BONNER METEOROLOGISCHE ABHANDLUNGEN

Heft 93 (2022) (ISSN 0006-7156)
Herausgeber: Andreas Hense

Sebastian Buschow

Spatial Verification with Wavelets

BONNER METEOROLOGISCHE ABHANDLUNGEN

Heft 93 (2022) (ISSN 0006-7156)
Herausgeber: Andreas Hense

Sebastian Buschow

Spatial Verification with Wavelets

Spatial Verification with Wavelets

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
M.Sc
Sebastian Buschow
aus
Gladbeck

Bonn, Februar, 2022

Diese Arbeit ist die ungekürzte Fassung einer der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn im Jahr 2022 vorgelegten Dissertation von Sebastian Buschow aus Gladbeck.

This paper is the unabridged version of a dissertation thesis submitted by Sebastian Buschow born in Gladbeck to the Faculty of Mathematical and Natural Sciences of the Rheinische Friedrich-Wilhelms-Universität Bonn in 2022.

Anschrift des Verfassers: Address of the author:

Sebastian Buschow
Institut für Geowissenschaften, Abt. Meteorologie
Universität Bonn
Auf dem Hügel 20
D-53121 Bonn

1. Gutachter: PD. Dr. Petra Friederichs,
Rheinische Friedrich-Wilhelms-Universität Bonn
2. Gutachter: Prof. Dr. Andreas Hense,
Rheinische Friedrich-Wilhelms-Universität Bonn
3. Gutachter: Prof. Dr. Henning Rust,
Freie Universität Berlin

Tag der Promotion: 20. Januar 2022

Abstract

Modern numerical weather prediction models can simulate atmospheric processes at a resolution of single kilometers. With growing complexity, the realism of these simulations becomes increasingly difficult to quantify. Simply put, more details also leave more room for diverse kinds of errors, especially in the hardly predictable locations of small-scaled features. Traditional scores that compare forecast and observation grid-point by grid-point tend to prefer smoother, less detailed predictions and fail to appraise the realism of the simulated spatial structure. This thesis explores novel verification methods based on image filters, which isolate components at individual spatial scales and locations. These so-called *wavelets* are widely used in image processing and computer vision. In the context of meteorological forecast verification, wavelets were previously employed to remove noise, or split up the overall error into small- and large-scale contributions. Pursuing a different direction, this study demonstrates how wavelets can extract specific information about the scale-structure, directedness and preferred orientation of the fields to be compared. The result is a series of scores which translate the abstract information resulting from the wavelet transform into robust, easily interpretable statements about the realism of the simulated correlation structure. Directional aspects in particular – predicted features being too linear, too round or oriented at the wrong angle – are not explicitly treated by most existing verification tools. In addition, it is shown how the wavelets' localized nature can be exploited to visualize the local correlation structure on a map, quantify spatially varying displacements, or correct structural errors in a simple post-processing algorithm. Unlike other popular approaches in the literature, the new techniques are not limited to the special case of precipitation verification. Provided that observations on a regular grid exist, wavelet-based scores can, in principle, be applied to any meteorological field of interest.

Contents

Chapter 1

Introduction

1.1 Motivation: The need for spatial verification

When Charney, Fjörtoft and von Neumann performed the first successful numerical weather prediction in 1950, they modeled the upper level flow over the entire North American continent on a 15×18 grid. Their one-day simulation of a single variable on a single vertical level at a horizontal resolution of 736 km was computed in approximately 24 hours, allowing them to "just [...] keep pace with the weather" (Charney et al., 1950). Seventy years later, the US National Centers for Environmental Prediction (NCEP) currently operate a regional model for a similar domain centered on the United States, which uses no fewer than 954×835 grid boxes on 51 vertical levels. The dynamical core of this "Rapid Refresh" system (RAP) simulates the evolution of pressure, temperature, wind velocity and moisture without the barotropic, quasi-geostrophic or hydrostatic approximations of its predecessors (Benjamin et al., 2016). Physical processes which occur on scales smaller than the 13 km grid spacing are represented by empirical algorithms, the so-called parametrizations. Similar models are operated by many weather services around the world.

It is undisputed that the progress in scientific understanding and raw computational resources has tremendously improved the realism of our simulations and the quality of our forecasts. But will this trend continue indefinitely into the future? Beyond RAP, NCEP employs an even higher resolved system called HRRR (high resolution rapid refresh), covering the contiguous United States with a 3 km grid. At this resolution, processes related to atmospheric convection, i.e., rapid buoyancy-driven vertical exchange in thunderstorms, leave the realm of parametrization and begin to be simulated explicitly by the dynamical core of the model. The ability to resolve such high-impact weather events is a main incentive for the development of so-called Meso-Scale models like HRRR. Has their goal been achieved? How much better are these computationally expensive ultra high-resolution models?

In the atmospheric sciences, the process of quantitatively evaluating a model against observations is called *verification*. To check that their experiment had succeeded, Charney et al. (1950) simply plotted the 24 h forecast beside the real, observed weather map and compared them by eye. This simple verification technique, while still popular today, is difficult to quantify or reproduce. On a complicated weather map, different experts will focus on different aspects of the forecast and may even change their mind when presented with the same verification task again. To address the need for an objective, repeatable verification procedure, researchers

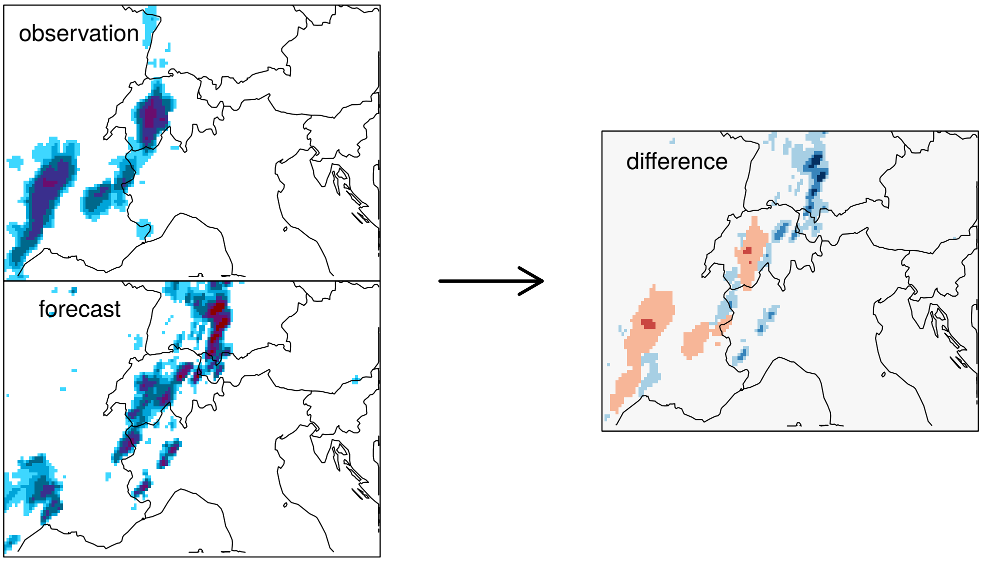


Figure 1.1: Point-wise verification of a precipitation forecast. Left: Predicted and observed hourly rain accumulation on 2007-07-21 16 UTC (MesoVICT case 2). Right: Difference image.

have developed numerous so-called scores, i.e., functions that quantify the difference between the predicted and observed values of one or more variables. The theoretical foundation of modern verification methods, based on the joint probability distribution of forecast and observation, was laid out by Murphy & Winkler (1987).

When the benefits of highly resolved modeling systems like HRRR are under review, verification studies often focus on precipitation forecasts for several reasons: Firstly, rain and snowfall events have a high impact on human activities including agriculture and aviation, as well as the safety and comfort of the general public. Secondly, precipitation location and intensity continue to be difficult to predict correctly. Many atmospheric processes from largest to smallest scales are involved in its simulation, making precipitation a good test of model performance as a whole. Lastly, rainfall is routinely observed as a highly resolved spatial field using weather radars – a big advantage over variables like wind and pressure which are mostly observed as point-wise measurements from weather stations, balloons and commercial airplanes.

Figure 1.1 schematically shows the verification of a high-resolution model forecast against an objective precipitation analysis. At a glance, the prediction looks mediocre: While the precipitation regions were mostly placed correctly, the forecast contains a spurious rain system over Germany and generally fails to reproduce the shape and extent of the observed features. When we take the difference of the two fields to compute an objective score, however, the forecast appears to be a complete failure. Almost none of the rain fields were placed *exactly* right, the mean absolute error is larger than the mean observed intensity. In this example, we would have improved our score by forecasting no rain at all. The reason for this disagreement between subjective and objective evaluation is the so-called *double penalty* effect: Any mis-placed precipitation system is punished by large difference values both in the location where it *should* have been predicted and in the location where the forecast placed it instead. Smoother, coarsely resolved forecasts are less prone to double penalties and therefore tend to receive better marks compared to the more realistic-looking competition.

When the first highly resolved mesoscale forecasting systems appeared, researchers quickly realized that the double penalty effect, while mostly harmless for smooth fields and at lower resolutions, dominates the objective scores in this regime. In an early survey of such experiments, Mass et al. (2002) state that “Decreasing grid spacing to less than 10-15 km generally improves the realism of the results but does not necessarily [...] improve the objectively scored accuracy of the forecasts.” Their assessment, mirrored by Done et al. (2004), encapsulates the frustration of model developers and forecasters who could not quantify the perceived gain in realism using traditional, gridpoint-by-gridpoint scores. In response to this issue, new verification scores were developed with the aim of avoiding double penalties and allowing for a fair assessment of highly resolved forecast fields. A detailed overview of the different strategies employed for this task will be given in chapter 3. Their common idea is to transform the observed and predicted fields from grid-point space to some other, abstract representation, for example via spatial filters, a change of basis function or segmentation into discrete objects. Some of these approaches aim to imitate the decision process of a human expert evaluating the forecast; many of them separate the overall error into different components like location, intensity and internal structure.

By 2007, the growing interest in spatial verification, as well as the large number of competing approaches, lead to the initiation of the first Spatial Forecast Verification Methods Intercomparison Project (Gilleland et al., 2009, ICP). The developers of various new techniques were invited to apply their methods to a standardized set of common test cases, ranging from simple geometric shapes via artificially perturbed observations, to realistic case studies. A main achievement of the project was the systematic classification of all new techniques into four categories, according to the kind of spatial transformation used: Neighborhood methods based on spatial smoothing filters (section 3.1.1), scale-separation methods based on spectral basis functions (3.1.2), feature-based methods using image-segmentation algorithms (3.1.3), and field deformation approaches which compute and optimize a cost function for the transformation from one field into the other (3.1.4).

The second phase of ICP launched in 2013 under the title “Mesoscale Verification Intercomparison over Complex Terrain” (Dorninger et al., 2018, MesoVICT). Besides updating the classification to include a newly identified fifth class (binary distance measures based on the distance transform, section 3.1.5), MesoVICT provided a range of new test cases centered on the European Alps and encouraged participants to consider probabilistic forecasts, as well as uncertain observations and model variables other than precipitation. The research collected in this thesis contributes to the MesoVICT project, as well as the overall effort to develop useful verification measures for state-of-the art forecast models, by introducing a novel verification technique based on wavelet transforms.

1.2 Research Plan: Wavelets

One of the earliest proposed solutions to the double penalty and related issues was the separation of the overall error into components corresponding to different spatial scales (Briggs & Levine, 1997). If, for example, a forecast simulates the motion of a cold front with high precision but fails to correctly locate individual storm cells in its wake, we should be able to evaluate the overall performance separately for those two spatial scales. In principle, this could be achieved via a two-dimensional Fourier transform which represents the meteorological field under consideration by a superposition of plane waves. Intuitively, we can see that this approach is not ideal by the simple fact that rain fields, like those shown in figure 1.1, do not resemble plane waves: Instead of a smooth, periodically repeating pattern, we typically find a limited number of discrete features with sharp edges, surrounded by uniform zero values. In addition, radar composites and regional model forecasts do not necessarily exhibit periodic boundary conditions. All of these aspects (edges, empty regions, aperiodic boundaries) are represented in Fourier space by a superposition of many different frequencies, thereby leaving the result difficult to interpret.

Wavelet transforms are an alternative to Fourier, which is better equipped to analyze images like those in figure 1.1. Instead of plane waves, the data is represented by a superposition of localized waveforms, which vary not only in frequency but also in location. To obtain a set of these small, localized waves (hence the diminutive wave*lets*), we select an appropriate function as the "mother"-wavelet and obtain her "daughters" by shifting and re-scaling. If, for example, the mother is $\psi(t)$, one of her daughters would be $\psi(t/s - t_0/s)/\sqrt{s}$. A more precise definition, as well as an introduction to all relevant aspects of wavelet transforms, is given in chapter 2. With wavelets, we can represent large, smooth regions in one part of the image by a few large daughter wavelets (large values of s) and capture localized features by a few smaller-scaled daughters at the appropriate locations t_0.

The basic wavelet-idea described above goes back to Haar (1910) and Gabor (1946). In geoscience, such functions were first popularized under the name "wavelets" by Goupillaud et al. (1984). The development of orthogonal wavelet bases, together with highly efficient algorithms for their computation (Mallat, 1989) lead to an explosion in wavelet applications across numerous fields of science and technology. Two-dimensional wavelet transforms were particularly successful in image processing and computer vision, i.e., in the automatic analysis, interpretation and enhancement of digital images. In their review of meteorological verification methods based on wavelets, Weniger et al. (2017) list image compression, segmentation, registration and fusion, as well as facial recognition and texture analysis among the ways in which the field of image processing has profited from wavelets. In contrast, forecast verification studies have almost exclusively focused on the wavelet's ability to decompose a single error into its components for different scales.

The overarching goal of this thesis is to explore the unused potential of wavelets for spatial forecast verification. More specifically, previous studies essentially use wavelets as a more convenient kind of Fourier transform but do not truly exploit the fact that wavelets are localized in space. We will investigate the unique opportunities granted by the local basis functions. It is demonstrated how wavelets can extract very specific kinds of structural forecast errors, allowing us to draw conclusions like "the spatial scale of the predicted rain field was too small" or "the simulated cold front was too linearly organized and rotated by an angle of 10°". In addition, we define a novel measure of displacement errors based on complex wavelets. A common advantage of these new scores is their potential applicability to variables other than precipitation.

The formulation of specific research objectives is postponed to section 3.4. First, we present a complete introduction to the relevant wavelet theory in chapter 2. We review the existing verification methods in section 3.1, elaborate on the various kinds of forecast errors (3.2) and survey ways in which the merits of a new verification tool can be assessed (3.3). Based on Weniger et al. (2017), section 3.4 describes the existing wavelet-based verification approaches and finally specifies five open research questions. In chapter 4, we briefly summarize the results published in Buschow et al. (2019), Buschow & Friederichs (2020), Buschow & Friederichs (2021a), as well as Buschow & Friederichs (2021b) (under review at time of writing) and Buschow (2021b) (manuscript in preparation), and explain how each of the research questions has been addressed. The full publications and drafts are attached in appendix A-E. The thesis ends with concluding remarks and suggestions for future research in chapter 5.

1.3 List of publications

The core results of this study have previously appeared in the following peer-reviewed publications:

> Buschow, S., Pidstrigach, J., & Friederichs, P. (2019). Assessment of wavelet-based spatial verification by means of a stochastic precipitation model (wv_verif v0.1.0). *Geoscientific Model Development*, *12*(8), 3401–3418

> Buschow, S., & Friederichs, P. (2020). Using wavelets to verify the scale structure of precipitation forecasts. *Advances in Statistical Climatology, Meteorology and Oceanography*, *6*(1), 13–30

> Buschow, S., & Friederichs, P. (2021a). SAD: Verifying the scale, anisotropy and direction of precipitation forecasts. *Quarterly Journal of the Royal Meteorological Society*

The basic ideas for all three of these studies were jointly developed with my supervisor Petra Friederichs, who also contributed to the revision and improvement of the final drafts. The design of the specific experiments, as well as all programming and most of the writing and visualization was carried out by me. Jakiw Pidstrigach independently worked on the wavelet selection procedure in Buschow et al. (2019) and lead the programming and writing for this part of the paper. A further publication is currently under review at Geoscientific Model Development, the preprint has been published as

> Buschow, S., & Friederichs, P. (2021b). Verification of Near Surface Wind Patterns in Germany using Clear Air Radar Echoes. *Geoscientific Model Development Discussions*

The original idea for this study was my own, Petra Friederichs contributed to the quality of the draft at later stages in the process.

A final manuscript, so far with no major contributions from other authors, is in preparation for future publication:

> Buschow, S. (2021b). Measuring Displacement Errors with Complex Wavelets. In preparation

Chapter 4 gives short summaries, the full articles are reproduced in appendix A-E. Appendix D differs from the archived pre-print in that the erroneously missing figure 4 is included. In addition, the software needed to perform the underlying wavelet transformation and the SAD verification method have been published as open source packages on the official CRAN archive:

> Buschow, S., Kingsbury, N., & Wareham, R. (2020). *dualtrees: Decimated and Undecimated 2D Complex Dual-Tree Wavelet Transform*. R package version 0.1.4

> Buschow, S. (2020). *sad: Verify the Scale, Anisotropy and Direction of Weather Forecasts*. R package version 0.1.3

The complete software needed for the reproduction of Buschow et al. (2019) and Buschow & Friederichs (2021b) has been permanently archived in a citable form:

> `https://doi.org/10.5281/zenodo.3257511` (Buschow, 2019)

> `https://doi.org/10.5281/zenodo.5036447` (Buschow, 2021a)

Chapter 2

Wavelets

This chapter is intended as a mostly self-contained introduction to the world of wavelets. As such, sections 2.1 and 2.2 introduce two of the most widely used wavelet-transforms from scratch, broadly following the first chapters of *Ten lectures on wavelets* (Daubechies, 1992). The more specialized techniques used in our verification-approach follow in 2.3 and 2.4, drawing mainly on Eckley et al. (2010) and Selesnick et al. (2005), respectively.

2.1 Continuous wavelets

The basic concepts of wavelet transforms are most easily understood using the example of one-dimensional time-series analysis. In this application, wavelets can analyze the frequencies in a signal at any given point in time. As an everyday example, we consider a piece of music, wherein the pitch of a note corresponds to the frequency of the sound wave. Figure 2.1 (a) shows the waveform associated with the main theme from Gabriel Fauré's *Pavane, Op.50*, played here by solo flute. From the amplitude $A(t)$ of the signal, we can see the volume changing over time as the musician (the author) attempts to emphasize the musical phrases and eventually runs out of breath. To observe the changes in pitch, we must zoom in on different parts of the signal (blue lines).

We can isolate the individual frequencies by applying the Fourier transform to the signal

$$\hat{A}(\omega) = \mathcal{F}\{A\}(\omega) = \int_{-\infty}^{\infty} A(t) \underbrace{e^{-2\pi i\omega t}}_{:=\overline{e_\omega(t)}} dt := \langle A(t), e_\omega(t) \rangle \,, \tag{2.1}$$

where we have introduced the notation

- $\hat{X}$ for the Fourier transform of a function X,
- $\overline{X}$ for the complex conjugate of X,
- $e_\omega(t) = e^{2\pi i\omega t}$ for the Fourier basis functions and
- $\langle X, Y \rangle = \int_{\mathbb{R}} X(t)\overline{Y(t)}dt$ for the scalar product between two (complex) functions on the real line.[1]

[1] A vector space (here the vector space of complex functions with the usual addition and scalar multiplication), together with such a scalar product, is called a *Hilbert space* if it fulfills an additional completeness requirement.

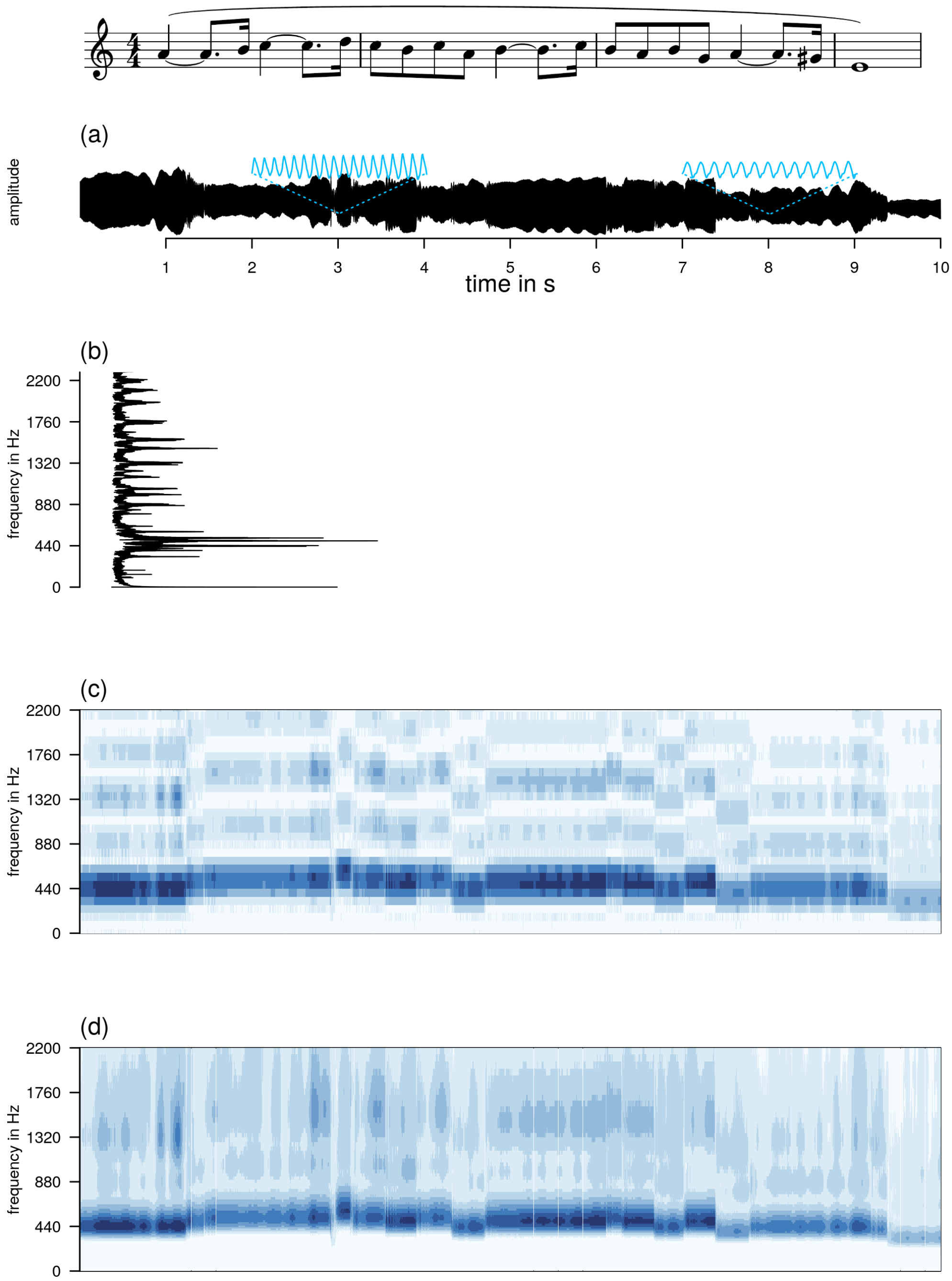


Figure 2.1: Theme from Gabriel Fauré's *Pavane, Op.50* represented as waveform (a, blue lines are zoomed sections of the signal), Fourier spectrum (b, power in arbitrary units on the x-axis), windowed Fourier transform (c) and continuous Morlet wavelet transform (d). Dark colors in (c,d) indicate high power.

