade-off between bias and variance error becomes visible. The medium size yields a smaller root mean squared error and a larger bias than the small size. The reduction in root mean squared error must originate from a reduction in variance error.

An optimal neighbourhood size appears to be somewhere in the range of the small and medium neighbourhood size. Verification measures do not indicate a clear superiority of either the small or the medium size and the large neighbourhood size does not lead to further improvements.

Before concluding this subsection it should be noted that the time extension of the large

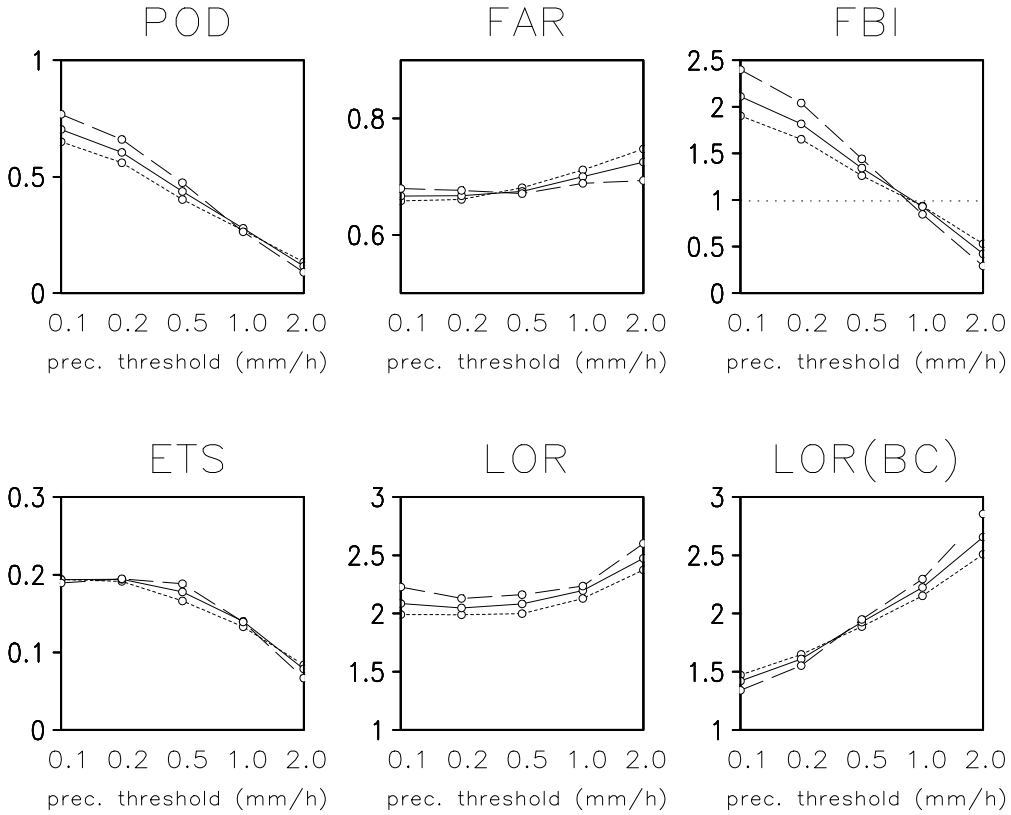

Figure 6.13: *Same as Fig. 6.1, but for different sizes of the neighbourhood: small (short dashed line), medium (solid line), large (long dashed line) (Tab. 4.1).*

neighbourhood (cf. Tab. 4.1) is inappropriately large when applied to 2m-temperature. A time window of 7 hours (cf. Fig. 4.10) allows for considerable variability of solar radiation. Solar radiation is an explanatory variable, i. e. it enhances predictability instead of contributing to random error. Currently, the neighbourhood method does not contain any correction for daily variability due to solar radiation. This may be a reason why the large neighbourhood does not attain better forecasts than the medium neighbourhood.

Expected value of precipitation

The three different neighbourhood sizes are also applied to the precipitation forecast. The resulting estimates of expected value are compared and evaluated by standard measures from the contingency table. As the results from different accumulations are very similar, only the 1h-accumulations are highlighted as an example (Fig. 6.13). The different choices of neighbourhood size result in slightly different outcomes of verification measures which are described in the following paragraph. It should be noted that the results need further confirmation by a larger verification data set. When the differences

in log odds ratio undergo a significance test, statistical significance is not attained.

Similarly to the 2m-temperature forecast, a variation of neighbourhood size amplifies or diminishes the post-processing effects that is described in the comparison between the post-processed precipitation forecast and the direct model output (cf. Fig. 6.1). Increasing the neighbourhood size generally leads to a worse FBI and corresponding changes in POD, FAR, ETS and LOR. When a large neighbourhood size is applied, the bias-corrected log odds ratio (LOR(BC)) indicates better forecast quality for high thresholds and lower forecast quality for low thresholds.

The deterioration of FBI with increasing neighbourhood size corresponds to the theoretical expectations, as the FBI represents the bias error. A simultaneous reduction of random error is difficult to detect, since most other verification measures are influenced by the FBI. Only the bias-corrected log odds ratio can give an indication of random error reduction. The result is ambiguous; the small neighbourhood size is favourable when small precipitation thresholds are considered and vice versa.

This behaviour can be a consequence of decreasing predictability with increasing precipitation amount. Optimal neighbourhood size is closely connected with the expected degree of deterministic predictability and large precipitation amounts often originate from convective processes that possess less deterministic predictability than large-scale processes. However, further investigations are still needed to give more confidence.

As a summary, an optimal universal neighbourhood size cannot be determined with respect to the expected value of precipitation. The effect of neighbourhood size does not only depend on precipitation amount, but also on the user's needs. As stated in Section 6.3.1 the value of the forecast depends on the user's ability of tolerating high false alarm rates or low hit rates.

Quantiles and exceedance probabilities

The question of an optimal neighbourhood size is also raised when estimating quantiles and exceedance probabilities. Again, the three different neighbourhood sizes are applied and resulting measures of forecast quality are compared. Temperature and precipitation quantiles are assessed via the reliability diagram (Fig. 6.14); exceedance probabilities of precipitation are assessed via the Brier skill score (Fig. 6.15), its decomposition, reliability diagrams and relative value (not shown).

For 2m-temperature and all precipitation accumulations the large neighbourhood size achieves best verification results. Saturation or decay of forecast quality with increasing neighbourhood size does not seem to be in sight. The decomposition of the Brier score (not shown) reveals that the superiority of the large size is due to an improvement in both resolution and reliability. The results of the reliability diagram reaffirm the great impact of neighbourhood size and the superiority of the large neighbourhood in terms of reliability. Relative value of the exceedance probability is less sensitive to

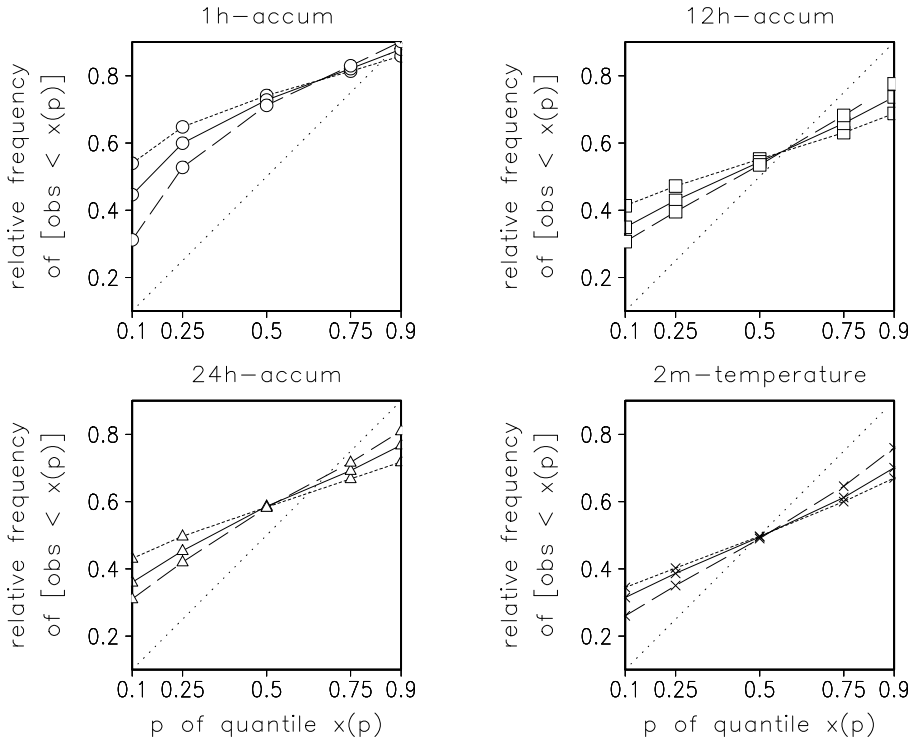

Figure 6.14: *Same as Fig. 6.3, but for different sizes of the neighbourhood: small (short dashed line), medium (solid line), large (long dashed line) (Tab. 4.1).*

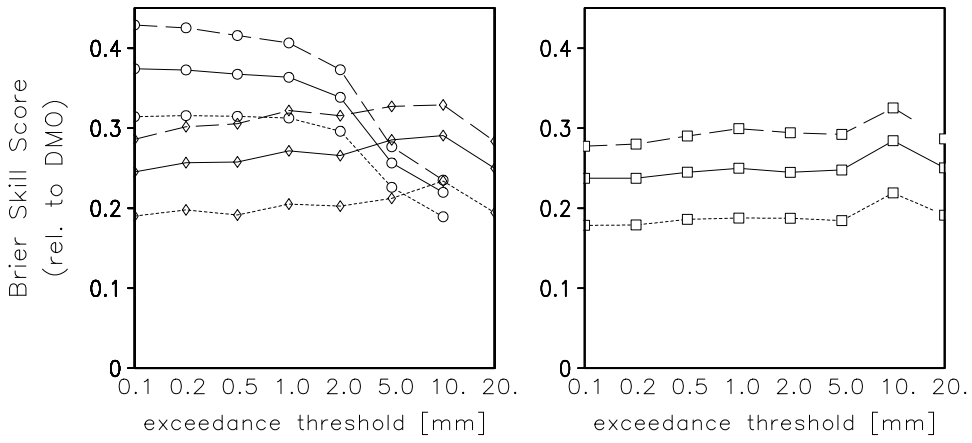

Figure 6.15: *Same as Fig. 6.5, but for different sizes of the neighbourhood: small (short dashed line), medium (solid line), large (long dashed line) (Tab. 4.1).*

neighbourhood size. For almost all cost-loss ratios the large neighbourhood reaches only marginally better forecast value than the medium and small neighbourhood.

On the one hand, it may be conceivable that the large neighbourhood sample represents forecast uncertainty best. The large neighbourhood possesses a spatial diameter of 140 km so this is in accordance with findings of Walser et al. (2004) who note that mesoscale predictability limitations may be critical even at scales exceeding 100 km.

On the other hand, this impression may be misleading. Recall that post-processed probability forecasts are generally underdispersive (Section 6.3). This means that the variance of the realisations within a medium-sized neighbourhood sample underestimates forecast uncertainty. Any measure that contributes to an increase of variance is likely to alleviate the problem so that the forecast achieves better verification scores. Increasing the neighbourhood size certainly has such an effect, no matter whether the realisations within the neighbourhood sample are still representative of the probability function to be estimated or not. Even if the large-sized neighbourhood yields a forecast that is less underdispersive, it is not guaranteed that the spread does mirror observed forecast uncertainty or even captures observed variations in forecast uncertainty. For example, it should be tested whether a small-sized neighbourhood sample with inflated variance yields better estimates than a large-sized neighbourhood sample.

An optimal neighbourhood size with respect to quantile estimation cannot be determined at this stage. The problem of underdispersiveness may superimpose any measurable effect of neighbourhood size on forecast reliability. Further investigations need to deal with the question whether the variance within the neighbourhood sample represents forecast uncertainty in a sufficient way.

The error concept of kernel smoothing may not be fully applicable to LM post-processing as it is deduced from the assumption that the available realisations exhibit realistic variance and no bias. In LM post-processing the underlying data from the raw model output probably do not possess realistic variance, as uncertainty in location and time represents only a fraction of overall forecast uncertainty. In addition, the data are biased if the large-scale forecast does not correspond to the observational data at all. Thus, an optimal neighbourhood size may not exist in the sense of the optimal kernel bandwidth.

Summary

The choice of neighbourhood size plays an important role in forecast quality of the resulting estimates. However, it is difficult to determine an optimal choice among the three sizes tested.

In terms of 2m-temperature, the small or medium size yields best expected values, but this may be due to an insufficient alleviation of the explanatory effect of solar radiation in the large neighbourhood. In terms of precipitation, an optimal neighbourhood size

cannot be found at all. The quality of expected value depends on precipitation amount and its value depends on the user's needs.

The quality of quantiles and exceedance probabilities seems to be best when the largest neighbourhood size is applied. Saturation or decay of forecast quality with increasing neighbourhood size does not seem to be in sight. An increase in neighbourhood size alleviates the problem of underdispersiveness, because it increases the variability within the neighbourhood sample. It could be concluded that the large size was optimal, but this conclusion may be premature. A large neighbourhood size probably does not address the origin of the post-processing deficiencies. Increasing the neighbourhood size can be seen as a way of hedging, as it can be used as a means of improving the score by forecasting something other than one's true beliefs.

Before concluding this section of neighbourhood size, it should be mentioned that a one-size-fits-all approach is most likely inadequate. A more elaborate methodology should allow for a variable size depending on surface forcing, flow, forecast lead time etc.

6.5.2 Consideration of explanatory variables

Introductory remarks

Several refinements of the neighbourhood are amended to the basic concept of neighbourhood post-processing (Chapter 4). Most of them refer to the effect of explanatory variables (Section 4.4) and consist in a linear modification of the model output (Section 4.4.2) or in changing the shape of the neighbourhood (Section 4.4.3).

A standard version of considering explanatory variables is defined in Tab. 4.2. The choice is justified by subjective assessment of a case study. In this subsection, objective verification of a larger data set reaffirms the optimality of the standard version. Variants of the standard version are defined and forecast quality of the respective results are measured. Recall that optimality is not solely attained through best quality of the post-processing forecasts, but through a best trade-off between the complexity of the post-processing procedure and the quality of the resulting forecasts.

2m-temperature

The standard version of temperature post-processing includes linear modification due to orographic height and neighbourhood cut-off due to snow cover and land-sea mask (cf. Tab. 4.2). Two test versions are defined here. The first one is a low benchmark; it does not take into account any explanatory variables. The second one is a high benchmark; it additionally includes neighbourhood distortion due to orographic height.

A comparison of the resulting verification measures (Tab. 6.8) shows that the version without consideration of explanatory variables performs considerably worse than the

Table 6.8: *Forecast quality measures of the expected value of 2m-temperature. Comparison of different considerations of explanatory variables in the post-processing procedure.*

Measure	Standard	Without Consideration of Explanatory Variables	With Additional Neighbourhood Distortion
rmse	1.81 K	2.00K	1.80 K
mae	1.33 K	1.48 K	1.33 K
bias	-0.05 K	-0.15 K	-0.05 K
stdv_fc	3.86 K	3.61 K	3.87 K
corr	0.90	0.87	0.90
skill_DMO	0.054	-0.045	0.059

standard version. In fact, forecast quality is markedly inferior to the forecast quality of the direct model output (cf. Tab. 6.4). This result clearly justifies the higher computational expenses of the standard version as compared to the simple version.

In terms of quantiles, however, the version without consideration of explanatory variables attains higher reliability (Fig. 6.16). Again, this is an effect of increased variance within the neighbourhood sample, similarly to an increase in neighbourhood size. The standard version yields an underdispersive forecast, so omitting explanatory variables tends to alleviate this problem. As discussed in the context of neighbourhood size (Section 6.5.1), the seeming improvement in forecast reliability is probably misleading. Any method that increases the sample variance is rewarded, no matter whether it addresses the origin of the deficiency or not.

An additional neighbourhood distortion due to orographic height results in a marginal improvement of the standard version (Tab. 6.8). Note that it simultaneously achieves a reduction in root mean squared error and an increase in forecast variability (*stdv_fc*). A future study should test whether these differences are statistically significant.

When selecting a version for an operational environment, increase in forecast quality needs to be related to the increase in computational expenses. Neighbourhood distortion requires a large amount of additional computing time. The slight increase in forecast quality does not justify the increased effort so that the standard version without neighbourhood distortion should still be preferred in an operational setting.

Precipitation

The standard version of precipitation post-processing does not consider any explanatory variables (cf. Tab. 4.2) although it is known that orography has a deterministic effect on the precipitation forecast. It is desirable to test whether the post-processed forecast would benefit from the additional consideration of orography as an explanatory variable. Therefore, a test version is defined which includes neighbourhood distortion due to

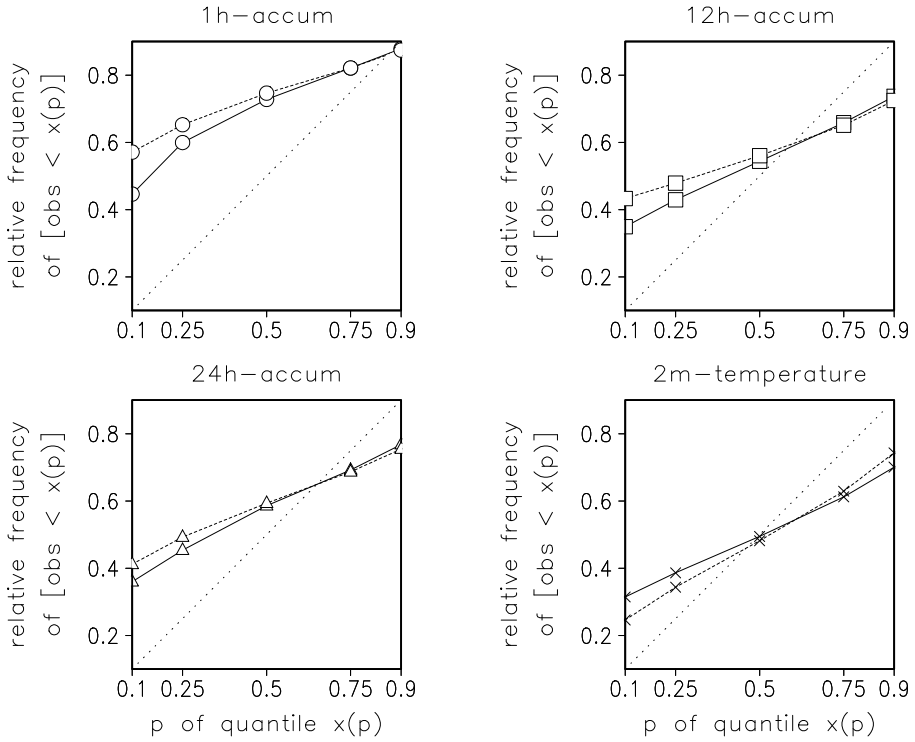

Figure 6.16: *Same as Fig. 6.3, but for different configurations of the neighbourhood procedure. Precipitation: standard version (solid line), version with neighbourhood distortion and linear modification of the output (dashed line). Temperature: standard version (solid line), version without consideration of explanatory variables (dashed line).*

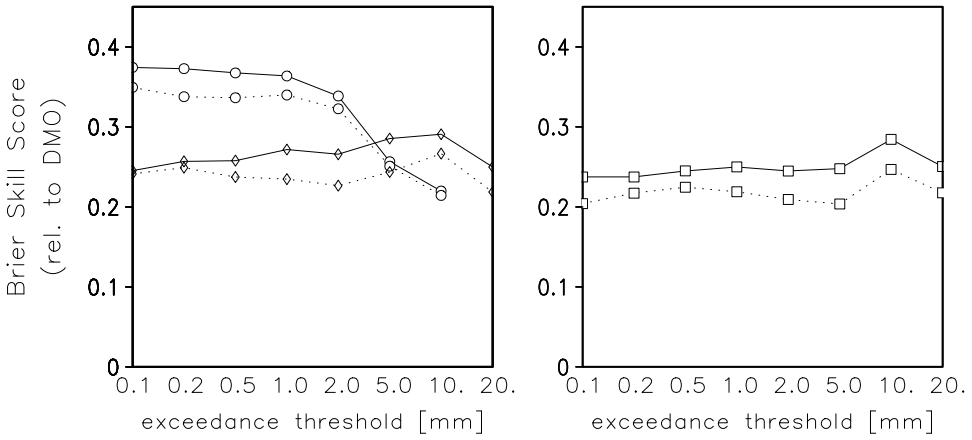

Figure 6.17: *Same as Fig. 6.5, but for different configurations of the neighbourhood procedure: standard version (solid line), version with neighbourhood distortion and linear modification of the output (dotted line).*

orographic height and linear modification of the model output due to orographic height and orographic slope.

The test version is used to derive estimates of expected value, quantiles and exceedance probability. Verification measures are calculated and compared to those derived from the standard version. Measures from the contingency table (not shown), quantile reliability (Fig. 6.16) and Brier skill score (Fig. 6.17) give unanimous results. In all cases the additional consideration of orography has a similar effect as decreasing the neighbourhood size (cf. Subsection 6.5.1). Thus, the standard version appears to be superior, at least in terms of quantiles and exceedance probability.

The similarity between the effect of neighbourhood size and the effect of considering explanatory variables is not surprising. Both the decrease in size and the additional consideration of explanatory variables reduces the variance within the neighbourhood sample. A reduction of variance results in a forecast that is less smooth and even more underdispersive. The latter is a clear disadvantage. Thus, the increased computational expense is not justifiable. Although theoretical arguments speak in favour of an additional consideration of orography, the neighbourhood procedure does not seem to have the ability to cope with this challenge. At this stage the standard post-processing version (Tab. 4.2) should be preferred to the test version for the time being, but further research is still needed so as to develop a better approach to considering explanatory variables in post-processing the precipitation forecast.

An additional neighbourhood distortion due to orographic height results in a marginal improvement of the standard version (Tab. 6.8). Note that it simultaneously achieves a reduction in root mean squared error and an increase in forecast variability (*stdv_fc*). A future study should test whether these differences are statistically significant.

Neighbourhood distortion requires a large amount of additional computing time. The

slight increase in forecast quality does not justify the increased effort so that the standard version without neighbourhood distortion should still be preferred in an operational setting.

6.6 Wavelet smoothing

This section uses objective verification to evaluate the results of linear and nonlinear wavelet smoothing (Chapter 5). Wavelet smoothing is used to estimate the expected value in this study. Thus, this section exclusively deals with the quality of categorical forecasts (cf. Appendix C.2).

6.6.1 Overview

Wavelet smoothing is another way to estimate the expected value. Several variants of wavelet smoothing are developed which use different criteria in the separation of signal from noise (Chapter 5). Scale-dependent wavelet smoothing is a linear procedure and uses the level criterion (Section 5.3). *Wavelet thresholding* is a nonlinear procedure and uses the absolute value criterion (Section 5.4). The additional field criterion (Section 5.5) accounts for correlations between different variables and is used as an augmentation to wavelet thresholding of the temperature field. All versions use shift averaging so as to attain more robustness in the estimate of expected value (Subsection 5.3.3).