For a discretely sampled signal of length n, equation 2.1 is replaced by the corresponding discrete Fourier transform, wherein the integral becomes a sum over a finite set of basis functions $\{e_\omega(t)|\omega \in \{0, \frac{1}{n}, \ldots, \frac{n-1}{n}\}\}$. Figure 2.1 (b) shows the resulting frequency spectrum $|\hat{A}|^2$ for our piece of music. The presence of certain notes is reflected by thin lines at the corresponding frequencies. The global maximum is close to $440\,Hz$ which corresponds to the concert pitch A, i.e., the first note of the piece.

While $\hat{A}$ gives us some idea about the different frequencies in the signal (confirming, for example, that the flute was properly tuned to $440\,Hz$), it contains no information on the timing of the individual notes. A natural idea to obtain such information is to apply the Fourier transform to short parts of the signal individually. This can be realized in the form of a windowed Fourier transform

$$\mathcal{WF}\{A\}(\omega, t_0) = \langle A(t), w(t_0 - t) \cdot e_\omega(t)\rangle := ((A \cdot \overline{e_\omega}) * w)(t_0)\,, \tag{2.2}$$

where $w(x)$ is a real-valued window function with values close to zero everywhere except for a small neighborhood around $x = 0$. Here, we have introduced the notation

$$\langle x(t), \overline{y(t_0 - t)}\rangle := (x * y)(t_0)$$

which denotes so-called *convolutions*: Intuitively, the function y, in our case the window w, is shifted to each time step; the result of the convolution is the product of the signal x with every shifted version of y. One useful property of convolutions is the fact that they become multiplications in Fourier space:

$$\mathcal{F}\{(x * y)(t_0)\}(\omega) = \mathcal{F}\{x\}(\omega)\mathcal{F}\{y\}(\omega) \tag{2.3}$$

Thanks to efficient Fourier algorithms, the quickest way to compute a convolution is often to transform to Fourier space, multiply and then transform back. Equation 2.3, and especially its counterpart in two dimensions, is therefore the key to acceptably fast implementation of not only the windowed Fourier transform, but also many of the wavelet methods described below.

Figure 2.1 (c) demonstrates that $\mathcal{WF}$ works as intended: Starting at $440\,Hz$, the maximum of the coefficients nicely traces the various changes of pitch throughout the piece. In addition to the main maximum, we recognize a range of weaker local maxima at discrete intervals above it: These are the overtones ($880\,Hz, 1320\,Hz$, $1760\,Hz$, $2200\,Hz$ for A), the relative presence of which determines the characteristic sound of the instrument.

Comparing panels (b) and (c), we observe that the lines at specific frequencies are much broader in the windowed transform. This is a classic example of Heisenberg's uncertainty principle (sometimes called Heisenberg-Gabor limit in signal processing), which states that a function cannot be perfectly localized in both the time and frequency domain. A more narrow window w improves the time localization at the cost of a lower resolution in frequency: The limit of an infinitely narrow window recovers figure 2.1 (a) (no frequency resolution), an infinitely wide window is the same as the regular Fourier transform (panel b, no time resolution).

To arrive at a useful analysis, w must therefore be chosen appropriately. Intuitively, we need larger windows to sample low frequencies while a smaller window would be more appropriate for high frequencies. Equation 2.2 with its fixed window for all ω is therefore generally not ideal to represent signals with multiple high and low pitches. To construct a more flexible alternative, we begin with a windowed waveform $\psi(t) = w(t)e_\sigma(t)$ and then simultaneously adapt frequency and location via the transform $t \to (t - t_0)/s$. This is the basic idea of wavelet transforms: Select a wave-function ψ as the so-called *mother wavelet* and generate a set of analyzing functions, her so-called *daughter wavelets*:

$$\psi(t|s, t_0) = \frac{1}{\sqrt{s}} \psi \left(\frac{t - t_0}{s} \right) ,$$

where the factor $1/\sqrt{s}$ re-scales all daughter wavelets to the same total energy $\langle \psi(.|s, t_0), \psi(.|s, t_0) \rangle = \langle \psi, \psi \rangle$. The parameter $s > 0$, called the *wavelet scale*, is closely related but not identical to the Fourier frequency ω. In analogy to $\mathcal{F}$ and $\mathcal{WF}$, we thus define the wavelet transform as

$$\mathcal{W}\{A\}(s, t_0) = \langle A(t), \psi(t|s, t_0) \rangle$$

A popular example of a mother with close relations to the Fourier transform is the *Morlet* wavelet (Mallat, 1999), given by

$$\psi(t) = C \cdot e^{-t^2/2}(e^{i\sigma t} - e^{-\sigma^2/2}) , \tag{2.4}$$

where C is a normalization constant. We recognize that ψ is just the Fourier basis function $e_{\sigma/(2\pi)}$ minus a constant, localized by a Gaussian window. The scale s can be related to an approximately equivalent Fourier frequency ω by finding the frequency where $\mathcal{F}\{\psi(.|s, t_0)\}$ is maximized (Torrence & Compo, 1998). For the Morlet wavelet, one can thus show that $\omega \approx (\sigma + \sqrt{2 + \sigma^2})/2s$. The resulting spectrum (figure 2.1) looks very similar to the windowed Fourier transform, albeit with a slightly better time- and slightly worse frequency resolution.

To re-capitulate the different signal representation from figure 2.1, consider the corresponding basis functions in figure 2.2:

(a) Time series: single time steps, all frequencies

(b) Fourier basis: single frequency, all time steps

(c) Windowed Fourier: fixed time-window for all frequencies

(d) Wavelets: A single mother function, scaled and shifted

In the next section, we will see how the adaptive nature of wavelets can be exploited to generate efficient representations of complicated signals by appropriately sampling the locations t_0 and scales s. Before moving on, one more fundamental point must be addressed, namely which functions ψ are appropriate as mother wavelets. The basic idea here is that we require the existence of an inverse wavelet transform, such that all information on A can be recovered from $\mathcal{W}\{A\}$. For simplicity, we assume here that A has zero mean. In practice, the mean value has to be treated separately. It can be shown (see for example Kaiser 2010) that an inverse transform is given by

$$A(t) = \int_0^\infty \int_{\mathbb{R}} \frac{1}{s^2} \psi(t|s, t_0) \mathcal{W}\{A\}(s, t_0) ds dt_0 ,$$

as long as the following two conditions hold:

$$\int_{\mathbb{R}} |\psi(t)|^2 dt < \infty \tag{2.5}$$

$$0 < \int_0^\infty \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < \infty \tag{2.6}$$

Condition 2.5 states that ψ must be square integrable and therefore has to decay to zero as $t \to \pm\infty$. The Hilbert-space of square integrable functions (with the scalar product defined above), will be referred to as $L^2(\mathbb{R})$. 2.6 is the so-called *admissibility condition*, which can only be fulfilled if $\hat{\psi}(0) = 0$ (otherwise the integrand would diverge at the origin). Since the zero frequency component is just the signal mean, it follows that

$$\hat{\psi}(0) = \int_{\mathbb{R}} \psi(t) dt = 0\,. \tag{2.7}$$

It can be shown that 2.7 is a sufficient condition for admissibility if $\hat{\psi}$ is also continuously differentiable (Mallat, 1999). For the Morlet wavelet, equation 2.7 is ensured by subtracting the constant $e^{-\sigma^2/2}$. By these definitions, wavelets are therefore indeed *little waves* that oscillate around zero (by 2.7) within some localized part of the real axis (by 2.5). Note that neither of the two conditions impose particularly strong constraints that would guarantee *good* localization in either time or frequency – they merely broadly define the realm of functions from which wavelets with useful properties can be selected.

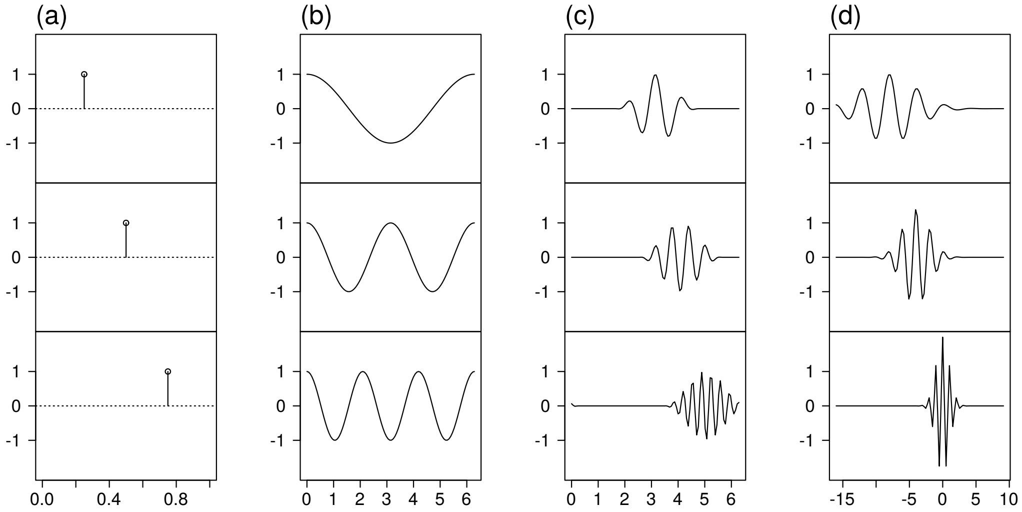


Figure 2.2: Real parts of the basis functions corresponding to the different signal representations in figure 2.1: Time series (a), Fourier (b), windowed Fourier (c) and Morlet wavelet (d).

2.2 Multi-Resolution Analysis

2.2.1 One Dimension

In the previous section, we have seen that both the continuous wavelet transform[2] $\mathcal{W}$ and the windowed Fourier transform $\mathcal{WF}$ represent our signal in a highly redundant way: A one-dimensional vector (figure 2.1 a) is transformed into a two-dimensional array of coefficients (figure 2.1 c, d). Furthermore, using equation 2.3, we need to compute three Fourier transforms for each scale of interest. The algorithm is thus far from computationally efficient.

We have already alluded to the idea that high frequencies need to be sampled more often from smaller sections of the signal, whereas a few long intervals should be sufficient to sample the low-frequency components. Formalization of this idea leads to the so-called Multi-Resolution Analysis (MRA). In his landmark paper that introduced the MRA, Mallat (1989) demonstrates that

- wavelets $\psi(.|s,t_0)$ can form an orthogonal basis of $L^2(\mathbb{R})$, if we allow only scales $s = 2^j$ for integers j and scale-dependent shifts $t_0 = 2^j n$
- orthogonal wavelet transforms can be implemented in an efficient recursive algorithm by repeated application of a special pair of filters
- the transform can be generalized to two dimensions by applying the filters to the rows and columns of a matrix in an alternating fashion

Here, a *filter* is an operator F which performs the convolution of a discrete signal $A_{1,\ldots,n}$ with a set of filter coefficients $f_{1,\ldots,n}$:

$$(FA)_k = \sum_{i=1}^{n} A_{i-k} f_i \tag{2.8}$$

The boundary conditions are treated periodically, i.e., $A_{n+i} = A_i$ and $A_{-i} = A_{n-i}$. A simple example filter would be $f_1 = f_2 = 1/2, f_{i>2} = 0$, which corresponds to a moving average.

To see how filters are related to efficient, orthogonal wavelet bases, we now introduce the Haar wavelet, which is given by

$$\psi(t) = \begin{cases} 1 & \text{if} \quad 0 \leq t < 1/2 \\ -1 & \text{if} \quad 1/2 \leq t < 1 \\ 0 & \text{otherwise}\,. \end{cases} \tag{2.9}$$

The function $\psi(t)$ is square integrable and integrates to zero. It can furthermore be shown that its Fourier transform $\hat{\psi}(\omega)$ decays quickly enough to fulfill Eq. 2.6, making this the simplest (and oldest, see Haar 1910) of all mother wavelets.[3] Her daughters are

$$\psi(t|s,t_0) = \begin{cases} 1/\sqrt{s} & \text{if} \quad t_0 \leq t < s/2 + t_0 \\ -1/\sqrt{s} & \text{if} \quad s/2 + t_0 \leq t < s + t_0 \\ 0 & \text{otherwise}\,. \end{cases} \tag{2.10}$$

[2] continuous in the sense that all shifts t_0 and scales s are allowed

[3] In chapter 3 we will see that she is also the most popular choice for forecast verification.

Following Mallat's suggestion, we now allow only those daughters whose scale and shift are $s = 2^j$, $t_0 = 2^j n$ and denote them as

$$\psi(t|2^j, 2^j n) := \psi_{j,n}(t) = 2^{-j/2}\psi\left(t/2^j - n\right) . \tag{2.11}$$

Since both scale and location are thus discretized, the resulting transformation is also called *discrete wavelet transform* (DWT). It is easy to see that two daughters with the same j but different n are orthogonal to each other since their support does not overlap:

$$\mathrm{supp}(\psi_{j,n}) = \{t \in \mathbb{R} | \psi_{j,n}(t) \neq 0\} = [2^j n, 2^j n + 2^j) \quad \Rightarrow \quad \langle \psi_{j,n}, \psi_{j,n' \neq n} \rangle = 0$$

Similarly, we can see from Eq. 2.10 that the positive and negative part of $\psi_{j+1,0}$ exactly cover the support of $\psi_{j,0}$ and $\psi_{j,1}$, respectively. The smaller daughters are thus supported on the constant parts of the next larger daughters which implies that $\psi_{j,n}$ and $\psi_{j' \neq j,n'}$ are orthogonal as well (their scalar products are a constant times the integral over the smaller daughter, i.e., zero).

We have thus seen that Mallat's choice of (s, t_0) generates an orthogonal set of Haar-wavelets which can therefore represent a signal of length 2^J ($J \in \mathbb{N}$) by exactly 2^J coefficients. A formal proof that the Haar-wavelets form a complete *basis* of $L^2(\mathbb{R})$ is given, for example, in Daubechies (1992). Here we are mainly interested in the efficient algorithm for the Haar-decomposition based on filters. To this end, consider a discretely sampled signal $A_1, A_2, A_3, \ldots$, which can be understood as a piece-wise constant function A on the real line. Apart from the scaling by $\sqrt{2}, 2, 2\sqrt{2}, \ldots$, the products of such a signal with the first few Haar daughters are given by

$$\begin{array}{ll}
j = 1: & (A_1) - (A_2),\ (A_3) - (A_4),\ (A_5) - (A_6),\ (A_7) - (A_8),\ \ldots \\
j = 2: & (A_1 + A_2) - (A_3 + A_4),\ (A_5 + A_6) - (A_7 + A_8),\ \ldots \\
j = 3: & (A_1 + A_2 + A_3 + A_4) - (A_5 + A_6 + A_7 + A_8),\ \ldots \\
& \ldots
\end{array}$$

We notice that each bracket expression is the sum of two bracket expressions from the previous scale. The coefficients can thus efficiently be calculated by algorithm 1.

Algorithm 1 Haar Multi Resolution Analysis (MRA) of Mallat (1989)

```
Input: signal $\vec{A} = (A_1, \ldots, A_{2^J})$
Output: wavelet coefficients and signal mean

for $j = 1$ to $J$ do
    split $\vec{A}$ into pairs $(A_1, A_2)$, $(A_3, A_4)$, etc.
    compute the wavelet coefficients at scale $j$ as the scaled differences of the pairs
    replace $\vec{A}$ by the scaled sums of the pairs
end for
```

By re-using the results from the previous scale, we have to compute half as many additions for scale 2, a quarter of the additions for scale three and so on. This recursive algorithm can be understood as a sequence of filters. Let G be the differencing filter with coefficients $g_1 = 1/\sqrt{2}, g_2 = -1/\sqrt{2}$ and H a moving average

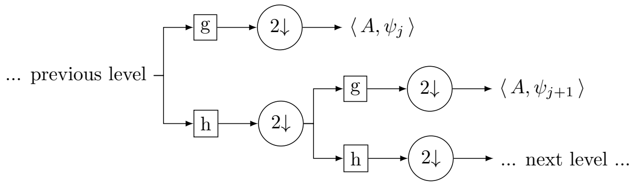


Figure 2.3: Schematic representation of the one-dimensional fast discrete wavelet transform: "g" and "h" mark the application of the high- and low-pass filter, "2 ↓" represents a down-sampling step where only every second sample is retained.

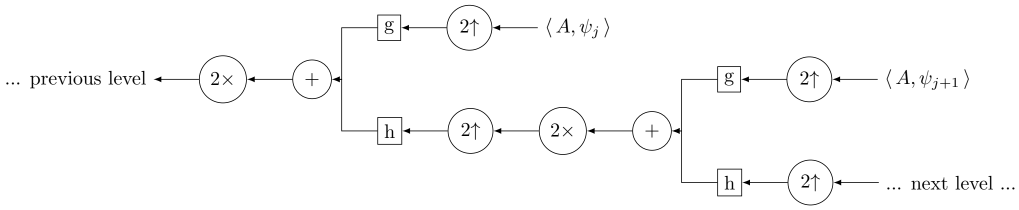


Figure 2.4: Inverse discrete wavelet transform: '+" denotes addition of the two inputs, "2×" is a multiplication by two and 2 ↑ indicates up-sampling of the inputs by a factor of two, i.e., inserting a zero after every sample.

filter with $h_1 = h_2 = 1/\sqrt{2}$. Algorithm 1 can thus also be represented by the diagram in figure 2.3.

The inverse transform can be realized in a similar fashion by inverting the direction of the diagram and replacing the downsampling by an upsampling step which inserts a zero after every value. This procedure is shown in figure 2.4: The wavelet coefficients are successively up-sampled and added up to reconstruct the original signal. This completes Mallat's multi-resolution analysis – a fast, orthogonal decomposition into a set of wavelet basis functions which are localized in both time and frequency. One strength of this approach lies in the possible generalization to other mother wavelets: Start by defining an appropriate averaging filter H by its non-zero coefficients $h_1, \ldots, h_N$, then generate the corresponding differencing filter via the relationship

$$g_i = (-1)^i h_{N-1-i} \, . \tag{2.12}$$

H corresponds to a low-pass filter while G, which is just an inverted version of H where every second sign is flipped, constitutes a high-pass filter. Applying H $j-1$ times, followed by one application of G (and downsampling after every filter as in figure 2.3) is equivalent to the projection onto a set of daughter wavelets $\psi_{j,n}$. Similarly, j-fold application of H corresponds to projection onto a *scaling function* $\phi_{j,n}$ which is itself a scaled, shifted version of some function ϕ, i.e.,

$$\phi_{j,n}(t) \sim \phi(t/2^j - n) \, . \tag{2.13}$$

ϕ is sometimes called the *father wavelet*, even though it integrates to a finite non-zero value and is thus not a wavelet in the sense described above.

Mathematically, the MRA is described as the nested Hilbert-spaces V_j, which are spanned by the $\phi_{j,n}$. We require that

(i) the spaces are *nested* in the sense that $V_j \supset V_{j+1}$ and
$f(t) \in V_j \Leftrightarrow f(t/2) \in V_{j+1}$

(ii) the shifts of $\phi_j(t) := \phi(2^{-j}t)$ are an orthogonal basis for V_j

(iii) the union of all V_j is a dense subspace of $L^2(\mathbb{R})$

(iv) the intersect of all V_j contains only $f(t) = 0$.

Condition three demands that every function in $L^2(\mathbb{R})$ is either in the union of the V_j or can be arbitrarily well approximated by a member of that union. The fourth condition puts a limit on how redundant the representation in terms of the $\phi_{j,n}$ can be by forbidding non-trivial functions from living in *all* subspaces. From conditions one and two, it follows that $\phi_{j+1}(t) = \phi_j(t/2)$ can be represented as a linear combination of the shifts of $\phi_j(t)$, i.e,

$$\begin{aligned} & \phi(t/2) = \int_{\mathbb{R}} \tilde{h}(n')\phi(t-n')dn' = (\tilde{h} * \phi)(t) \qquad | \,\langle \phi(t-n), \ldots \rangle \\ \Leftrightarrow \quad & \langle \phi(t-n), \phi(t/2) \rangle = \tilde{h}(n)\,, \end{aligned}$$

where we have used the orthogonality of the shifted ϕ to obtain the second line. The convolution of a signal A with a father wavelet at scale j can thus be written as the repeated convolution with $\tilde{h}$, i.e.,

$$\begin{aligned} (A * \phi_j)(t) &= (A * (\tilde{h} * \phi_{j-1}))(t) = (A * (\tilde{h} * \tilde{h} * \ldots * \phi_0))(t) \\ &= ((\ldots((A * \tilde{h}) * \tilde{h}) * \ldots) * \phi_0)(t)\,. \end{aligned}$$

We can thus completely define the father wavelet by setting $\phi_0 := \tilde{h} = h$ which shows that the convolution with the scaled ϕ can indeed be realized by iteratively applying a filter h as in the algorithm described above. If all four conditions above hold, an orthogonal wavelet basis of $L_2(\mathbb{R})$ can then be created from ϕ via equation 2.12. This recipe has been used to design numerous wavelet bases with different properties. Filters with few non-zero coefficients (like the Haar) have good localization in time whereas longer filters can achieve better localization in frequency. To quantify this trade-off, we introduce the notion of *vanishing moments*: A function $\psi(t)$ has n vanishing moments if

$$\forall\ 0 \leq q < n: \quad \int_{\mathbb{R}} t^q \psi(t)dt = 0\,.$$

All wavelets satisfy this condition for $n = 0$ by definition. More vanishing moments indicate better frequency localization: It can easily be shown that multiplication by a polynomial in time becomes a derivative with respect to frequency in Fourier space, i.e., $\mathcal{F}\{tX(t)\} = \frac{i}{2\pi}\frac{\partial}{\partial\omega}\hat{X}$. A larger number of vanishing moments therefore corresponds to a flatter $\hat{X}$ near the origin (higher-order root), thereby limiting the bandwidth from below. Furthermore, if a signal A behaves like a polynomial of degree $q < n$ within an interval of length 2^j, then all wavelet coefficients for scales $< j$ vanish in this interval if ψ has n vanishing moments. ψ thus gives a very sparse

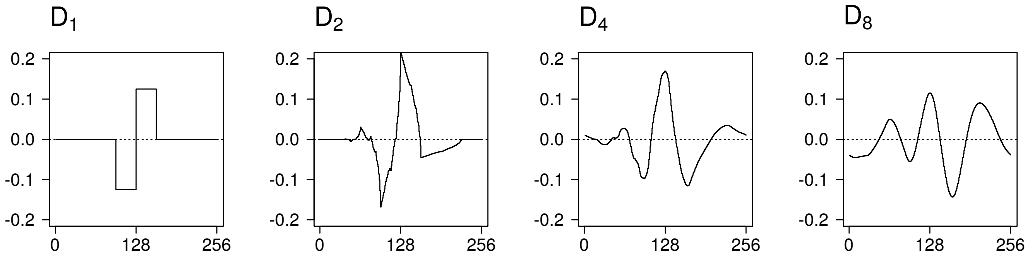


Figure 2.5: Daubechies wavelets with 1, 2, 4 and 8 vanishing moments. Shown are the daughters at scale $j = 6$ for a signal of length 256. Each daughter was shifted so that its maximum is near the center of the plot.

representation of signals which locally behave like smooth polynomials. Daubechies (1988) proved that the smallest possible support length for a wavelet with n vanishing moments is $2n$, and constructed a family of wavelets with optimally short support. We will refer to the nth Daubechies wavelet (n vanishing moments) as D_n.

Figure 2.5 shows a few examples from this family. D_1 is identical to the Haar wavelet from before ($h_1 = h_2 = 1/\sqrt{2}$, see equation 2.9), which has one vanishing moment. Higher-order wavelets become increasingly smooth with a greater number of sign changes while increasing the length of their support.

2.2.2 Two Dimensions

Efficient algorithms are desirable in time series analysis, but almost indispensable in higher dimensions. An important application is the field of image processing which is concerned with the analysis and efficient representation of two-dimensional arrays – the grey-scale intensities in a photograph or, in our case, values of a meteorological variable on a regular grid.

To extend the MRA algorithm to two dimensions, we simply apply one of the two filters along the rows of the matrix and then another filter along the columns of the result. This decomposition, schematically shown in figure 2.6, thus projects an image onto three directional daughter wavelets at each scale:

$$\begin{aligned}
\psi_{j,90^\circ} &= \phi_j(x)\psi_j(y) \qquad \textit{(vertical)}\\
\psi_{j,0^\circ} &= \psi_j(x)\phi_j(y) \qquad \textit{(horizontal)}\\
\psi_{j,45^\circ} &= \psi_j(x)\psi_j(y) \qquad \textit{(diagonal)}
\end{aligned}$$

Figure 2.7 shows that $\phi_j(x)\psi_j(y)$ (panel a) and $\psi_j(x)\phi_j(y)$ (b) are mirrored versions of each other and act similar to directional derivatives in the x- and y-direction.[4] The diagonal daughter $\psi_j(x)\psi_j(y)$ (c) differs form her sisters in several ways. Being the product of two high-pass filters, these basis functions clearly represent an overall smaller scale than their sisters with the same j (count the number of wave-crests). In addition, their orientation is not unique, the "checkerboard" pattern is equally aligned along either of the two image diagonals. We will discuss the consequences of these properties in section 2.4.

[4] For the Haar wavelet in panel (a), they are exactly that: Finite difference approximations of ∂_x and ∂_y, calculated after some level of spatial averaging.

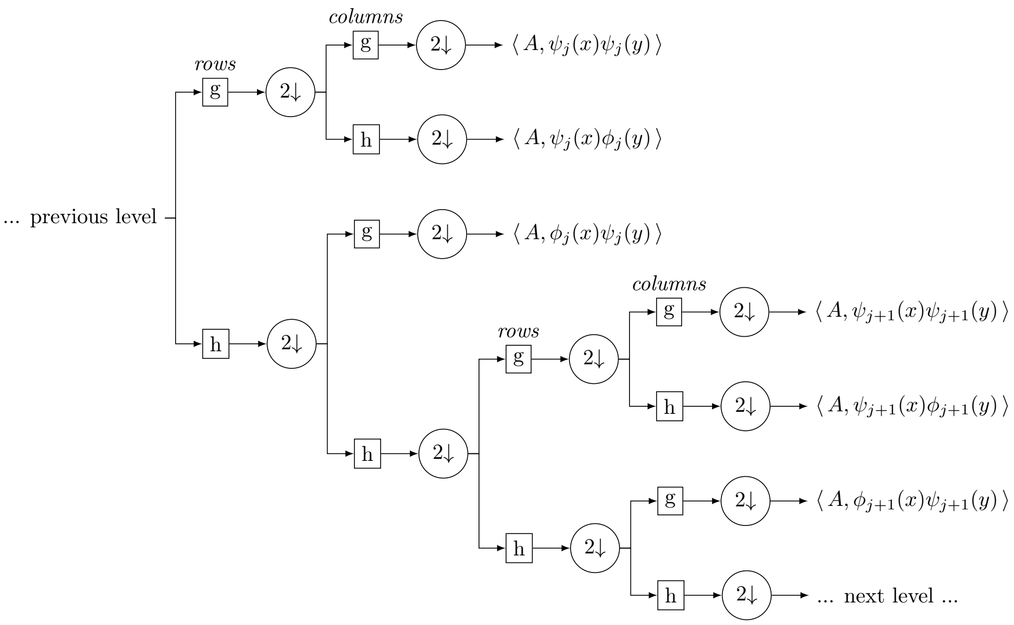

columns
rows
g
2↓
h
$\langle A, \psi_j(x)\psi_j(y) \rangle$
$\langle A, \psi_j(x)\phi_j(y) \rangle$
$\langle A, \phi_j(x)\psi_j(y) \rangle$
... previous level
$\langle A, \psi_{j+1}(x)\psi_{j+1}(y) \rangle$
$\langle A, \psi_{j+1}(x)\phi_{j+1}(y) \rangle$
$\langle A, \phi_{j+1}(x)\psi_{j+1}(y) \rangle$
... next level ...

Figure 2.6: Two-dimensional discrete wavelet transform.

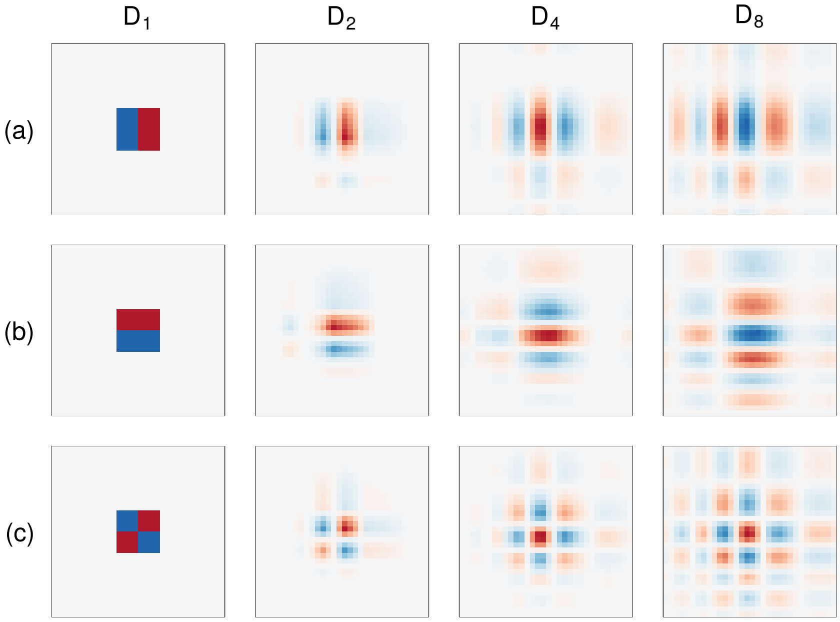


Figure 2.7: Vertical (a), horizontal (b) and diagonal (c) daughters corresponding to the Daubechies wavelets from figure 2.5. Negative values are shown in blue, positive in red.

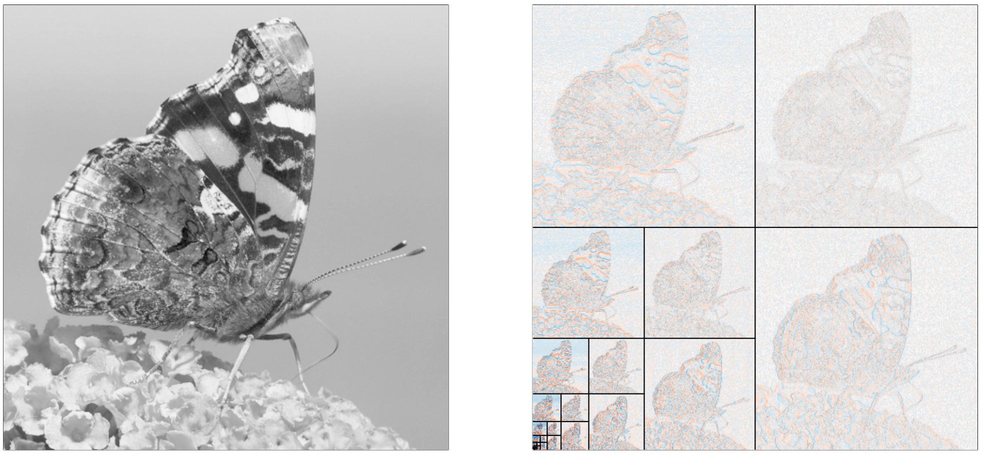


Figure 2.8: Haar-MRA applied to a photo of a butterfly: Horizontal, diagonal and vertical coefficients for $j = 1$ shown in the top-left, top-right and bottom-right quadrant. The bottom left quadrant likewise contains those coefficients for $j = 2$ and so on. Blue and red colors indicate positive and negative values, all coefficients were scaled by $x \to \operatorname{sign}(x)\sqrt[4]{x}$ for clearer visualization.

The 2D MRA represents an input image of size 512×512 by three 256×256 images at scale $j = 1$, three 128×128 images at scale $j = 2$ and so on. Figure 2.8 visualizes the decomposition in the style of Mallat (1989) such that all coefficients appear in one plot; the value of the father wavelet at $j = 9$ is represented by the pixel in the bottom left corner. Note that both images in this figure consist of 512×512 pixels – the decomposition is orthogonal and thus contains no redundant information. Comparing the bottom-right to the top-left quadrants, we observe increased positive and negative coefficients along the vertical and horizontal edges of the pattern inside the wing. Increased absolute values within the featureless background of the image appear only on large scales, i.e., near the bottom left of the plot. We furthermore notice that the diagonal coefficients appear to be weaker overall.

To conclude this introduction to the widely used methods of continuous and discrete wavelet transforms, we give an example for the *power* of wavelet transforms. A further noteworthy feature of figure 2.8 is that most coefficients are close to zero (we applied a forth-root transform to render more of them visible) – the representation in wavelet space is *sparse*. In figure 2.9, we have implemented a naive image compression algorithm by simply setting the smallest $p\,\%$ of all coefficients to zero before transforming back. For $p = 75\,\%, 90\,\%$, we can hardly discern any differences between the original and compressed image; only at very strong compression, artifacts begin to emerge. If we simply store the indices of the retained coefficients along with their values, the representation in the bottom left panel ($90\,\%$ compression) requires only one fifth of the original image size; the reconstruction procedure is quick thanks to the recursive MRA algorithm.

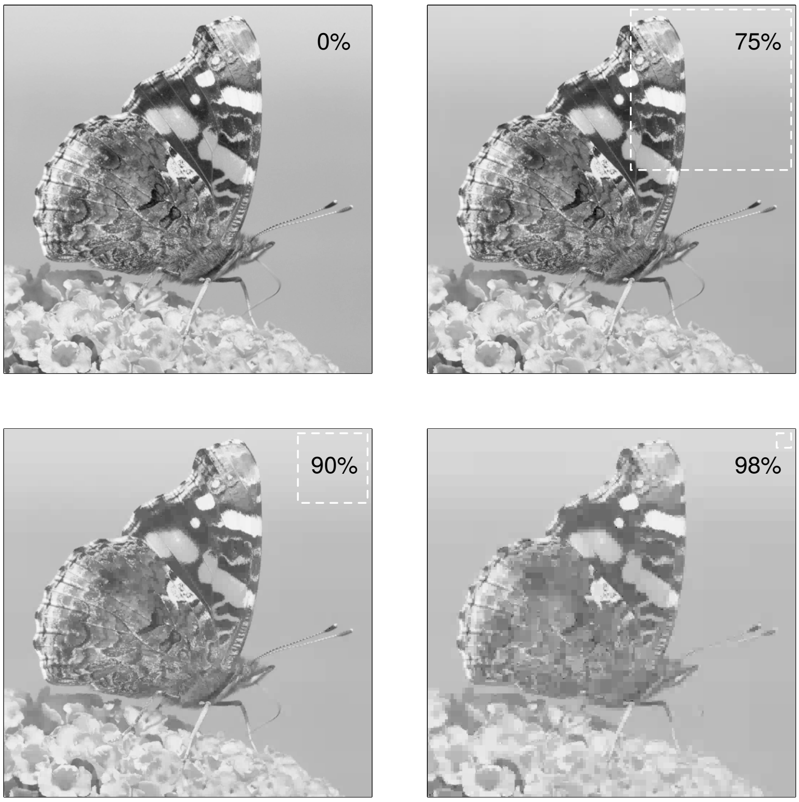


Figure 2.9: Butterfly image reconstructed from all coefficients (top left) and after removing the smallest 75 %, 90 % or 98 %. The dashed square represents the number of retained coefficients as the equivalent pixel area.

2.3 The redundant discrete wavelet transform

As discussed in chapter 1, our goal is to analyze, and later verify, the structure of an image. Before formalizing this idea in terms of spatial auto-covariances, we can intuitively see how the discrete wavelet transform might allow us to distinguish between images with small- large-scaled patterns: For a smooth image with large features, the coefficients of the larger daughter wavelets are big and those corresponding to smaller daughters become negligible; the reverse is true for images with a very fine-grained texture. Similarly, the separation into the directions could be used to differentiate linear from isotropic structures.

A simple MRA-based verification strategy would be to 1) perform the two-dimensional DWT of forecast and observation 2) compute the total "energy", i.e. the sum over all coefficients squared, for each combination of scale and direction and 3) compare the two resulting $3 \times J$ matrices. It is, however, easy to see that the resulting score would be sensitive not only to differences in spatial scale and orientation, but also to differences in the locations of the features: If, for example, the edge of a particular feature happens to be aligned with the sign-change of a Haar daughter wavelet, it will contribute strongly to that daughter's coefficient; shifting that edge across the support of the wavelet will decrease the coefficient.

Figure 2.10 demonstrates that these effects can have a substantial magnitude for the kinds of images we want to study: As we shift the rain field from figure 1.1 across the grid of the daughter wavelets at scale j, the corresponding energy oscillates with a period of 2^j. The impact is limited for those daughters which are much smaller than the dominant features in our example ($j < 3$). The large scales, however, vary strongly: Initially $j = 5$ is almost tied with $j = 4$ in terms of total energy; after a shift by 20 pixels, it drops almost to the level of $j = 2$. Which scales dominate to what extend thus heavily depends on the location of the features.

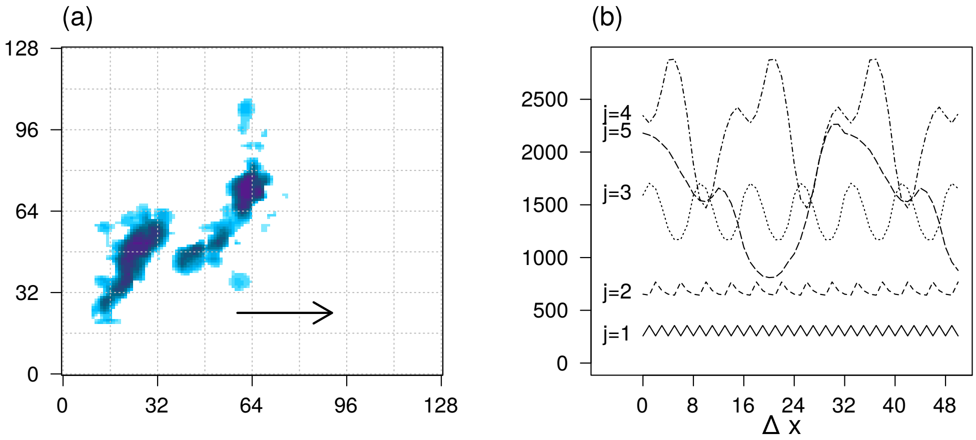


Figure 2.10: Impact of shifts in the input image on the 2D MRA: Rain field from figure 1.1 (a) and resulting sum over squared wavelet coefficients (b) for the horizontally aligned Haar daughter wavelets at scale $j = 1, ..., 5$ as a function of the shift Δx applied to the image in (a).

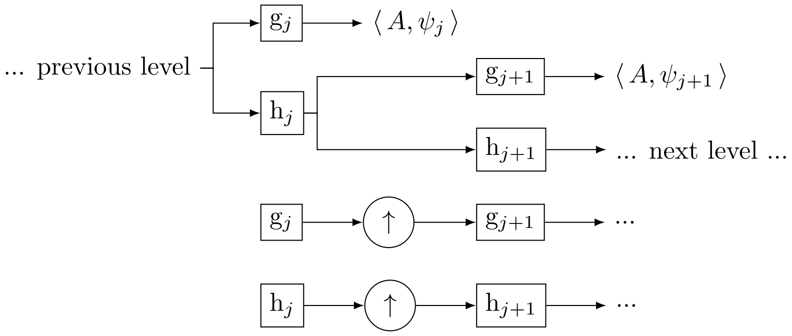


Figure 2.11: Redundant discrete wavelet transform in one dimension, realized as an *algorithme à trous*: Filter coefficients are up-sampled before each level of the decomposition, the signal is not down-sampled.

A natural, albeit computationally costly, solution to this problem is to repeat the analysis for every possible shift and average over the results. This is equivalent to dropping the restriction to shifts $t_0 = 2^j n$ (equation 2.11) and instead allowing all integer displacements in both directions. This is the so-called *redundant discrete wavelet transform* (henceforth RDWT) which represents an input image of size $2^J \times 2^J$ by $3 \times J \times 2^J \times 2^J$ coefficients.

Recall, however, that the wavelets associated with the MRA are not generally given by continuous functions ψ but only in terms of their filter coefficients h_i.[5] How, then, can the redundant transform be implemented? The answer was first given by Holschneider et al. (1990) who introduced the *algorithme à trous* (French for "hole-algorithm") which is shown in figure 2.11: Instead of down-sampling the signal, both filters are up-sampled by inserting additional zeros ("holes") before each convolution. The resulting wavelet transform is more well-behaved under shifts: A displacement of the input signal simply results in an equal shift of all coefficient fields. As a consequence, the spatial mean over the coefficients is invariant under shifts.

Besides removing the problem shift-variance, the RDWT is attractive for our purposes because it can be linked to the spatial structure of the image in a mathematically well-defined framework. This is the theory of the *locally stationary wavelet process* (henceforth LSW) developed by Nason et al. (2000) and Eckley et al. (2010). To describe the ideas behind the LSW, we must first introduce some basic concepts from spatial statistics. Let $Z(\mathbf{r})$ denote a stochastic field of real-valued random variables on a regular discrete grid $\mathbf{r} = (x, y)^T \in \mathbb{Z}^2$. The probability of observing a certain subset of the state space $S \subset \mathbb{R}$ at some location $\mathbf{r}_1$ is given by

$$Prob(Z(\mathbf{r}_1) \in S) = \int_S d\mu_{\mathbf{r}_1}(z)\,,$$

where $\mu_{\mathbf{r}_1}$ is the probability measure associated with the process at $\mathbf{r}_1$. The expec-

[5] With the exception of the Haar wavelet, none of the Daubechies wavelets have an explicit functional representation.

tation value is defined as

$$E[Z(\mathbf{r}_1)] = \int_{\mathbf{R}} z \; d\mu_{\mathbf{r}_1}(z) \,,$$

provided that this integral exists. The relationship between the values at two locations $\mathbf{r}_1, \mathbf{r}_2$ can be described by the autocovariance function

$$Cov(Z(\mathbf{r}_1), Z(\mathbf{r}_2)) = E\left[(Z(\mathbf{r}_1) - E[Z(\mathbf{r}_1)]) \cdot (Z(\mathbf{r}_2) - E[Z(\mathbf{r}_2)])\right] \,.$$

How strongly the values at two different parts of the grid correlate with each other shapes the pattern of the resulting images. Cov thus formally describes the spatial structure of the process.

We will call Z a *stationary* process if $E[Z]$ is the same at every location $\mathbf{r}$, and the covariance depends only on the distance $\mathbf{r}_1 - \mathbf{r}_2 := \boldsymbol{\tau}$. Under this assumption, information on the covariance structure can be obtained from a single image since several pairs of locations belong to the same $\boldsymbol{\tau}$. We will refer to the stationary covariance function as

$$c(\boldsymbol{\tau}) = Cov(Z(\mathbf{r}), Z(\mathbf{r} - \boldsymbol{\tau})) \,. \tag{2.14}$$

For a sufficiently well-behaved process, we can equivalently study c's Fourier transform, the so-called *spectrum*

$$\Gamma(\boldsymbol{\omega}) = \mathcal{F}\{c\}(\boldsymbol{\omega}) \,,$$

which describes the distribution of the total variance across frequencies. Here, $\mathcal{F}$ denotes the two-dimensional Fourier transform and $\boldsymbol{\omega}$ the vector of frequencies. According to equation 2.14, an estimate of $c(\boldsymbol{\tau})$ is given by a convolution of some realization Z_1 onto itself. By equation 2.3, the squared Fourier transform $|\hat{Z}_1(\boldsymbol{\omega})|^2$ can therefore serve as an estimate of Γ.

Assuming global stationarity, the structure of a spatial field can thus conveniently and efficiently be studied based on the Fourier transform. For meteorological fields, this fully stationary setting is likely too restrictive, for example when multiple weather systems create different precipitation patterns throughout the domain. Nason et al. (2000) lift the strong stationarity assumption by replacing the waves of Fourier by wavelets and demanding that stationarity only holds *locally*. Here, we directly introduce the two-dimensional version of their theory, which was developed by Eckley et al. (2010). Assume that $Z(\mathbf{r})$ lives on some finite-sized regular grid, i.e.,

$$\mathbf{r} \in \mathcal{R} = \{0, 1, ..., N_x - 1\} \times \{0, 1, ..., N_y - 1\} \subset \mathbb{Z}^2$$

and can be written as a random superposition of two-dimensional Daubechies daughter wavelets:

$$Z(\mathbf{r}) = \sum_{j=1}^{\infty} \sum_{d=1}^{3} \sum_{\mathbf{u} \in \mathcal{R}} \psi_{j,d,\mathbf{u}}(\mathbf{r}) \cdot w_{j,d,\mathbf{u}} \cdot \xi_{j,d,\mathbf{u}} \tag{2.15}$$

Here, we have adopted the notation $\mathbf{u} = (u_x, u_y)^T \in \mathcal{R}$ for the shift vector and $d = 1, 2, 3$ for the orientation (horizontal, vertical or diagonal) of the daughter wavelets at scale j. Each daughter has an associated fixed weight w and is multiplied by a normally distributed random variable ξ with zero mean and unit variance. We further assume that all of the ξ for different scales, locations and orientations are uncorrelated.