Scale-dependent wavelet smoothing with shift averaging is equivalent to kernel smoothing (cf. Section 5.3.3). Wavelet thresholding is a step beyond kernel smoothing; the amount of smoothing is not only spatially variable, but also adapts to the data to be smoothed.

In this section, objective verification is used to evaluate and compare the resulting estimates of the different wavelet smoothing procedures to other post-processing procedures. Emphasis is laid on the forecast of 2m-temperature, because the case study in Chapter 5 already indicates that wavelet smoothing is more promising when applied to 2m-temperature instead of precipitation.

Tab. 6.9 lists the usual measures of forecast quality (cf. Appendix C.2) for the temperature forecast. Simple kernel smoothing and different versions of wavelet smoothing are applied, respectively. The results are discussed in the following subsections.

In terms of precipitation, objective verification merely shows that the precipitation field does not benefit from wavelet thresholding and that scale-dependent wavelet smoothing leads to the same effect than simple spatial smoothing.

Table 6.9: *Forecast quality measures of the expected value of 2m-temperature. Comparison of rectangular kernel smoothing and various wavelet smoothing procedures.*

Measure	Kernel K_{rect} (Eq. 5.2) $h_x = h_y = 2$	Level Criterion $l_{\text{cutlevel}} = -2$	Wavelet Thresholding	Wavelet Thresholding with Additional Field Criterion
rmse	1.93 K	1.93 K	1.89 K	1.88 K
mae	1.42 K	1.42 K	1.39 K	1.39 K
bias	-0.12 K	-0.10 K	-0.06 K	-0.07 K
stdv_fc	3.76 K	3.79 K	3.86 K	3.84 K
corr	0.88	0.88	0.89	0.89
skill_DMO	-0.005	-0.005	0.015	0.017

6.6.2 Scale-dependent wavelet smoothing of 2m-temperature

Scale-dependent smoothing with the wavelet level criterion is compared to two low benchmarks: the direct model output and simple kernel smoothing with the rectangular kernel (Eq. 5.2).

As expected from theory (cf. Section 5.3.3), the level criterion attains the same degree of forecast quality as the rectangular kernel (Tab. 6.9). In comparison to kernel smoothing, the only advantage of linear wavelet smoothing is the increase in computational efficiency.

In comparison to the direct model output (Tab. 6.4), linear wavelet smoothing and simple kernel smoothing attain slightly poorer forecast quality. This is due to the neglect of explanatory variables and stresses the enormous relevance of their consideration.

6.6.3 Wavelet thresholding of 2m-temperature

Comparison to scale-dependent wavelet smoothing and to the DMO

The visual inspection of the case study (Chapter 5) indicates that wavelet thresholding yields better forecast fields than the level criterion (Fig. 5.15 and 5.12). This impression is now confirmed by objective verification.

Compared to kernel smoothing and linear wavelet smoothing, wavelet thresholding attains a reduction in root mean squared error (Tab. 6.9). This is even more valuable when considering that wavelet thresholding causes a much smaller amount of spatial smoothing than kernel smoothing (see *stdv_fc*). Moreover, the bias and the mean absolute error are reduced as well. The correlation between forecast and observations is also slightly improved.

This is a great achievement when recalling the simplicity of wavelet thresholding. The

computational cost is very low and the method is exclusively based on statistical considerations. Recall that no physical considerations enter the method and there is no explicit consideration of explanatory variables. Note, however, that the neglect of explanatory variables prevents wavelet thresholding from a large forecast improvement compared to the direct model output (Tab. 6.4 and Tab. 6.9). This lack of physical guidance becomes even more obvious in the next paragraph which compares wavelet thresholding to the neighbourhood method.

Comparison to the neighbourhood method

The comparison between wavelet thresholding and the neighbourhood method allows for a direct comparison between statistical guidance and physical guidance. Both methods allow for a spatially variable amount of local smoothing. In neighbourhood smoothing physical considerations guide the amount of local smoothing, whereas in wavelet thresholding statistical properties of the data itself determine the amount of local smoothing.

A comparison of wavelet thresholding with the standard neighbourhood method is not straightforward, because it is difficult to choose a neighbourhood size so that the average amount of smoothing corresponds to that of wavelet thresholding. Further, the neighbourhood is a three-dimensional smoothing procedure, whereas wavelet thresholding is a two-dimensional smoothing procedure. Nevertheless, a rough comparison should be possible. Among the three neighbourhood sizes it is the small size that fits the average amount of smoothing in wavelet thresholding best.

Results from the small neighbourhood with consideration of explanatory variables (cf. Tab. 6.7) clearly outperform the results of wavelet thresholding (Tab. 6.9). This indicates that physical considerations are more valuable in temperature post-processing than statistical guidance only. The explicit consideration of explanatory variables is a key mechanism when post-processing 2m-temperature. Note, however, that it involves a much larger computational cost when compared to the purely statistical concept of wavelet thresholding.

Although wavelet thresholding is currently based on statistical guidance only, it is quite possible to additionally include physical considerations into the existing procedure. For example, wavelet thresholding could include the field of orographic elevation by the additional field criterion (Eq. 5.4) so that there is less smoothing in regions of complex orography. Such a combination of physical considerations and statistical guidance may yield best results. Wavelet thresholding still bears a potential for another great leap in forecast quality.

Table 6.10: *Forecast quality measures of the expected value of 2m-temperature. Similarly to Tab. 6.9, various wavelet smoothing procedures are applied, but shift averaging is omitted here.*

Measure	Level Criterion $l_{\text{cutlevel}} = -2$	Wavelet Thresholding	Wavelet Thresholding with Additional Field Criterion
rmse	1.97 K	1.96 K	1.95 K
mae	1.44 K	1.44 K	1.44 K
bias	-0.10 K	-0.07 K	-0.06 K
stdv_fc	3.82 K	3.90 K	3.87 K
corr	0.88	0.88	0.88
skill_DMO	-0.026	-0.020	-0.020

6.6.4 Refinements of wavelet thresholding

The additional field criterion

Wavelet thresholding of 2m-temperature is augmented by the additional field criterion in order to account for the interrelation between different variables (cf. Section 5.5). The additional field is the precipitation forecast.

Objective verification shows that the additional field criterion has very little effect on forecast quality (Tab. 6.9). The increased effort is not rewarded, at least not when precipitation is chosen as the additional field.

As mentioned in the previous subsection, the inclusion of orographic elevation may be more beneficial. As the field of orography is static, the additional field criterion then becomes equivalent to the location criterion.

Shift averaging

Shift averaging is introduced so as to attain more robustness in the estimate of expected value (Subsection 5.3.3). Objective verification strongly confirms the decision to include shift averaging in all variants of wavelet smoothing.

The beneficial effect of shift averaging on objective verification becomes obvious when comparing Tab. 6.10 with Tab. 6.9. Shift averaging improves the root mean squared error, the mean absolute error and the correlation between forecasts and observations. Forecast bias is not affected. Surprisingly, shift averaging increases the degree of spatial smoothing only to a small extent, as forecast variability (*stdv_fc*) is only slightly reduced when compared to the smoothed forecast that does without shift averaging (Tab. 6.10). Therefore, the reduction in root mean squared error does not simply originate from an increased spatial smoothness, but indicates a true enhancement of the forecast. In fact,

the application of shift averaging is absolutely essential in terms of wavelet thresholding. Only in combination with shift averaging, the wavelet thresholding procedure attains an improvement of the direct model output (see *skill_ DMO* in Tab. 6.9 and 6.10).

6.7 Experimental ensemble with stochastic parametrisation

Overview

In addition to the post-processed forecasts, a stochastic parametrisation scheme is implemented into the LM and experimental ensemble forecasts are produced for the verification period July 10–24, 2002. The ensemble members are exclusively perturbed by the stochastic parametrisation scheme. Section 3.4 describes the specific configuration of the ensemble experiments and the input noise.

The experimental ensemble is partly motivated by the desire to create another benchmark forecast for the development and verification of the post-processing methods (cf. Chapters 1 and 3). The outcome of the ensemble was projected to be an appropriate guidance for LM post-processing, because the ensemble and the post-processing presumably address a similar portion of total forecast error.

In this section, quality measures of the LM ensemble forecast are presented. The ensemble results are juxtaposed with the quality of the raw model output of the unperturbed simulation, kernel smoothing, wavelet smoothing and the standard neighbourhood method. Estimates of the expected value and of the exceedance probability of precipitation are investigated. The ensemble estimates are derived by assigning equal weights to each of the 10 ensemble members. The ensemble mean is an estimator of the expected value; the ensemble mean of the indicator data $\tilde{\delta}^c(z)$ (Eq. 4.4) is an estimator of the exceedance probability.

As the experimental ensemble does not aim at the representation of overall forecast uncertainty, it is likely that the comparison between ensemble forecasts and observations makes the forecast look underdispersive. Nevertheless, the ensemble is expected to increase the quality of a single simulation to a certain degree.

Ensemble mean of 2m-temperature

Tab. 6.11 shows the verification measures for the ensemble mean of 2m-temperature. Several other estimates of the expected value are added for comparison and listed in the order of their computational cost. The results of the ensemble mean are almost identical to those of the deterministic LM forecast, i. e. the unperturbed simulation. Compared to the effect of post-processing, the effect of the stochastic parametrisation scheme is

Table 6.11: *Forecast quality measures of the expected value of 2m-temperature. Comparison of the original forecast, kernel smoothing, wavelet smoothing, the standard neighbourhood method and the ensemble mean. Recall that kernel smoothing with kernel K_{ball} is equivalent to neighbourhood post-processing without explanatory variables.*

Measure	Original LM Forecast	Kernel K_{ball} (Eq. 4.10) (Small Size)	Wavelet Thresholding	Standard Neighbourh. (Small Size)	Ensemble Mean
rmse	1.92 K	1.91 K	1.89 K	1.82 K	1.91 K
mae	1.40 K	1.41 K	1.39 K	1.35 K	1.40 K
bias	-0.01 K	-0.11 K	-0.06 K	-0.03 K	-0.01 K
stdv_fc	4.04 K	3.72 K	3.86 K	3.92 K	4.03 K
corr	0.89	0.89	0.89	0.90	0.89
skill_DMO	0.000	0.006	0.015	0.048	0.003

negligible. This is in accordance with the findings of Chapter 3 which already stated that the effect of stochastic parametrisation on the forecast is surprisingly small.

Ensemble mean of precipitation

The ensemble mean of precipitation is assessed by forecast quality measures from the contingency table (cf. Section 6.3.1 and Appendix C). The results (not shown) imply again that there is no discernible difference between the quality of the ensemble mean and the quality of the deterministic LM forecast.

Relative value (Subsection 6.4.2 and Appendix C) of the ensemble mean is also assessed and compared to other estimates of the expected value (not shown). It turns out that forecast value of the ensemble mean and forecast value of the unperturbed simulation are identical. While the neighbourhood estimate of the expected value improves the raw model output to a very small extent (cf. Fig. 6.11), the ensemble mean does not lead to any improvement.

The only difference between the ensemble mean and the unperturbed simulation consists in a slightly larger number of light precipitation events (not shown). This originates from a small degree of spatial smoothing which is caused by ensemble averaging. Spatial smoothing of the unperturbed simulation becomes also visible in the ensemble mean of the 2m-temperature forecast (see *stdv_fc* in Tab. 6.11). Many other ensemble forecasting systems exhibit this behaviour, too (e. g. Eckel and Walters, 1998; Hamill and Colucci, 1998).

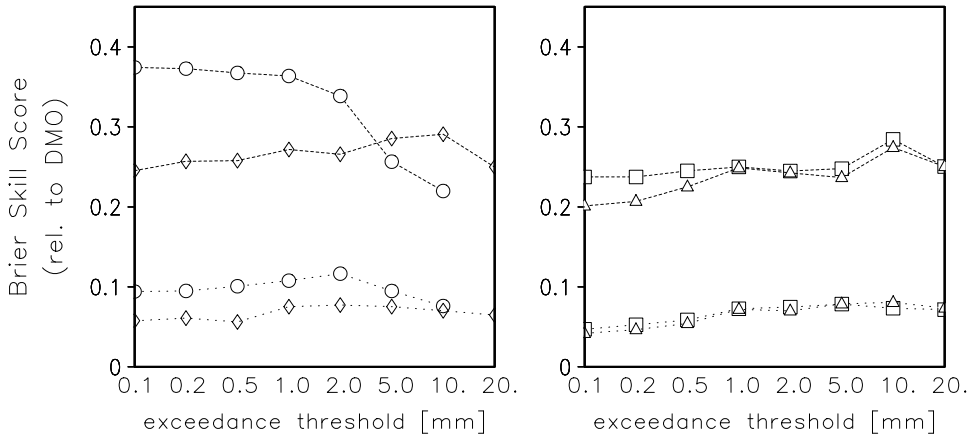

Figure 6.18: *Same as Fig. 6.5, but for the ensemble forecast and neighbourhood post-processing: exceedance probability from the standard neighbourhood method/medium size (dashed line); exceedance probability from the ensemble forecast (dotted line).*

Exceedance probability of precipitation

The exceedance probability from the ensemble forecast is assessed by the Brier skill score (Section 6.3.4 and Appendix C.3.1). The skill score uses the deterministic simulation as a reference forecast. Fig. 6.18 shows that the ensemble forecast still possesses some positive skill, but the amount of skill is very small compared to that attained by standard neighbourhood post-processing. One might speculate on a similar amount of skill when simply complementing the deterministic LM forecast by a climatological spread. The generally poor Brier scores of the LM ensemble result from an extreme underdispersion. Although underdispersiveness is expected to some extent, the spread is still surprisingly poor.

An investigation of forecast value leads to similar results as the investigation of forecast quality. The ensemble forecast attains a slightly larger relative value than the single simulation, but standard neighbourhood post-processing outperforms ensemble forecasting by far (Fig. 6.19).

6.8 Summary and outlook

Objective verification supplements the subjective investigations of the previous chapters. Verification measures are estimated from a data base of forecasts and corresponding observations. The data cover a summer period of two weeks. The various forecast systems are investigated in terms of forecast quality, consistency and value.

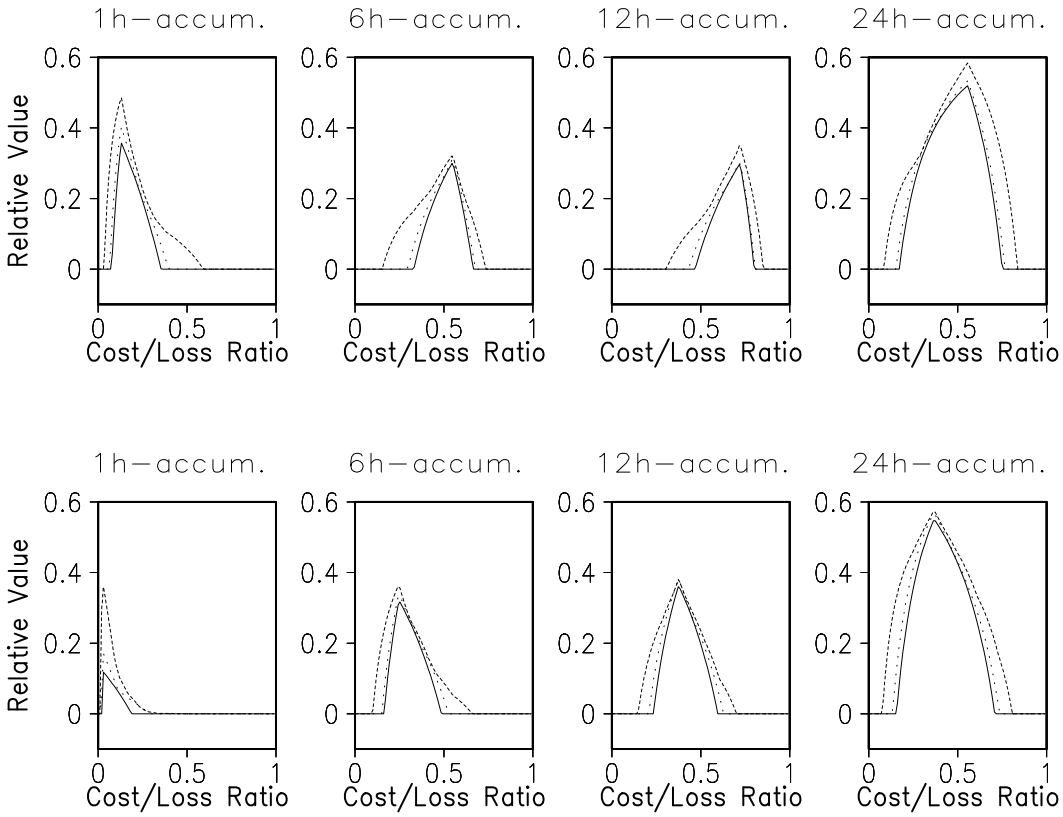

Figure 6.19: *Same as Fig. 6.11, but for the original forecast (solid line), the exceedance probability from standard neighbourhood post-processing/medium size (dashed line) and the exceedance probability from the ensemble forecast (dotted line).*

Forecast goodness of the neighbourhood methodology

Neighbourhood post-processing derives estimates of the following quantities: the expected value of temperature and precipitation, quantiles of temperature and precipitation and exceedance probabilities of precipitation. The standard neighbourhood version (Subsection 4.4.4) with a *medium* size (Tab. 4.1) is investigated.

Benefits of the post-processed forecast

Neighbourhood post-processing is able to improve the direct model output in terms of all three aspects of forecast goodness: quality, consistency and value. However, the improvements are only substantial if the user has the possibility to choose an appropriate forecast according to her demands of expected false alarm rates and hit rates. In other words, only if the user receives a probability or a quantile forecast, i. e. a range of forecasts that somehow characterises the probability function of the forecast, the user has the possibility to take full advantage of the neighbourhood method. Then the user can base her decisions on the forecast according to her individual needs and the role of the forecaster and of the decision maker can be clearly separated.

This result stresses the need for a concerted effort to educate users so that the acceptance of probability forecasts is increased and so that users are able to integrate the weather forecast into their individual decision making process.

Small improvements are also attained in the categorical forecast, namely in the expected value of 2m-temperature. The improvement originates from a reduction in random error and is accompanied by a slight increase in bias. Such a trade-off between bias and variance error is in accordance with theoretical expectations.

Necessary reduction in forecast sharpness

The improvement of the direct model output is accompanied by a reduction in forecast sharpness. The direct model output possesses good sharpness, because it contains very precise information. However, such preciseness exceeds the amount of information which is deterministically predictable. As a consequence, the direct model output exhibits poor reliability and resolution. A reduction in forecast sharpness is necessary in order to regain good forecast quality.

Shortcomings of the post-processed forecast

When compared to a perfect forecast, verification scores indicate two shortcomings: forecast bias and underdispersiveness. Forecast bias is a problem of the precipitation forecast and does not originate from the post-processing procedure. It is already present in the direct model output and is simply transmitted to the post-processed forecast.

Underdispersiveness may result from the fact that neighbourhood post-processing exclusively focusses on small-scale forecast uncertainty whereas the comparison to observational data makes the overall forecast error visible. In addition, the neighbourhood method assumes that small-scale error is exclusively associated with uncertainty in timing and location. In reality, this is only a portion of the total error.

A new verification strategy for the 90%-quantile

A new verification strategy is developed with respect to the assessment of the 90%-quantile. In order to assess forecast value, the standard decision making problem is reformulated. In addition, another low benchmark is defined: the *inflated* DMO.

When looking at 1h-accumulations of precipitation, relative value indicates that the neighbourhood estimate of the 90%-quantile is superior to the inflated DMO. This result must be interpreted with caution. In case of the 90%-quantile, negative effects of forecast bias and underdispersiveness cancel each other out. If only one of the two shortcomings is partly removed, verification measures may still record a deterioration of forecast goodness in the 90%-quantile.

Forecast calibration: A potential remedy?

Underdispersiveness and forecast bias are the main factors that deteriorate forecast reliability. As a supplementary calibration procedure may be able to improve the reliability of a forecast, calibration of the forecast appears promising. However, the setup of a calibration procedure would require a historical data set of model simulations which is not available at the moment.

Optimal configuration of the neighbourhood methodology

The neighbourhood methodology contains several free parameters such as neighbourhood size or the consideration of explanatory variables. In order to decide on an operational configuration of the neighbourhood method, these parameters need to be optimised in some way.

Neighbourhood size

Neighbourhood size is probably the key parameter in the post-processing procedure. Although theoretical considerations point to the existence of an optimal neighbourhood size, its identification turns out to be difficult.

In terms of 2m-temperature, objective verification suggests an optimal neighbourhood size which possesses a diameter of 40–85 km in space and a diameter of 3–4 h in time.

However, this result is not entirely trustworthy. Further investigations need to include a larger number of different options and need to take the diurnal cycle into account.

In terms of precipitation, best verification results are attained by the largest neighbourhood size among those tested. A point of maximum forecast quality or a point of saturation does not come into sight. The results may be blurred by the general problem of underdispersiveness so that any increase in spread may have a beneficial effect on verification measures, irrespective of its true appropriateness.

Consideration of explanatory variables

Subsection 4.4.4 suggests a standard version of considering explanatory variables on the basis of subjective investigations. Objective verification reaffirms its optimality. Optimality is not solely attained through best quality of the post-processing forecasts, but through a best trade-off between the complexity of the post-processing procedure and the quality of the resulting forecasts.

When post-processing 2m-temperature, the linear modification due to orographic elevation is indispensable. An additional neighbourhood distortion is set aside, because the improvement is too small when compared to the associated increase in computational cost.

In terms of precipitation post-processing, the consideration of orographic elevation does not enhance forecast quality at all. This does not imply that a proper treatment of orography were unnecessary. It rather implies that the current methods are still insufficient and that even more complex methods are required than those tested in this study.

Wavelet smoothing of 2m-temperature

Objective verification of wavelet smoothing reveals most interesting results when looking at the expected value of 2m-temperature. Estimates from level-dependent wavelet smoothing and wavelet thresholding are compared to neighbourhood estimates and to the direct model output. In this comparison, the influence of smoothing intensity is excluded to the greatest possible extent. Level-dependent wavelet smoothing, simple kernel smoothing and the neighbourhood method without consideration of explanatory variables are theoretically equivalent. The investigation results in the following ranking of forecast quality, starting with the best result:

1. standard neighbourhood post-processing with explanatory variables
2. wavelet thresholding (nonlinear smoothing)
3. direct model output
4. level-dependent wavelet smoothing (linear smoothing)

Linear wavelet smoothing is not able to improve the direct model output. This is due to the neglect of explanatory variables and demonstrates that the consideration of explanatory variables is very relevant when post-processing 2m-temperature. With this fact in mind, the success of wavelet thresholding is all the more resounding, as wavelet thresholding attains improvements of the raw model output *without* an explicit physical consideration of explanatory variables. The use of a translation-invariant wavelet transform is absolutely essential here.

The superiority of the neighbourhood method shows that physical considerations are more valuable in temperature post-processing than statistical guidance only. However, the inclusion of explanatory variables into wavelet thresholding is still possible, for example via the additional field criterion. Therefore, wavelet thresholding is a very promising post-processing technique, because it may attain another leap in forecast quality if explanatory variables are included successfully.

It should be noted that the differences between results from the neighbourhood method, wavelet thresholding and the direct model output are very small. The current conclusions are preliminary and need further confirmation by tests of statistical significance and a larger data set.