We call Z a *locally stationary wavelet process* (LSW) if the following two conditions hold:

(i) Every $w_{j,d,\mathbf{u}}$ has an associated Lipschitz continuous[6] function $W_{j,d}(x,y)$ defined on $(x,y) \in (0,1) \times (0,1)$ which it approaches in the limit $N_x, N_y \to \infty$ in the following sense:

$$\sup_{\mathbf{u}} |w_{j,d,\mathbf{u}} - W_{j,d}(u_x/N_x, u_y/N_y)| \leq C_{j,d}/\max(N_x, N_y)$$

where the sum over all $C_{j,d}$ is finite.

(ii) The Lipschitz constants $L_{j,d}$ for $W_{j,d}$ satisfy

$$\exists K \in \mathbb{R}: \;\; \forall j,d : L_{j,d} \leq K$$
$$\sum_{j=1}^{\infty} \sum_{d=1}^{3} 2^{2j} L_{j,d} < \infty$$

To understand the idea behind this definition, imagine that Z generates the image of the butterfly in figure 2.8 at increasingly high resolutions as $N_x, N_y \to \infty$. The further we zoom in, i.e., the greater the resolution, the closer the coefficients $w_{j,d,\mathbf{u}}$ have to approach the continuous functions $W_{j,d}$ whose variation in space must vanish across infinitely small intervals – the process approaches stationarity in small regions if we zoom in far enough.

The locally stationarity counterpart to the classic relationship between the stationary auto-covariance and the Fourier transform is then given by

$$\sum_{d} \sum_{j=1}^{\infty} |W_{j,d}(\mathbf{r})|^2 \Psi_{j,l}(\boldsymbol{\tau}) = \lim_{N_x, N_y \to \infty} Cov(Z(\mathbf{r} \circ \mathbf{N}), Z(\mathbf{r} \circ \mathbf{N} + \boldsymbol{\tau})), \tag{2.16}$$

where $\mathbf{r} \circ \mathbf{N}$ is the element-wise product $(xN_x, yN_y)^T$ and $\Psi_{j,l}$ denotes the so-called *autocorrelation wavelet*

$$\Psi_{j,l}(\boldsymbol{\tau}) = \sum_{\mathbf{u} \in \mathbf{Z}^2} \psi_{j,d,\mathbf{u}}(\mathbf{0}) \psi_{j,d,\mathbf{u}}(\boldsymbol{\tau}), \tag{2.17}$$

i.e., the spatial autocorrelation of the redundant discrete daughter wavelets themselves. Eckley et al. (2010) prove that this representation is unique, meaning that the process is uniquely defined by the local covariances *or* the limiting functions W. In analogy to the Fourier case, the set of these squared coefficients $|W_{j,d}(\mathbf{r})|^2$ is called the *local wavelet spectrum*. We can obtain an estimator of this spectrum by computing the RDWT of a realization Z_1 and squaring the wavelet coefficients, i.e., $I_{j,d,\mathbf{u}} = \langle Z_1, \psi_{j,d,\mathbf{u}} \rangle^2$. The expectation value of $I_{j,d,\mathbf{u}}$ is asymptotically biased. Uniting the indices (j,d) into a single index η for simpler notation, this result can be written as

$$E[I_{\eta,\mathbf{u}}] = \sum_{\eta'} \underbrace{\langle \Psi_\eta, \Psi_{\eta'} \rangle}_{:=A_{\eta\eta'}} |W_{\eta'}(u_x/N_x, u_y/N_y)|^2 + \mathcal{O}(\min(N_x, N_y)^{-1}).$$

[6] $f(\mathbf{r})$ is *Lipschitz* if there is a non-negative Lipschitz constant L such that $|f(\mathbf{r}_1) - f(\mathbf{r}_2)| \leq L||\mathbf{r}_1 - \mathbf{r}_2||$

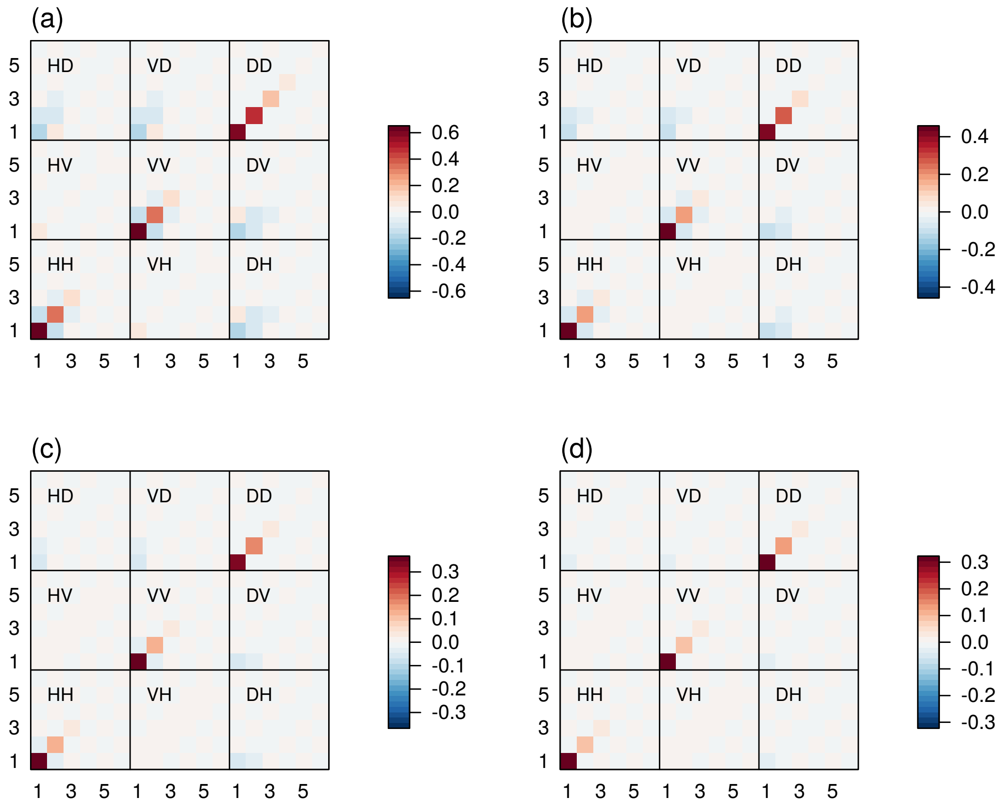


Figure 2.12: Bias correction matrices A^{-1} for the Haar wavelet (a), D_2 (b), D_4 (c) and D_8 (d) on a domain of size 64×64

Our naive estimator of the coefficient for scale j and direction d thus contains information from the true limit coefficients at all (j, d). How much information is “leaking” from one pair of coefficients to another is determined by the matrix A, containing the scalar products of all pairs of autocorrelation wavelets. The bias can be removed multiplying I_η at every location by the inverse matrix A^{-1}:

$$E\left[(A^{-1}I_\eta)_{\mathbf{u}}\right] = |W_\eta(u_x/N_x, u_y/N_y)|^2 + \mathcal{O}(\min(N_x, N_y)^{-1})$$

Figure 2.12 shows some examples of the bias correction operator A^{-1} which depends on the choice of mother wavelet ψ and resolution (N_x, N_y). The values on the main diagonal of A^{-1} are all positive and decrease with increasing scale. The correction thus mainly decreases the large-scale coefficients relative to the smaller scales, thereby compensating for the greater degree of redundancy of the larger daughter wavelets. This effect decreases as we move from coarse wavelets with short support (figure 2.12 a) to larger smoother basis functions (d). As a secondary effect, the first side diagonals corresponding to the horizontal and vertical daughters contain negative values, indicting that information is leaking between neighboring scales. Conversely, for the diagonal direction, these entries are closer to zero but some of the cross-terms between the diagonal daughters and their horizontal and vertical sisters are negative. This reflects the special role of the diagonal direction which is not a rotation of the other two (cf. figure 2.7).

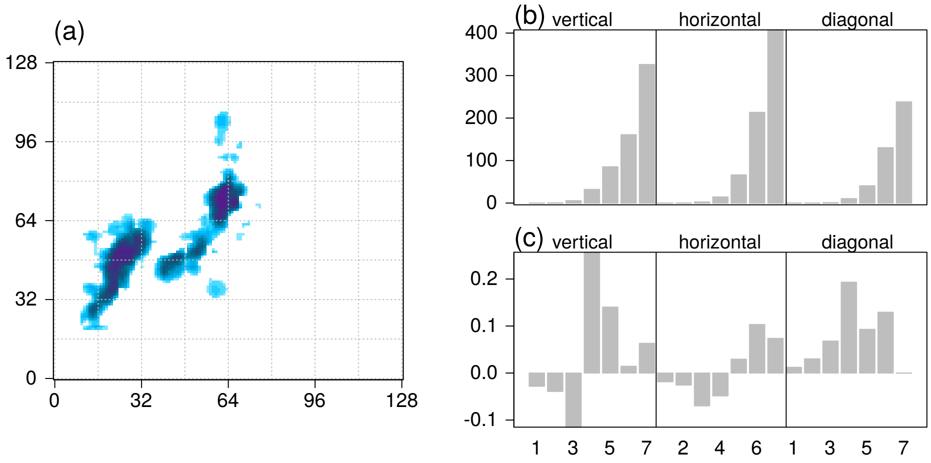


Figure 2.13: Example rain field from figure 2.10 (a), corresponding uncorrected spatial mean Haar-wavelet spectrum (b) and the result after bias correction with A^{-1} (c).

We can thus summarize the overall covariance structure of an image by computing the RDWT, applying the bias correction and averaging the squared coefficients over space to obtain one value for each combination of scale and direction. These coefficients can be interpreted as a measure of the variability represented by each pair (j, d). This analysis is shown for our example rain field in figure 2.13. Before the bias correction, the coefficients grow monotonically with scale due to the increasing degree of redundancy (panel b). A^{-1} transfers much of that energy to the intermediate scales (3-5) and strongly increases the small-scale diagonal entries at the expense of the other two directions. The resulting spectrum (panel c) has a peak in the vertical direction at scale four (roughly 16 pixels for the Haar wavelet). The diagonal variance component has increased values on larger scales, corresponding to the fact that the rain features are elongated most strongly along the diagonal. Both the horizontal and vertical spectrum contain substantial negative values which pose a challenge to our interpretation in terms of "variability" which should intuitively be greater than zero. Such questions are addressed in chapter 4.

2.4 The dual-tree complex wavelet transform

Besides the distribution of an image's variance across a range of spatial scales, directional aspects are an important part of the spatial structure. Is the pattern round or linear? Are there sharp edges? What are the preferred directions? Figure 2.14 (b) demonstrates that the orthogonal DWT from section 2.2 is not very well suited to answer such questions: If we rotate the input image, one would intuitively expect that each group of directional daughter wavelets becomes dominant at some angle where the pattern is perfectly aligned with the horizontal, vertical or diagonal direction. In reality, the diagonal daughters always capture less variability than their vertical and horizontal sisters who, as a result, each dominate across a broad range of angles. From this analysis, one would conclude that the original image was fairly anisotropic, whereas the same image rotated by 70° is nearly isotropic (all three directions have roughly the same energy).

The issue is partly remedied if we use the bias corrected RDWT (figure 2.14 c) instead. Thanks to the improved behavior under shifts, all curves look considerably smoother. As discussed above, the bias correction moves energy from the horizontal and vertical to the diagonal direction, which is now almost on par with the others. Notice, however, that 45° is the dominant direction for two shorter intervals while the other two directions appear only once for a larger range of angles. This effect is due to the "checkerboard" pattern of the diagonal daughters (figure 2.7) whose direction is both +45° and −45°. Using this transform, it is therefore impossible to decide between the two distinct diagonal orientations. In addition, the overall degree to which one direction dominates, i.e., the analyzed anisotropy, still depends considerably on the orientation of the image.

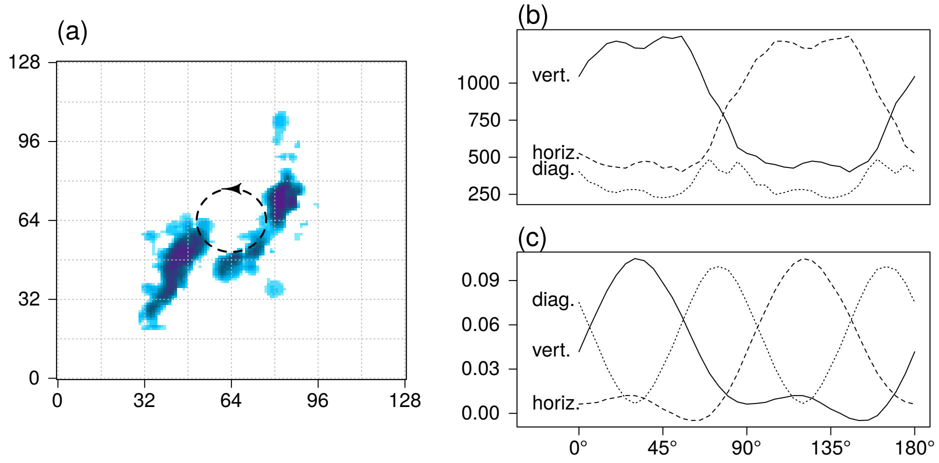


Figure 2.14: Wavelet spectra averaged over all locations and scales as a function of the angle by which the image in (a) is rotated. Panel (b) corresponds to the orthogonal Haar DWT, (c) shows the corresponding bias corrected RDWT.

The poor directional properties of the classic DWT were one of the key motivations behind the *Dual-Tree Complex Wavelet Transform* (henceforth DTCWT) of Kingsbury (1999)[7]. The following overview of their approach is primarily based on the tutorial paper by Selesnick et al. (2005). As implied in the name, the basic idea of the DTCWT is to replace the real-valued mother and father wavelets ψ and ϕ by complex functions

$$\begin{aligned}\psi(x) &= \psi_r(x) + i\psi_i(x)\\ \phi(x) &= \phi_r(x) + i\phi_i(x)\,,\end{aligned}$$

where (ψ_r, ψ_i) and (ϕ_r, ϕ_i) are pairs of real-valued wavelets and scaling functions, respectively, which are out of phase by 90°[8]. This phase-shift is analogous to the Fourier basis functions whose real and imaginary part are sine and cosine-waves. It can be shown that the resulting complex parent functions then have only positive frequency components, which will be the key to overcoming the poor directionality.

In practice, the projection onto the real and imaginary parts of the wavelet can be realized by two separate MRAs with mother and father filters g_r, g_i, h_r, h_i which are designed such that the two resulting functions satisfy the phase-shift condition. The corresponding diagram is identical to two copies of figure 2.3 – hence the "Dual-Tree" part of the name. The double redundancy of this transform (twice as many coefficients as input values) ameliorates the shift-sensitivity of the classic DWT: Each support contains two wavelets which are out of phase with each other. A small shift of the input will result in a transfer of energy from the imaginary to the real part (or vice versa). The *phase* of $\langle A, \psi\rangle$ changes while the amplitude remains nearly constant.

As in the real-valued DWT, we can create a two-dimensional transform by applying one filter to the rows and another to the columns of and image. One complex diagonal daughter is then given by

$$\begin{aligned}\psi(x)\psi(y) &= (\psi_r(x) + i\psi_i(x)) \cdot (\psi_r(y) + i\psi_i(y))\\ &= \psi_r(x)\psi_r(y) - \psi_i(x)\psi_i(y) + i(\psi_r(x)\psi_i(y) + \psi_i(x)\psi_r(y))\,.\end{aligned} \tag{2.18}$$

To see how the phase-shift impacts the direction of the resulting wavelet, let us imagine, for now, that ψ is simply a Haar wavelet with support length 4 and a phase-shift by 90° corresponds to a shift of the wavelet by one unit[9]. According to equation 2.18, the real part of $\psi(x)\psi(y)$ then schematically looks like this:

$$\underbrace{\begin{bmatrix} & & & & \\ - & - & + & + & \\ - & - & + & + & \\ + & + & - & - & \\ + & + & - & - & \end{bmatrix}}_{\psi_r(x)\psi_r(y)} - \underbrace{\begin{bmatrix} & - & - & + & + \\ & - & - & + & + \\ & + & + & - & - \\ & + & + & - & - \\ & & & & \end{bmatrix}}_{\psi_i(x)\psi_i(y)} = \underbrace{\begin{bmatrix} & + & + & - & - \\ - & & + & & - \\ - & - & & + & + \\ + & & - & & + \\ + & + & - & - & \end{bmatrix}}_{\Re(\psi(x)\psi(y))}$$

For better visibility, we have marked positive values by $+$ and negative values by $-$. We note that the orientation of $\psi_r(x)\psi_r(y)$ and $\psi_i(x)\psi_i(y)$ is ambivalent while their

[7] Notably, lacking shift invariance is also among the issues that the DTCWT is meant to address.

[8] More precisely, they are a *Hilbert pair*, i.e., $\psi_i = \mathcal{H}(\psi_r) = (\frac{1}{\pi t} * \psi_r)$.

[9] This is not the Hilbert transform, just a simplified example!

difference is uniquely oriented at 45° to the left of the vertical. For the imaginary part, we find

$$\underset{\psi_r(x)\psi_i(y)}{\begin{bmatrix} - & - & + & + & \\ - & - & + & + & \\ + & + & - & - & \\ + & + & - & - & \\ & & & & \end{bmatrix}} + \underset{\psi_i(x)\psi_r(y)}{\begin{bmatrix} & & & & \\ & - & - & + & + \\ & - & - & + & + \\ & + & + & - & - \\ & + & + & - & - \end{bmatrix}} = \underset{\Im(\psi(x)\psi(y))}{\begin{bmatrix} - & - & + & + & \\ - & - & & + & + \\ + & & - & & + \\ + & + & & - & - \\ & + & + & - & - \end{bmatrix}}$$

The imaginary part thus has the same orientation and is phase shifted with respect to the real part (notice how empty the diagonal in $\Re(\psi(x)\psi(y))$ is filled with negative values in $\Im(\psi(x)\psi(y))$). A daughter wavelet for the other diagonal is given by

$$\begin{aligned} \psi(x)\overline{\psi(y)} &= (\psi_r(x) + i\psi_i(x)) \cdot (\psi_r(y) - i\psi_i(y)) \\ &= \psi_r(x)\psi_r(y) + \psi_i(x)\psi_i(y) + i(\psi_i(x)\psi_r(y) - \psi_r(x)\psi_i(y)) \, . \end{aligned} \tag{2.19}$$

Following the same scheme as equations 2.18 and 2.19, we obtain a total of six complex daughter wavelets with six distinct orientations:

$$\begin{aligned} \psi(x)\psi(y) &= \psi_r\psi_r - \psi_i\psi_i + i(\psi_i\psi_r + \psi_r\psi_i) \\ \psi(x)\overline{\psi(y)} &= \psi_r\psi_r + \psi_i\psi_i + i(\psi_i\psi_r - \psi_r\psi_i) \\ \psi(x)\phi(y) &= \psi_r\phi_r - \psi_i\phi_i + i(\psi_i\phi_r + \psi_r\phi_i) \\ \psi(x)\overline{\phi(y)} &= \psi_r\phi_r + \psi_i\phi_i + i(\psi_i\phi_r - \psi_r\phi_i) \\ \phi(x)\psi(y) &= \phi_r\psi_r - \phi_i\psi_i + i(\phi_i\psi_r + \phi_r\psi_i) \\ \phi(x)\overline{\psi(y)} &= \phi_r\psi_r + \phi_i\psi_i + i(\phi_i\psi_r - \phi_r\psi_i) \end{aligned} \tag{2.20}$$

Here, we have dropped the arguments of the functions on the right hand sides in the interest of shorter notation, the first function is always implicitly of x and the second of y. The complex conjugates of these wavelets ($\overline{\psi(x)}\psi(y)$ etc.) would add no new directions since the imaginary parts are merely inverted. The two-dimensional DTCWT can be realized as four separate DWTs with the four different combinations of real and imaginary filters applied to the rows and columns. This procedure, including the re-combination into the six daughters according to equation 2.20, is shown in figure 2.15. Despite the seemingly complicated diagram, the algorithm is actually easy to understand and implement as it merely consists of four completely independent DWTs, the results of which can simply be added and subtracted at the very end.

Instead of the Haar wavelet from our crude example above, other filters with better frequency localization are used in practice. For an overview of different filter-design approaches for the DTCWT, we refer to Selesnick et al. (2005) and references therein. As a further technical detail, it must be mentioned that the first stage of the DTCWT is usually treated by a different set of filters than all others. This is related to the fact that the desired 90° phase-shift can only approximately be realized for a finite-sized filter. In particular, it is violated for the first stages where the daughter wavelets are short. This effect can be compensated by using a different set of filters for the first level, chosen such that g_i is equal to g_r, shifted by one sample (and likewise for h_r and h_i).