Ensemble experiments with stochastic parametrisation

The investigations in Chapter 3 state that the ensemble concept of this study is not mature enough to serve as a guidance in the development of LM post-processing until the specific configuration of input perturbations is justified. Objective verification presents an additional reason. Forecast quality of the experimental ensemble is not considerably better than the quality of a single simulation. Moreover, it is even far below the quality of the post-processed forecast.

The poor verification results of the stochastic parametrisation scheme either indicate that the stochastic effect of sub-grid scale processes is not a key player in small-scale predictability or indicate that the ensemble fails to simulate the effect realistically. This result confirms the urgent need for more research in ensemble forecasting and in the simulation of model uncertainties. Pertinent issues for future investigations are pointed out in Chapter 3.

Potential enhancements of the verification procedure

Observational data

As the verification period only comprises a summer time period of two weeks, the verification procedure covers a limited range of weather regimes and the results are not representative of a general case. In addition, the limited number of observations raises

the question of statistical significance, especially since the various forecast systems yield verification measures that differ only marginally when compared to the natural variability in forecast measures. This study examines the statistical significance only sporadically. Additional confidence arises from the fact that the findings of this study generally agree with those derived from the DWD verification periods whenever a direct comparison is possible. Nevertheless, there is further need for a comprehensive investigation of statistical significance.

Furthermore, using observational data as comparative data may result in misleading verification results. The comparison to observational data implicitly addresses the whole spectrum of forecast errors while the post-processing method is designed to account for small-scale random error only. An instance of such misleading results may be the seemingly never-ending improvement in Brier score with increasing neighbourhood size (Section 6.5.1). It would be desirable to establish a comparative data base that addresses the same part of the overall forecast uncertainty as the post-processing method does. The experimental ensemble forecast was envisaged to provide such a data base (Chapter 1). However, it turns out that further research is needed until the ensemble may serve as a useful guidance in the evaluation of the post-processing methods.

Variety of verification measures

As a specific verification measure reflects only a certain aspect of the joint distribution of forecasts and observations, it is desirable to apply a large variety of forecast measures. Although this study already looks into a very large number of common measures, it is not guaranteed that all information affecting predictability are already captured.

Most obviously, resolution of the quantile forecasts is not assessed due to a lack of a corresponding forecast measure. Moreover, it should be tested whether neighbourhood post-processing is able to correctly reproduce day-to-day variations in predictability and whether it can correctly distinguish forecasts with small uncertainty from forecasts with large uncertainty. This issue is very interesting, because even for operational ensemble prediction systems the ability to forecast the forecast skill is still a challenge (Hamill and Colucci, 1998).

Matching pairs of forecasts and observations

Another important issue is the appropriate matching of forecasts and observations. The current approach of choosing the nearest grid point is probably the simplest matching procedure among those currently known. Although one could argue that the nearest grid point approach represents the user's perspective best, the approach can still be improved.

Comparing point observations with model grid boxes is a general problem in forecast verification, especially when dealing with precipitation which is a highly intermittent

variable. The spatial representativeness of the observation is unknown and may not correspond to the model grid box. More elaborate procedures should take care of this problem, for example by estimating the spatial covariance of precipitation observations (see Chilès and Delfiner, 1999). Some existing methods apply bilinear interpolation or directly take field properties into account, e. g. by scale decomposition (Casati et al., 2004) or pattern matching (Ebert and McBride, 2000).

Last but not least, more attention should be paid to the fact that an observation is only one realisation of the true state of the atmosphere. Forecast verification should treat observational data in a probabilistic way as well.

7 Summary and conclusions

Overview

In order to take full advantage of short-range forecasts from deterministic high-resolution NWP models, the direct model output must be addressed in a probabilistic framework. A promising approach is mesoscale ensemble prediction. However, its operational use is still hampered by conceptual deficiencies and large computational costs.

This study tackles two relevant issues: (1) the representation of model related forecast uncertainty in mesoscale ensemble prediction systems and (2) the development of statistical post-processing procedures that retrieve additional probabilistic information from a single model simulation.

Representation of model related forecast uncertainty The insufficient representation of model related forecast uncertainty is one of the main problems in mesoscale ensemble forecasting. Theoretical arguments speak in favour of stochastic parametrisation, but only few studies have carried out experiments with stochastic parametrisation schemes in complex weather prediction models. By implementing stochasticity in an operational high-resolution model, this study constitutes a novel contribution to current research in ensemble forecasting and provides a basis for future research in this area.

Statistical post-processing of a single simulation Statistical post-processing fills the current gap between the need for probabilistic forecasting and its operational constraints. The study designs two schemes: the *neighbourhood method* and *wavelet smoothing*. The schemes are intended to derive probabilistic information in computing environments that currently do not allow for sufficient ensemble simulations. Both approaches fall under the framework of estimating a large array of statistical parameters on the basis of a single realisation on each parameter. In this study, the statistical post-processing schemes derive probabilistic information from a single simulation, but the schemes may also be used as a supplement to an existing ensemble. Compared to ensemble forecasting, the statistical post-processing schemes require a very small amount of computing time and are suitable for immediate implementation. The schemes do not rely on historical error statistics so that they are ready to use if such data are not at hand. Moreover, the source code of the schemes satisfies operational requirements at the German Weather Service, i. e. it is portable, it is tailored to the respective data formats, it includes error-handling schemes and it is thoroughly documented.

Special emphasis is laid on mesoscale forecast uncertainty of summer precipitation and 2m-temperature in Europe. The choice is motivated by the desire to address forecast variables that are especially relevant to a standard user in Germany. Source of forecast guidance is the deterministic high-resolution model Lokal-Modell (LM) of the German Weather Service.

Conclusions of this study are based on subjective assessment of a case study (10 July 2002) and on objective verification of 15 short-range forecasts. Objective verification condenses forecast goodness into a few verification measures which are derived from statistical comparison between forecasts and observations. The observational data encompasses the SYNOP data of Germany during a time period of two weeks. Forecast goodness is examined in terms of quality, consistency and value.

Experimental ensembles with stochastic parametrisation

A stochastic parametrisation scheme is implemented into the LM in an attempt to simulate the stochastic effect of sub-grid scale processes on the resolved scales. The scheme is similar to the operational scheme at ECMWF, but contains several upgrades such as an energetically consistent perturbation of the turbulence scheme, a perturbation of roughness length and a correction for energetic inconsistency. In order to reveal the effect of the stochastic parametrisation scheme on the forecast, experimental ensemble simulations are produced. The results are investigated by eye-ball inspection, statistical analysis and objective verification.

This study, being one of the first to systematically analyse the benefits of stochastic parametrisation in operational short-range forecasting, clearly points out previously undefined issues and therefore provides the basis for essential future research. Subjective eye-ball inspection of case studies shows that the stochastic parametrisation scheme has a substantial effect on precipitation amount. However, objective verification demonstrates that the ensemble does not attain better forecast goodness than a single LM simulation. Current stochastic parametrisation schemes do not yet benefit short-range high-resolution ensemble simulations in the mid-latitudes. Either the stochastic effect of sub-grid scale processes is not a key player in model related uncertainty or the scheme does not yield a realistic simulation.

These results point the way to relevant future research. Stochastic parametrisation may produce more credible results and provide better insight into small-scale predictability if the following issues are tackled. First, the specific configuration of the input perturbations needs better justification. Its choice is still somewhat arbitrary, but sensitivity studies of this study demonstrate that they have an enormous influence on the ensemble outcome. Secondly, the numerical integration scheme needs reconsideration. Current numerical schemes are not properly sanctioned for use in combination with stochastic perturbation and may cause errors that generally do not become obvious by eye-ball inspection.

Neighbourhood method

The neighbourhood method is based on the notion of small-scale spatio-temporal ergodicity including explicit corrections for so-called *explanatory variables* which are suspected to enhance small-scale predictability. It looks in the spatio-temporal neighbourhood of a point to get a set of forecasts and creates a *surrogate ensemble* for the forecast at the central point of the neighbourhood. While the idea of a neighbourhood has already started to emerge in the context of forecast verification, this study is one of the first to apply the concept to the derivation of additional forecast products.

The neighbourhood method derives estimates of quantiles, exceedance probabilities and expected values at each grid point of the LM. While the estimation of exceedance probabilities and expected value simply applies the respective maximum-likelihood estimator, quantile estimation involves the concepts of a ranking formula and of kernel quantile regression so as to preclude a large variance error in the estimate. Technically, the neighbourhood method satisfies requirements of operational use so that it is ready for immediate implementation into the operational forecasting system of DWD.

Although the neighbourhood concept bears similarities to geostatistics, it is still unique, because there is no geostatistical approach which combines the estimation of expected value, quantiles and exceedance probabilities into a single concept. Especially the problem of quantile estimation requires a new approach, because it is not covered by standard geostatistical methods. A consultation of the geostatistical perspective turns out very useful, as it helps to clarify the underlying statistical assumptions of the neighbourhood and to interpret errors in the resulting estimates.

This study does not only develop and apply a novel post-processing concept, but also evaluates the goodness of the resulting forecasts in a systematic way. Objective verification demonstrates that neighbourhood post-processing is able to improve the direct model output. Improvements in forecast quality are associated with a reduction of detailed forecast information on the small scales. The reduction is necessary, because the direct model output contains very precise forecast information which exceeds the amount that is deterministically predictable. Such information contributes to the noise level and deteriorates forecast quality. Verification measures show that post-processing reduces the preciseness and regains good forecast quality.

The improvement in forecast goodness varies with the statistical quantity and the forecast variable under consideration. Improvements are only substantial if the post-processed forecast is formulated as an exceedance probability or as a quantile. The estimates of exceedance probabilities and quantiles clearly outperform the direct model output in terms of quality, consistency and value.

When compared to a perfect forecast, some deficiencies become visible in the estimates of exceedance probabilities and quantiles. The precipitation forecasts exhibit a forecast bias. The bias does not originate from the post-processing procedure, but is already present in the direct model output and is simply transmitted to the post-processed

forecast. Furthermore, all estimates are underdispersive, i. e. they capture a certain aspect of forecast uncertainty which is not representative of the full range.

If the post-processed forecast is formulated as the expected value, the neighbourhood method gives less advantage. The expected value of 2m-temperature attains a small but consistent improvement in forecast quality. The improvement originates from a reduction in random error and is accompanied by a slight increase in bias. This trade-off is in accordance with theoretical expectations from geostatistics. The expected value of precipitation does not achieve an overall improvement; improvements depend on precipitation amount and the user's ability to tolerate low hit rates or high false alarm rates.

The main advantage of the post-processed forecast is its ability to convert the deterministic model output into probabilities or quantiles. Their benefit consists in the variety of forecasts for different users and their different decision making processes. The expected value is at a disadvantage with probabilities and quantiles, because it is a categorical forecast which is "tailored" to only one specific type of user. This result stresses the need for a concerted effort to educate users so that the acceptance of probability forecasts is increased and so that users are able to integrate the weather forecast into their individual decision making process.

This study points to issues of future research that may lead to a further increase of the benefits already attained. Neighbourhood size is a key parameter of the post-processing procedure. Theoretical considerations indicate the existence of an optimal size, but the identification of this size needs further investigation. Furthermore, the inclusion of explanatory variables benefits the post-processed temperature forecast, but does not yet lead to a further enhancement of the precipitation forecast. The post-processed precipitation forecast may further gain in quality if future research succeeds in developing cost-effective methods that describe and capture the complex effect of explanatory variables on precipitation. Last but not least, the post-processed forecast may additionally profit from a supplementary calibration procedure. This plan may be tackled as soon as a historical data set of model simulations becomes available. Generally, best results may be obtained if the neighbourhood method is combined with other approaches to probabilistic forecasting such as ensemble prediction or probabilistic model output statistics.

Wavelet smoothing

The concept of wavelet smoothing originates from the field of *image denoising* and includes concepts of *multiresolution analysis* and *non-parametric regression*. The use of wavelets is considered attractive because of their connection to statistical multi-scale modeling, their ability to represent data in a parsimonious way and their noise-whitening property. In this study, the method is used to produce estimates of the expected value, but it may be easily extended to the additional estimation of exceedance probabilities.

Two main types of wavelet smoothing procedures are applied to LM output fields: *scale-dependent wavelet smoothing* and *wavelet thresholding*. In order to avoid artefacts in the post-processed field, the study augments the standard smoothing procedure by introducing the *translation-invariant* wavelet transform. Incidentally, it is shown that scale-dependent wavelet smoothing with the translation-invariant transform is equivalent to simple kernel smoothing. In the context of wavelet thresholding, the application of the translation-invariant wavelet transform proves to be absolutely essential.

Wavelet thresholding is especially interesting, because it is a *nonlinear* smoothing technique that automatically adapts the amount of spatial smoothing to local properties of the underlying data. This study is the first to apply wavelet thresholding to "real" meteorological forecast data and to derive an additional forecast product by doing so.

As opposed to the neighbourhood methodology, the current version of wavelet thresholding does not contain an explicit consideration of explanatory variables. Instead, it merely relies on a clever data representation via the wavelet transform and on the validity of statistical assumptions. In comparison to the neighbourhood methodology, wavelet thresholding is almost effortless, both in terms of its development and in terms of its computational costs. From a technical viewpoint, wavelet smoothing is even better suited for operational use than the neighbourhood methodology.

The study reveals that the success of wavelet thresholding largely depends on the statistical properties of the underlying data. The relevance of the underlying data is impressively demonstrated by two contrasting examples. Wavelet thresholding proves very useful for post-processing 2m-temperature, but is unsuitable for post-processing summer precipitation. Obviously, precipitation data do not satisfy the statistical assumptions of wavelet thresholding so that the procedure does not manage to filter those patterns that are deterministically predictable. The particular cause probably consists in the fact that the noise is heavy-tailed and that the signal-to-noise ratio is weak.

Wavelet thresholding is much more interesting in terms of 2m-temperature, because the smoothing procedure apparently preserves deterministically predictable temperature patterns and removes others. Verification shows that wavelet thresholding achieves a slight improvement of the raw model output even though explanatory variables are not explicitly considered. This is a resounding success when compared to the dramatic deterioration in forecast quality that results from simple kernel smoothing without explanatory variables.

In spite of this success, wavelet thresholding does not reach the same forecast quality as the neighbourhood method yet. The explicit consideration of explanatory variables appears indispensable when post-processing 2m-temperature. The study already presents a concept for a cost-effective inclusion of explanatory variables into the wavelet thresholding procedure. Additional research along this avenue has the potential to attain another leap in forecast quality which might be competitive with the quality achieved by neighbourhood post-processing. A change from neighbourhood post-processing to wavelet thresholding would be associated with an enormous reduction of complexity and com-

putational costs. Thus, this study has developed a novel approach to post-processing which is very promising.

Final remarks

In summary, the study investigates and advances pertinent issues in probabilistic forecasting. The investigations of this study are multi-faceted, delving both into theoretical and into application-oriented issues. The study develops and carefully evaluates pragmatic short-term solutions and also moves towards ambitious long-term solutions. The implementation and evaluation of the stochastic parametrisation scheme provides a basis for further progress in the ensemble approach. The development and extensive evaluation of novel post-processing techniques demonstrates that they are immediately able to fill the gap between the need for probabilistic forecasting and its operational constraints. Thus, the study fosters the sensible use of deterministic models in cases of limited predictability. All in all, the study adds another tessera to the mosaic of probabilistic prediction in weather and climate.

A Wavelets and the lifting scheme

In order to make this study self-contained, a short introduction to wavelets is given here. Section A.1 starts with some fundamentals of the wavelet transform. In particular, the difference between continuous and discrete wavelets is explained. Section A.2 continues with the definition of multiresolution analysis and states that there is a direct link to discrete wavelets if they are (bi-)orthogonal. Section A.3 generalises these explanations to the two-dimensional case. This is not trivial due to the introduction of a diagonal direction in addition to the vertical and horizontal. Section A.4 introduces the lifting scheme, the particular wavelet transform used in this study. Thomas Gerstner (Institute for Numerical Simulation, University of Bonn) kindly provided the corresponding software of the wavelet transform algorithm.

A.1 Fundamentals of the wavelet transform

A.1.1 Overview

The concept of wavelet transforms was formalized in early 1980's and has been extensively applied in a wide area of subjects since then, for example image compression, signal processing, time frequency analysis or accelerating numerical methods, to list only few.

Introductions to wavelet theory can be found in Chui (1992), Daubechies (1992) or Kumar and Foufoula-Georgiou (1994), for example. An overview of wavelet based multiresolution analyses is presented by Jawerth and Sweldens (1994). Introductory texts with regard to applications in geophysics are contained in Kumar and Foufoula-Georgiou (1993a) and Kumar and Foufoula-Georgiou (1997). A mathematical overview on wavelets in statistics is presented by Härdle et al. (1998) and corresponding texts for the non-specialist with emphasis on applications is presented by Ogden (1997) and Vidakovic (1999). The content of the following subsections is taken from this literature.

Wavelet applications take advantage of various nice properties inherent in wavelet transforms (Vidakovic, 1999). Wavelets generate local bases that provide more adaptive and parsimonious representations of real phenomena with complicated time and frequency behaviours; wavelets enable sub-band decomposition; wavelets dis-balance energy in data so that a signal can be well described by only a few energetic components; wavelets

whiten data so that time correlated signals become almost uncorrelated in the wavelet domain; wavelets detect self-similar phenomena by being self-similar themselves.

The choice of a wavelet is not unique. On the contrary, there is a great variety in commonly used wavelet functions, as different applications make different demands on the wavelet transform. It would not be reasonable to present an exhaustive list here. The wavelets used in this study are *discrete*, *bi-orthogonal* and *two-dimensional*.

In the following subsections, wavelets are formally introduced. In order to emphasize the difference between discrete and continuous wavelets, the continuous form is described first and then the discrete form is introduced.

A.1.2 Continuous form

The first definitions of wavelets are given in the Fourier domain: A wavelet ψ is a $L^2(\mathbb{R})$ function for which the Fourier transformation Υ satisfies

$$\int_0^\infty |\Upsilon(t\omega)|^2 \frac{dt}{t} = 1 \quad \text{for almost all } \omega.$$

A wavelet has (1) zero mean and (2) compact support, or sufficiently fast decay, to obtain localisation in space. It is because of these two properties that the function ψ is called a wavelet. The first property ensures that ψ has a wiggle, and the second ensures that it is not a sustaining wave.

A *wavelet transform* of a function $f(x) \in L^2(\mathbb{R})$ is defined as the integral transform

$$Wf(x) = \int_{-\infty}^{\infty} f(u)\overline{\psi}_{\mu,x}(u)du \quad \mu > 0 \tag{A.1}$$

with a family of functions

$$\psi_{\mu,x}(u) \equiv \frac{1}{\sqrt{\mu}}\psi\left(\frac{u-x}{\mu}\right)$$

where the functions $\psi_{\mu,x}(u)$ are wavelets and $\overline{\psi}_{\mu,x}(u)$ denotes their complex conjugates. The function f is mapped from one index x to two indices consisting of the scale parameter μ and the location parameter x. The wavelets are translates and dilates of the *mother wavelet* $\psi_{1,0}(u)$. Changing the value of μ has the effect of dilating or contracting the mother function and changing the value of x has the effect of changing the position of the mother function in space. The normalising constant $\frac{1}{\sqrt{\mu}}$ is chosen so that

$$\int \mid \psi_{\mu,x}(u) \mid^2 du = 1$$

for all scales μ.

The integral transform (Eq. A.1) can also be looked upon as convolution of $f(u)$ with the function $\overline{\psi}_{\mu,x}(u)$, and the wavelet transform can be looked upon as a sub-band decomposition or sub-band filtering operation. Whereas the Fourier transform allows exact

localisation in frequency and using the Dirac delta function allows exact localisation in space, the wavelet transform is a compromise between localisation in frequency and localisation in space. There is a trade-off between the degree of precise localisation in frequency and precise localisation in space, following Heisenberg's uncertainty principle. Wavelets with a relatively precise localisation in space generally have a small support, so that the localisation in frequency is less precise.

A.1.3 Discrete form and wavelet basis

In the *discrete* wavelet transform the parameters μ and x are discretised, whereas the *continuous* wavelet transform is characterised by continuous parameters μ and x. One can choose $\mu = \mu_0^l$ and $x = mx_0 \cdot \mu_0^l$ where l and m are integers, $x_0 > 0$ and $\mu_0 > 1$ are fixed translation and dilation steps, respectively. The indices l, m replace the indices μ, x of the wavelet $\psi_{\mu,x}$. Often $\mu_0 = 2$ and $t_0 = 1$ are chosen:

$$\psi_{l,m}(x) = 2^{l/2}\psi(2^l x - m) \quad l, m \in \mathbb{Z}.$$

Discrete wavelet transforms can be constructed so as to constitute a *basis* of $L^2(\mathbb{R})$. They give a *non-redundant* representation of a function f without loss of information. Non-redundant transforms are more efficient and parsimonious, whereas redundant representations are more robust to noise and can give representations close to the continuous case.

A.2 (Bi-)orthogonal wavelets and multiresolution analysis

If a discrete wavelet basis satisfies certain conditions, it qualifies as a *wavelet frame*, as *bi-orthogonal* or even *orthogonal*. Bi-orthogonal and orthogonal wavelet transforms offer an elegant key to *multiresolution analysis* (MRA). The framework of MRA allows for an especially parsimonious representation of data so that only few components achieve a good description of the signal.

A family of closed subspaces $V_l \subset L^2(\mathbb{R}), l \in \mathbb{Z}$ is called a *dyadic multiresolution analysis* (e. g. Madych, 1992; Jawerth and Sweldens, 1994) if

$$\forall l \in \mathbb{Z} : V_l \subset V_{l+1}, \tag{A.2}$$

$$\overline{\lim_{l\to\infty} V_l} = \overline{\bigcup_{l\in\mathbb{Z}} V_l} = L^2(\mathbb{R}), \tag{A.3}$$

$$\lim_{l\to-\infty} V_l = \bigcap_{l\in\mathbb{Z}} V_l = \{0\}, \tag{A.4}$$

$$f(x) \in V_l \Leftrightarrow f(2x) \in V_{l+1}, l \in \mathbb{Z} \quad \textit{(scale invariance)}, \tag{A.5}$$

$$f(x) \in V_0 \Leftrightarrow f(x+m) \in V_0, m \in \mathbb{Z} \quad \textit{(shift invariance)}, \tag{A.6}$$

$$\exists \phi(x) \in V_0 : \{\phi(x-m)\}_{m\in\mathbb{Z}} \quad \textit{is a} \text{ Riesz } \textit{basis for } V_0. \tag{A.7}$$

The MRA framework is neither unique nor arbitrary; several multiresolution frameworks can be constructed. Projecting a given function f into a subspace V_l is like using a telescope: every time we proceed to a finer scale $l+1$, we add more details to the approximation. Furthermore, we look with the same eyes at all scales (Eq. A.5) and in all directions (Eq. A.6).

The function $\phi(x)$ is commonly referred to as the *scaling function* and defines the original space V_0. The scaled versions V_l of the original space V_0 can be expressed by

$$V_l = \text{span}\{\phi_{l,m} | m \in \mathbb{Z}\}$$

where $\phi_{l,m} = 2^{l/2}\phi(2^l - m)$. The function ϕ is also called *father wavelet*.

In practice, the scaling function ϕ itself is generally not used for constructing the subspaces V_l or projecting a given function f into a subspace V_l. Instead, orthogonal or bi-orthogonal wavelet transforms are used, as they are also directly related to MRA. Every approximation space V_{l+1} can be written as the sum of a coarser approximation V_l and a detail space O_l given by the (bi-)orthogonal wavelets $\psi_{l,m}$ ($m \in \mathbb{Z}$).

$$V_{l+1} = V_l + O_l \quad \text{with} \quad O_l = \text{span}\{\psi_{l,m} | m \in \mathbb{Z}\}.$$

If orthogonal wavelets $\psi_{l,m}$ are chosen, the detail space O_l and the approximation space V_l are even orthogonal. However, under the orthogonal wavelet multiresolution framework the only real and compactly supported symmetric or antisymmetric wavelet is the Haar wavelet. The generalisation to bi-orthogonal wavelets has been considered to allow for more flexibility when smoother wavelets or wavelets with better frequency localisation are needed. Bi-orthogonal wavelets consist of two sets of wavelets, of which one set is used for decomposition and the other set is used for reconstruction of a function f (Cohen, 1992).

A.3 The two-dimensional counterpart

Now consider a two-dimensional function that is square integrable over the real *plane*: $f(x,y) \in L^2(\mathbb{R}^2)$. The simplest way of constructing a wavelet basis for $L^2(\mathbb{R}^2)$ is to take the simple product of two uni-dimensional wavelets:

$$\Psi_{l_1,l_2,m_1,m_2}(x,y) = \psi_{l_1,m_1}(x) \cdot \psi_{l_2,m_2}(y).$$

This approach does not retain the nice features of two-dimensional multiresolution analysis, though. A two-dimensional multiresolution analysis is defined through six properties that are analogue to those above, but the one-dimensional spaces $V_l \subset L^2(\mathbb{R})$ are replaced by two-dimensional subspaces $\boldsymbol{V}_l \subset L^2(\mathbb{R}^2)$:

$$\begin{aligned}
&\forall l \in \mathbb{Z} : \boldsymbol{V}_l \subset \boldsymbol{V}_{l+1}, \\
&\overline{\lim_{l\to\infty} \boldsymbol{V}_l} = \overline{\bigcup_{l\in\mathbb{Z}} \boldsymbol{V}_l} = L^2(\mathbb{R}^2), \\
&\lim_{l\to-\infty} \boldsymbol{V}_l = \bigcap_{l\in\mathbb{Z}} \boldsymbol{V}_l = \{0\}, \\
f(x,y) \in \boldsymbol{V}_l &\Leftrightarrow f(2x,2y) \in \boldsymbol{V}_{l+1}, l \in \mathbb{Z} \quad \textit{(scale invariance)}, \\
f(x,y) \in \boldsymbol{V}_0 &\Leftrightarrow f(x+m_1, y+m_2) \in \boldsymbol{V}_0, m_1, m_2 \in \mathbb{Z} \quad \textit{(shift invariance)}, \\
\exists \Phi(x,y) \in \boldsymbol{V}_0 &: \{\Phi(x-m_1, y-m_2)\}_{m_1,m_2\in\mathbb{Z}} \quad \textit{is a} \text{ Riesz } \textit{basis for } \boldsymbol{V}_0.
\end{aligned}$$

Note that the two-dimensional multiresolution analysis involves two translation indices m_1, m_2, but allows only a single dilation index l.

Two-dimensional wavelet functions are sought that fit into the two-dimensional multiresolution framework. The wavelet functions should span the two-dimensional detail spaces $\boldsymbol{O}_l$:

$$\boldsymbol{V}_{l+1} = \boldsymbol{V}_l + \boldsymbol{O}_l.$$

Let us extend the one-dimensional case to the two-dimensional case by defining the two-dimensional space $\boldsymbol{V}_0$ as the tensor product of two one-dimensional spaces:

$$\boldsymbol{V}_0 = V_0 \otimes V_0 = \overline{\text{span}\{f(x)g(y), f, g \in V_0\}}$$

and define $\boldsymbol{V}_l = V_l \otimes V_l$ for any l. Also, define the two-dimensional scaling function $\Phi(x,y) \equiv \phi(x)\phi(y)$. This set of subspaces inherits the two-dimensional versions of the multiresolution analysis properties. Breaking down the two-dimensional subspace $\boldsymbol{V}_{l+1}$ into its one-dimensional parts gives insight into the nature of the two-dimensional detail space $\boldsymbol{W}_j$:

$$\begin{aligned}
\boldsymbol{V}_{l+1} &= V_l \otimes V_l \\
&= (V_l + W_l) \otimes (V_l + W_l) \\
&= V_l \otimes V_l + \big((V_l \otimes W_l) + (W_l \otimes V_l) + (W_l \otimes W_l)\big) \\
&= \boldsymbol{V}_l + \boldsymbol{W}_l.
\end{aligned}$$

It can be seen that the detail space $\boldsymbol{W}_l$ is made up of three subspaces. Bases for these three sub-detail spaces are the corresponding tensor products of their components, so define the following two-dimensional wavelets:

$$\begin{aligned}
\Psi^h(x,y) &= \phi(x,y)\psi(x,y) \\
\Psi^v(x,y) &= \psi(x,y)\phi(x,y) \\
\Psi^d(x,y) &= \psi(x,y)\psi(x,y).
\end{aligned}$$

They fit the two-dimensional multiresolution framework by satisfying

$$\boldsymbol{W}_l = \text{span}\{\Psi^h_{l,m_1,m_2}(x,y), \Psi^v_{l,m_1,m_2}(x,y), \Psi^d_{l,m_1,m_2}(x,y) | m_1, m_2 \in \mathbb{Z}\}$$

where

$$\Psi^\times_{l,m_1,m_2}(x,y) = 2^l \Psi^\times(2^l x - m_1, 2^l y - m_2) \quad \text{and} \quad \times = h, v, d.$$

Note that the two-dimensional case exhibits several mother wavelets instead of only one. They correspond to specific spatial orientations. As already indicated by the upper indices, the wavelet Ψ^h corresponds to the horizontal direction, the wavelet Ψ^v corresponds to the vertical direction and the wavelet Ψ^d corresponds to the diagonal direction. The wavelet transform acts like spatially oriented frequency channels.

A.4 Constructing the wavelet transform: Lifting scheme

A.4.1 Introductory remarks and motivation

In this study, the *lifting scheme* (Sweldens, 1997) is chosen among the various techniques to construct wavelet bases. The original motivation for developing the lifting scheme in the 1990's was to build *second generation wavelets*, i. e. wavelets adapted to situations where Fourier techniques are no longer available, for example in non-Euclidean spaces. A construction using the lifting scheme is entirely spatial. It is a simple and fast scale-by-scale transform. The lifting scheme is not confined to constructing second generation wavelets; it also provides a framework for the construction of first generation wavelets, which are all translates and dilates of one or a few basic shapes. When putting the transformation algorithm in the wavelet framework, it turns out to correspond to the moment-preserving, bi-orthogonal CDF(2,2) wavelet transform of Cohen et al. (1992) which was originally constructed using Fourier arguments.

With regard to this study, the lifting scheme has the following advantages:

- The wavelets are directly defined on a spatial lattice, so discretisation is unnecessary.
- The original function can be perfectly reconstructed.
- There are no artefacts at the boundaries of a bounded domain.
- The wavelet transform is very efficient in terms of computing time and memory.

A.4.2 One-dimensional case

Sweldens (1996) briefly describes the one-dimensional version of the two-dimensional wavelet transform algorithm that is used in this study. Consider a function $f(x)$ defined

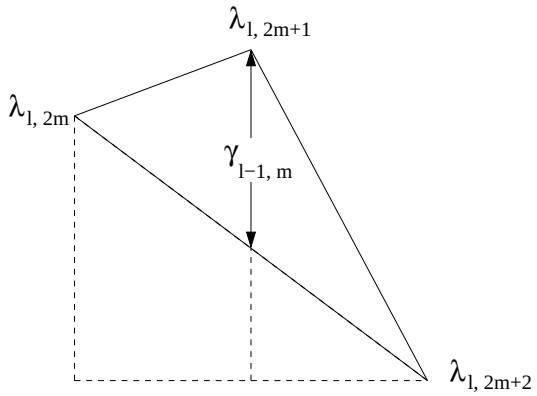

Figure A.1: *The wavelet coefficient is the failure to be linear (after Sweldens, 1996).*

on $2^n + 1$ equidistant grid points. Its values at the grid points $m = 0, \ldots, 2^n$ are denoted as $\lambda_{0,m} := f(m)$. The lifting scheme consists of three steps: *split*, *predict* and *update*.

Split The signal is split into two disjoint sets of samples. One set contains the values $\lambda_{0,m}$ at the odd indexed grid points m and the other set contains the values $\lambda_{0,m}$ at the even indexed grid points m and is defined as follows

$$\lambda_{-1,m} := \lambda_{0,2m} \quad \text{for} \quad m = 0, \ldots, 2^{n-1}.$$

Predict If the signal has a local correlation structure, then $\lambda_{0,2m+1}$ will be approximated by the mean of its neighbours $\lambda_{-1,m}$ and $\lambda_{-1,m+1}$. The difference between the original value $\lambda_{0,2m+1}$ and its approximation is contained in the *wavelet coefficient*

$$\gamma_{-1,m} := \lambda_{0,2m+1} - \frac{1}{2}(\lambda_{-1,m} + \lambda_{-1,m+1}) \quad \text{for} \quad m = 0, \ldots, 2^{n-1}.$$

The wavelet coefficient contains the information to *predict* the odd set using the even set. They measure to which degree the original function is a piecewise linear function (Fig. A.1). In terms of frequency channels the wavelet coefficients $\gamma_{-1,m}$ represent high frequencies, whereas $\{\lambda_{-1,m}\}$ represent low frequencies.

Update One could imagine to continue by iterating the two steps above: Starting from $\{\lambda_{-1,m}\}$ one could find the values $\{\lambda_{-2,m}\}$ and $\{\gamma_{-2,m}\}$ on the next level -2 and so on. The iteration terminates when the coarsest level $-n$ has been reached. However, this procedure introduces considerable aliasing. We would like the average of the coefficients $\{\lambda_{l,m}\}$ to be the same on each level: $\sum_m \lambda_{l-1,m} = \frac{1}{2}\sum_m \lambda_{l,m}$ for all $l = 0, \ldots, -n$. In order to ensure this, a third step is introduced: *update* or *lifting* the $\lambda_{-1,m}$ with the help of the wavelet coefficients $\gamma_{-1,m}$:

$$\tilde{\lambda}_{-1,m} = \lambda_{-1,m} + \frac{1}{4}(\gamma_{-1,m-1} + \gamma_{-1,m}).$$

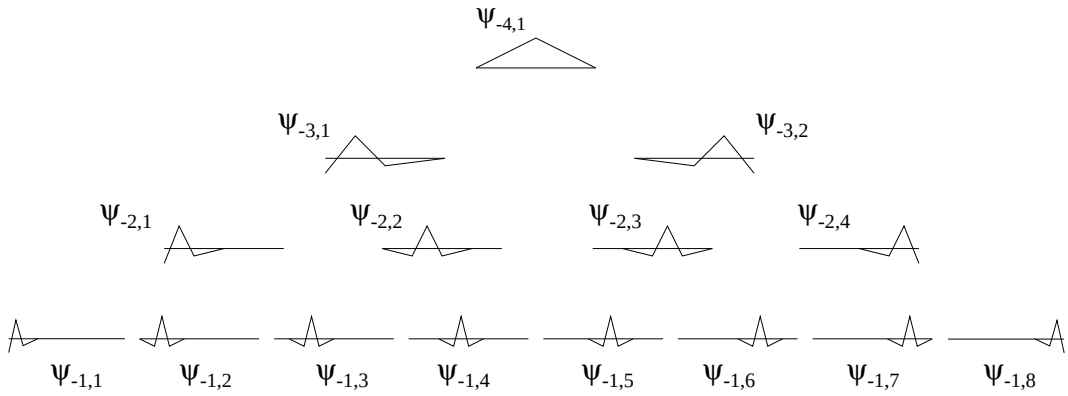

Figure A.2: *Bi-orthogonal wavelet basis constructed by the lifting scheme with* $n = 4$ *(after Gerstner, 1998).*

The inverse transform can immediately be found by inverting the operations above: (1) undo update by sign inversion, (2) undo predict by adding in the failure to be linear, (3) then merge the even and odd samples.

When the lifting scheme is put into the wavelet framework, the corresponding wavelet basis functions $\psi_{l,m}$ can be identified. Fig. A.2 shows the wavelet basis functions pertaining to a function that is defined on $2^4 + 1$ equidistant grid points. The interior wavelet functions $\phi_{l,m}$ $(l = -1, -2)$ $(m = 2, \ldots, 2^{n+l} - 1)$ are translates and dilates of one mother wavelet.

The boundary wavelets illustrate the boundary handling of the lifting scheme. As the function is defined on a bounded domain, the conventional update step sometimes interferes with a boundary. In such a case the update step is modified and yields modified wavelet functions near the boundary of the domain (Fig. A.2). The modification is constructed in way that preserves the first moment of the function in every approximation space.

A.4.3 Two-dimensional case

The transformation algorithm can be generalised to a two-dimensional function $f(m_1, m_2)$ defined on a finite number of equidistant grid points $\{0, \ldots, 2^n\} \times \{0, \ldots, 2^n\}$. While the location index m of the wavelet coefficients γ is extended to the horizontal location index m_1 and the vertical location index m_2, there is still only one level index l.

The three steps of the lifting scheme are carried out in a similar manner as in the one-dimensional case. The splitting step becomes more complex and in the prediction and update step the horizontal, vertical and diagonal directions are considered separately. In the following paragraph, the first iteration step is described.

Split The signal $\lambda_{0,m_1,m_2} \equiv f(m_1, m_2)$ is split into four disjoint sets of samples. The even set contains the values $\{\lambda_{0,2m_1,2m_2} | (m_1, m_2 = 0, \ldots, 2^{n-1}\}$, the odd set contains the values $\{\lambda_{0,2m_1+1,2m_2+1}\}$ and the two mixed sets contain the values $\{\lambda_{0,2m_1+1,m_2}\}$ and $\{\lambda_{0,2m_1,m_2+1}\}$, respectively. The even set is renamed as follows

$$\lambda_{-1,m_1,m_2} := \lambda_{0,2m_1,2m_2}$$

and is carried over to the next iteration whereas the two mixed sets and the odd set are replaced by the wavelet coefficients γ^h, γ^v, γ^d. The upper indices refer to the horizontal, vertical and diagonal directions, respectively.

Predict (horizontally) If the signal has a local correlation structure, the elements of the odd set and the two mixed sets will be approximated by the mean of its neighbours. The two mixed sets will be approximated by the mean of their two horizontal neighbours and two vertical neighbours, respectively, and the odd set will be approximated by the mean of two opposite diagonal members (Fig. A.3). Let us first consider the horizontal approximation of the mixed set $\{\lambda_{0,2m_1+1,2m_2}\}$. The difference between the original value and its approximation is contained in the wavelet coefficient

$$\gamma^h_{-1,m_1,m_2} = \lambda_{0,2m_1+1,2m_2} - \frac{1}{2}(\lambda_{-1,m_1,m_2} + \lambda_{-1,m_1+1,m_2})$$

for $m_1 = 0, \ldots, 2^{n-1} - 1$ and $m_2 = 0, \ldots, 2^{n-1}$.

Predict (vertically) The other mixed set $\{\lambda_{0,2m_1,2m_2+1}\}$ is vertically approximated so that the corresponding wavelet coefficient is

$$\gamma^v_{-1,m_1,m_2} = \lambda_{0,2m_1,2m_2+1} - \frac{1}{2}(\lambda_{-1,m_1,m_2} + \lambda_{-1,m_1,m_2+1})$$

for $m_1 = 0, \ldots, 2^{n-1}$ and $m_2 = 0, \ldots, 2^{n-1} - 1$.

Update (horizontally and vertically) Again, the average of the coefficients $\{\lambda_{l,m_1,m_2}\}$ are preserved on each level: $\sum_{m_1}\sum_{m_2}\lambda_{l-1,m_1,m_2} = \frac{1}{4}\sum_{m_1}\sum_{m_2}\lambda_{l,m_1,m_2}$ for all $l = 0, \ldots, -n$. Both the elements of the even set $\{\lambda_{-1,m_1,m_2}\}$ and the elements of the odd set $\{\lambda_{0,2m_1+1,2m_2+1}\}$ are lifted with the help of the wavelet coefficients from the two mixed sets:

$$\begin{aligned}
\tilde{\lambda}_{-1,m_1,m_2} &= \lambda_{-1,m_1,m_2} + \tfrac{1}{12}\big(\gamma^v_{-1,m_1,m_2-1} + \gamma^v_{-1,m_1,m_2} \\
&\qquad + \gamma^h_{-1,m_1-1,m_2} + \gamma^h_{-1,m_1,m_2} \big) \\
\tilde{\lambda}_{0,2m_1+1,2m_2+1} &= \lambda_{0,2m_1+1,2m_2+1} + \tfrac{1}{12}\big(\gamma^h_{-1,m_1,m_2} + \gamma^h_{-1,m_1,m_2+1} \\
&\qquad + \gamma^v_{-1,m_1,m_2} + \gamma^v_{-1,m_1+1,m_2} \big).
\end{aligned}$$

Predict (diagonally) Before heading on to the next iteration step on level -2, the odd set $\{\lambda_{0,2m_1+1,2m_2+1}\}$ is also expressed in terms of wavelet coefficients. The $\lambda_{0,2m_1+1,2m_2+1}$ is approximated by the mean of two opposite diagonal neighbours; the pair is chosen out of two possibilities and the choice is based on the position m_1,

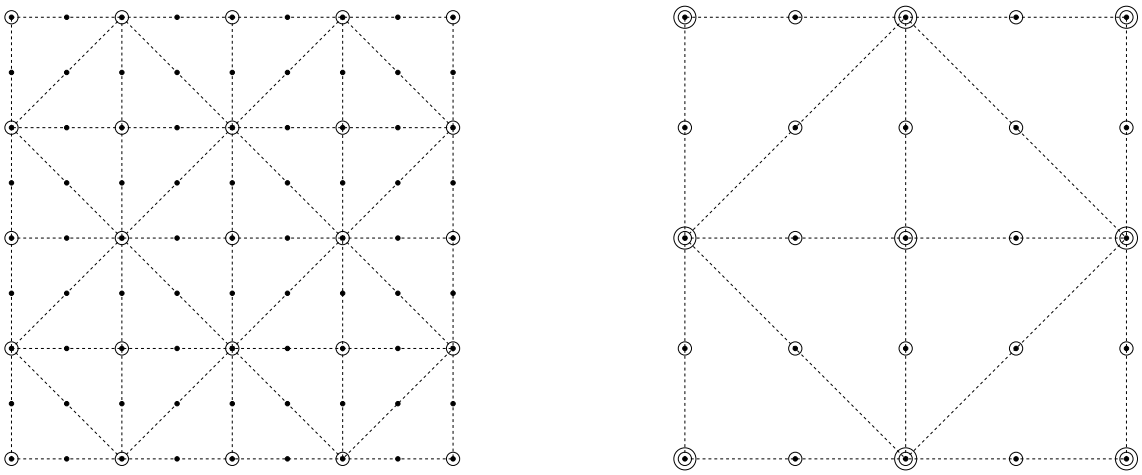

Figure A.3: *Schematic illustration of the lifting scheme on a two-dimensional grid of* $(2^3+1) \times (2^3+1)$ *grid points. Left: The first iteration. Right: The second iteration. The first iteration (left) derives wavelet coefficients* γ *at the locations of the odd set and the two mixed sets (marked by* $\cdot$*). The coefficients are the differences between the original function value and its linear interpolation of those two neighbours that are connected by the dashed line. The function values* λ *of the even set (marked by* $\odot$*) are lifted and carried over to the next iteration step. The second iteration (right) starts from the even set of the first iteration. Again, the set is split into an even set (marked by the double circle), an odd set and two mixed sets (marked by* $\odot$*). Wavelet coefficients are derived for the locations* $\odot$ *by linear interpolation of the two neighbours that are connected by the dashed line. The values of the even set (marked by the double circle) are lifted and carried over to the next iteration step.*

m_2 of $\lambda_{0,2m_1+1,2m_2+1}$. If $(m1+m2) \in \{2h|h=0,\ldots,2^{n-1}\}$ the difference between the original value and its approximation is contained in the wavelet coefficient

$$\gamma^d_{-1,m_1,m_2} = \tilde{\lambda}_{0,2m_1+1,2m_2+1} - \frac{1}{2}(\tilde{\lambda}_{-1,m_1,m_2} + \tilde{\lambda}_{-1,m_1+1,m_2+1})$$

and if $(m1+m2) \in \{2h+1|h=0,\ldots,2^{n-1}-1\}$ the difference is contained in the wavelet coefficient

$$\gamma^d_{-1,m_1,m_2} = \tilde{\lambda}_{0,2m_1+1,2m_2+1} - \frac{1}{2}(\tilde{\lambda}_{-1,m_1,m_2+1} + \tilde{\lambda}_{-1,m_1+1,m_2})$$

for $m_1 = 0,\ldots,2^{n-1}-1$ and $m_2 = 0,\ldots,2^{n-1}-1$.

Update (diagonally) The $\tilde{\lambda}_{-1,m_1,m_2}$ are lifted with the help of the wavelet coefficients from the odd set:

$$\tilde{\tilde{\lambda}}_{-1,m_1,m_2} = \tilde{\lambda}_{-1,m_1,m_2} + \tfrac{1}{12}\big(\gamma^d_{-1,m_1-1,m_2-1} + \gamma^d_{-1,m_1-1,m_2} + \gamma^d_{-1,m_1,m_2-1} + \gamma^d_{-1,m_1,m_2} \big).$$

The steps split, predict, update are iterated to coarser levels $l < 0$ until the level $-n$ has been reached. The grid spacing in the prediction and update steps is increased to $2^{-(l+1)}$ $(l < 0)$ in each iteration (cf. Fig. A.3).

When the lifting scheme is put into the wavelet framework, the corresponding two-dimensional wavelet functions are equivalent to the *k-disc wavelet* with $k = 1$ (Stollnitz et al., 1996). Four mother wavelets can be identified, representing the horizontal, the vertical and the two diagonal directions, respectively. All interior wavelet functions are translates and dilates of one of these four mother wavelets. Each mother wavelet resembles the function in Fig. A.4. The horizontal and vertical mother wavelets are rotated by 90°, respectively, and the diagonal wavelets are rotated by 45° and dilated by a factor of $\sqrt{2}$. Due to the various update steps in the lifting scheme, the mother wavelet has a hexagonal support.