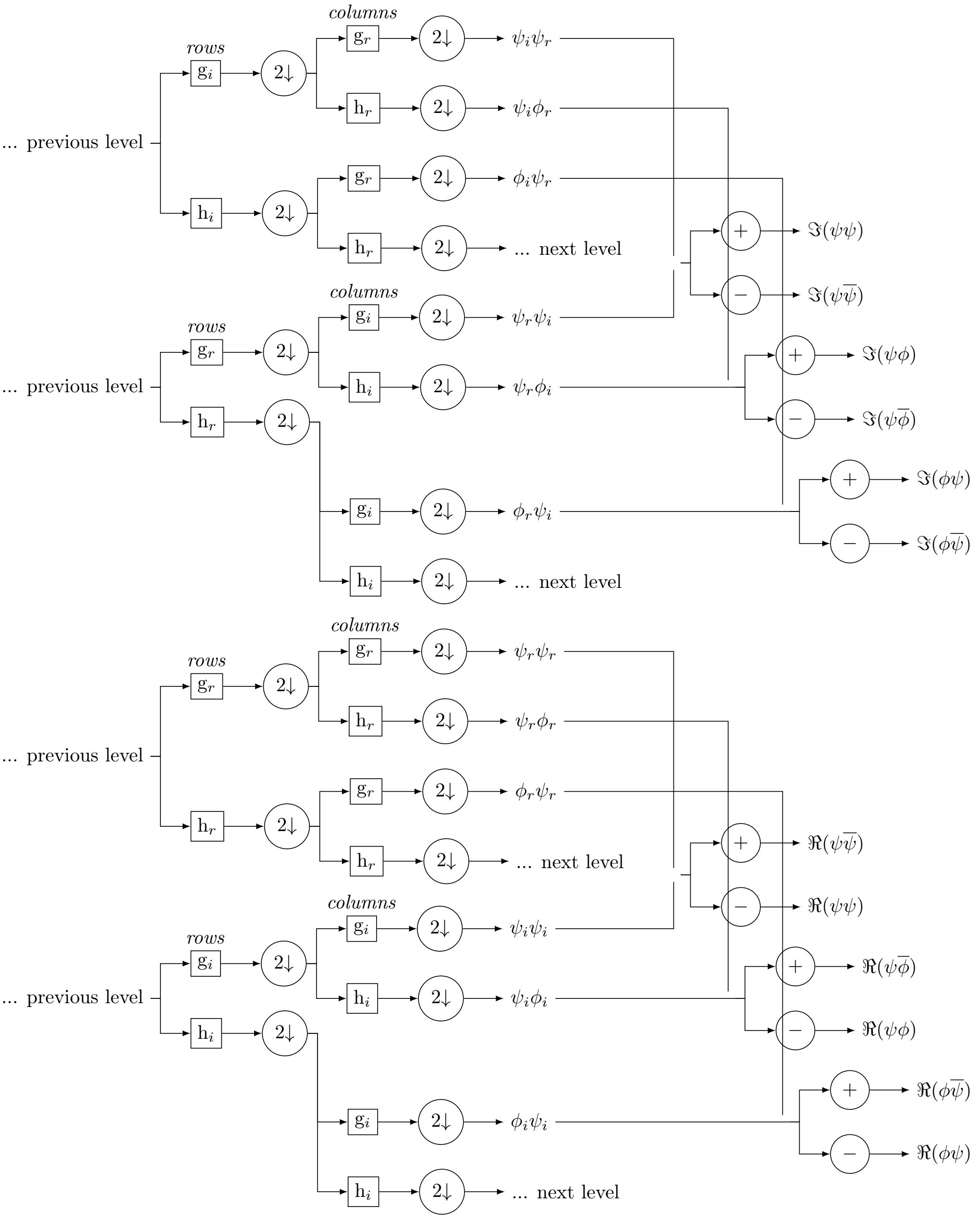


Figure 2.15: One level of the two-dimensional DTCWT. g_r, g_i, h_r, h_i are the filters corresponding to $\psi_r, \psi_i, \phi_r, \phi_i$. The split, followed by + and −, indicates that the two inputs are added or subtracted.

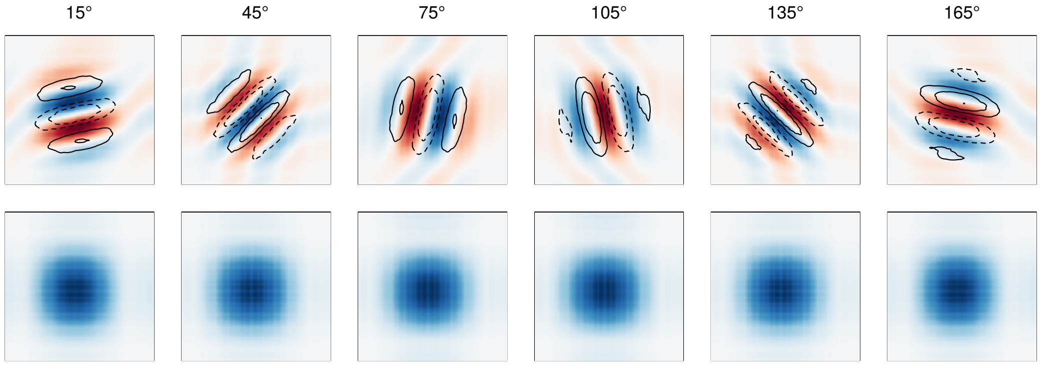


Figure 2.16: DTCWT daughter wavelets corresponding to the `q_shift_b` and `near_sym_b` filters from the `dualtrees` R-package. Top row: Real and imaginary part as colors and contours. Bottom: Modulus squared.

Figure 2.16 shows the resulting daughter wavelets at scale $j = 5$ using a nearly symmetrical biorthogonal wavelet with 19 wavelet filter coefficients for the first level and "quarter-shift" wavelet filters with 14 coefficients on all other levels. As desired, the real and imaginary part are shifted by 90°. The overall amplitude $|\psi|^2$, shown the bottom row, follows a smooth bell-shape without oscillating. This is another advantage of the DTCWT as it allows for an unambiguous localization of features.

The horizontal and vertical direction are replaced by 15°, 165°, 75° and 105°. As in our simple example above, the two diagonals are now distinctly oriented at 45° and 135°. Notice, however, that these two daughters still differ qualitatively from the other four (compare the number of wave-crests) because they are still generated from two high-pass filters. Kingsbury (2006) address this by applying special filters $g_r^{\text{diag}}, g_i^{\text{diag}}$ in the diagonal branches of the transform. In the resulting DTCWT, all directions are (nearly) equal. As a caveat, the inverse transform can no longer perfectly re-construct the input image.

To demonstrate the benefits of the DTCWT, we repeat our experiments on shifted and rotated images (figures 2.14 and 2.10, respectively). Figure 2.17 a) demonstrates that the shift invariance is nearly perfect. The RDWT achieves the same result at the cost of a $3 \times J$-fold redundancy (for an input image of size $2^J \times 2^J$) – the DTCWT is redundant by a factor of only four. The directional behavior (panel b) is also clearly improved compared to the DWT (figure 2.14 b), each direction has one distinct maximum. The outsider role of the diagonal daughters, however, remains clearly visible. Using the corrected diagonal filters of Kingsbury (2006) (panel d), we can achieve a near-perfect balance between the six orientations. A bias-correction as in figure 2.14 (c) is not needed. Due to the increased number of sampled directions, we can also expect that measures of anisotropy derived from this transform should not depend strongly on the orientation. More details are given in chapter 4, as well as Buschow & Friederichs (2021a) (in appendix C), where such a measure is explicitly defined.

Despite the already nearly perfect shift-invariance of the DTCWT, a fully redundant version of this transform is nonetheless needed when we want to quantify the spatial structure at *every* individual location. The redundant DTCWT can be realized by replacing each of the four DWTs by the corresponding RDWT, i.e.,

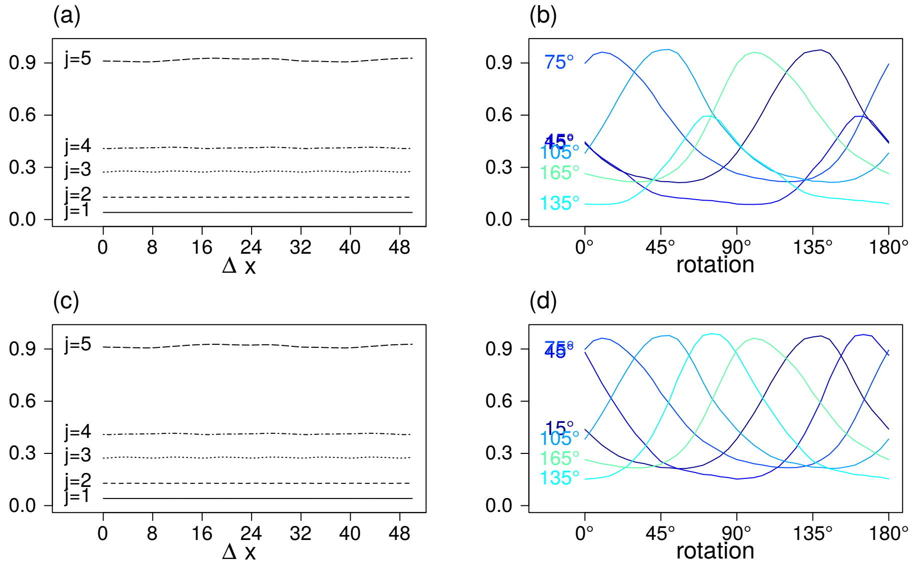


Figure 2.17: Shift- and rotation sensitivity of the DTCWT. (a) and (c) correspond to the shift-experiment from figure 2.10, (b) and (d) to the rotations from figure 2.14. The top row (a,b) shows results for the wavelets shown in figure 2.16, the diagonal-corrected wavelets following Kingsbury (2006) are shown below (c,d).

removing the downsampling step and upsampling the filters instead. Nelson et al. (2018) present the extension of the LSW framework to the case of DTCWT wavelets. Their definition is identical to equation 2.15, except that ψ and w are both complex valued. The stationarity conditions and the need for a bias correction matrix A^{-1} are unchanged as well. There is, at the time of writing, no proof that A is invertible for any wavelets other than the Daubechies family used by Nason et al. (2000) and Eckley et al. (2010). Nelson et al. (2018) report that, in practice, A^{-1} can be computed for the DTCWT case. This can be achieved by (1) numerically calculating the autocorrelation wavelets via the complex counterpart to equation 2.17 (taking the complex conjugate of the first wavelet in the expression), (2) calculating the matrix of their inner products A and (3) numerically inverting it, if possible.

As an example, figure 2.18 shows how the bias-corrected redundant DTCWT can be used to locally analyze the preferred orientations in an image. By marking the regions where the directional components are largest, we have essentially performed a crude multi-scale edge detection algorithm: The legs and antennae are very slim and show up only on the smallest scales. Fine horizontal and diagonal lines in the bottom half of the wing show up on this scales as well. At $j = 2, 3$, the left edge of the wing is traced by large values for various directions following the shape of the edge; for the two largest scales, the broad light spots in the upper half of the wing become the main feature. Only the highest contrast along the left side of the wing is intense enough to show up in the top 1 % of coefficients at scale 5.

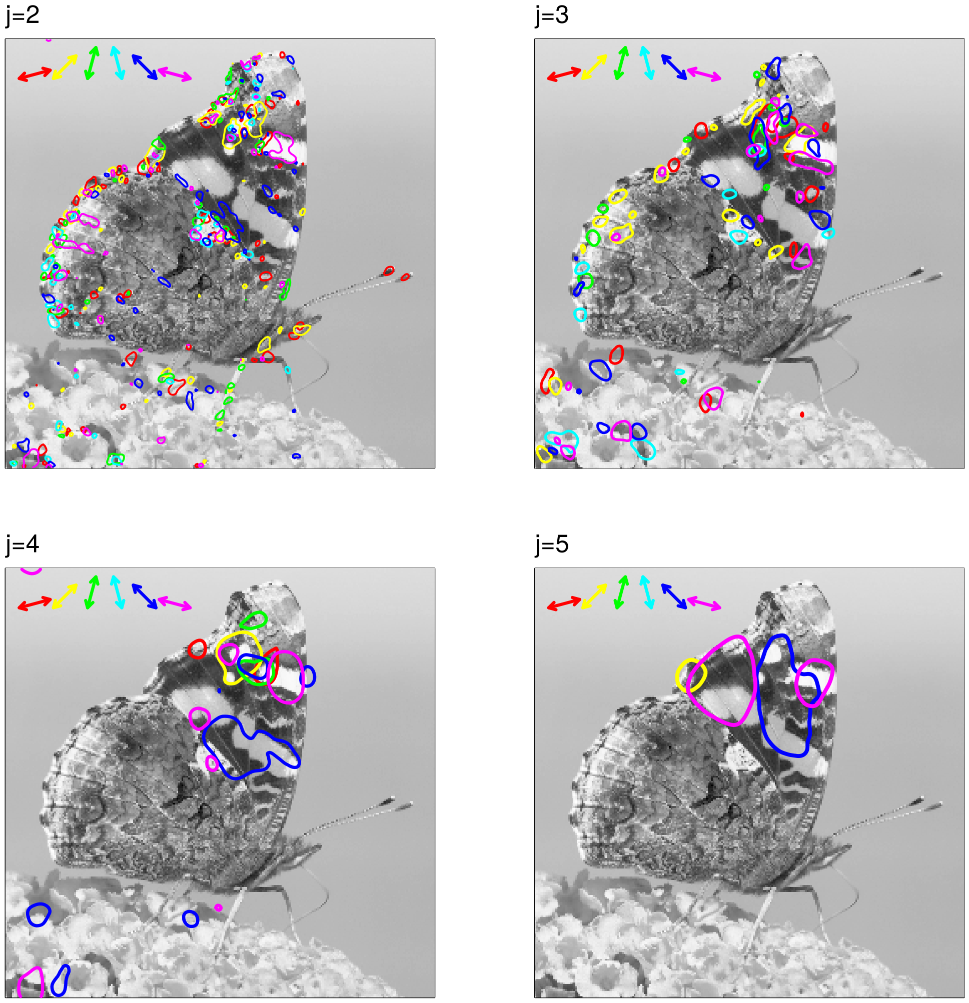


Figure 2.18: Redundant DTCWT analysis of the butterfly test image: Contours contain pixels where the squared coefficients for a particular direction are in the top 1 % of all values for a particular scale $j = 2, 3, 4, 5$.

Chapter 3

Spatial Verification

This chapter introduces the main approaches to the problem of spatial forecast verification found in the literature. We begin with the canonical classification of verification techniques (Gilleland et al., 2009; Dorninger et al., 2018) based on their technical implementation. For each of the five classes, one popular example is described in some detail. Next, we survey the various kinds of forecast errors and see the reactions of the existing scores. The penultimate section of this chapter discusses different ways of assessing the merits of a verification technique. In the final section, we revisit the current state of the art in wavelet-based forecast verification in light of the preceding discussions and identify open research questions.

3.1 The five styles of spatial verification

One main result of the ICP project was the successful categorization of the numerous existing spatial verification techniques into four classes (Gilleland et al., 2009), namely neighborhood, feature-based, field-deformation and scale-separation methods. Binary distance metrics, originally seen as a type of field-deformation method, were later re-classified into their own, fifth category (Dorninger et al., 2018). This already hints at the imperfections of the system (as is rightfully acknowledged Dorninger et al. 2018): Some approaches combine multiple styles (Lack et al., 2010; Yano & Jakubiak, 2016; Yu et al., 2020), others are hard to classify at all (Marzban & Sandgathe, 2009; Hou & Wang, 2019). The five-class framework is nonetheless widely used and very helpful to get an overview of the many different ways in which the spatial verification problem has been tackled. The subsequent sections therefore describe each class in turn and introduce one example score in detail. These five scores will serve as illustrative examples for sections 3.2 and 3.3, and help put the new verification strategy introduced in chapter 4 into context.

3.1.1 Neighborhood methods: FSS

A very straightforward solution to the double-penalty problem is to relax the requirement that forecast and observation should agree at each individual grid-point. Instead, we may consider a neighborhood around each grid-point and ask whether the mean properties in this neighborhood are in agreement. If the data are given as gridded fields, this is equivalent to applying some form of smoothing filter.

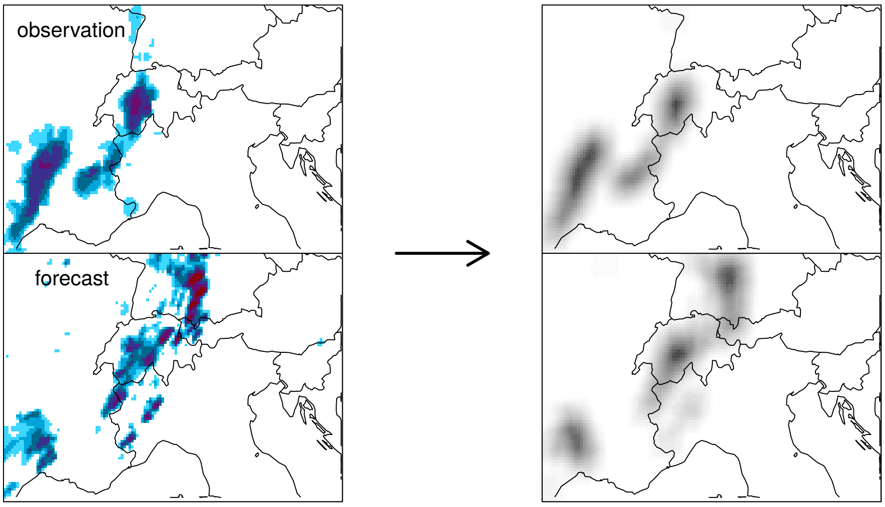


Figure 3.1: Example of the neighborhood verification approach: Forecast and observation are thresholded, the result is smoothed via a moving average.

Figure 3.1 shows why this approach is sometimes called *fuzzy* verification: The information whether it rained at any particular pixel is smeared out across a neighborhood of size L, such that displacement errors smaller than L are not punished. Here, the smoothing was not applied to the original intensities, but to the thresholded fields of ones and zeros, indicating whether some threshold T was exceeded at each point. This is the approach of the Fractions Skill Score (FSS) introduced by Roberts & Lean (2008), which is summarized in algorithm 2.

Algorithm 2 Fractions Skill Score of Roberts & Lean (2008)

Input: forecast Y, reference X, thresholds T, neighborhood sizes L
Output: FSS as a function of T and L

```
for all $T$,$L$ do
    calculate the fraction of $X$ and $Y$ exceeding $T$ in a square neighborhood of size $L$ around each grid-point.
    calculate the mean-square error (MSE) between the two fields of fractions
    calculate the reference $\mathrm{MSE}_{\mathrm{ref}}$ as the maximum possible MSE in the case where forecast and observation are uncorrelated in space
    calculate $\mathrm{FSS}_T(L) = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE}_{\mathrm{ref}}}$
end for
```

The result of the FSS-analysis is a group of curves $\mathrm{FSS}_T(L)$ indicating, for each threshold, the forecasts skill as a function of the spatial scale below which errors are neglected. The limit value $\mathrm{FSS}(L \to \infty)$ is equal to one if the frequency $f_o(T)$ with which T is exceeded in the observations was predicted correctly. Positive (negative) values of the asymptotic FSS indicate a positive (negative) frequency bias. To aid the interpretation of the FSS, Roberts & Lean (2008) furthermore introduce the concept of a minimum skillful scale l_{min}, defined as the smallest neighborhood size with $\mathrm{FSS}_T(l) > 0.5 + f_o(T)$. The threshold $0.5 + f_o(T)$ is the value of FSS one would obtain at $L = 1$ if the predicted fraction at every grid-point were $f_o(T)$. This serves as a baseline skill which a useful forecast should ideally exceed. The ability to make statements like *"forecast X can skillfully predict precipitation locations within a range of $l_{min} \approx ... km$"* has contributed to FSS's considerable popularity. For an overview of several other neighborhood techniques, we refer to Ebert (2008).

3.1.2 Scale-separation methods: ISS

It is interesting to note that Ebert (2008) includes the wavelet-based Intensity Scale Skill-Score (ISS) of Casati et al. (2004) in their review of fuzzy methods. This technique was only later re-classified into the scale-separation category. The similarity and possible confusion between the two classes is also evidenced by the full title of Roberts & Lean (2008) which describes FSS as *"scale selective"*. The canonical distinction between neighborhood and scale-separation is that methods in the former category apply some form of low-pass filter (i.e. smoothing) to remove all variability below a certain spatial scale, while the latter class of techniques relies on high-pass filters to split the total variability up into a spectrum of scales. To see the fundamentally different kinds of information that are generated by such a spectral decomposition, we consider the aforementioned ISS, given by algorithm 3.

Algorithm 3 Intensity-Scale Skill-Score of Casati et al. (2004)

```
Input: forecast $Y$, reference $X$, thresholds $T$
Output: ISS as a function of $T$ and and scale $j$

1: apply noise to all rainy pixels in $Y$ and $X$ to avoid discretization errors
2: replace non-zero values in both fields by their binary logarithm (set zero values
   to $-6$) to render the distribution more nearly normal
3: re-calibrate $Y$ to the marginal distribution $F_X(x)$ via $Y' := F_X^{-1}(F_Y(Y))$
4: for all $T$ do
5:    convert $X$ and $Y'$ into binary images by thresholding at $T$
6:    compute the difference $X - Y'$
7:    compute the observed fraction of threshold exceedances $f_o(T)$
8:    decompose $X - Y'$ into components $Z_j$ at scales $j = 1, \ldots, J$ via a Haar-MRA
9:    for all scales $j$ do
10:      compute the spatial mean $\overline{Z_j^2}$
11:      compute $ISS(j,T) = 1 - \overline{Z_j^2} / (2 \cdot f_o(T) \cdot (1 - f_o(T)) \cdot J)$
12:   end for
13: end for
```

The reference score in the ISS (step 11) corresponds to a Poisson process. Due to the orthogonality of the MRA, the individual $\overline{Z_j^2}$ add up to the overall MSE between the thresholded fields. An example of this error decomposition is shown in figure 3.2. The key difference to fuzzy methods is that the information on different scales does not overlap: Consider an observed feature of size 2^j which was completely missed by the forecast. In this case, ISS will indicate decreased skill at scale j but not at

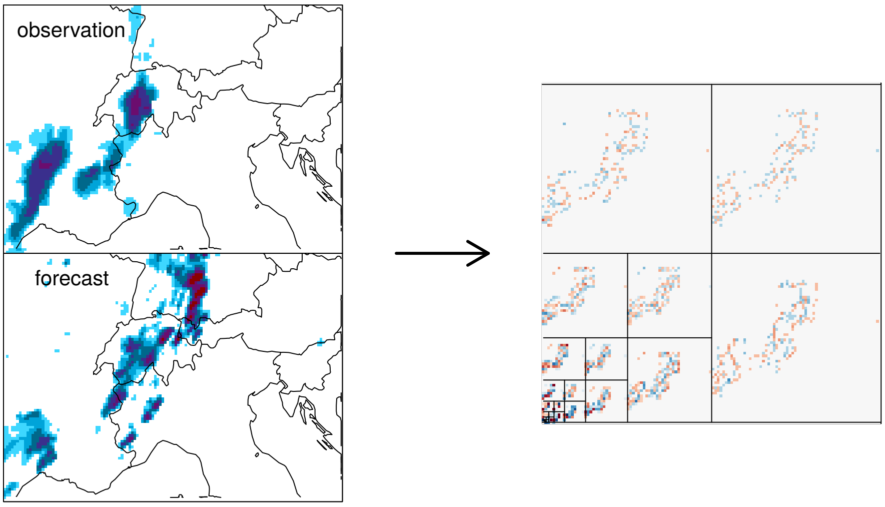


Figure 3.2: Illustration of the ISS: Separation of the binary difference between forecast and observation into different scales via a Haar-MRA.

$j+1, j-1$, etc.[1] In contrast, all FSS values are lowered by the missing feature until L is large enough to both contain the missing observed object and an equal amount of predicted precipitation. The observation that this L may well be much larger than j further indicates that scale and neighborhood size relate to different length scales: The former is more closely related to the size of the erroneous objects while the latter can be seen as a measure of their spatial displacement[2].