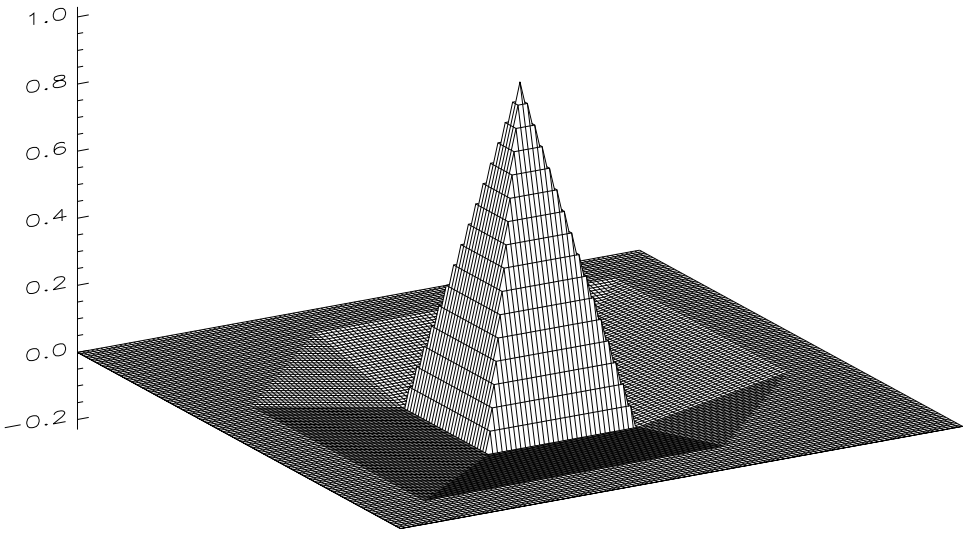

Figure A.4: *One of the mother functions of the wavelet basis, resulting from the two-dimensional lifting scheme.*

B Translation-invariant wavelet smoothing

B.1 Illustration of the problem

The resulting wavelet estimate at a specific location considerably depends on grid details which are determined by the particular choice of the reduced fraction of the domain. The reduced section is required to contain $(2^n + 1) \times (2^n + 1)$ grid points (cf. Subsection 5.2.2). If the detailed choice of this section is made in an arbitrary way the resulting estimate is also arbitrary to a certain extent.

B.1.1 Effect of choosing the origin

Before the wavelet smoothing procedure (Fig. 5.1) is started, a fraction of the LM domain is selected which contains $(2^n + 1) \times (2^n + 1)$ grid points (cf. Section 5.2). The reduced domain is chosen to be as large as possible, i. e. the parameter n satisfies $(2^n + 1) \leq \min(N_x, N_y) < (2^{n+1} + 1)$ if the size of the original LM domain is $N_x \times N_y$ grid points. The position of the reduced domain within the original domain is determined by the co-ordinates of the south-western corner point of the reduced domain. This point is also referred to as the *origin* (i_s, j_s). There is a number of $(N_x - 2^n) \cdot (N_y - 2^n)$ possibilities to position the origin (cf. Fig. B.1).

At first glance, the *a priori* choice of origin seems to be irrelevant. However, the following example illustrates that the choice of origin can have a surprisingly large effect on the post-processed forecast. Fig. B.2 is a result of curtailing the two finest levels according to Eq. 5.1 with $l_{\text{curtail}} = -2$. The post-processing procedure is the same than the one leading to Fig. 5.8, except for the choice of origin. The origins differ by a westward shift by one grid point and northward shift by one grid point. The shifted origin does not just result in a shifted version of the post-processed forecast, but appears to affect the entire post-processing procedure and also the resulting wavelet estimator of the expected value (cf. Fig. 5.8 and Fig. B.2).

Note that the choice of origin does not necessarily have a larger effect when the distance of the shift is increased. For example, a northward shift by eight grid points does not have any effect and the post-processed forecast is indeed just a shifted version of Fig. 5.8.

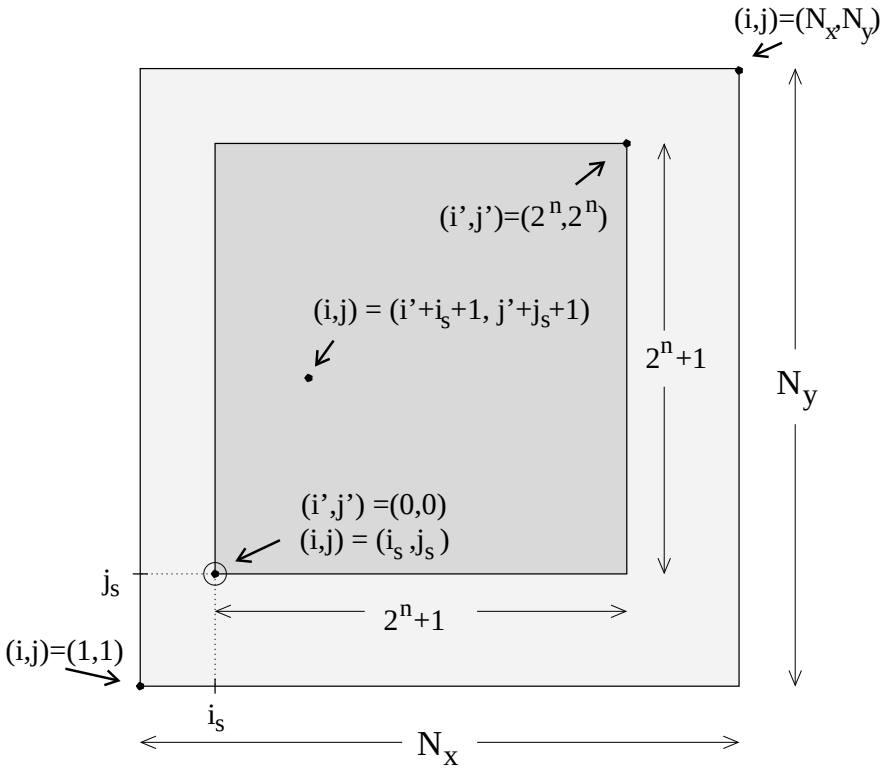

Figure B.1: *Position of the reduced domain within the original domain. The original domain comprises $N_x \times N_y$ grid points and the reduced domain comprises $(2^n + 1) \times (2^n + 1)$ grid points. The position of the reduced domain is determined by the co-ordinates (i_s, j_s) of the origin, i. e. the south-western corner point of the reduced domain. The origin is marked by the symbol $\odot$.*

Dealing with the effect of origin calls for some additional notation. The original forecast is still denoted z_{ij} where the indices $i = 1, \ldots, N_x$, $j = 1, \ldots, N_y$ still refer to the original LM domain. A forecast field that has been reduced to a size of $(2^n+1) \times (2^n+1)$ and has been post-processed by wavelet smoothing is now denoted as $\tilde{z}_{i'j'}$ where the indices $i'j'$ refer to the co-ordinates of the reduced field: $i' = 0, \ldots, 2^n$, $j' = 0, \ldots, 2^n$. Furthermore, the post-processed field $\tilde{z}_{i'j'}$ is a function of the choice of origin: $\tilde{z}_{i'j'}(i_s, j_s)$. The origin is represented by its spatial co-ordinates $i_s = 0, \ldots, (N_x - 2^n - 1)$, $j_s = 0, \ldots, (N_y - 2^n - 1)$ (cf. Fig. B.1). The post-processed forecast $\tilde{z}_{i'j'}(i_s j_s)$ is an estimator of the expected value μ_{ij} where the indices of μ refer to the original LM domain and are calculated from $i = i' + i_s + 1$, $j = j' + j_s + 1$.

At one and the same geographical location $i = (i' + i_s + 1)$, $j = (j' + j_s + 1)$ the smoothed forecasts $\tilde{z}_{i',j'}(i_s, j_s)$ can differ considerably for different choices of shifts i_s, j_s. As the choice of origin is arbitrary, the wavelet estimator of the expected value μ_{ij} is far from robust. Ways must be found how to discover the best choice of origin or how to combine different choices of origin. In order to achieve this, the role of the origin in wavelet smoothing must be understood first.

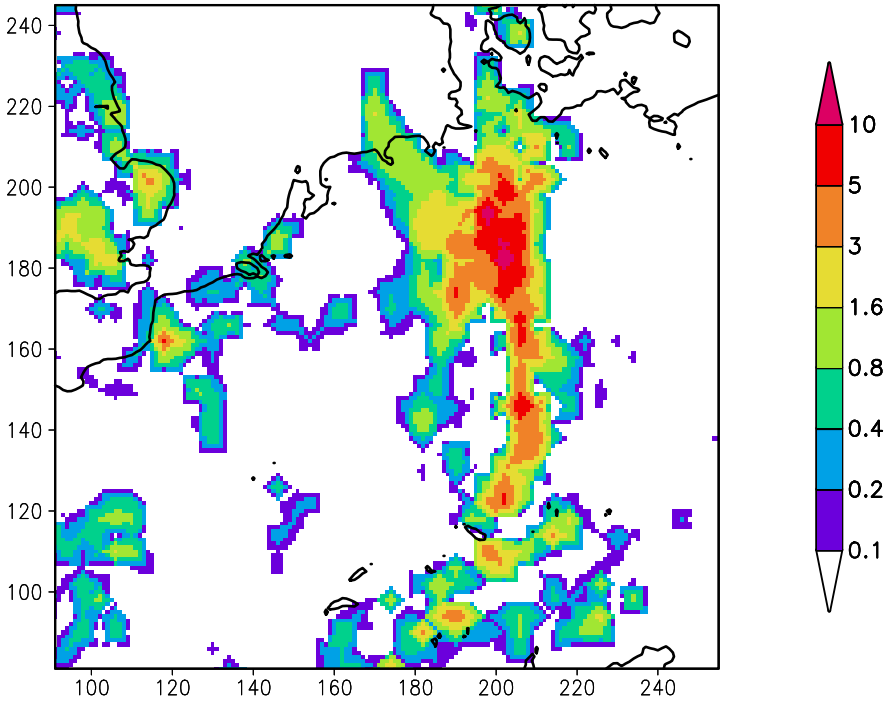

Figure B.2: *Post-processed forecast from a slightly shifted version of the original precipitation forecast in Fig. 5.4. Unit: mm. Reconstruction after curtailing the two finest levels. The full domain of the post-processed forecast comprises* 257×257 *grid points; here a section of* 165×165 *grid points is shown which is the same as the section shown in Fig. 5.8.*

B.1.2 How does the effect come about?

The effect of a shifted origin is due to the dependence of the wavelet transform on the lattice. The lifting scheme introduces a hierarchy to the points of the lattice (cf. Fig. A.3). Positioning the lattice defines the rank of a certain grid point within the hierarchy. The choice of origin determines whether a certain grid point belongs to the even set, to the odd set or to a mixed set in the splitting operation of the lifting scheme (cf. Appendix A). For example, if wavelet smoothing consists in curtailing the finest level, members of the odd set and the mixed sets will be replaced by their predicted values, whereas members of the even set will merely be lifted (cf. Appendix A).

The dependence of the wavelet transform on the choice of origin has been taken up in the literature, initially in the context of first-generation wavelets. Coifman and Donoho (1995) state that wavelet denoising of finite data sets exhibits pseudo-Gibbs phenomena because of unfortunate mis-alignments between features in the signal and features in the wavelet basis.

B.2 Solution of the problem: Averaging shifts

B.2.1 General methodology

Optimisation with respect to shifts of the origin is not possible in general situations (Coifman and Donoho, 1995). Therefore, different smoothing results due to different shifts can be seen as equally valid. Coifman and Donoho (1995) suggest to apply a range of shifts, and average over the several results so obtained (Fig. 5.10):

$$\tilde{z}_{ij}^{\text{invar}} = \frac{1}{|\mathcal{O}|} \sum_{(i_s,j_s)\in\mathcal{O}} \tilde{z}_{i'j'}(i_s, j_s) \tag{B.1}$$

where the indices i', j' of the reduced domain satisfy

$$i' = (i - i_s - 1) \quad \text{and} \quad j' = (j - j_s - 1)$$

so that they refer to the same geographical location than the point (i, j) within the original LM domain (cf. Fig. B.1). The range of different origins (i_s, j_s) is given by the set $\mathcal{O}$. The set $\mathcal{O}$ must satisfy

$$\mathcal{O} \subseteq \Big\{(i_s, j_s)\Big|(i_s, j_s) \in \{1, \ldots, N_x - 2^n\} \times \{1, \ldots, N_y - 2^n\}\Big\} \tag{B.2}$$

so that the shifted reduced domain does not extend outward beyond the boundaries of the original domain. The number of elements in $\mathcal{O}$ is denoted by $|\mathcal{O}|$. The averaged field $\tilde{z}_{ij}^{\text{invar}}$ (Eq. B.1) is only derived for those points (i, j) that are contained in each of the shifted reduced domains:

$$(i,j) \in \bigcap_{(i_s,j_s)\in\mathcal{O}} \Big\{(i' + i_s + 1, j' + j_s + 1)\Big|(i', j') \in \{0, \ldots, 2^n\} \times \{0, \ldots, 2^n\}\Big\}. \tag{B.3}$$

The approach of averaging shifts is also called *translation-invariant* wavelet denoising (Coifman and Donoho, 1995); that is why the resulting field $\tilde{z}_{ij}^{\text{invar}}$ is denoted by the upper index *invar*. The post-processed field $\tilde{z}_{ij}^{\text{invar}}$ is expected to yield a more robust wavelet estimator of the expected value μ_{ij}.

B.2.2 Optimum range of shifts

Even if the set $\mathcal{O}$ satisfies Eq. B.2, it is not necessarily the optimum choice for the derivation of the averaged field $\tilde{z}_{ij}^{\text{invar}}$ (Eq. B.1). In the following, an optimum range of shifts is sought.

Criteria of optimality

On the one hand, the range of shifts $\mathcal{O}$ should be large enough to cover all possible outcomes of wavelet smoothing which are due to the dependence of the wavelet transform on the lattice. On the other hand, the range of shifts should be as small as possible in order to minimise the increase in computational effort and to maximise the number of points contained in each of the shifted domains. Every additional shift reduces the number of points (i, j) that belong to the overlap of the applied shifts (cf. Eq. B.3) and that can be averaged to the translation-invariant estimator $\tilde{z}_{ij}^{\text{invar}}$ (Eq. B.1).

Redundancy in the outcome of different shifts

As already indicated in Subsection B.1.1 there is a redundancy in the outcome of different shifts. The redundancy occurs if wavelet coefficients of coarse levels remain untouched in the curtailing step of wavelet post-processing. Let us assume that only coefficients of the finest level are modified. Then the curtailing step of the wavelet smoothing procedure only affects those wavelet coefficients that are calculated in the first iteration step of the transformation algorithm. The role of the lattice within the first iteration step is illustrated by Fig. A.3. When the lattice is horizontally shifted by four grid points, the transformation algorithm treats the respective grid points in an identical way again. The same is true for vertical shifts by four grid points and diagonal shifts by two grid points.

A non-redundant range of shifts

Redundancy comes to the rescue when trying to limit the number of shifts in $\mathcal{O}$. Let us assume that wavelet coefficients of levels coarser than l_{mod} ($l_{\text{mod}} \in \{-n+1, \ldots, -1\}$) are not modified in the smoothing procedure. Then any range of shifts

$$\mathcal{O}_{\text{nonred}}(i_a, j_a) = \\ \left\{(i_s, j_s) \middle| (i_s, j_s) \in \{i_a, \ldots, i_a + 2^{-l_{\text{mod}}+1} - 1\} \times \{j_a, \ldots, j_a + 2^{-l_{\text{mod}}} - 1\}\right\} \quad i_a, j_a \in \mathbb{N}_0 \tag{B.4}$$

that still fits into the allowed range (Eq. B.2) yields a non-redundant set of smoothed values $\tilde{z}_{ij}(i_s, j_s)$. The set is complete, in the sense that it attains the full range of possible values $\tilde{z}_{ij}(i_s, j_s)$ that may occur due to different shifts. The point (i_a, j_a) provides an arbitrary spatial offset for the shifts (cf. Fig. B.3).

The number of shifts in $\mathcal{O}_{\text{nonred}}(i_a, j_a)$ is

$$|\mathcal{O}_{\text{nonred}}(i_a, j_a)| = 2^{-2l_{\text{mod}}+1}.$$

The overlap of the shifted domains is a rectangular field with side lengths that consist of $(2^n + 1) - (2^{-l_{\text{mod}}+1} - 1)$ grid points and $(2^n + 1) - (2^{-l_{\text{mod}}} - 1)$ grid points, respectively.

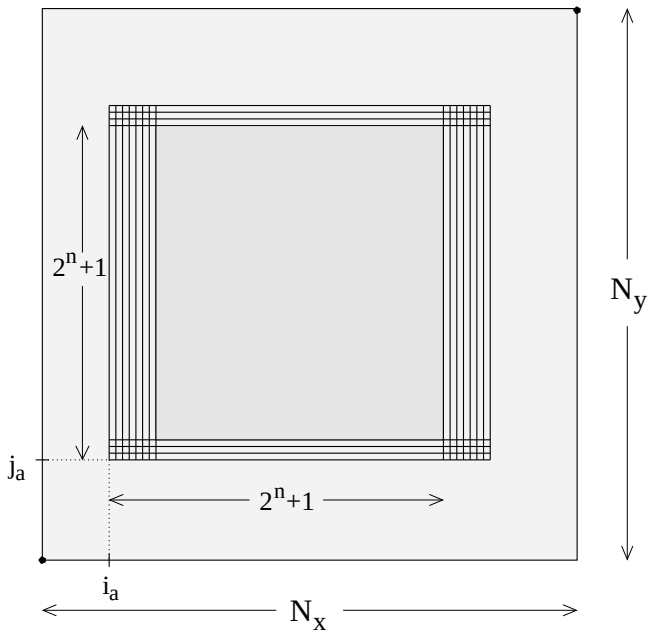

Figure B.3: *A range of shifts that yields a non-redundant set of smoothed values $\tilde{z}_{ij}(i_s, j_s)$ according to Eq. B.4. The set is complete, in the sense that it attains the full range of possible values $\tilde{z}_{ij}(i_s, j_s)$ that may occur due to different shifts. The point (i_a, j_a) provides an arbitrary spatial offset for the shifts.*

The lower left corner point of the overlap is situated at $(i_a + 2^{-l_{\mathrm{mod}}+1} - 1, j_a + 2^{-l_{\mathrm{mod}}} - 1)$ (cf. Fig. B.3). When applying the range of shifts $\mathcal{O}_{\mathrm{nonred}}(i_a, j_a)$ to Eq. B.1, the translation-invariant estimator $\tilde{z}_{ij}^{\mathrm{invar}}$ can be derived for locations within this overlap.

If the size parameter n of the reduced domain is chosen to be as large as possible (i. e. $(2^n + 1) \leq \min(N_x, N_y) < (2^{n+1} + 1)$) there is a possibility that some of the shifts in $\mathcal{O}_{\mathrm{nonred}}(i_a, j_a)$ (Eq. B.4) do not fit into the allowed range of the original domain (cf. Eq. B.2). There might not be enough space between the boundaries of the original field $N_x \times N_y$ and the reduced field $(2^n + 1) \times (2^n + 1)$ for shifting the reduced field around (cf. Fig. B.3). This case becomes more likely the coarser the level l_{mod}. This study circumvented the case by choosing a reduced field that is four times smaller than the maximum size possible. Then the size parameter n satisfies the condition $(2^{n+1} + 1) \leq \min(N_x, N_y) < (2^{n+2} + 1)$.

An alternative with less computational costs

In practice, the number $|\mathcal{O}_{\mathrm{nonred}}(i_a, j_a)|$ of shifts might still turn out prohibitively large. For example, there is already a number of 32 shifts in $\mathcal{O}_{\mathrm{nonred}}(i_a, j_a)$ if only the two finest levels are modified and there is a number of 128 shifts if the three finest levels are modified.

If all shifts in $\mathcal{O}_{\text{nonred}}(i_a, j_a)$ are applied, the computational cost of the post-processing procedure will rise substantially and the translation-invariant estimator $\tilde{z}_{ij}^{\text{invar}}$ will be derived at a smaller number of points (i, j). Computational efficiency had been one of the decisive advantages of wavelet smoothing and is in danger of getting lost now.

As a remedy it is suggested to randomly select only a few shifts from $\mathcal{O}_{\text{nonred}}(i_a, j_a)$, derive the respective smoothed fields $\tilde{z}_{i'j'}$ and take their average. This still contributes to increased robustness of the estimate while keeping the computational effort within the bounds of feasibility.

B.3 Is there a relation between shift variability and forecast uncertainty?

As different smoothing results due to different shifts are treated as equally valid, they might be able to serve as a means to synthesize noise and derive an estimate of forecast uncertainty ε_{ij} (Eq. 4.6) in addition to the estimate of the expected value μ_{ij}. Does the variability within the set of smoothed forecasts estimate anything?

B.3.1 Is there an equivalence?

Let us first clarify that the variability between the differently shifted smoothed forecasts is not a direct estimator of forecast uncertainty, because the set of smoothed forecasts is not a sample of the forecast itself. This becomes obvious when the set of smoothed forecasts is juxtaposed with an existing realisation of the forecast, namely the original non-smoothed forecast. The non-smoothed forecast remarkably sticks out of the set of smoothed forecasts which illustrates the fact that they do not belong to the same sample.

Equivalently, neither the differences between the shifted smoothed forecasts and the original non-smoothed forecast are suitable for a direct estimate of forecast uncertainty. The variability between those differences is exactly the same as the variability between the smoothed forecasts themselves.

B.3.2 Is there a proportionality?

If at all, it is more conceivable to treat the differently shifted smoothed forecasts as *iid* realisations of the expected value estimator. This would be of interest if the variance between different realisations of the estimator were associated with the unknown variance between different forecast realisations, i. e. forecast uncertainty or noise. For example, such an association exists in case of the maximum-likelihood estimator of the expected

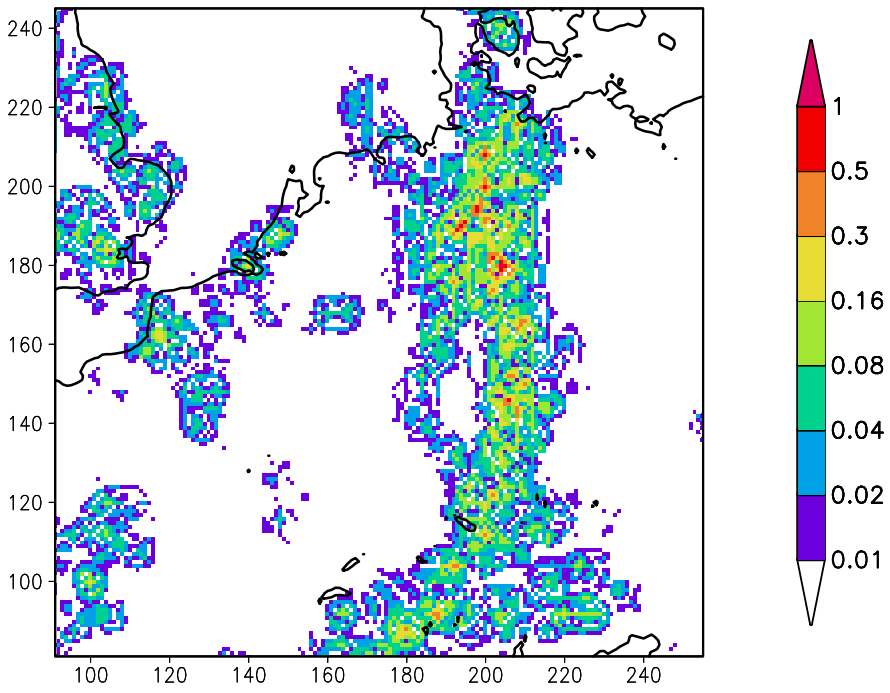

Figure B.4: *Square root $\hat{\sigma}_{ij}^{\text{shift}}$ of the estimated variability between resulting estimates (Eq. B.5) when the whole range of different shifts is applied. Unit: mm. The two finest levels are curtailed in the smoothing procedure. Fig. 5.8 and B.2 are two examples of estimates that contribute to the variability estimate.*

value (cf. Eq. 4.3) which indeed possesses a variance proportional to the variance of the realisations themselves.