Other scale-separation techniques need not aim to decompose error-images into scales, but they all share the technical basis of isolating individual scales of variability; examples include Willeit et al. (2015) and Wong & Skamarock (2016) who rely on Fourier transforms and the variogram-based approaches of Marzban & Sandgathe (2009), Scheuerer & Hamill (2015) and Ekström (2016). An overview of other wavelet-based techniques is given at the end of this chapter.

3.1.3 Feature-based methods: SAL

The basic idea of feature-based spatial verification techniques is that a human expert, tasked with visually evaluating the forecast shown in figure 3.3, might analyze the images in terms of discrete objects: The observation may be decomposed into two large, elongated precipitation regions in the western half of the domain. In the forecast, the eastern feature looks similar to the observed one, only displaced to the North-East while the other object is nearly at the right location but has a completely different shape.

In terms of image processing, this corresponds neither to a high- nor low-pass filter but rather an image segmentation algorithm, as shown in the right part of figure 3.3: Forecast and observation are thresholded at some value T which may be

[1] This is only *exactly* true if the precipitation errors are shaped like Haar wavelets.

[2] Skok & Roberts (2018) exploit this fact to measure displacement errors using FSS

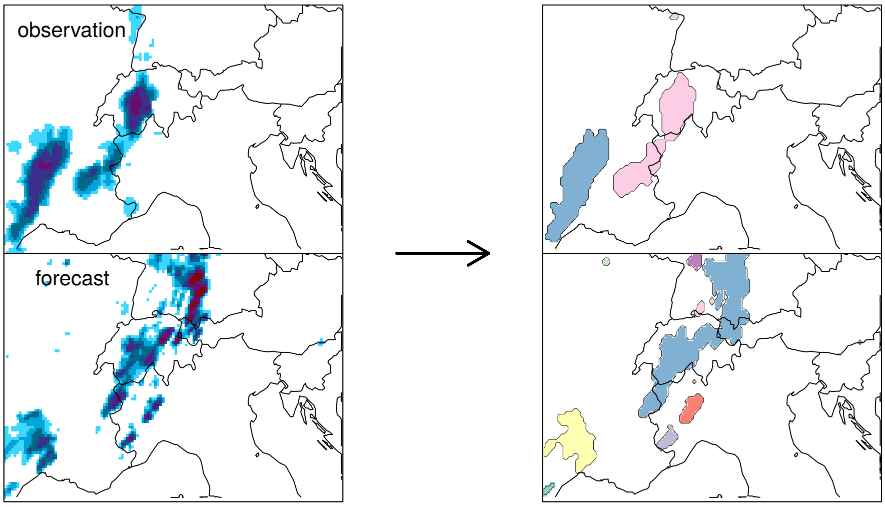


Figure 3.3: Example of an object decomposition procedure.

the same or different for the two images. In the resulting binary image, connected regions of ones are identified as discrete objects. To obtain fewer, more reasonable-looking objects, the segmentation is typically preceded by a spatial smoothing step (Davis et al., 2006). Depending on the specific method, a number of properties like location, shape, orientation or intensity are calculated for each resulting object.

As a relatively simple and popular example, we consider the Scale, Amplitude and Location method (Wernli et al., 2008, SAL) which records only the total rain intensity R_i, centroid location $\mathbf{r}_i$ and the ratio V_i between total and maximum intensity in each object i. We let $\langle . \rangle$ denote the average over all objects in a field, weighted by their intensities R_i. SAL then computes two object based characteristics for forecast and observation separately: The "peakedness" $V_{\text{obs, for}} = \langle V_i \rangle$ and the scattering of objects $\mathbf{r}_{\text{obs, for}} = \langle |\mathbf{r}_i - \langle \mathbf{r}_i \rangle| \rangle$, which result in two object-based scores:

$$S = 2 \cdot \frac{V_{\text{for}} - V_{\text{obs}}}{V_{\text{for}} + V_{\text{obs}}}$$
$$L_2 = \frac{2}{d_{\text{max}}} \cdot |\mathbf{r}_{\text{for}} - \mathbf{r}_{\text{obs}}| \,,$$

where d_{max} is the maximum distance between two points in the domain. These are combined with two non-object scores,

$$A = 2 \cdot \frac{R_{\text{for}} - R_{\text{obs}}}{R_{\text{for}} + R_{\text{obs}}}$$
$$L_1 = |\langle \mathbf{r}_i \rangle_{\text{for}} - \langle \mathbf{r}_i \rangle_{\text{obs}}| / d_{\text{max}} \,,$$

to obtain the full SAL analysis. Here, $R_{\text{obs, for}}$ denotes the total observed and forecast intensity. A is thus a simple relative intensity error, S compares the forecast and observed structure in terms of number and peakedness of objects, and $L = L_1 + L_2$ quantifies location errors.

Note that SAL uses the object-decomposition to calculate the properties of the two fields but does not compare individual features to each other. Other popular

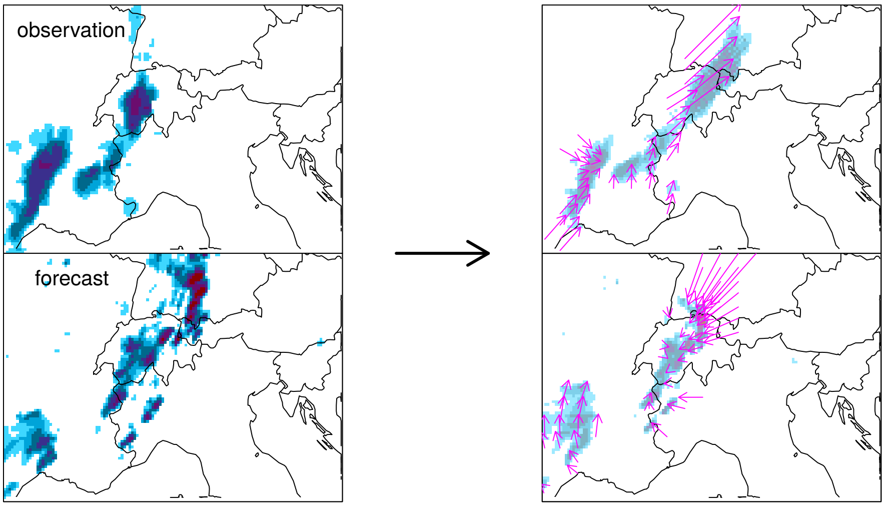


Figure 3.4: Schematic representation of a field deformation method. Top right: Optical flow of the observed into the forecast field. Bottom right: vice versa. Arrows go from the original grid points to the corresponding locations in the morphed image (pale colors).

techniques like CRA (Ebert & McBride, 2000) and MODE (Davis et al., 2006, 2009) emulate the human reasoning described in the beginning of this section more closely by attempting to match predicted to observed features. If successful, this step allows for a very precise quantification of displacement and other errors. Unfortunately, the question which parts of the two fields are supposed to represent the same feature is often difficult to answer objectively and may sometimes be entirely ill-posed (see discussion in section 3.2). CRA therefore often requires a human expert to inspect individual verification results (Mariani & Casaioli, 2018) while MODE calculates a total interest function by comparing every forecast object to each observed object, thereby largely circumventing the need to find individual matches.

3.1.4 Field-deformation methods: DAS

Another strategy to derive meaningful error measures between two images is to find an algorithm which explicitly corrects the errors and then quantify the amount of change made by that algorithm. This is the basic idea of field-deformation methods, which compute a vector field that transforms one image into the other (figure 3.4). In the field of computer vision, this type of vector field is known as an *optical flow* (Lucas & Kanade, 1981; Horn & Schunck, 1981). Our example for this class of verification scores is the Displacement and Amplitude Score (Keil & Craig, 2007, 2009, DAS), which employs the pyramid-matching scheme in algorithm 4.

Displacement errors are quantified by the mean absolute value of the estimated vector field, the point-wise error after application of the flow can be seen as an amplitude error. In order to properly treat false alarms and misses, Keil & Craig (2009) proposed to consider the flow separately at pixels with predicted precipitation

Algorithm 4 Pyramid matching optical flow of Keil & Craig (2007)

Input: forecast Y, reference X, maximum averaging length 2^N
Output: vector field approximately transforming Y into X

```
set $n = N$
while $n > 0$ do
    coarse-grain $X$ and $Y$ by averaging over $2^n$ pixels to obtain $X', Y'$
    shift $Y'$ by $\pm 2$ pixels in each direction
    apply a Gaussian smoothing kernel to $X' - Y'_{\text{shifted}}$
    for each coarse-grained pixel, select the shift which minimizes the local error
    interpolate the optimal shifts to the resolution of $Y$
    apply the optimal shifts to $Y$
    set $n = n - 1$
end while
compute the sum over all intermediate shift vector fields
```

(flow from the forecast into the observations, top row of figure 3.4) and at pixels with observed precipitation (from the observation into the forecast, bottom row of the figure).

Han & Szunyogh (2016, 2018) developed several adjustments to the original DAS-approach, including a separation of the error after morphing into an amplitude and a residual structure component. Other field-deformation studies include the optical flow implementations of Gilleland et al. (2010b), Marzban & Sandgathe (2010). More recently, Farchi et al. (2016) and Stucki et al. (2020) have applied the earth mover's distance (Rubner et al., 2000) to compute the minimum cost of transforming one field into the other exactly.

3.1.5 Distance Measures: Baddeley's Δ

Like the optical flow algorithms, wavelets and image segmentation, the final class of spatial verification techniques comes from the realm of image processing / computer vision. The basic idea of this approach is to compare two binary images by computing how far the non-zero pixels in one image are to the nearest non-zero pixel in the other. This is realized by thresholding forecast and observation and applying a distance transform (figure 3.5). One example score based on this so-called distance map is Baddeley's Δ (Baddeley, 1992). Let A and B denote the sets of locations $\mathbf{r}$ where forecast and observation have non-zero values (after thresholding) and let $d(\mathbf{r}, X)$ be the distance from location $\mathbf{r}$ to the next element of of the set X, i.e., the distance map shown in figure 3.5. Then Baddeley's Δ is given by

$$\Delta_{p,w}(A, B) = \left[\frac{1}{N} \sum_{i=1}^{N} |w(d(\mathbf{r_i}, A)) - w(d(\mathbf{r_i}, B))|^p \right]^{1/p} , \tag{3.1}$$

where the sum runs over all pixels i in the domain and w is a weighting function that can be used to weaken the influence of small outlying objects. Intuitively, we compare the distance to the nearest *predicted* rain pixel to the distance to the nearest *observed* rain pixel, thereby measuring the average displacement between non-zero regions in

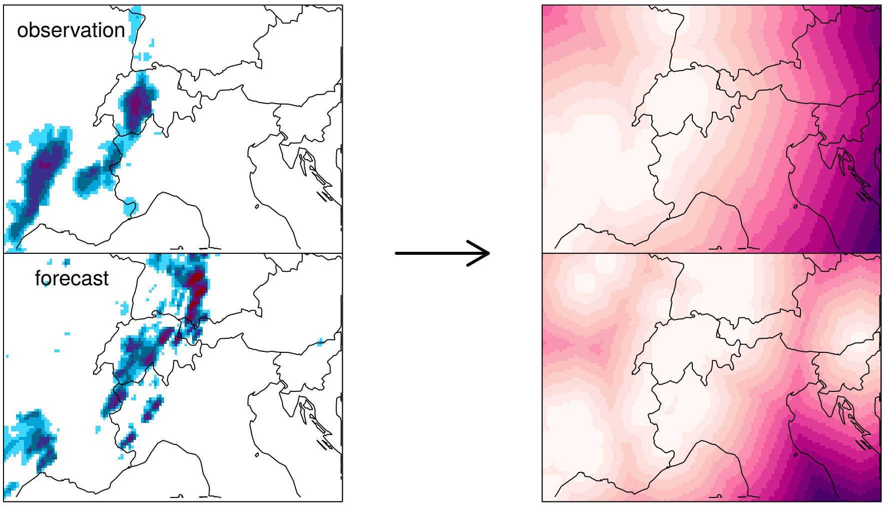


Figure 3.5: Distance transform of forecast and observations: Colors in the right panels indicate the distance to the nearest non-zero pixel in the respective image on the left (zero in white, darker colors correspond to greater distance).

the two binary images. It can be shown that Δ is a metric in the mathematical sense (positive, symmetric, triangle inequality), which is a desirable property since it guarantees that forecasts can be consistently ordered by their Δ-score. It was first applied to meteorological verification problems by Gilleland (2011). Other scores derived form the distance maps can be asymmetrical with respect to forecast and observation, thereby potentially differentiating misses from false alarms. For a recent review and inter-comparison of several distance measures, we refer to Gilleland et al. (2020). Further new developments, including a combined measure of distance and intensity errors, are described by Gilleland (2021).

	shift	rotation	bias	margin	autocor.
FSS	3	3	3	3	1
ISS	2	2	1	1	3
S	1	1	2	2	4
A	1	1	4	1	1
L	4	1	2	2	3
D$_{KC}$	4	3	1	1	3
A$_{KC}$	1	1	3	3	3
BD	4	2	2	2	3

1: invariant
2: abrupt change
3: smooth change
4: quantified

Table 3.1: Reaction of our example scores to shifts, rotations, biases, other changes to the marginal distribution and changes to the spatial auto-correlation. D$_{KC}$ and A$_{KC}$ denote the displacement and amplitude component of DAS as defined by Keil & Craig (2009).

3.2 Types of errors

In essence, the classification presented in section 3.1 is based on the type of data transformation used by each score. This approach does not necessarily tell us which scores can be expected to give similar results: Displacement errors, for example, can be quantified by FSS (neighborhood), BD (distance metric), SAL (objects) and DAS (deformation). To identify points of comparison for new wavelet-based techniques, we therefore now discuss the various kinds of possible forecast errors and how existing scores react to them. A non-exhaustive list of possible forecast errors includes shifts in space, rotations, additive and multiplicative biases, other changes to the marginal distribution and changes to the spatial auto-correlation structure. Scores can either (1) be invariant under these transformations, (2) react abruptly or (3) smoothly to them or (4) quantify them explicitly. For visual reference throughout this section, an artificial example of each error type is schematically shown in figure 3.6.

Table 3.1 roughly categorizes how each of our example scores reacts to the different kinds of errors. A similar overview including some additional scores is given by table 2 in Gilleland et al. (2009). Table 1 in Gilleland et al. (2010a) attempts to broadly assign kinds of errors to the different verification styles from the previous section. As these authors acknowledge, a one to one map between the technical basis of a score and the errors it quantifies, is not generally possible. Below, we discuss the behavior of our example methods under the various kinds of errors in some detail.

Shifts By default, deformation approaches like DAS allow for non-uniform optical flow fields which incorporate shifts of the entire image, shifts of objects within the image and arbitrary deformations of those objects into one final score. The same is true for BD and other distance measures for which the shape of the objects is as important as their location. Displacement and structure errors are thus not separated. A similar confusion occurs in SAL's location component: One half of this score is simply given by the distance between the two fields' centroids, which is invariant under any re-arrangement of the individual objects as long as the center of mass is unchanged. The other half measures the difference in scattering of individual objects around the centroids, which is arguably a matter of spatial correlation struc-

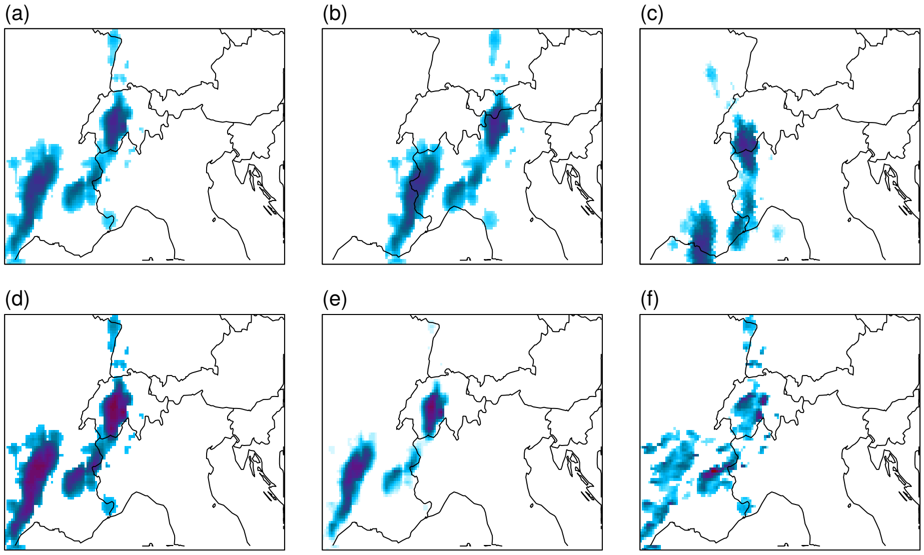


Figure 3.6: Different kinds of errors artificially applied to our example field (a): Shift (b), rotation (c), additive bias (d), altered marginal distribution (e) and erroneous correlation structure (f). The marginal error in (e) was realized by squaring the intensities, for the structural error in (f) we used the post-processing introduced in Buschow & Friederichs (2021a).

ture rather than location. The original formulation of FSS gives no explicit measure of displacement errors. Instead, its value at neighborhood size L is invariant under shifts $< L$. Skok & Roberts (2018) demonstrated that this behavior can be exploited to derive an explicit measure of displacement errors, but only if the area above the threshold is nearly the same in both fields. Our final example score ISS does react strongly to misplaced features but cannot explicitly distinguish between forecasts with additional, missing, mis-scaled or misplaced objects at a particular scale. Due to the orthogonal wavelet transform, shifts from one wavelet support to another can result in abrupt changes of the scores.

Rotations can be seen as a weak point of our five metrics as none of them can explicitly quantify cases where a predicted object has the wrong orientation. SAL, for example, is invariant under rotations of one or both complete fields and rotations of individual features around their center. Even worse, to the MRA used by ISS, diagonally oriented patterns look fundamentally different from vertical and horizontal ones (see section 2.4), resulting in potentially abrupt reactions to rotation. For DAS, rotation errors can lead to unintuitive flow fields because the rigid rotation that a human expert might recognize is not necessarily the optimal solution to the optimization problem. Marzban & Sandgathe (2010) nonetheless derive information on consistent directional errors from a similar optical flow approach. Other methods that can measure rotations explicitly include the object-based MODE, the 2D-Fourier approach of Wong & Skamarock (2016) and the very recent variogram-method introduced in Bellier et al. (2020).

Biases and other errors in the marginal distribution can easily be studied with univariate verification scores (an example being SAL's A component). It can therefore often be desirable to separate such errors from the spatial characteristics measured by multi-variate techniques. Any score that relies on thresholding the fields (FSS, BD, ISS, SAL and basically all other object-based approaches) will react non-smoothly to changes in the marginal distribution as features can shrink, disappear or fracture into multiple pieces when parts of the image cross the threshold. Wernli et al. (2008) describe the delightfully named *camel effect* where a feature with two peaks can split in twain when the threshold (or equivalently the mean of the data) changes. While these authors suggest that the phenomenon should be rare and harmless (hence the friendly name), Weniger & Friederichs (2016) demonstrate that this assumption heavily depends on the data to which SAL is applied. Other authors suggest to remove the forecast's marginal biases by selecting individual thresholds based on quantiles (Skok & Roberts, 2018) or calibrating the forecast to the observed marginal distribution (Casati et al., 2004). Besides SAL, the only explicit measure of marginal errors among our example scores is given by DAS where the point-wise error after morphing is used.

A further noteworthy aspect of the marginal distribution is the tail behavior. In a probabilistic, lower-dimensional setting, the issue of verifying extreme events was discussed in Lerch et al. (2017). None of the spatial methods mentioned above were designed with such considerations in mind: Thresholding methods like FSS, ISS and BD completely ignore extreme events, unless they are used in the definition of the threshold value. SAL uses the extreme value within each object in the definition of S, thereby including the marginal extremal behavior in the (spatial) structure component. How optical flow algorithms react to individual extremes depends on the details of the underlying algorithm.

Correlations The final class of errors discussed here deals with the spatial correlation structure. In the context of rain fields, this includes properties like the sizes of individual rain cells, the presence of linear structures like fronts or squall lines and the organization of these structures in space. As discussed above, SAL includes scale aspects in the S-component (together with the marginal tail behavior) while the spatial organization is part of the location score and the linear or isotropic shape of the features is ignored. The remaining four examples scores all feel changes to these properties but do not separate them from other errors. In principle, any number of structural characteristics can be assessed in a more general feature-based framework like MODE, but the results are always conditional on the definition of the objects: Depending on the threshold and smoothing filter, a string of convective cells along a squall line might constitute a single, large, an-isotropic pattern or many small, circular features. Spectral approaches like those of Marzban & Sandgathe (2009), Willeit et al. (2015) and Wong & Skamarock (2016), on the other hand, can study structures on multiple scales simultaneously. Their theoretical relationship with the auto-covariance function (under the right assumptions) makes them a natural choice for structure-verification. Most of our work presented in chapter 4 concerns the development of useful structure scores based on (wavelet-)spectral techniques.

3.3 Scoring the scores

The question how one should evaluate the merits of a verification strategy is of central importance to our investigation. In general, there are two ways of addressing the issue: Firstly, many desirable properties can be identified a priori from the definition of a particular score. Secondly, scores can be subjected to a number of empirical tests. In the following sections, we give an overview of both approaches and how they have been used in the literature.

3.3.1 Intrinsic merits

Mathematical properties This aspect underlies many of the other points discussed below as it determines our ability to study a score without explicitly calculating it. Besides pleasing anyone with a degree in mathematics, scores with a mathematically simple definition allow users to reason, on paper, about such aspects as the behavior in limiting cases (empty or full fields, individual non-zero pixels, random fields), the range of possible values (normalizability) and the response to noisy observations. On one end of the scale, BD is a mathematical metric, the definition of which fits in a single formula (equation 3.1). The simple form of FSS has allowed researchers to analytically compute its value for single pixels and idealized displaced rain-bands (Skok, 2015, 2016) and eventually derive an explicit measure of displacement errors from it (Skok & Roberts, 2018). Weniger & Friederichs (2016) exploited the similarly simple definition of SAL's L-component to design indicators of its sensitivity to parameter changes and observational noise. At the high end of mathematical complexity, field deformation techniques like DAS or the approaches of Gilleland et al. (2010b) or Marzban & Sandgathe (2010) rely on numeric optimization schemes, leaving their behavior potentially hard to predict a priori.

Sensitivities and invariances In section 3.2, we have already seen the variety possible forecasts errors. Which of these errors a score does and does not recognize determines who it may be useful for. Generally, abrupt changes ("2" in table 3.1) are undesirable. Some users wish to isolate specific kinds of errors and seek a score with a single "4" (explicit quantification) and many invariances ("1"s). Others require a single summary measure like FSS which smoothly ("3") incorporates all kinds of errors.

A related point is the sensitivity to equal changes in both forecast and observation. This is a potential weakness of the MRA-based ISS where the lacking shift invariance of the underlying wavelet transform can lead to abrupt changes in the score when features in *both* images are shifted from the support of one daughter wavelet to the next. Casati (2010) recommend repeating the calculation with various shifts to assess the issue. Similarly, some of the distance measures discussed in Gilleland et al. (2020), including BD, can change considerably depending on where the objects are located with respect to the domain boundary. BD furthermore depends on the size of the verification domain – an issue it shares with CRA (Mariani & Casaioli, 2018), SAL and any other score that uses the domain size to normalize the measured location errors.

Intuitiveness One of the most important factors determining the popularity of a verification strategy is the ease or difficulty with which one can understand how it arrives at its judgment. On the one hand, scores can be difficult to understand due to the sheer complexity of the underlying algorithm (arguably the case for the "region trees" of Hou & Wang (2019) and the optical flow approach of Gilleland et al. 2010b). One the other hand, scores with very simple definitions like the point-wise RMSE can be nearly uninterpretable due to the many possible, potentially compensating error sources combined in a single value. Intuitiveness is often mentioned as a strength of feature-based methods like MODE (Davis et al., 2009) and CRA (Ebert & McBride, 2000): The identification and comparison of objects is easy to visualize and reason about, without having to understand the technical implementations. This argument only holds when well-defined objects are known to exist in the data. Similarly, the matching of predicted and observed objects can lead to very intuitive verification results, as long as the two fields can be assumed to contain slightly different versions of the same objects. FSS is another example of a score that owes at least some of its popularity to the fact that the core concept can be explained and understood easily without having to delve into the details of its implementation.

Number of free parameters The free parameters associated with a score can be considered a double-edged sword: Fundamentally, the outcome of a verification study should depend on the qualities of the forecast verified, not the settings of the verification procedure. On the other hand, many degrees of freedom allow researchers to focus on very particular attributes of a prediction. MODE is a notorious example of this: The total interest function is calculated as an arbitrarily weighted average over the differences in any number of attributes selected by the user. In Davis et al. (2006), the overall score is based on 24 % centroid distance separation, 35 % minimum distance between object borders, 12 % orientation angle difference, 17 % area ratio and 12 % intersection area. While this may be seen as great flexibility, the choice of these percentages requires a lot of care, hinders comparability between different studies and, in the worst case, opens the door to abusive practices where the parameters could be tuned to achieve any verification result one desires. Numerous free parameters are thus overall undesirable. Other popular methods like FSS, SAL and ISS merely require users to select one or more thresholds and spatial scales (wavelet scales in ISS, neighborhood sizes in FSS, smoothing filter in object decomposition algorithms), both of which may have a simple physical interpretation.

Applicability Our last point concerns the kind of data required to apply a particular score. For example, forecast and observations must be given as matrices on the same regular grid in order to apply any of the five example methods from section 3.1. Appropriate treatment of missing data is relatively straightforward for FSS but considerably more troublesome for the wavelet-based ISS or feature-based methods like SAL. As discussed above, methods based on object decomposition must furthermore assume that the data actually contains at least one well-defined object. This assumption may hold for many rain fields[3] but is questionable for other meteorological variables like temperature or wind. ISS and FSS, on the other hand, can always be computed as long as a sensible choice of thresholds is known.

[3] not all: what if a small verification domain is completely covered by a single rain band or uniformly filled with small convective cells?

3.3.2 Testing strategies

Throughout the spatial verification literature, the most popular way of demonstrating the merits of a new verification method consist of applying it to one or more test cases, where the desired outcome of the verification is known a priori.

Geometric shapes The simplest way of generating a forecast with known error characteristics is to start with a simple geometric shape (e.g. a circle or oval with value one inside and value zero outside) as the observation and applying some shift, rotation, addition, or other type of distortion to create a faulty prediction. A set of such geometric test cases was created for the first phase of the ICP project and verified by almost all participating scores. A second, more extensive test suite was developed for MesoVICT (Gilleland et al., 2020), other authors experimented with different simple shapes appropriate to their methodology (Skok, 2015, 2016; Wernli et al., 2008; Han & Szunyogh, 2016). Such idealized tests can be very helpful in identifying a score's basic behavior and detecting fundamental weaknesses. The covariance structure and marginal distribution of these test images is, however, very unlike the real case, limiting their applicability and informative value.

Random fields If a spatial verification method is meant to recognize errors in the predicted spatial auto-covariance function, simple tests can be constructed by simulating a random process where the covariance function is prescribed. While popular in the context of (uni-variate) ensemble verification, this obvious idea has so far rarely been applied to spatial verification methods (exceptions include Marzban & Sandgathe (2009), Scheuerer & Hamill (2015) and Jacobson et al. 2020). One obvious reason is that pre-determined displacement errors are not straightforward to realize in a random framework. It should, however, be noted that scattered fields with small, randomly placed objects are not an uncommon occurrence in precipitation forecasting. The behavior of spatial verification metrics under such circumstances is of some interest, especially for those methods which attempt to identify matching objects in forecast and observation – a problem with no well-defined solution in effectively random situations.

Realistic case studies Some papers introducing a new verification technique state that their goal is to better match the subjective judgment of a human expert (Davis et al., 2009; Wernli et al., 2008). Many others implicitly assume the same by comparing the objective scores to their subjective, visual judgment of a few selected forecasts (Casati et al., 2004; Keil & Craig, 2007; Roberts & Lean, 2008; Yano & Jakubiak, 2016; Hou & Wang, 2019). While clearly helpful to illustrate a score's behavior in a realistic setting, this is hardly a sufficient proof of usefulness. On the one hand, the authors of such studies are, consciously or subconsciously, compelled to cherry-pick cases where their score works as intended. If the automatic verification result is known before writing the subjective assessment, it is furthermore hard to arrive at an unbiased, independent judgment. In an effort to circumvent at least the second issue, Ahijevych et al. (2009) compiled the opinions of 26 experts for each of the realistic test cases used within the ICP project. Their experiment revealed further issues with the idea of simulating human expert opinions: When simply asked to rank forecasts from worst to best, experts with different professional and educational backgrounds will value different aspects of forecast performance. Unsurprisingly, no individual score was found to consistently agree with the panel of experts.

As an intermediate step between real and idealized tests, simple well-defined errors like shifts, additive biases or deletion of features can be applied to an observed field to generate a synthetic forecast with a known error, as shown in figure 3.6. Several such perturbed cases were provided and used within the ICP.

An alternative approach consists of comparing all combinations of fields from a large data-base, where the members of some groups are known to be, on average, more similar to each other than to members of the other groups. Weniger & Friederichs (2016) and Radanovics et al. (2018) applied such tests to the SAL method. Both of these studies compared the scores obtained by a set of realistic forecasts verified against the corresponding observations to the sores obtained after randomly re-arranging the forecasts in time. This is a test of the scores discriminatory ability: If the forecasting system can be assumed to have any skill at all, we can be sure that individual forecasts are somewhat similar to the observations at the time for which the forecast is valid, and less similar (on average) to the observations at all other times. A score which cannot differentiate between original and shuffled data is clearly not helpful for the data at hand. Kapp et al. (2018) extended this statistical testing approach by comparing not only forecasts to observations but also individual predictions to the other members of the same ensemble forecast: Can the score be used to match a) individual forecasts and b) the observations to the correct forecast ensemble? This can be seen as a form of the random field approach discussed above, since ensemble members represent, at least conceptually, multiple realizations from the predicted distribution of possible outcomes.

Score inter-comparison A final, very natural test of new verification techniques consists of systematically comparing their judgment to established scores from the literature. If certain scores like SAL or FSS are well-tested and generally believed to give useful results for a certain class of cases, they can partly replace the human expert as the point of reference. The question which scores give the same or complementary information is central to ICP and MesoVICT. Both of these projects have provided freely available standardized test cases, thereby greatly facilitating the integration of new and old techniques into the spatial verification toolkit.

3.4 Wavelet-based verification

In preparation for the project that eventually became the present thesis, Weniger et al. (2017) conducted a review of spatial verification methods based on wavelets. Their main findings can be summarized as follows:

- Spatial wavelet-based verification evolved from initial ideas of Kumar & Foufoula-Georgiou (1993), via the approach of Briggs & Levine (1997) into the ISS of Casati et al. (2004) which is by far the most popular scale-separation technique to date.
- Most studies use wavelets to either remove unwanted "noise" (point-measure enhancement methods like Briggs & Levine 1997) or decompose the overall error into multiple scales (classic scale-separation like ISS).
- Almost all spatial verification studies rely on a two-dimensional MRA based on the Haar wavelet.

For the sake of completeness, it must be mentioned that Yano & Jakubiak (2016) explored an almost completely different approach to wavelet-based spatial verification, which is missing from Weniger et al. (2017): Emphasizing the previously under-used localization capabilities of wavelets, these authors present an unusual object-identification procedure in wavelet-space. Unlike all other approaches, these authors rely on a two-dimensional orthogonal Meyer wavelet decomposition with separate scales for the x- and y-direction. To the best of our knowledge, their innovative approach has not been tested beyond the single example case presented in the original publication.

In parallel to Yano & Jakubiak (2016), Weniger et al. (2017) and Kapp et al. (2018) developed a novel wavelet based verification approach. Based on the literature review, they realized that none of the existing techniques were designed to isolate specific aspects of forecast performance. In particular, any score based on an

Algorithm 5 Wavelet-based verification of Kapp et al. (2018)

Input: forecast ensemble $Y_1, \dots, Y_m$, reference X, largest scale J, number of LDA-vectors
Output: logarithmic score in reduced wavelet-space

```
Linearly smooth the edges of $Y_i$ and $X$
pad all fields with zeros to obtain square images of size $2^J \times 2^J$
Compute the two redundant Haar-wavelet transforms.
Apply the bias correction matrix.
Apply soft wavelet thresholding to the coefficient fields to smooth them.
Remove the smallest and largest two scales.
Average the observed and predicted coefficient fields in space.
Apply LDA to the ensemble of spectra, reduce the dimension.
In LDA-space, estimate the mean and covariance matrix of the ensemble.
Assuming multivariate normality, calculate the log-likelihood that the observation is drawn from the predicted distribution.
```

MRA will inevitably include displacement errors. Their goal was thus to develop a pure, shift-invariant structure score which would be similar to SAL's S-component but without the troublesome aspects of thresholding and object detection found in Weniger & Friederichs (2016). In addition, no structure-score available at the time was appropriate for ensemble forecasts.[4] To achieve this target, Kapp et al. (2018) employ the texture classification method of Eckley et al. (2010) which is based on the redundant discrete wavelet transform and locally stationary wavelet processes introduced in section 2.3. Their final methodology, with which they verify hourly rainfall accumulations from COSMO-DE-EPS against COSMO-REA2 reanalysis data, is summarized in algorithm 5.

Based on the discussion in section 3.3, we can identify the strengths, as well as limitations of this approach. From a mathematical point of view, the framework of Eckley et al. (2010) is attractive because of the theoretically guaranteed relationship with the spatial covariance matrix. Fourier- and variogram-transforms enjoy similar properties only under the strong assumption of global stationary, whereas the LSW approach is valid as long as the correlations vary slowly in space. The formulation in terms of continuous operators furthermore avoids the potentially erratic behavior associated with thresholding operations. Because the method makes no assumptions about the nature of the underlying fields aside from local stationarity, its applicability is not limited to rain fields or, more generally, fields with well-defined objects. In terms of a posteriori tests, Kapp et al. (2018) demonstrate that the method has strong discriminatory abilities, i.e., rain fields on different days can rather easily be distinguished from one another based on the spatial structure alone.

Concerning the other points discussed in section 3.3, we recognize several directions for improvement which serve as the jumping-off point for the studies summarized in chapter 4. Low intuitiveness likely constitutes the biggest weakness of the texture-based verification: Both Weniger et al. (2017) and Kapp et al. (2018) recognize that the physical interpretation of the three directional spectra is very difficult – a problem which is aggravated by the further data reduction via LDA. In addition, the bias correction step produces spectra with negative values, the interpretation of which is unclear. Concerning free parameters, the method has four main degrees of freedom which need to be addressed systematically: Boundary conditions, wavelet choice, smoothing procedure and the selection of scales to use. With the exception of smoothing, Kapp et al. (2018) give some explanation of their chosen settings but real sensitivity tests are nonetheless needed. One technical issue concerns the use of the RDWT which, as discussed in section 2.4, has poor directional properties. In terms of applicability, Kapp et al. (2018) consider only the case of ensemble forecasts; their method cannot be used in a deterministic setting. A final, more general point is raised in the outlook of Kapp et al. (2018) and, independently, by Yano & Jakubiak (2016): The true advantage of wavelets over other transformations like Fourier and variograms is the capability for localization. This property is not really utilized by a score derived from the spatial mean spectra.

[4] In the meantime, an ensemble version of SAL has been developed by Radanovics et al. (2018).

Based on these considerations, we can identify five research questions:

Q1 How can the wavelet spectra be interpreted?

Q2 How can the method be adapted to the verification of individual forecasts?

Q3 Can we further exploit the localization capability of the wavelet transform?

Q4 Is the redundant discrete Haar-wavelet transform the best tool for the job?

Q5 How can wavelets be used to verify variables other than precipitation?

In addition to answering these questions, the publications summarized in chapter 4 apply the new and adapted approach to a wide variety of test cases including geometric tests, random fields, the standardized MesoVICT cases and a complete year of real ensemble forecasts from COSMO-DE-EPS. The new scores are furthermore compared to other approaches from the literature which measure similar characteristics. A series of sensitivity experiments assesses the importance of boundary conditions, smoothing and selection of scales, and establishes best practices.

Chapter 4

Summary of Results

This chapter briefly summarizes the main results of the publications attached in appendices A, B and C, as well as the preprint in appendix D and the unpublished manuscript in appendix E. The notation throughout this section follows Buschow & Friederichs (2021a).

4.1 First test on a stochastic precipitation model

Published as “Assessment of wavelet-based spatial verification by means of a stochastic precipitation model (wv_verif v0.1.0)” (Buschow et al. 2019, appendix A).

The first publication contained in this thesis is primarily concerned with the questions Q1-Q3 listed above. Like Weniger et al. (2017) and Kapp et al. (2018), we rely on the bias corrected RDWT (see section 2.3), which yields a shift-invariant measure of the local covariance structure around each location $\mathbf{u}$, in the form of the local wavelet spectrum $e_{j,d,\mathbf{u}}$. If the input has dimensions $2^J \times 2^J$, the transform yields $3 \times J$ values *at every grid point.* The greatly inflated amount of data can only be interpreted and used for verification after some form of aggregation. Here, we propose a simple and intuitive approach, which is the nucleus for our structure-verification strategy: Dropping the explicit dependence on the location $\mathbf{u}$ for now, we denote the direction-averaged spectrum by $e_1, e_2, ..., e_J$. Now treat these energies like point masses, located at position $z = 1, ..., J$ along a line and define the center of mass of this arrangement as

$$z_c = \frac{1}{\sum_{j=1}^{J} e_j} \sum_{j=1}^{J} j \cdot e_j \,. \tag{4.1}$$

This *central scale* of the wavelet spectrum is a value between 1 and J which summarizes the distribution of energy across scales: If the image is dominated by small-scaled patterns, z_c approaches 1. Conversely, when most of the total variance resides on larger scales, z_c is closer to J. This transformation from three directional spectra to a single easy to interpret number is our first answer to Q1. When z_c is calculated from the spatial mean spectrum ($e_{j,d,\mathbf{u}}$ averaged over all $\mathbf{u}$), we condense the scale-structure of each forecast and observation into one scalar quantity. This number can then be verified by any appropriate uni-variate verification measure. Most simply, the difference $dz = z_c^{(for)} - z_c^{(obs)}$ is a deterministic structure score which tells us

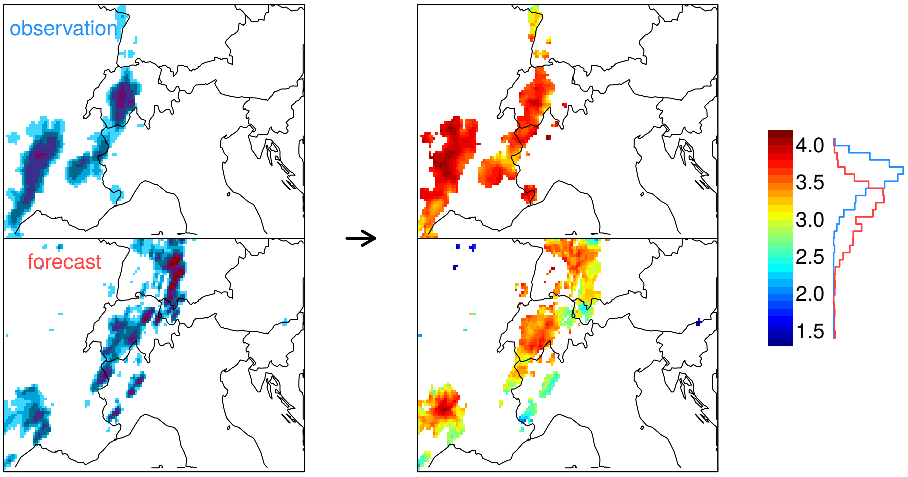


Figure 4.1: Transformation from rain fields to the map of central scales. Blue and red curve beside the color bar indicate the distribution of scales in the observed and predicted field, respectively.

whether an individual forecast field was too large or too small in scale. We have thereby addressed Q2.

The second main innovation of Buschow et al. (2019) is the way in which the newly defined scores are objectively tested. Instead of relying on expert opinions in a limited number carefully selected case studies, we use the physically based stochastic rain model of Hewer et al. (2017) to generate semi-realistic test cases with pre-defined stationary covariance structure. This allows for an ideal test of our structure scores: We randomly draw an "observation" from a specific model configuration and compare it to a number of "forecasts", only one of which has the correct covariance structure. Does our score correctly reward the best forecast? The answer is yes, as long as all errors in the parameters of the covariance model have the same effect on the scale of resulting fields. If, however, these errors exhibit compensating effects, the simplification to a single number z_c is too drastic and the best forecast can no longer be confidently determined. To solve this problem and pass the more difficult test, we replace dz by the so-called Earth Mover's Distance (EMD, Rubner et al. (2000)) between the forecast and observed mean spectra. This score, henceforth named *semd*, measures not only the shift towards smaller or larger scales, but the total amount of work needed to transform one spectrum into the other. Unlike $|dz|$, which is a lower bound on *semd*, the more complex score can correctly distinguish the best forecast in a majority of cases throughout all of our experiments. Its performance clearly beats the object-based S of Wernli et al. (2008) and comes close to that of the stationary variogram score adapted from Scheuerer & Hamill (2015), which explicitly exploits the global stationarity of our test data.

Lastly, the central scale also allows us to further study the *local* correlation structure (Q3). Instead of averaging the spectra in space, we can also compute z_c at every grid point individually. The resulting *map of central scales* is shown in figure 4.1 for our example from before. We have thus visualized one important aspect of the $2^J \times 2^J \times 3 \times J$ wavelet coefficients in a single image. The interpretation of these plots is straightforward: The observed field is almost entirely dominated by two large-scaled features ($z > 3.5$). The forecast, on the other hand, contains three regions with intermediate scales $z \approx 3$ and a number of smaller, intense precipitation cells which are correctly recognized as small-scale variability ($z \leq 2.5$). For the purpose of objective verification, we can compare the histograms of observed and predicted centrals scales (blue and red curve in figure 4.1) via the EMD. In contrast to the spatial mean spectrum, which measures how the total *variance* is distributed across scales, the histogram of z_c determines the fraction of the total *area* dominated by each scale. In the synthetic test cases, the resulting score, henceforth called *hemd*, performed similarly to *semd*. Recall, however, that these tests rely on globally stationary structures, meaning that the probability distribution of central scales does not vary in space. The true potential of the local approach is discussed in more detail in the next section.

4.2 The scale structure of precipitation forecasts

Published as “Using wavelets to verify the scale structure of precipitation forecasts” (Buschow & Friederichs 2020, appendix B).

With the introduction of z_c and the scores based on the EMD, we have answered, at least in part, the first three of our five research questions. The focus of the second publication is on applying these new techniques to real numerical weather forecasts and radar observations. In doing so, we address a number of implementation issues and investigate the sensitivity to the remaining free parameters of the method.

The basis for these experiments is a set of ensemble precipitation forecasts from DWD’s COSMO-DE-EPS for the complete year 2011. 14 hand selected cases from this data-set were studied in Kapp et al. (2018). These authors explicitly refrained from using radar images as validation data in order to avoid missing data. Upon closer consideration, the issue of gaps in the radar images is just one aspect of the broader problem of selecting proper boundary conditions for the wavelet transform: Any field which does not cover the entire globe is effectively *missing* data beyond its outer edge. If the only boundaries are at the edges of a rectangular model domain, we conclude that reflective boundaries are likely the most natural and convenient way of extending the data. Radar images are more challenging because the boundaries are irregularly shaped and can intersect parts of the image. Here, we follow the simplest route and replace any grid point with missing radar data by zeros in both the observation and the forecast. In a sensitivity test, we repeat the verification without removing radar gaps from the forecast. The impact of the different boundary conditions turns out to be moderate, the scores obtained in the two experiments remain highly correlated. A further test reveals that the choice of Daubechies mother wavelet is even less impactful. Switching from radar to reanalysis as validation data, on the other hand, does make a significant difference.

Beyond these sensitivity analyzes, we confirm that the wavelet spectra, and thus the central scales z_c, are easily capable of distinguishing between rain fields generated by convective and frontal weather situations. To test the scores’ discriminatory abilities, we conduct one of the experiments for realistic case studies discussed in section 3.3.2: All scores are computed for each combination of observed and predicted

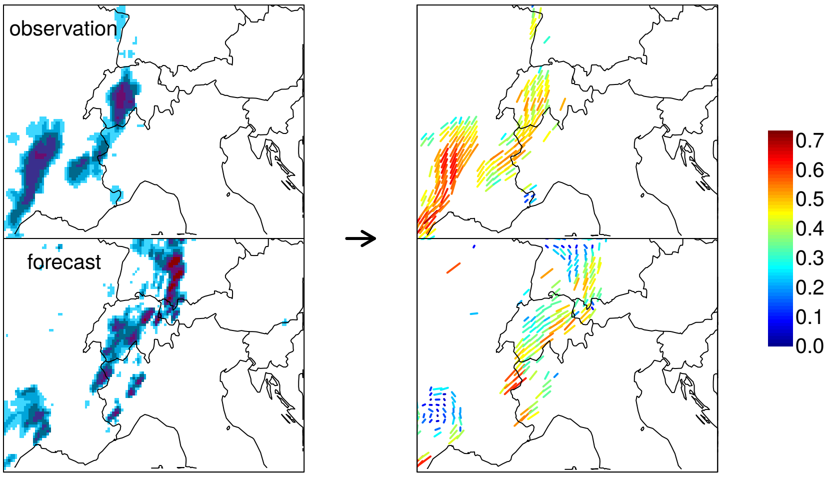


Figure 4.2: Transformation from rain fields to the map of anisotropy (color, arrow length) and direction (arrow orientation).

field from all days in the data-set. A useful score with strong discrimination should typically give the best ratings to the forecasts that match the day of the observation. Instead of the radar observations, we furthermore validate all forecasts against each other, thereby eliminating model errors from the experiment. If the forecast ensemble is considered as a set of realizations from a daily changing probability distribution, this test is equivalent to the random field experiments from Buschow et al. (2019). Confirming the results from the idealized tests, we find that *semd* and *hemd* are approximately as discriminatory as the established variogram alternative and substantially better than S. The two wavelet-based scores are also highly correlated with each other in these realistic tests. Since plots like figure 4.1 prove to be quite helpful in the analysis and verification of individual case studies, we therefore conclude by recommending the use of *hemd* as a wavelet based structure score for precipitation verification.