Let us start with a visual inspection of the variability between the estimates that originate from differently shifted smoothing procedures. The square root of the variability between those estimates is expressed by the term

$$\hat{\sigma}_{ij}^{\text{shift}} = \sqrt{\frac{1}{|\mathcal{O}|} \sum_{(i_s,j_s)\in\mathcal{O}} \left(\tilde{z}_{ij}(i_s,j_s) - \tilde{z}_{ij}^{\text{invar}}\right)^2}. \tag{B.5}$$

Fig. B.4 shows an example of the variability between the resulting estimates when the whole range of possible shifts is applied. Fig. 5.8 and B.2 are two examples of estimates that contribute to the variability estimate in Fig. B.4.

The spatial structures of the variability field indeed bear some similarities to those of the neighbourhood and the ensemble estimators of forecast uncertainty. It seems as if the variability between differently shifted smoothing results would only have to be inflated by some factor until it could be interpreted in terms of forecast noise.

However, the existence and optimal choice of such an inflation factor is closely tied to some fundamental questions that have not been answered yet. First, it is doubtful that the differently shifted smoothed forecasts can be viewed as *iid* realisations. Secondly, it is not clear whether the differently shifted smoothed forecasts really cover the phase space we seek information about. For example, unlike the maximum-likelihood estimator, the wavelet estimator may also be afflicted with a bias. The bias is not necessarily the same for each shift so that the variability between differently shifted smoothed forecasts may not only reflect forecast uncertainty, but also the difference between systematic errors of different estimators. Furthermore, smoothed forecasts might not be independent because every smoothed forecast is derived from the same forecast and via the same principle of smoothing.

These issues must be tackled first in order to arrive at a statistically sound interpretation of the inflated variability field. For the time being, the questions above and the statistical interpretation of the inflated variability field are set aside and remain for future investigations.

C Measures in forecast verification

C.1 Overview

This appendix briefly explains some measures in objective forecast verification. Only a few aspects are highlighted here, just as much as is required to understand the measures used in the verification sections of this report. For a more extensive review see Stanski et al. (1989) or Wilks (1995), for example.

Forecast goodness can be divided into three types: *consistency*, *quality* and *value* (Murphy, 1993). Consistency is the correspondence between forecasters' judgments and their forecasts; quality is the correspondence between the forecasts and the respective observations; value is the benefit gained by decision makers through the use of the forecasts. Whereas consistency is difficult to measure objectively, objective measures of quality and value exist. Sections C.2 and C.3 deal with measures of forecast quality and Section C.4 deals with measures of forecast value.

Furthermore, forecasts can be categorised into *categorical* forecasts and *probability* forecasts. A categorical forecast consists of a flat statement that one and only one of a set of possible events will occur. Probability forecasts contain an expression of uncertainty, in distinction to categorical forecasts.

Probability forecasts can be formulated for *continuous* predictands and *discrete* predictands. In case of discrete predictands, the set of possible events is finite. The probability forecast ranges from 0 to 1, with 0 for a certain "no" and 1 for a certain "yes". The exceedance probability is a probability forecast for a predictand that has been discretised beforehand.

Probability forecasts for continuous predictands can be formulated by quantiles. The quantile specifies the value for which there is a probability p that the corresponding observation will be less than this value.

Different measures of quality are used, depending on whether the forecast is a categorical or a probability forecast and whether the predictand is discrete or continuous. Section C.2 deals with quality measures for categorical forecasts and Section C.3 deals with quality measures for probability forecasts. The measure of forecast value in Section C.4 can be applied to both types of forecasts.

Table C.1: *Two-by-two contingency table for forecast and occurrence of binary events. The numbers of cases in the different categories are denoted a, b, c and d.*

		Observed	
		Yes	No
Forecast	Yes	a	b
Forecast	No	c	d

C.2 Quality of categorical forecasts

Different quality measures of categorical forecasts are used, depending on the set of possible events. The set of possible events may either comprise all values on a relevant portion of the real line (*continuous* predictand) or a finite set of possible events (*discrete* predictand). If the finite set consists of only two events, i. e. "yes" or "no", the predictand is *dichotomous*. First, forecasts of dichotomous predictands are considered, then forecasts of continuous predictands.

C.2.1 Forecasts of dichotomous predictands

Quality measures of a dichotomous forecast are usually derived from the contingency table (Tab. C.1). Some of the common measures of forecast quality are the probability of detection (POD), the false alarm rate (FAR), the frequency bias (FBI), the equitable threat score (ETS) and the logarithm of the odds ratio (LOR).

The POD (Wilks, 1995) only looks in cases where the event was observed and relates the number of correct forecasts to the total number of observed events. That is, the POD is the conditional probability that the event would be forecast, given that it occurs. It is computed as

$$\mathrm{POD} = \frac{a}{a+c}.$$

The FAR only looks in cases where the event was forecast and relates the number of false alarms to the total number of forecasts. That is, the FAR is the conditional probability of a false alarm, given that the event is forecast. The FAR is defined as

$$\mathrm{FAR} = \frac{b}{a+b}.$$

The frequency bias or bias ratio (Wilks, 1995) does not represent a conditional probability. It is defined as the ratio of "yes" forecasts to the number of "yes" observations:

$$\mathrm{FBI} = \frac{a+b}{a+c}.$$

The equitable threat score is similar to the threat score (Wilks, 1995), but accounts for forecasts that verify by chance. It is defined as

$$\mathrm{ETS} = \frac{a - a_{\mathrm{chance}}}{a + b + c - a_{\mathrm{chance}}}$$

where a_{chance} is the number of hits which would be achieved by a random forecast:

$$a_{\mathrm{chance}} = \frac{(a+b)\cdot(a+c)}{a+b+c+d}.$$

The odds ratio is defined as the ratio of the odds of making a good forecast (a hit) to the odds of making a bad forecast (a false alarm) (Stephenson, 2000). It can be expressed by the "cross-product-ratio":

$$\mathrm{OR} = \frac{ad}{bc}.$$

The logarithm of the odds ratio (LOR) ranges from zero to infinity. It is equal to zero when the forecasts and the observations are independent. The log odds ratio can easily be tested for statistical significance, because it is asymptotically Gaussian distributed with a variance of $(1/a + 1/b + 1/c + 1/d)$ (Stephenson, 2000).

It should be noted that the odds ratio benefits from a large frequency bias. In order to correct for this misleading effect a bias-corrected odds ratio OR(BC) is defined:

$$\mathrm{OR(BC)} = \frac{a'd'}{b'c'}$$

where the numbers from the contingency table are modified so that the forecast has zero frequency bias:

$$a' = a - \alpha a \quad c' = c + \alpha a \quad b' = b - \alpha b \quad d' = d + \alpha b.$$

The correction parameter α is defined as

$$\alpha = \frac{b-c}{a+b}.$$

C.2.2 Forecasts of continuous predictands

Some measures of categorical forecast quality are the root mean squared error (*rmse*), the mean absolute error (*mae*), the bias (*bias*) and the correlation (*corr*) between forecasts

and observations. The measures are defined as

$$rmse = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - o_i)^2}$$

$$mae = \frac{1}{N}\sum_{i=1}^{N}|y_i - o_i|$$

$$bias = \frac{1}{N}\sum_{i=1}^{N}(y_i - o_i)$$

$$corr = \frac{1}{(N-1)}\frac{\sum_{i=1}^{N}(y_i - \overline{y})(o_i - \overline{o})}{stdv\_fc \quad stdv\_obs}$$

where N is the number of observations, o_i are the observations, y_i the corresponding forecasts and the overline denotes the mean. The standard deviation of the forecast data ($stdv\_fc$) and the standard deviation of the observational data ($stdv\_obs$) are defined as:

$$stdv\_fc = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y})^2}$$

$$stdv\_obs = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(o_i - \overline{o})^2}.$$

The root mean squared error $rmse$ is related to the bias, the standard deviation of the forecast data, the standard deviation of the observational data and the correlation between observational data and forecast data:

$$(rmse)^2 = (bias)^2 + (stdv\_fc)^2 + (stdv\_obs)^2 - 2\,(stdv\_fc)\cdot(stdv\_obs)\cdot corr. \quad \text{(C.1)}$$

The standard deviation of the observational data $stdv\_obs$ is independent of the forecast. An improved $rmse$ can be exclusively achieved through a reduction in $stdv\_fc$ which does not necessarily imply a true improvement in forecast quality (Taylor, 2001). Therefore, it is important to keep an eye on $stdv\_fc$ when the $rmse$ is interpreted.

The skill relates the root mean squared error to a reference forecast and the perfect forecast:

$$skill = \frac{rmse - rmse_{\text{reference}}}{rmse_{\text{perfect}} - rmse_{\text{reference}}}$$

Positive skill indicates that the forecast under investigation is more accurate than the reference forecast. When the skill is related to the direct model output of the single unperturbed LM simulation, it is denoted $skill\_DMO$.

C.3 Quality of probability forecasts

Before describing the particular verification measures in detail, a brief remark should be made on the general procedure that is applied to the verification of probability forecasts in this study. Currently, it is common practice to compare the forecast probability to the series of observed events having the categorical outcomes "yes" or "no" (e. g. Wilks, 1995). As this study does not aim at a revision of common verfication procedures, it simply conforms to this practice. However, it would be more appropriate to compare the forecast probabilities to the *probability* of the respective event instead of the series of its categorical outcomes. This may be achieved by hierarchical Bayesian space-time modelling (Wikle et al., 1998), but remains for future research.

C.3.1 Forecasts of dichotomous predictands

In this study the exceedance probability is a probability forecast of a dichotomous predictand. Some quality measures of the exceedance probability are the Brier score, the ranked probability score, the reliability diagram (Wilks, 1995), the relative operating characteristic (Stanski et al., 1989) or the rank histogram (cf. e. g. Hamill and Colucci, 1998). The Brier score and the Brier skill score are by far most common. This section presents their definition including the concepts of reliability and resolution. Furthermore, the reliability diagram and the relative operating characteristic are introduced.

Brier score

The Brier score (e. g. Wilks, 1995) is essentially the mean squared error of probability forecasts:

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

where y_k denotes a forecast probability and o_k denotes the subsequent observations with $o_k = 1$ if the event occurs and $o_k = 0$ otherwise. The index k indicates a numbering of the n forecast/event pairs. The Brier skill ranges from 0 to 1, with perfect forecasts exhibiting $\mathrm{BS} = 0$. Less accurate forecasts receive higher Brier scores.

The Brier score can be decomposed into the sum of three terms related to reliability, resolution and uncertainty (e. g. Wilks, 1995):

$$\mathrm{BS} = \mathrm{BS}_{\mathrm{rel}} - \mathrm{BS}_{\mathrm{res}} + \mathrm{BS}_{\mathrm{unc}}.$$

Reliability measures the correspondence between the forecast probability and the conditional probability of an observed event, given it was forecast. Resolution is a measure of the degree to which the forecast frequencies differ from the sample climatology. Uncertainty is independent of the forecast system and expresses the inherent uncertainty

in forecasting the event. It attains a maximum when the overall sample climatology is 50%.

Brier skill score

The Brier skill score relates the Brier score to a reference forecast:

$$\mathrm{BSS} = \frac{\mathrm{BS} - \mathrm{BS}_{\mathrm{reference}}}{\mathrm{BS}_{\mathrm{perfect}} - \mathrm{BS}_{\mathrm{reference}}}.$$

Often the reference forecast is chosen to be the climatological forecast. For climatological forecasts the uncertainty term and the Brier score are equivalent (e. g. Wilks, 1995), so the Brier skill score in reference to a climatological forecast reduces to:

$$\mathrm{BSS} = \frac{\mathrm{BS} - \mathrm{BS}_{\mathrm{climate}}}{\mathrm{BS}_{\mathrm{perfect}} - \mathrm{BS}_{\mathrm{climate}}} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{unc}}} = \frac{\mathrm{BS}_{\mathrm{res}} - \mathrm{BS}_{\mathrm{rel}}}{\mathrm{BS}_{\mathrm{unc}}}.$$

If the climatological forecast is chosen as a reference, the Brier skill score is only positive when the uncertainty term is greater than the Brier score and when the resolution term is greater than the reliability term.

Reliability diagram

Reliability is often illustrated by a reliability diagram (Wilks, 1995). The reliability diagram shows the conditional frequency of the event given it was forecast as a function of the forecast probability. Thus, the diagram shows forecast quality broken down to the different forecast probabilities. The unconditional frequency of forecast probabilities should be visualized in some way, too. Additionally, the reliability diagram relates the forecast to a perfectly reliable forecast, a forecast without skill and a forecast without resolution.

Relative operating characteristic

The relative operating characteristic (ROC) (e. g. Stanski et al., 1989) can be used to compare probabilistic and deterministic forecasts. The ROC can be derived empirically from the 2×2 contingency table (Tab. C.1). For a probability forecast, a set of contingency tables is constructed, one for each probability category. For example, in the contingency table for forecast probability category 10%, an event is considered as forecast, if the forecast probability is 10% or higher.

From each contingency table a hit rate $\mathrm{HR}_{\mathrm{ROC}} = a/(a+c)$ and a false alarm rate $\mathrm{FAR}_{\mathrm{ROC}} = b/(b+d)$ is derived. A low forecast probability is expected to attain a large hit rate and a large false alarm rate. A high forecast probability is expected to attain a low hit rate and a low false alarm rate. The ROC curve is obtained by plotting the

hit rate $\mathrm{HR}_{\mathrm{ROC}}$ (ordinate) versus the false alarm rate $\mathrm{FAR}_{\mathrm{ROC}}$ (abscissa) over the entire range of probability categories. The point (0,0) is added for the case "never forecast" and the point (1,1) is added for the case "always forecast".

A perfect forecast system yields a $\mathrm{HR}_{\mathrm{ROC}} = 1$ and a $\mathrm{FAR}_{\mathrm{ROC}} = 0$, so the closer the forecasts come to the upper left corner, the better the forecast. If $\mathrm{HR}_{\mathrm{ROC}} = \mathrm{FAR}_{\mathrm{ROC}}$, the forecast is not able to distinguish between event occurrence and non-occurrence. Therefore, the diagonal represents the line of "no skill". The area underneath the ROC curve is a measure of forecast quality. The larger the area the better the forecast. The ROC is qualitatively similar to resolution in the sense that it assesses the ability of the forecast system to discriminate between occurrences and non-occurrences.

In addition to the empirical estimation described above, the ROC curve and the area underneath may also be parametrised by the odds ratio. The odds ratio (cf. Section C.2) can be used to obtain the hit rate $\mathrm{HR}_{\mathrm{ROC}}$ as a function of the false alarm rate $\mathrm{FAR}_{\mathrm{ROC}}$ (Stephenson, 2000). In this function the odds ratio plays the role of a fixed parameter. Such a parametrisation is only possible if the odds ratio has a constant value for all probability categories that are defined. As this is not the general case, this study refrains from a parametrisation of the ROC curve. Stephenson (2000) circumvents this problem by defining only one probability category in addition to "never forecast" and "always forecast".

C.3.2 Forecasts of continuous predictands

Quantiles are probabilistic forecasts of a continuous predictand. As the definition of quantiles is mathematically closely related to the exceedance probability, their concepts of forecast quality are similar. Criteria such as resolution and reliability also apply to quantiles. Reliability diagrams can also be produced for quantile forecasts. They show how frequently a quantile of probability p exceeds the corresponding observation and relate this frequency to the probability p itself.

Measuring resolution or the accuracy of quantiles poses a problem, because neither the Brier skill score nor the root mean squared error are applicable to quantiles. Alternative standard measures for quantiles do not exist in the literature yet.

In terms of resolution, Bremnes (2004) looks at empirical distributions of quantiles that were forecast for a specific probability p. If the variance of the distribution is large, the forecasts tend to deviate from climatology and supposedly possess high resolution.

In terms of accuracy it might be possible to transform the weighted mean absolute error (WMAE) into a forecast measure. The WMAE is defined as

$$\mathrm{WMAE}(p) = \sum_{i=1}^{N} \rho\Big(q_i(p), o_i\Big)$$

$$\text{with} \quad \rho\Big(q_i(p), o_i\Big) = \begin{cases} (1-p) \cdot |q_i(p) - o_i| & : \quad q_i(p) < o_i \\ p \cdot |q_i(p) - o_i| & : \quad q_i(p) \geq o_i \end{cases}$$

where N is the number of observations, o_i $(i = 1, \ldots, N)$ denotes the observations, $q_i(p)$ denotes the corresponding quantile forecasts for a probability p. Perfect quantile forecasts have zero WMAE. The WMAE attains a minimum when the quantile forecasts are perfectly reliable. This can easily be seen by taking directional derivatives. So far, the WMAE has been used as an optimisation term in quantile regression, a method to derive conditional quantiles from a historical data set (Koenker and Hallock, 2001; Bremnes, 2004). The transformation into a forecast measure remains for future research.

C.4 Forecast value

C.4.1 Categorical forecasts

Forecast value (Richardson, 2000; Zhu et al., 2002; Mylne, 2002) is based on the idea that a user suffers a loss L if an event occurs and the user is not protected. If the user takes preventive action to guard against this potential loss, the user will incur a cost C. The cost-loss scenario (Tab. 6.5) is proposed by Thompson and Brier (1955) and has recently gained increased attention.

The user will experience a mean expense E_{forecast}, if she takes protective action whenever the event is forecast:

$$E_{\text{forecast}} = (a+b)C - cL$$

where the symbols a, b and c refer to the contingency table in Tab. C.1. Forecast value represents the saving the user gains by using the forecast, so the mean expense E_{forecast} has to be related to a baseline reference expense. This baseline is inferred from climatology. If no forecast is available, the user will always protect or never protect, depending on her cost/loss ratio and on the climatological frequency of the event (Murphy, 1977). The mean expense is denoted E_{climate} in this case. When a perfect forecast system is available, the mean expense E_{perfect} is not necessarily zero, since protective action might only mitigate the loss L. So the relative value V of a forecast system is defined as the following relation of expenses E_{forecast}, E_{climate} and E_{perfect} (e. g. Richardson, 2000):

$$V = \frac{E_{\text{climate}} - E_{\text{forecast}}}{E_{\text{climate}} - E_{\text{perfect}}}.$$

The relative value ranges from zero to one, with $V = 0$ for a climatological forecast and $V = 1$ for a perfect forecast. It is often assumed that the climatological frequency is equal to the sample frequency when E_{climate} is estimated. This leads to an underestimation of E_{climate} (Atger, 2001).

C.4.2 Exceedance probability

If the forecast is expressed in terms of probabilities, the user has to define a threshold probability p_t for her decision making process. The process is as follows: When the forecast probability is equal or greater than p_t, the user decides to take preventive action. It can be shown that the optimal decision level of a particular user is equal to her cost/loss ratio if the forecast is perfectly reliable (e. g. Murphy, 1977). The value of a probability forecasts is always greater than the value of a categorical forecast if the respective probability forecast is perfectly reliable (Murphy, 1977).

C.4.3 Quantiles

If the forecast is expressed in terms of quantiles, relative value can also be estimated under certain assumptions. Let us focus on the 90%-quantile of precipitation. The user has to define a precipitation threshold $y_{0.9}$ for her decision making process: When the forecast 90%-quantile is equal or greater than $y_{0.9}$, the user decides to take preventive action. This decision making process implies that the user's cost-loss ratio is a decreasing function of precipitation amount y and that it attains 10% at the threshold $y_{0.9}$.

In case of categorical forecasts or exceedance probabilities, relative value is usually displayed as a function of the cost-loss ratio. In case of 90%-quantiles, the cost-loss ratio of 10% is fixed, but the user-specific precipitation threshold $y_{0.9}$ can be varied.

D List of symbols

Formal Symbolism	Meaning	defined in Section
$\mathbb{R}$	real numbers	
$\mathbb{R}^+$	positive real numbers	
$\mathbb{R}_0^+$	non-negative real numbers	
$\mathbb{N}$	positive integers	
$\mathrm{E}[\cdot]$	expected value of a random variable $\cdot$	
$\mathrm{VAR}[\cdot]$	variance of a random variable $\cdot$	
$\Pr[\cdot]$	probability of an event $\cdot$	
$\mathrm{NV}(\mu, \sigma^2)$	Gaussian distribution with parameters μ and σ^2	
ϵ	realisation of a Gaussian white noise process	
μ	expected value of a Gaussian random variable	
σ^2	variance of a Gaussian random variable	
$\hat{\cdot}$	statistical estimator of a parameter $\cdot$	
i,j	integer variables referring to location	
k	integer variable referring to a point of time	
x, y	continuous variables referring to location	
t	continuous variable referring to a point in time, for example model integration time	
$\ddagger$	continuous variable referring to model output time	
Δt	time step in the LM model integration (40 s)	
Δx	grid spacing of the model LM, eastward direction (7 km)	
Δy	grid spacing of the model LM, northward direction (7 km)	
$\Delta\ddagger$	time step in the LM model output (1 h)	
ω	spatial or spatio-temporal weights	
$\mathbf{e}(t)$	vector of prognostic variables in the LM	3.2.1
$\mathbf{A}\big(\mathbf{e}(t)\big)$	non-parametrised processes of the LM	3.2.1
$\mathbf{P}\big(\mathbf{e}(t);\alpha\big)$	parametrised processes of the LM; net diabatic forcing	3.2.1
α	fixed parameters of a parametrisation scheme	3.2.1
$\mathbf{P}'\big(\mathbf{e}(t);\alpha\big)$	perturbed parametrised processes	3.2.2
$\big\langle x(r,t)\big\rangle_{D,T}$	random number in the stochastic parametrisation scheme	3.2.2
r	grid point of the model	3.2.2
D, T	spatio-temporal auto-correlation of the random numbers	3.2.2

Formal Symbolism	Meaning	defined in Section
$\mathbf{e}_j$	prognostic variables within an ensemble member	3.2.2
$\langle x_j(r,t)\rangle_{D,T}$	series of random numbers that is unique to $\mathbf{e}_j$	3.2.2
$\mathbf{P}_{\mathrm{m/c/r}}(\mathbf{e}(t);\alpha)$	diabatic forcing from a selection of parametrised processes: microphysics, convection, radiation	3.2.2
$\mathbf{P}_{\mathrm{turb}}(\mathbf{e}(t);\alpha)$	diabatic forcing from the turbulence scheme	3.2.3
a	amplitude parameter of the input noise	3.2.4
Q	total radiative flux	3.2.5
Q_{sol}	short-wave radiation flux	3.2.5
Q_{th}	long-wave radiation flux	3.2.5
$\frac{\partial T_{\mathrm{sol}}}{\partial t}$	temperature tendency due to solar radiation	3.2.5
z	height	3.2.5
Q^v_{sol}	virtual short-wave radiation flux which is energetically consistent with the perturbed simulation	3.2.5
$\mathbf{P}_{\mathrm{rough}}(\mathbf{e}(t);\alpha)$	diabatic forcing from turbulent vertical mixing governed by roughness length	3.2.6
$z_0(r)$	original LM roughness length at grid point r	3.2.6
$z_0'(r)$	perturbed roughness length at grid point r	3.2.6
$\langle y(r)\rangle$	random number	3.2.6
$\tilde{r}$	location near r	3.2.6
y_t	Gaussian random number used in the stochastic parametrisation scheme	3.4.1
x_t	uniformly distributed random number used in the stochastic parametrisation scheme	3.4.1
η^τ	auto-correlation for time lag τ	3.4.1
T	e-folding time of the auto-correlation	3.4.1
$\mathcal{N}$	all points of the discrete model domain $\mathcal{N} = \{1,\ldots,N_x\}\times\{1,\ldots,N_y\}\times\{1,\ldots,N_t\}$	4.2
N_x, N_y, N_t	number of grid points in space and time	4.2
$\mathcal{U}$	spatio-temporal neighbourhood, set of points	4.2
ρ_{xy}, ρ_x, ρ_y, ρ_t	size of neighbourhood	4.2
Z	random variable	4.2
z	realisation of a random variable direct model output of the predictand	4.2
$\{z_1,\ldots,z_N\}$	*surrogate* ensemble of Z from the neighbourhood	4.2
$\mathcal{S}$	sample of Z, $\mathcal{S} = \{z_1,\ldots,z_N\}$	4.2
N	number of points within $\mathcal{U}$, sample size of $\mathcal{S}$	4.2
π^c	exceedance probability with threshold c: $\Pr[Z > c]$	4.2
π	exceedance probability	4.2
π^c	exceedance probability related to threshold c	4.2
$\tilde{\delta}^c$	function $\tilde{\delta}^c : \mathbb{R} \to \{0,1\}$ referring to a threshold c	4.2
c	threshold value	4.2

Formal Symbolism	Meaning	defined in Section
z°	random variable of a discrete distribution with two possible outcomes	4.2
F	probability distribution function	4.2
$\mathcal{A}$	event, subspace of the probability space	4.2
$\bar{\mathcal{A}}$	event, so that $\mathcal{A}$ and $\bar{\mathcal{A}}$ are mutually exclusive	4.2
ζ	(unknown) model-output function without noise	4.3
x_i, y_j, t_k	spatial and temporal grid locations on a continuously defined model output field	4.3
K_{ball}	spherical Kernel function, analogue to simple neighbourhood	4.3
u, v, w	auxiliary variables in the integration of the kernel	4.3
V_{ball}	volume	4.3
$\breve{Z}$	refined version of Z	4.4
$\breve{\mathcal{U}}$	refined version of $\mathcal{U}$	4.4
$\breve{\mathcal{S}}$	refined version of $\mathcal{S}$	4.4
$\boldsymbol{d}$	vector of explanatory variables	4.4
N_d	number of explanatory variables	4.4
g	function or relation between explanatory variables and model output	4.4
$\boldsymbol{\beta}$	parameter vector describing the explanatory effect	4.4
β°	offset parameter	4.4
γ_{expl}	smoothing parameter in the regression of the explanatory effect	4.4
$\breve{\mathcal{U}}$	refined version of the neighbourhood	4.4
$\mathcal{U}^{\text{cutoff}}$	neighbourhood with a cut-off	4.4
$\mathcal{U}^{\text{distort}}$	distorted neighbourhood	4.4
$c_{s,s'}$	criterion for the distortion of the neighbourhood	4.4
$o_{s,s'}$	orographic term in the criterion for the distortion	4.4
$R_{s,s'}$	distortion matrix	4.4
$\tilde{R}_{s,s'}$	preliminary distortion matrix	4.4
$\boldsymbol{e}_1$, $\boldsymbol{e}_2$	eigenvalues of the matrix $R_{s,s'}$	4.4
$\tilde{\boldsymbol{e}}_1$, $\tilde{\boldsymbol{e}}_2$	eigenvalues of the matrix $\tilde{R}_{s,s'}$	4.4
λ_1, λ_2	eigenvalues of the matrix $R_{s,s'}$	4.4
$\tilde{\lambda}_1$, $\tilde{\lambda}_2$	eigenvalues of the matrix $\tilde{R}_{s,s'}$	4.4
ϖ_{intpol}	spatial weighting parameter for bilinear spatial interpolation at points with missing values	4.4
p	probability $p \in [0, 1] \subset \mathbb{R}$	4.5
$Z_{(l)}$	lth order statistic	4.5
$z_{(l)}$	realisation of lth order statistic	4.5
q	quantile function, inverse of the distribution F	4.5
q_e	empirical quantile function	4.5

Formal Symbolism	Meaning	defined in Section
ε	random number in the definition of the empirical quantile function	4.5
α	parameter in the plotting position formula	4.5.2
p_l	discrete values of probability p laid down by the plotting position formula	4.5.2
κ_l	interpolating series of p_l	4.5.2
K	kernel function in quantile regression	4.5.3
t	kernel argument with $t \in [-1, 1]$	4.5.3
s	integrating variable, corresponding to the p-axis	4.5.3
K_r	boundary kernel function	4.5.3
r	parameter of the boundary kernel	4.5.3
h_p	kernel bandwidth in quantile regression	4.5.3
n	number of realisations $z > 0$	4.5.5
F°	discrete distribution function with two possible outcomes	4.5.5
F^+	continuous distribution function defined on $\mathbb{R}_0^+$	4.5.5
q^+	quantile function of non-zero precipitation, inverse of the distribution F^+	4.5.5
n	integer variable related to the number of spatial grid points on the model domain: $(2^n + 1) \times (2^n + 1)$	5.2
l	level (= scale) in wavelet smoothing; dilation index $(l = 0, \ldots, -n)$	5.2 and A
m	location in wavelet smoothing; translation index $(m = 0, \ldots, 2^n)$	5.2 and A
m_1, m_2	location referring to one of the two horizontal directions, respectively	5.2 and A
h, v, d	indices referring to the horizontal, diagonal, vertical direction, respectively	5.2 and A
γ_{l,m_1,m_2}	wavelet coefficient	5.2 and A
$\gamma^{\text{mod}}_{l,m_1,m_2}$	modified (= de-noised) wavelet coefficient	5.2
$\Psi_{m_1,m_2,l}$	wavelet function	5.2 and A
$\Phi_{m_1,m_2,l}$	scaling function	5.2 and A
V_l, W_l, O_l	subspaces in multiresolution analysis	5.2 and A
Υ	Fourier transformation of a function	A
ω	frequency	A
Wf	Wavelet transform of a function f	A
$\psi_{\mu,x}(u)$	continuous wavelet function	A
l_{curtail}	curtailing level	5.3
K_{rect}	rectangular Kernel function (2D)	5.3
$\mathcal{O}$	range of different origins; range of shifts	B
i_s	choice of origin	B
j_s	choice of origin	B

Formal Symbolism	Meaning	defined in Section
$\tilde{z}_{ij}(i_s, j_s)$	result of wavelet de-noising as a function of the origin	B
$\tilde{z}_{ij}^{\text{invar}}$	translation invariant estimator	B
$\mathcal{O}_{\text{nonred}}$	range of non-redundant origins	B
$\lvert\mathcal{O}_{\text{nonred}}\rvert$	number of non-redundant origins	B
i_a, j_a	spatial offset of shifts	B
$\hat{\sigma}_{ij}^{\text{shift}}$	square root of the estimated variability between differently shifted smoothing results	B
ν	threshold value in wavelet thresholding	5.4
ν_l	level-dependent threshold in wavelet thresholding	5.4
$\tilde{\nu}$	threshold that depends on an additional field	5.4
γ_{l,m_1,m_2}^{z}	wavelet coefficients of an additional field	5.5
ETS	equitable threat score	6.3.1 and C
FAR	false alarm rate	6.3.1 and C
FBI	frequency bias	6.3.1 and C
OR	log odds ratio	6.3.1 and C
LOR(BC)	log odds ratio with bias correction	6.3.1 and C
OR	odds ratio	6.3.1 and C
POD	probability of detection	6.3.1 and C
a	number of hits in contingency table	C
b	number of false alarms in contingency table	C
c	number of misses in contingency table	C
d	number of correct rejections in contingency table	C
rmse	root mean square error	6.3.2 and C
corr	correlation between observations and forecasts	6.3.2 and C
stdv_fc	standard deviation of the forecast data	6.3.2 and C
stdv_obs	standard deviation of the observational data	6.3.2 and C
mae	mean absolute error	6.3.2 and C
skill	skill score that relates the *rmse* to a reference	6.3.2 and C
skill_DMO	skill score that relates the *rmse* to the DMO	6.3.2 and C
BS	Brier score	6.3.4 and C
BS_{rel}	reliability term of the BS	6.3.4 and C
BS_{res}	resolution term of the BS	6.3.4 and C
BS_{unc}	uncertainty term of the BS	6.3.4 and C
$\text{BS}_{\text{climate}}$	BS of the climate forecast	6.3.4 and C
BSS	Brier skill score	6.3.4 and C
p_t	threshold probability in the decision making process	6.4.2 and C
C	cost	6.4.2 and C
L	loss	6.4.2 and C
E	expenses in cost-loss scenario	C
V	relative value of the forecast	C

E List of acronyms

Acronym	Meaning	defined in Chapter
2m-temperature	temperature in 2 m height above the ground	1
2D	two-dimensional	6
3D	three-dimensional	6
BS	Brier score	6
BS_{rel}	reliability term of the BS	6
BS_{res}	resolution term of the BS	6
BS_{unc}	uncertainty term of the BS	6
$BS_{climate}$	BS of the climate forecast	6
$BS_{reference}$	BS of a reference forecast	C
BSS	Brier skill score	6
CAPE	convective available potential energy	3
corr	correlation between observations and forecasts	6
COSMO	COnsortium for Small Scale MOdeling	1
DM	high-resolution model, predecessor of the LM at DWD	2
DMO	direct model output	4
DWD	German Weather Service (Deutscher Wetterdienst)	1
E	geographical longitude (east)	6
ECMWF	European Centre for Medium-Range Weather Forecasts	3
ETS	equitable threat score	6
FAR	false alarm rate	6
FBI	frequency bias	6
GME	global NWP model of DWD	2
HHH	parameter setting of experimental ensemble: *high-high-high*	3
HLL	parameter setting of experimental ensemble: *high-low-low*	3
iid	independent and identically distributed	5
LAM	limited-area model	1
LLL	parameter setting of experimental ensemble: *low-low-low*	3

Acronym	Meaning	defined in Chapter
LEPS	limited-area ensemble prediction system	1
LM	Lokal-Modell	1
LOR	log odds ratio	6
LOR(BC)	log odds ratio with bias correction	6
mae	mean absolute error	6
MLL	parameter setting of experimental ensemble: *medium-low-low*	3
MISE	mean integrated squared error	4
MOS	model output statistics	1
MPI	message passing interface	2
MRA	multiresolution analysis	5 and A
N	geographical latitude (north)	6
NAG	Numerical Algorithms Group Ltd	3
NCEP	National Centres for Environmental Prediction	1
neighbourhood PP	neighbourhood postprocessing	6
NWP	numerical weather prediction	1
obs	observation	6
OR	odds ratio	6
pdf	probability density function	1
POD	probability of detection	6
rmse	root mean squared error	6
ROC	relative operating characteristic	6
RR	precipitation at the ground	6
skill	skill score based on the root mean squared error of continuous forecasts	6
skill_DMO	*skill* that relates to the direct model output of a single simulation as a reference	6
SREF	short-range ensemble forecasting	1
stdv_fc	standard deviation of the forecast data	6
stdv_obs	standard deviation of the observational data	6
SVAT	soil-vegetation-atmosphere-transfer	2
SYNOP	surface synoptic observation	6
T	temperature in 2m height	6
UTC	co-ordinated universal time	3
WMAE	weighted mean absolute error	C
h	hour	
hPa	hectopascal	
K	Kelvin	
km	kilometer	
m	meter	
mm	millimeter	
s	second	

Bibliography

Abramovich, F. and Benjamini, Y. (1995). Thresholding of wavelet coefficients as multiple hypothesis testing procedure. *in* A. Antoniadis and G. Oppenheim (eds), *Wavelets and Statistics*. Vol. 103 of *Lecture Notes in Statistics*. Springer. New York.

Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998): Wavelet thresholding via a Bayesian approach. *J. Roy. Statist. Soc.* **B 60**, 725–749.

Abry, P., Flandrin, P., Taqqu, M. S. and Veitch, D. (2000). Wavelets for the analysis, estimation and synthesis of scaling data. *Self Similar Network Traffic Analysis and Performance Evaluation*. Wiley. New York.

Ament, F. (2001). Initialisierung von Wolken im Lokal-Modell aus Meteosat-Messungen. Diploma thesis. Meteorological Institute, University of Bonn, Germany.

Anderson, C. J., W. A. Gallus, J., Arritt, R. W. and Kain, J. S. (2002). Impact of adjustments in the Kain-Fritsch convective scheme on QPF of elevated convection. *Preprints, 15th Conference on NWP, San Antonio, TX, Aug 12–16*. 23–24.

Anthes, R. A., Y.-H. Kuo, D. P. B., Errico, R. M. and Bettge, T. W. (1985). Predictability of mesoscale atmospheric motions. *Issues in Atmospheric and Oceanic Modeling*. Vol. 28 Part B of *Advances in Geophysics*. Academic Press. 159–202.

Applequist, S., Gahrs, G. E. and Pfeffer, R. L. (2002): Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Wea. Forecasting* **17**, 783–799.

Arakawa, A. (1966): Computational design for long term numerical integrations of the equations of fluid motion: 1. Two-dimensional incompressible flow. *J. Comp. Phys.* **1**, 119–143.

Arritt, R. W., Anderson, C. J. and W. A. Gallus, J. (2001). Ensemble forecasts using perturbed physics in a multidimensional parameter space. *Preprints, 14th Conference on Numerical Weather Prediction, Ft. Lauderdale, FL*.

Atger, F. (2001): Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlin. Proc. Geophys.* **8**, 401–417.

Balzer, K. (1998). Aktuelle Herausforderungen und erste Antworten. *in* K. Balzer, W. Enke and W. Wehry (eds), *Wettervorhersage*. Springer. Berlin. 71–92.

Benjamini, Y. and Hochberg, Y. (1995): Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc.* **57**, 289–300.

Bernardet, L. R., Grasso, L. D., Nachamkin, J. E., Finley, C. A. and Cotton, W. R. (2000): Simulating convective events using a high-resolution mesoscale model. *J. Geophys. Res.* **105**(D11), 14963–14982.

BMBF (2004). *Research for climate protection and protection from climate impacts: A contribution to the BMBF framework programme "Research for Sustainability"*. Federal Ministry of Education and Research (BMBF), Publications and Website Division. Bonn, Berlin.

Bremnes, J. B. (2004): Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.* **132**, 338–347.

Briggs, W. M. and Levine, R. A. (1997): Wavelets and field forecast verification. *Mon. Wea. Rev.* **125**, 1329–1341.

Bright, D. R. and Mullen, S. L. (2002): Short-range ensemble forecasts of precipitation during the Southwest Monsoon. *Wea. Forecasting* **17**, 1080–1100.

Brooks, H. E. and Doswell III, C. A. (1993): New technology and numerical weather prediction: A wasted opportunity? *Weather* **48**, 173–177.

Brooks, H. E., Tracton, M. S., Stensrud, D. J., DiMego, G. and Toth, Z. (1995): Short-range ensemble forecasting: Report from a workshop, 25–27 July 1994. *Bull. Amer. Meteor. Soc.* **76**(9), 1617–1624.

Buizza, R., Miller, M. and Palmer, T. N. (1999): Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Q. J. R. Meteorol. Soc.* **125**, 2887–2908.

Cacciamani, C., Cesari, D., Grazzini, F., Paccagnella, T. and Pantone, M. (2000): Numerical simulation of intense precipitation events south of the Alps: Sensitivity to initial conditions and horizontal resolution. *Meteorol. Atmos. Phys.* **72**, 147–159.

Cam, L. L. (1961). A stochastic description of precipitation. *in* J. Neyman (ed.), *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability.* Vol. 3. Berkeley, CA. 165–186.

Casati, B., Stephenson, D. B. and Ross, G. (2004): A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorol. Appl.* **11**, 141–154.

Chang, S. G., Yu, B. and Vetterli, M. (2000): Wavelet thresholding for multiple noisy image copies. *IEEE Transactions on Image Processing* **9**(9), 1631–1635.

Chilès, J.-P. and Delfiner, P. (1999): *Geostatistics: Modeling Spatial Uncertainty.* Wiley series in probability and statistics. Wiley. New York.

Chomé, F., Vannitsem, S. and Nicolis, C. (2002): Intrinsic dynamics of the Eta regional model: Role of the domain size. *Meteorol. Z.* **11**, 403–408.

Chui, C. K. (ed.) (1992): *An Introduction to Wavelets*. Vol. 1 of *Wavelet Analysis and its Applications*. Academic Press. San Diego.

Cohen, A. (1992). Biorthogonal wavelets. *in* C. K. Chui (ed.), *Wavelets: A Tutorial in Theory and Applications*. Vol. 2 of *Wavelet Analysis and its Applications*. Academic Press. San Diego.

Cohen, A., Daubechies, I. and Feauveau, J. (1992): Biorthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math.* **45**, 485–560.

Coifman, R. R. and Donoho, D. L. (1995). Translation-invariant de-noising. *in* A. Antoniadis and G. Oppenheim (eds), *Wavelets and Statistics*. Vol. 103 of *Lecture Notes in Statistics*. Springer. New York.

Colle, B. A., Mass, C. F. and Westrick, K. J. (2000): MM5 precipitation verification over the Pacific Northwest during the 1997-99 cool seasons. *Wea. Forecasting* **15**, 730–744.

Craig, P. S., Goldstein, M., Rougier, J. C. and Seheult, A. H. (2001): Bayesian forecasting for complex systems using computer simulators. *J. Amer. Statist. Assoc.* **96**, 717–729.

Cressie, N. (1991): *Statistics for Spatial Data*. Wiley. New York.

Cunnane, C. (1978): Unbiased plotting positions - a review. *J. Hydrol.* **37**, 205–222.

Daubechies, I. (1992): *Ten Lectures on Wavelets*. SIAM. Philadelphia.

David, H. A. (1981): *Order Statistics*. Wiley Series in Probability and Mathematical Statistics. Wiley. New York.

Davies, H. C. (1976): A lateral boundary formulation for multilevel prediction models. *Q. J. R. Meteorol. Soc.* **102**, 405–418.

Davis, C. and Carr, F. (2000): Summary of the 1998 workshop on mesoscale model verification. *Bull. Amer. Meteor. Soc.* **81**, 809–819.

de Elía, R., Laprise, R. and Denis, B. (2002): Forecasting skill limits of nested, limited-area models: A perfect-model approach. *Mon. Wea. Rev.* **130**, 2006–2023.

Diggle, P. (1985): A kernel method for smoothing point process data. *Appl. Statistics* **34**, 138–147.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998): Model-based geostatistics. *Appl. Statistics* **47**, 299–350.

Do, M. N., Dragotti, P. L., Shukla, R. and Vetterli, M. (2001). On the compression of two-dimensional piecewise smooth functions. *IEEE International Conference on Image Processing (ICIP), Thessaloniki, Greece, 2001*.

Doms, G. and Schättler, U. (1999). The nonhydrostatic limited-area model LM (Lokal-Modell) of DWD. Part I: Scientific documentation. *Technical report*. German Weather Service (DWD), Research Department, P. O. 100465, 63004 Offenbach, Germany.

Donoho, D. L. and Johnstone, I. M. (1994): Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81**, 425–455.

Donoho, D. L. and Johnstone, I. M. (1998): Minimax estimation via wavelet shrinkage. *Ann. Statist.* **26**(3), 879–921.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995): Wavelet shrinkage: Asymptotia? *J. Roy. Statist. Soc.* **B 57**, 301–369.

Du, J., Mullen, S. L. and Sanders, F. (1997): Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.* **125**, 2427–2459.

Duynkerke, P. G. (1992): The roughness length for heat and other vegetation parameters for a surface of short grass. *J. Appl. Meteorol.* **31**, 579–586.

Ebert, E. E. and McBride, J. L. (2000): Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.* **239**, 179–202.

Eckel, F. A. and Walters, M. K. (1998): Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting* **13**, 1132–1147.

Ehrendorfer, M. (1997): Predicting the uncertainty of numerical weather forecasts: a review. *Meteorol. Z.* **6**, 147–183.

Epanechnikov, V. A. (1969): Nonparametric estimation of a multidimensional probability density. *Theor. Probab. Appl.* **14**, 153–158.

Evans, R. E., Graham, R. J., Harrison, M. S. J. and Shutts, G. J. (1998). Preliminary investigations into the effect of stochastic backscatter to the Unified Model. *Forecasting Research Technical Report 241.* The Met Office, UK.

Ewald, B., Penland, C. and Temam, R. (2003): Accurate integration of stochastic climate models with application to El Niño. *Mon. Wea. Rev.* **132**, 154–164.

Farge, M. (1992): Wavelet transforms and their applications to turbulence. *Ann. Rev. Fluid. Mech.* **24**, 395–457.

Flandrin, P. (1992): Wavelet analysis and synthesis of fractional Brownian motion. *IEEE Transactions on Information Theory* **38**, 910–917.

Folland, C. and Anderson, C. (2002): Estimating changing extremes using empirical ranking methods. *J. Climate* **15**, 2954–2960.

Fournier, A. (2000): Introduction to orthonormal wavelet analysis with shift invariance: Application to observed atmospheric blocking spatial structure. *J. Atmos. Sci.* **57**, 3856–3879.

Fritsch, J. M., Houze Jr., R. A., Adler, R., Bluestein, H., Bosart, L., Brown, J., Carr, F., Davis, C., Johnson, R. H., Junker, N., Kuo, Y.-H., Rutledge, S., Smith, J., Toth, Z., Wilson, J. W., Zipser, E. and Zrnic, D. (1998): Quantitative precipitation forecasting: Report of the Eighth Prospectus Development Team, U.S. Weather Research Program. *Bull. Amer. Meteor. Soc.* **79**(2), 285–299.

Frogner, I.-L. and Iversen, T. (2002): High-resolution limited-area ensemble predictions based on low-resolution targeted singular vectors. *Q. J. R. Meteorol. Soc.* **128**, 1321–1341.

Gardiner, C. W. (1997): *Handbook of Stochastic Methods*. Springer. Berlin.

Gasser, T. and Müller, H.-G. (1984): Estimating regression functions and their derivatives by the kernel method. *Scand. J. Stat.* **11**, 171–185.

Gasser, T., Kneip, A. and Kohler, W. (1991): A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc.* **86**, 643–652.

Gerstner, T. (1998). Adaptive hierarchical methods for landscape representation and analysis. *in* S. Hergarten and H. J. Neugebauer (eds), *Process Modelling and Landform Evolution*. Springer. Berlin.

Gigerenzer, G. and Edwards, A. (2003): Simple tools for understanding risks: From innumeracy to insight. *British Medical Journal* **327**, 741–744.

Glahn, H. R. and Lowry, D. A. (1972): Use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.* **11**, 1203–1211.

Göber, M., Wilson, C. A., Milton, S. F. and Stephenson, D. B. (2004): Fairplay in the verification of operational quantitative precipitation forecasts. *J. Hydrol.* **288**, 225–236.

Göttelmann, J. (1999): Compression of atmospheric data by spherical wavelets. *Contr. Atmos. Phys.* **72**(1), 95–104.

Grasso, L. D. (2000): The differentiation between grid spacing and resolution and their application to numerical modeling. *Bull. Amer. Meteor. Soc.* **81**, 579–580.

Grimit, E. P. and Mass, C. F. (2002): Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting* **17**, 192–205.

Großmann, C. and Roos, H. G. (1994): *Numerik partieller Differentialgleichungen*. Teubner. Stuttgart.

Hamill, T. M. and Colucci, S. J. (1997). Perturbations to the land-surface conditions in short-range ensemble forecasts. *Preprints, 12th Conference on Numerical Weather Prediction (Phoenix)*. American Meteorological Society. 273–276.

Hamill, T. M. and Colucci, S. J. (1998): Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.* **126**, 711–724.

Hamill, T. M. and Wilks, D. S. (1995): A probabilistic forecast contest and the difficulty in assessing short-range forecast uncertainty. *Wea. Forecasting* **10**, 620–631.

Hamill, T. M., Hansen, J. A., Mullen, S. L. and Snyder, C. (2005). Workshop on ensemble forecasting in the short to medium range. Submitted to *Bull. Amer. Meteor. Soc.*

Hansen, J. A. (2002): Accounting for model error in ensemble-based state estimation and forecasting. *Mon. Wea. Rev.* **130**, 2372–2391.

Härdle, W. (1990): *Applied Nonparametric Regression*. Cambridge University Press. Cambridge.

Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998): *Wavelets, Approximation and Statistical Applications*. Vol. 129 of *Lecture Notes in Statistics*. Springer. New York.

Harris, D. and Foufoula-Georgiou, E. (2001): Multiscale statistical properties of a high-resolution precipitation forecast. *J. Hydrometeor.* **2**, 406–418.

Hasselmann, K. (1976): Stochastic Climate Models. Part I: Theory. *Tellus* **28**, 473–485.

Heise, E. (2002). Parametrisierungen. *Die neue Modellkette des DWD I*. Vol. 27(3/4) of *promet*. DWD. 130–141.

Hodges, K. I. (1999): Extension of spherical nonparametric estimators to nonisotropic kernels: An oceanographic application. *Mon. Wea. Rev.* **127**, 214–227.

Hoffmann, R. N., Liu, Z., Louis, J.-F. and Grassotti, C. (1995): Distortion representation of forecast errors. *Mon. Wea. Rev.* **123**, 2758–2770.

Högström, U. (1988): Non-dimensional wind and temperature profiles in the atmospheric surface layer: a re-evaluation. *Boundary-Layer Meteorol.* **42**, 55–78.

Hou, D., Kalnay, E. and Droegemeier, K. K. (2001): Objective verification of the SAMEX '98 ensemble forecasts. *Mon. Wea. Rev.* **129**, 73–91.

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

Imkeller, P. and von Storch, J.-S. (eds) (2001): *Stochastic Climate Models*. Vol. 49 of *Progress in Probability*. Birkhäuser. Basel.

Jansen, M. (2001): *Noise Reduction by Wavelet Thresholding*. Vol. 161 of *Lecture Notes in Statistics*. Springer. New York.

Jawerth, B. and Sweldens, W. (1994): An overview of wavelet based multiresolution analysis. *SIAM Rev.* **36**(3), 377–412.

Jensen, A. and la Cour-Harbo, A. (2001): *Ripples in Mathematics*. Springer. Berlin.

Judd, K. and Smith, L. A. (2001): Indistinguishable states II: Imperfect model scenario. *Physica D* **151**, 125–141.

Katul, G. and Vidakovic, B. (1998): Identification of low-dimensional energy containing/flux transporting eddy motion in the atmosphere surface layer using wavelet thresholding methods. *J. Atmos. Sci.* **55**(3), 377–389.

Kessler, E. (1969). On the distribution and continuity of water substance in atmospheric circulations. *Meteorological Monographs*. Vol. 10.

Klemp, J. B. and Wilhelmson, R. B. (1978): The simulation of three-dimensional convective storm dynamics. *J. Atmos. Sci.* **35**, 1070–1096.

Kloeden, P. E. and Platen, E. (1992): *Numerical Solution of Stochastic Differential Equations*. Springer. Berlin.

Koenker, R. and Hallock, K. F. (2001): Quantile regression. *Journal of Economic Perspectives* **15**, 143–156.

Krishnamurti, T. N., Kishtawal, C. M., LaRow, T., Bachiochi, D., Zhang, Z., Williford, C. E., Gadgil, S. and Surendran, S. (2000): Multimodel superensemble forecasts for weather and seasonal climate. *J. Climate* **13**, 4196–4216.

Krzysztofowicz, R. and Sigrest, A. A. (1999): Calibration of probabilistic quantitative precipitation forecasts. *Wea. Forecasting* **14**, 427–454.

Kumar, P. and Foufoula-Georgiou, E. (1993a): A multicomponent decomposition of spatial rainfall fields. 1. Segregation of large- and small-scale features using wavelet transforms. *Water Resour. Res.* **29**(8), 2515–2532.

Kumar, P. and Foufoula-Georgiou, E. (1993b): A multicomponent decomposition of spatial rainfall fields. 1. Self-similarity in fluctuations. *Water Resour. Res.* **29**(8), 2533–2544.

Kumar, P. and Foufoula-Georgiou, E. (1993c): A new look at rainfall fluctuations and scaling properties of spatial rainfall using orthogonal wavelets. *J. Appl. Meteor.* **32**, 209–222.

Kumar, P. and Foufoula-Georgiou, E. (1994). Wavelet analysis in Geophysics: An introduction. *in* E. Foufoula-Georgiou and P. Kumar (eds), *Wavelets in Geophysics*. Vol. 4 of *Wavelet Analysis and its Applications*. Academic Press. San Diego.

Kumar, P. and Foufoula-Georgiou, E. (1997): Wavelet analysis for geophysical applications. *Rev. Geophys. Res.* **35**, 385–412.

Laprise, R., Varma, M. R., Denis, B., Caya, D. and Zawadzki, I. (2000): Predictability of a nested limited-area model. *Mon. Wea. Rev.* **128**, 4149–4145.

Lin, J. W.-B. and Neelin, J. D. (2000): Influence of a stochastic moist convective parameterization on tropical climate variability. *Geophys. Res. Lett.* **27**(22), 3691–3694.

Lin, J. W.-B. and Neelin, J. D. (2002): Considerations for stochastic convective parameterization. *J. Atmos. Sci.* **59**, 959–975.

Lorenz, E. N. (1963): Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 409–418.

Lorenz, E. N. (1969): The predictability of a flow which possesses many scales of motion. *Tellus* **21**, 289–307.

Louis, J. F. (1979): A parametric model of vertical eddy fluxes in the atmosphere. *Boundary Layer Meteorol.* **17**, 187–202.

Madych, W. R. (1992). Some elementary properties of multiresolution. *in* C. K. Chui (ed.), *Wavelets: A Tutorial in Theory and Applications*. Vol. 2 of *Wavelet Analysis and its Applications*. Academic Press. San Diego.

Marsigli, C., Montani, A., Nerozzi, F., Paccagnella, T., Tibaldi, S., Molteni, F. and Buizza, R. (2001): A strategy for high-resolution ensemble prediction. II: Limited-area experiments in four Alpine flood events. *Q. J. R. Meteorol. Soc.* **127**, 2095–2115.

Mason, S. J. (2004): On using "climatology" as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.* **132**, 1891–1895.

Masry, E. (1993): The wavelet transform of stochastic processes with stationary increments and its application to fractional Brownian motion. *IEEE Transactions on Information Theory* **39**, 260–264.

Mass, C. and Kuo, Y.-H. (1998): Regional real-time numerical weather prediction: Current status and future potential. *Bull. Amer. Meteor. Soc.* **79**, 253–263.

Mass, C. F., Ovens, D., Westrick, K. and Colle, B. A. (2002): Does increasing horizontal resolution produce better forecasts? *Bull. Amer. Meteor. Soc.* **83**(3), 407–430.

Mellor, G. L. and Yamada, T. (1974): A hierarchy of turbulence closure models for planetary boundary layers. *J. Atmos. Sci.* **31**, 1791–1806.

Misra, V. and Yau, M. K. (2001): An ensemble strategy for high-resolution regional model forecasts. *Meteorol. Atmos. Phys.* **78**, 61–74.

Mölder, M. (1998). Roughness lengths and roughness sublayer corrections in partly forested regions. *Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 345*.

Mölders, N. (2001): On the uncertainty in mesoscale modeling caused by surface parameters. *Meteorol. Atmos. Phys.* **76**, 119–141.

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. (1996): The ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119.

Moon, Y.-I. and Lall, U. (1994): Kernel quantile function estimator for flood frequency analysis. *Water Resour. Res.* **30**(11), 3095–3103.

Mullen, S. L. and Buizza, R. (2000). Quantitative precipitation forecasts over the United States by the ECMWF Ensemble Prediction System. *Technical Memorandum 307*. ECMWF.

Müller, H.-G. (1991): Smooth optimum kernel estimators near endpoints. *Biometrika* **87**(3), 521–530.

Murphy, A. H. (1977): The value of climatological, categorical and probabilistic forecasts in the cost-loss situation. *Mon. Wea. Rev.* **105**, 803–816.

Murphy, A. H. (1979): Probabilistic temperature forecasts: The case for an operational program. *Bull. Amer. Meteor. Soc.* **60**, 12–19.

Murphy, A. H. (1993): What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting* **8**, 281–293.

Mylne, K. R. (2002): Decision-making from probability forecasts based on forecast value. *Meteorol. Appl.* **9**, 307–315.

Mylne, K. R., Evans, R. E. and Clark, R. T. (2002): Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting. *Q. J. R. Meteorol. Soc.* **128**, 361–384.

Nason, G. P. (1995). Choice of the threshold parameter in wavelet function estimation. *in* A. Antoniadis and G. Oppenheim (eds), *Wavelets and Statistics*. Vol. 103 of *Lecture Notes in Statistics*. Springer. New York.

Nason, G. P. and Silverman, B. W. (1995). The stationary wavelet transform and some statistical applications. *in* A. Antoniadis and G. Oppenheim (eds), *Wavelets and Statistics*. Vol. 103 of *Lecture Notes in Statistics*. Springer. New York.

Nicolis, C. (2003): Dynamics of model error: Some generic features. *J. Atmos. Sci.* **60**, 2208–2218.

Nutter, P. A. (2003). *Effects of nesting frequency and lateral boundary perturbations on the dispersion of limited-area ensemble forecasts*. PhD thesis. University of Oklahoma, School of Meteorology. Norman, OK.

Ogden, R. T. (1997): *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser. Boston.

Orrell, D., Smith, L., Barkmeijer, J. and Palmer, T. (2001): Model error in weather forecasting. *Nonlin. Proc. Geophys.* **8**, 357–371.

Palmer, T. N. (2001): A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Q. J. R. Meteorol. Soc.* **127**, 279–304.

Palmer, T. N., Brankovic, C. and Richardson, D. S. (2000): A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Q. J. R. Meteorol. Soc.* **126**, 2013–2034.

Papanicolaou, G. and Kohler, W. (1974): Asymptotic theory of mixing stochastic ordinary differential equations. *Comm. Pure Appl. Math.* **27**, 641–668.

Penland, C. (2003): A stochastic approach to nonlinear dynamics. *Bull. Amer. Meteor. Soc.* **84**, ES43–ES52.

Perica, S. and Foufoula-Georgiou, E. (1996): Linkage of scaling and thermodynamic parameters of rainfall: Results from midlatitude mesoscale convective systems. *J. Geophys. Res.* **101**(D3), 7431–7448.

Pielke, R. A. (1991): A recommended specific definition of "resolution". *Bull. Amer. Meteor. Soc.* **72**, 1914.

Press, W., Teukolsky, S., Vetterling, W. and Flannery, B. (1992): *Numerical Recipes in FORTRAN: The art of scientific computing*. Cambridge University Press. Cambridge.

Prudhomme, C. and Reed, D. W. (1999): Mapping extreme rainfall in a mountainous region using geostatistical techniques: A case study in Scotland. *Int. J. Climatol.* **19**, 1337–1356.

Raible, C. C., Bischof, G., Fraedrich, K. and Kirk, E. (1999): Statistical single station short-term forecasting of temperature and probability of precipitation: Area interpolation and NWP combination. *Wea. Forecasting* **14**, 203–214.

Rajagopalan, B. and Lall, U. (1995): A kernel estimator for discrete distributions. *Nonparametric Statistics* **4**, 409–126.

Raschendorfer, M. (2001). The new turbulence parameterization of LM. *in* G. Doms and U. Schättler (eds), *COSMO Newsletter No. 1.* 89–97.

Reiss, R.-D. (1993): *A Course on Point Processes*. Springer. Berlin.

Richardson, D. S. (2000): Skill and economic value of the ECMWF Ensemble Prediction System. *Q. J. R. Meteorol. Soc.* **126**, 649–668.

Richardson, D. S. (2001a): Ensembles using multiple models and analyses. *Q. J. R. Meteorol. Soc.* **127**, 1847–1864.

Richardson, D. S. (2001b): Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Q. J. R. Meteorol. Soc.* **127**, 2473–2489.

Riphagen, H. A., Bruyere, C. L., Jordan, W., Poolman, E. R. and Gertenbach, J. D. (2002): Experiments with the NCEP Regional Eta Model at the South African Weather Bureau, with emphasis on terrain representation and its effect on precipitation predictions. *Mon. Wea. Rev.* **130**, 1246–1263.

Ritter, B. and Geleyn, J.-F. (1992): A comprehensive radiation scheme for numerical weather prediction models with potential applications in climate simulations. *Mon. Wea. Rev.* **120**, 303–325.

Rodriguez-Iturbe, I. and Eagleson, P. S. (1987): Mathematical models of rainstorm events in space and time. *Water Resour. Res.* **23**, 181–190.

Roulston, M. S. and Smith, L. A. (2003): Combining dynamical and statistical ensembles. *Tellus* **55 A**, 16–30.

Salsón, S. and Garcia-Bartual, R. (2003): A space-time rainfall generator for highly convective Mediterranean rainstorms. *Natural Hazards and Earth System Sciences* **3**, 103–114.

Scheirer, R. and Macke, A. (2001): On the accuracy of the independent column approximation in calculating the downward fluxes in the UVA, UVB, and PAR spectral ranges. *J. Geophys. Res.* **106**(D13), 14301–14312.

Schraff, C. and Hess, R. (2002). Datenassimilation für das LM. *Die neue Modellkette des DWD I.* Vol. 27(3/4) of *promet*. DWD. 156–164.

Schraff, C. H. (1997): Mesoscale data assimilation and prediction of low stratus in the alpine region. *Meteorol. Atmos. Phys.* **64**, 21–50.

Shutts, G. (2003). Some interpretations of stochastic physical parametrizations. *ECMWF Seminar Proceedings 2001: Key Issues in the Parameterisation of Subgrid Physical Processes*. 111–126.

Shutts, G. (2004). A stochastic kinetic energy backscatter algorithm for use in ensemble prediction systems. *Technical Memorandum 449*. ECMWF.

Silverman, B. W. (1986): *Density Estimation for Statistics and Data Analysis*. Chapman and Hall. London, New York.

Simonoff, J. S. (1996): *Smoothing Methods in Statistics*. Springer Series in Statistics. Springer. New York.

Smith, L. A. (2000). Disentangling uncertainty and error: On the predictability of nonlinear systems. *in* A. I. Mees (ed.), *Nonlinear Dynamics and Statistics*. Birkhäuser. Boston. 31–64.

Smith, L. A. (2003). Predictability past predictability present. *ECMWF Seminar Proceedings 2002: Predictability of weather and climate*. 219–242.

Smith, L. A., Ziehmann, C. and Fraedrich, K. (1999): Uncertainty dynamics and predictability in chaotic systems. *Q. J. R. Meteorol. Soc.* **125**, 2855–2886.

Stanski, H. R., Wilson, L. J. and Burrows, W. R. (1989). Survey of common verification methods in meteorology. *World Weather Watch Technical Report No. 8 WMO/TD No. 358*. WMO. Geneva.

Stensrud, D. J., Bao, J.-W. and Warner, T. T. (2000): Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.* **128**, 2077–2107.

Stensrud, D. J., Brooks, H. E., Du, J., Tracton, M. S. and Rogers, E. (1999): Using ensembles for short-range forecasting. *Mon. Wea. Rev.* **127**, 433–445.

Stephenson, D. B. (2000): Use of the "odds ratio" for diagnosing forecast skill. *Wea. Forecasting* **15**, 221–232.

Steppeler, J., Doms, G., Schättler, U., Bitzer, H. W., Gassmann, A., Damrath, U. and Gregoric, G. (2003): Meso-gamma scale forecasts using the nonhydrostatic model LM. *Meteorol. Atmos. Phys.* **82**, 75–96.

Stollnitz, E. J., DeRose, T. D. and Salesin, D. H. (1996): *Wavelets for Computer Graphics: Theory and Applications*. Morgan Kaufmann Publishers, Inc. San Francisco.

Stull, R. B. (1988): *An Introduction to Boundary Layer Meteorology*. Kluwer Academic Publishers. Dordrecht, Boston, London.

Sweldens, W. (1996): Wavelets and the lifting scheme: A five minute tour. *Z. Angew. Math. Mech.* **76 (Suppl. 2)**, 41–44.

Sweldens, W. (1997): The lifting scheme: A construction of second generation wavelets. *SIAM J. Math. Anal.* **29**(2), 511–546.

Taylor, K. E. (2001): Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.* **106**(D7), 7183–7192.

Theis, S., Hense, A., Damrath, U. and Renner, V. (2002). Ensemble prediction and statistical postprocessing of weather parameters for the LM. *in* G. Doms and U. Schättler (eds), *COSMO Newsletter No. 2*. 152–161.

Thompson, J. C. and Brier, G. W. (1955): The economic utility of weather forecasts. *Mon. Wea. Rev.* **83**, 249–254.

Tiedtke, M. (1989): A comprehensive mass flux scheme for cumulus parametrization in large-scale models. *Mon. Wea. Rev.* **117**, 1779–1800.

Torrence, C. and Compo, G. P. (1998): A practical guide to wavelet analysis, with significance and confidence testing. *Bull. Amer. Meteor. Soc.* **79**, 61–78.

Toth, Z. and Kalnay, E. (1993): Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.* **74**, 2317–2330.

Toth, Z. and Vannitsem, S. (2001). Model errors and ensemble forecasting. *Proceedings of the 8th ECMWF Workshop on Meteorological Operational Systems, Reading, UK, Nov 12–16*. 146–154.

Toth, Z., Zhu, Y. and Marchok, T. (2001): The use of ensembles to identify forecasts with small and large uncertainty. *Wea. Forecasting* **16**, 463–477.

Verhoef, A. (1997): Some practical notes on the parameter κB^{-1} for sparse vegetation. *J. Appl. Meteorol.* **36**, 560–572.

Vidakovic, B. (1998): Nonlinear wavelet shrinkage with Bayes rule and Bayes factors. *J. Amer. Statist. Assoc.* **93**, 173–179.

Vidakovic, B. (1999): *Statistical Modeling by Wavelets*. Wiley. New York.

Vidakovic, B. D., Katul, G. G. and Albertson, J. D. (2000): Multiscale denoising of self-similar processes. *J. Geophys. Res.* **105**, 27049–27058.

Wackernagel, H. (1998): *Multivariate Geostatistics: An Introduction with Applications*. Springer. Berlin.

Walser, A., Lüthi, D. and Schär, C. (2004): Predictability of precipitation in a cloud resolving model. *Mon. Wea. Rev.* **132**, 560–577.

Wandishin, M. S., Mullen, S. L., Stensrud, D. J. and Brooks, H. E. (2001): Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.* **129**, 729–747.

Waymire, E., Gupta, V. K. and Rodriguez-Iturbe, I. (1984): A spectral theory of rainfall intensity at the meso-β scale. *Water Resour. Res.* **20**, 1453–1465.

Wen, L., Yu, W., Lin, C. A., Beland, M., Benoit, R. and Delage, Y. (2000): The role of land surface schemes in short-range, high spatial resolution forecasts. *Mon. Wea. Rev.* **128**, 3605–3617.

Wikle, C. K., Berliner, L. M. and Cressie, N. (1998): Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics* **5**, 117–154.

Wilks, D. S. (1995): *Statistical Methods in the Atmospheric Sciences*. Academic Press. San Diego, London.

Wilks, D. S. (2002): Smoothing forecast ensembles with fitted probability distributions. *Q. J. R. Meteorol. Soc.* **128**, 2821–2836.

Xu, K. M., Arakawa, A. and Krueger, S. K. (1992): The macroscopic behaviour of cumulus ensembles simulated by a cumulus ensemble model. *J. Atmos. Sci.* **49**, 2402–2420.

Yano, J.-I., Moncrieff, M. W. and Wu, X. (2001): Wavelet analysis of simulated tropical convective cloud systems. Part II: Decomposition of convective-scale and mesoscale structure. *J. Atmos. Sci.* **58**, 868–876.

Zepeda-Arce, J., Foufoula-Georgiou, E. and Droegemeier, K. K. (2000): Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *J. Geophys. Res.* **105**, 10129–10146.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D. and Mylne, K. (2002): On the economic value of ensemble based weather forecasts. *Bull. Amer. Meteor. Soc.* **83**, 73–83.

Ziehmann, C. (2000): Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus A* **52**, 237–251.

Acknowledgements

At the Meteorological Institute of the University of Bonn (MIUB) I am very thankful to my supervisor Prof. Dr. Andreas Hense who initiated the ideas of this work. I thank him for many fruitful discussions, for his scientific advice and for his trust in my ability. Many thanks go to Prof. Dr. Jin-Song von Storch (Max Planck Institute for Meteorology) and to Prof. Dr. Heini Wernli (Institute for Atmospheric Physics, University of Mainz) who kindly accepted to be co-examiners of my PhD thesis.

The funding of this work was mainly provided by the German Weather Service (DWD). I thank Prof. Dr. Andreas Hense (MIUB) and Dr. Volker Renner (DWD) for initiating and preparing the proposals. The continuous collaboration with Dr. Volker Renner and Dr. Ulrich Damrath (DWD) had a formative influence on my work. I also thank DWD for giving me free access to the model LM, the computing and data facilities and for the technical support.

I would like to acknowledge Dr. Thomas Gerstner (Institute for Numerical Simulation of the University of Bonn) who kindly provided the source code of the wavelet transformation and helped me to understand its underlying ideas.

The friendly working atmosphere at MIUB allowed for many informal and creative discussions that fostered my scientific way of thinking. Especially the "LM group" enabled mutual assistance in all matters concerning the LM. I am much obliged to the staff at MIUB for the help I received regarding library, computer and administrative matters. Special thanks go to my friends at MIUB and to my brother Jochen who kindly proofread the manuscript of my thesis.

I wish to thank Matthias for his love, support and patience during my PhD studies. I am much devoted to my parents who gave me complete freedom in my decision making and always supported my aims without reservation.