4.3 Verifying Scale, Anisotropy and Direction (SAD)

Published as "SAD: Verifying the scale, anisotropy and direction of precipitation forecasts" (Buschow & Friederichs 2021a, appendix C).

The condensation of the wavelet spectra into a single number z_c was the key to developing a verification method which is intuitively interpretable and allows us to utilize the local wavelet spectra instead of the spatial mean. Both of the previous publications mention in their respective outlook that the most important aspect neglected in this manner is the directionality. As discussed in detail in section 2.4, the averaging over the three directions is necessary because of the poor directional properties of the classic RDWT. In the third publication, we address this issue (and thereby Q4) and replace the Daubechies RDWT by the DTCWT of Kingsbury (1999). The resulting local spectra have six instead of three directions and allow us to extend the idea of a central scale to include information on the anisotropy

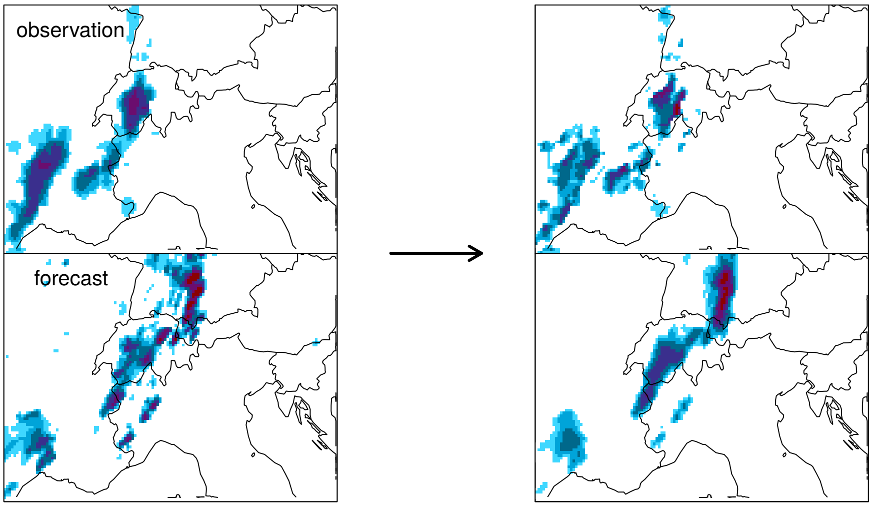


Figure 4.3: Structural bias correction. Left: Original observation and forecast. Right: Observation adjusted to the predicted mean spectrum and forecast adjusted to the observed mean spectrum.

and orientation of the local covariance structure. Instead of averaging the $e_{j,d}$ over the directions and placing them along a line, we arrange the six directional energies $(d = 1, ..., 6)$ for each scale at equal distances from each other along a circle in the $x - y$-plane. The scale j determines the z-component as before. The center of mass of this arrangement of point masses, located along the edges of a regular hexagonal prism (see Fig. 4 in appendix C), has three components. z_c is the same as before and continues to measure the spatial scale. x and y are transformed into polar coordinates (ρ, θ). The radius ρ allows us to measure the degree of anisotropy: When energy is distributed equally among the six directions, we have $\rho = 0$ and covariance is isotropic. Conversely, when ρ is large, one orientation dominates. The dominant angle of orientation itself can be inferred from θ as $\varphi = 15° + \theta/2$.

In figure 4.2, we have visualized the local anisotropy and direction via colored arrows. As expected from the discussion and tests in section 2.4, these arrows smoothly follow along the edges of the individual precipitation features and correctly distinguish between linear and round patterns. We observe that the western precipitation field is strongly oriented in the observation and nearly circular in the forecast. The precipitation region over the western Alps, on the other hand, is forecast only slightly too isotropic and has nearly the correct orientation.

In analogy to the scale error dz, we can now define the anisotropy error $d\rho$ and the orientation error $d\varphi$ which have similarly simple interpretations in terms of “too round” or “too linear” and “rotated by ...°”. Since $d\varphi$ is meaningless for very small values of ρ, we furthermore define the combined error dxy as the euclidean distance between the centers in the x-y-plane. Lastly, all structural error characteristics can be combined into a single score $semd_d$ which consists of the EMD between the two directional spectra, arranged at the locations within the hexagonal geometry described above. The effectiveness of the new scores is demonstrated using both geometric and real test cases from the MesoVICT project. We find that the overall error $semd_d$ can typically be explained as a linear combination of dxy and dz with different contributions depending on the forecast model and the weather situation.

As hinted in section 2.4, we furthermore show that the decimated DTCWT can be used in place of the redundant version when only global characteristics are of interest. Besides massively saving computational costs, the non-redundant transform has the advantage of being (almost perfectly) invertible. We exploit this property to define an experimental bias correction algorithm: After transforming forecast and observation, we apply a multiplicative correction to the predicted local spectra such that the resulting spatial mean spectrum is identical to the observed one. By transforming back to image space, we obtain a new precipitation field which combines the spatial placement of the forecast with the structure of the observation. This procedure is exemplified in figure 4.3, where we apply the correction both ways: When the predicted structure is imposed on the observed field (top right panel), the large and mostly homogeneous precipitation regions receive additional internal structure and lose the clear orientation. Conversely, the transformed forecast field (shown on the bottom right of figure 4.3) is rendered smoother and more coherent while the intensity of the scattered precipitation cells is reduced. By visualizing what an improved forecast would look like, the bias correction algorithm can assist in understanding the verdict of the verification method (Q1) while exploiting the unique localization properties of the wavelet transform (Q3). As a side-note, we have already used this technique to produce the forecast with the erroneous correlation structure in figure 3.6 f. Another, admittedly narrow, application for our algorithm is thus the generation of fields with artificially perturbed spatial correlations as test cases for other verification methods.

4.4 Verification of near surface wind patterns

Under review as “Verification of Near Surface Wind Patterns in Germany using Clear Air Radar Echoes” (Buschow & Friederichs 2021b, appendix D).

Throughout the three papers summarized above, we have repeatedly mentioned that the wavelet-approach can be applied to any variable of interest because it requires neither the existence of objects nor meaningful thresholds. Indeed, the underlying wavelet algorithms were originally developed for the analysis of photographs and make no assumptions about the structure of the image at all. While this constitutes a potentially major advantage over the popular object-based verification measures, the actual usefulness of our approach in this context is an open research question (Q5).

A series of recent papers including Skinner et al. (2016); Skok & Hladnik (2018); Zschenderlein et al. (2019); Schlager et al. (2019) documents growing interest in spatially verifying wind forecasts in particular. This is also one of the mission statements of MesoVICT (Dorninger et al., 2018). The example of wind fields raises three main questions concerning the applicability of our approach: What is the spatial structure of typical wind fields? How can it be observed? How should the vector nature of the wind fields be treated? Here, we address these questions by focusing on a relatively narrow range of weather phenomena, which can be spatially observed in an unexpected way: On warm, sunny days, Rayleigh-Bénard-like convective cells

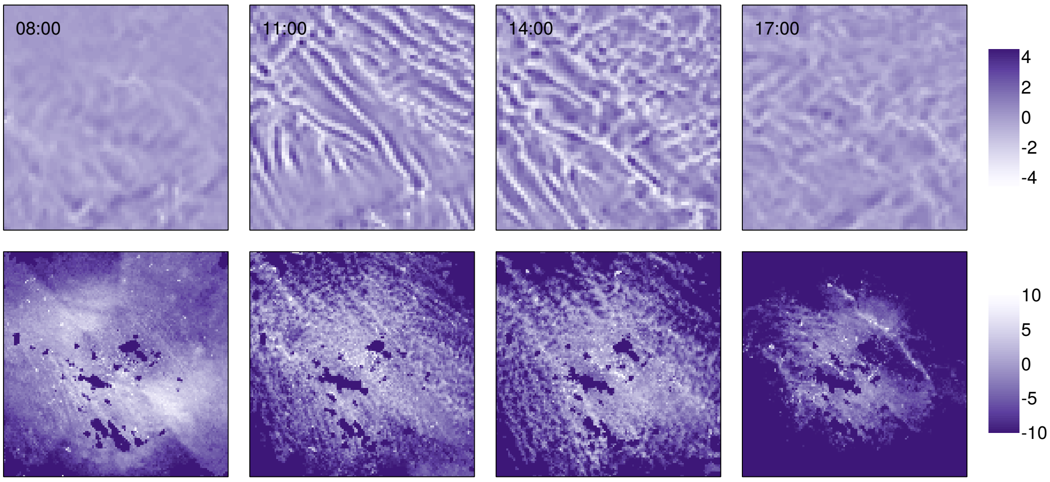


Figure 4.4: COSMO-REA2 10 m divergence (top row in $m/(s \times \text{grid-point})$) and RADOLAN-RX reflectivity in dBz (bottom row) around the Dresden Radar station on 2009-07-29. Color scales were chosen such that strong convergence (negative divergence) and increased reflectivity have the same color.

and rolls are common occurrences throughout the atmospheric boundary layer over simple terrain. The convergent regions of this circulation attract swarms of insects which consequently visualize the spatial pattern on radar scans (bottom row in figure 4.4). Studies like Banghoff et al. (2020) exploit these "clear-air" radar echoes to examine the development of boundary-layer wind patterns; Thurston et al. (2016) even qualitatively compare them to model data. With the wavelet-based structure scores, we can go one step further and quantify the spatial structures in the observations and model data. Here, we compare clear-air radar echoes from the German RADOLAN radar composite to 10 m divergence fields from COSMO-REA2. We have thereby limited the range of expected spatial structures, found a way to observe them, and limited the vector-valued wind field to its scalar, divergent component. The two quantities, model divergence and concentration of reflecting insects, are made comparable by abstracting their spatial structure in terms of scale, anisotropy and direction.

Comparing over 20000 individual images, we find that the scale-structure at rural radar stations follows a well-defined diurnal cycle with a peak of small-scale activity in the early afternoon. This behavior is surprisingly well reproduced by the model: Studies including Zhou et al. (2014), Ching et al. (2014) and Poll et al. (2017) suggest that meso-scale models like COSMO-DE should produce an unrealistically delayed pattern on too-large scales. Part of the explanation is that our analysis is limited to scales larger than the 1 km resolution of RADOLAN: For a fair comparison, the smallest scale is removed from the 1 km resolution radar data. Smaller circulations, which would initiate earlier in the day, are represented in neither data-set; on the observed scales, the model fares well, as least as far as z is concerned. The direction φ of the patterns in both data sets is primarily along the model wind direction, but the anisotropy ρ does not match: The model is generally more strongly directed and has a preference for linear features before noon, which is not observed.

These results are illustrated in figure 4.4, where several snapshots from an observed and modeled diurnal cycle are shown side by side. Here, no smoothing was applied to the radar data, which consequently exhibits fine-scale patterns below the resolution of COSMO. The timing of the roll-like structures, as well as their overall linear character and orientation, is nonetheless decently reproduced. We furthermore see that the stripes are far more regular (and thus anisotropic) in the model than the real world.

4.5 Measuring displacement errors

> Unpublished manuscript, working title "Measuring Displacement Errors with Complex Wavelets" (Buschow 2021b, appendix E).

Our fifth and final study aims to make further use of the wavelet's localized nature (Q3) to define a measure of displacement errors. With SAD, we have essentially developed a more robust, specific and widely applicable alternative to SAL's S-component. The definition of a complementary location score with the same benefits of the wavelet approach is thus a natural next step. We achieve this using the previously neglected phase information of the complex wavelet coefficients (recall that SAD is defined using the modulus squared). In analogy to the Fourier transform, a shift of the underlying data translates linearly into a scale-dependent change of phase. The phase difference $\Delta\Phi$ between transformed forecast and observation can be averaged over directions and locations (weighted by the squared amplitude to focus on regions with non-zero variance) to obtain a scale-wise measure of location errors. This idea is schematically shown in figure 4.5, where the (weighted) phase differences in our example case were plotted at $j = 3, 4, 5, 6$. Notice how the largest displacements (shown in dark red) have different locations depending on the scale: At $j = 5$, the displaced feature in the South-West corner dominates; the shift of the other rain fields is more relevant on the largest scale.

An estimate of the displacement's magnitude in grid-points can be obtained by multiplying the phase at scale j by 2^j. When a single summary score is desired, we simply take the average over all scales to get the maximum estimated displacement between the images. In our example, we find a displacement of $\Delta\Phi(6) \cdot 2^6 \approx 16$ grid points (roughly 128 km), which corresponds to the distance between two dashed lines in figure 4.5.

The new location score is mathematically convenient because it uses the same wavelet transform as SAD and exploits exactly the information discarded in the structure analysis. A few simple experiments with artificially shifted images, as well as the geometrical tests from Gilleland et al. (2020), confirm that the new method can recover the approximately correct shifts even when different features move in different directions. This is a consequence of the localized basis functions which allow us to handle spatially varying shifts appropriately. With the possible exception of Yano & Jakubiak (2016), who do not explicitly define a score, this is the first and only example of a displacement measure based on scale-separation. From a practical point of view, this means that, unlike SAL, CRA, BD or MODE, it can be applied to non-intermittent fields without well-defined objects. As a proof of concept, we verify forecasts of precipitation, wind speed, potential temperature and equivalent potential temperature. Our data-set includes one- two- and three-day forecasts from the

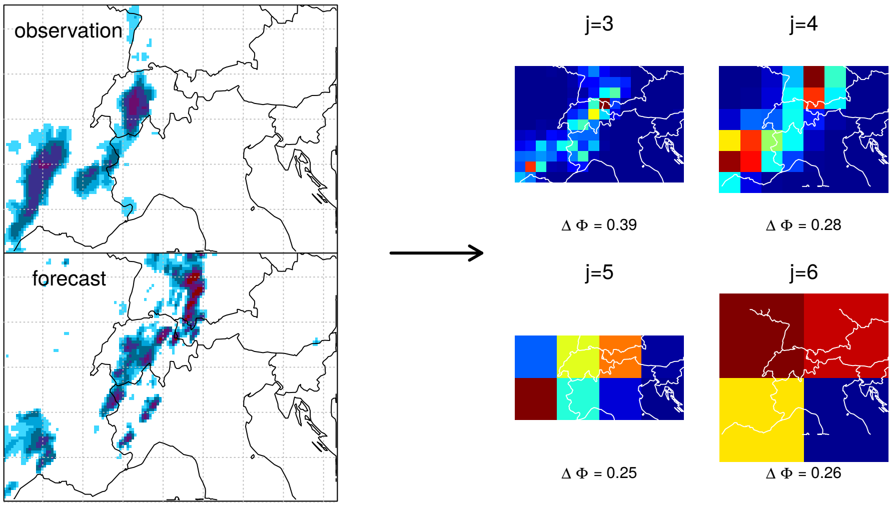


Figure 4.5: Calculation of the phase-based location score: Forecast and observation (left) are wavelet-transformed, the phase differences are weighted and averaged over the six directions to obtain one map of displacements per scale j. The sum over each field gives the phase error $\Delta\Phi(j)$.

Italian MOLOCH and BOLAM models which are verified against the station-based VERA analysis (Bica et al., 2007) during the MesoVICT period. For precipitation, we find a clear decrease in skill with lead-time (which cannot be seen in RMSE or SAL's L), as well as slight advantages for the higher-resolved MOLOCH model in convective situations. The other variables show overall better scores, especially at small scales. This is due to the constant presence of non-zero variance near spatially fixed features such as mountains and coasts. For the two temperatures, the VERA analysis data includes additional fine-scale information beyond the resolution of the station network, resulting in slight systematic advantages for the finer model. Lastly, we can average the phase errors over time, rather than space, to obtain scale-dependent maps of regions with consistently good or bad localization.

Chapter 5

Conclusion and Outlook

5.1 Concluding remarks

When the first highly resolved weather prediction models appeared in the early 2000s, researchers quickly discovered that the perceived improvement in realism failed to be rewarded by traditional measures of verification. To address the issue, a large toolkit of more sophisticated, *spatial* verification techniques has been developed. The research collected in this thesis has added another useful tool to the box.

Among the five well-known examples of spatial methods we discussed in chapter 3, the fractions skill score of Roberts & Lean (2008) is perhaps the most popular and widely used. One likely reason for the success of FSS is its relative simplicity – it is easy to understand what is measured and how, without diving deeper into the underlying mathematics. Chapter 2 has shown that wavelet transforms, which are the technical heart of our new method, have many useful properties; ease of understanding is not one of them. Despite harmless-sounding introductory texts like *Wavelets for Kids* (Vidakovic & Mueller, 1994), *A Friendly Guide to Wavelets* (Kaiser, 2010), or indeed *A really friendly guide to wavelets* (Valens, 1999), the topic initially seems daunting to all but the most mathematically inclined readers.

Consequently, a key objective of this work was to condense the information contained in the wavelet transform of meteorological fields into intuitively understandable quantities. While understanding the inner workings of our final SAD method requires some background knowledge on discrete wavelet transforms, the three structural aspects it measures, namely Scale, Anisotropy and Direction, can easily be explained to any layperson. Our technique, expanding on the groundwork laid by Weniger et al. (2017) and Kapp et al. (2018), is not an extension or adaptation of any existing score, but represents an entirely new way of exploiting the unique properties of wavelets for the purpose of spatial forecast verification.

Concerning the wish-list of desirable properties laid out in section 3.3, we have retained some of the mathematical advantages that originally drew Weniger et al. (2017) to wavelets (no discontinuous operators, connection to covariance matrix), while greatly increasing the intuitiveness of the resulting scores. In terms of sensitivities and invariances (table 3.1), the SAD structure scores quantify rotations and changes in correlation-length and -anisotropy separately, while being almost perfectly invariant under shifts. Conversely, the new location score quantifies shifts and largely ignores structure differences, as evidenced by the good performance of the fine-scaled MOLOCH model compared to coarse observations. Both scores are insensitive to the marginal distribution of the data since the wavelet spectra are normalized to unit energy. To confirm the practical usefulness of the new approach, all of the different testing strategies from section 3.3.2 have been employed. In terms of discriminatory ability, SAD compares favorably to its direct inspiration SAL. The two approaches tend to agree on the sign of the errors, indicating that (1) SAL's "structure" component primarily measures spatial scales and (2) such errors can robustly be identified, irrespective of the technical implementation of the score.

More generally, we have seen that the classic discrete wavelet transforms, which were used by practically all previous wavelet-based scores, are not ideal due to their shift sensitivity and poor directional properties. The observation that the dual-tree complex wavelet transform avoids these issues at a moderate computational cost, is relevant not only to the specific case of forecast verification, but to any aspect of meteorological image processing. For example, the wavelet-based organization index of Brune et al. (2018) originally relied on the RDWT but switched to the DTCWT (Brune et al., 2021) due to the enhanced directionality. For the same reason, a switch to dual-trees would improve the method of Kapp et al. (2018) in future applications as well. Post-processing approaches like the wavelet-based smoothing described in Theis (2005) could profit from the nearly perfect shift invariance of the decimated transform. In principle, a dual-tree version of the ISS could be defined as well, thereby mostly eliminating the undesirable shift-properties discussed in Casati (2010). The loss of orthogonality, however, would mean that the result is no longer a true decomposition of the point-wise MSE. Whether or not this is worthwhile, may be a topic of future research. To facilitate further applications, a new open source implementation of the DTCWT in the statistical programming language R was developed and published on the official CRAN repository (Buschow et al., 2020, `dualtrees`).

Recalling the large number of existing verification methods mentioned in chapter 3, critical readers are likely to ask whether it was really necessary to add yet another technique to the list. In other words, who could actually benefit from the new approach? The first group that comes to mind are the developers of dynamical or statistical models: Does a new convection parametrization produce too much or too little small-scale variability? Are cold fronts simulated with the proper degree of orientation and coherence? Is the spatial structure of my post-processed precipitation field accurate? Such questions are closely related to the notion of "realism" which Mass et al. (2002) found themselves unable to validate using point-wise error measures. The only two widely known scores which explicitly quantify this aspect of forecast quality are SAL's structure component S and the variogram-score of Scheuerer & Hamill (2015). Our experiments have shown that, besides the known pitfalls of thresholding-based object identification (Weniger & Friederichs, 2016), S is lacking in discriminatory ability. With the correct parameter settings, the method of Scheuerer & Hamill (2015) can be very good at detecting the best-matching spatial structure but its judgment is not very specific or interpretable – the spatial correlations were either well represented or not. Neither of the two established scores explicitly quantifies directional aspects of structure, S ignores them entirely.[1]

[1] For the sake of completeness, it should be mentioned that Bellier et al. (2020) have very recently developed a directional version of the variogram score. In their introduction, they mention the poor directional properties of classic wavelet transforms as a motivation to use variograms instead. All of their well-justified criticisms have been addressed by the switch to the DTCWT in Buschow & Friederichs (2021a).

Another group with potential interest in the merits of the predicted spatial structures are researchers attempting to model other components of the climate system (surface or underground hydrology for example) who rely on precipitation fields as input. In such applications, the precise location of rainfall may be less relevant than the question whether water is introduced in intense, localized bursts or smoothly across larger regions (see Wernli et al. (2008), Lobligeois et al. (2014), Loritz et al. (2021) and references therein). Bellier et al. (2020) point out the particular importance of the rain field's fine-scale directional structure for hydrological modeling. This could also be a use-case for the structural bias correction described in Buschow & Friederichs (2021a), which represents another unique advantage of the wavelet-approach: With the exception of field deformation techniques, none of the other spatial verification methods present an obvious opportunity to correct the errors they have measured. With SAD, we can produce a post-processed field which combines the predicted spatial locations with the correlation structure of the observations.

A further "selling point" of the wavelet-approach is its direct applicability to variables other than precipitation. Our study of clear-air boundary layer circulations has showcased that meaningful and interpretable information can be extracted from spatial observations other than rainfall – whenever we wish to quantify whether the spatial structures in two images *look similar*, wavelets can help. Our experiments have, however, also highlighted the two main difficulties when spatially verifying non-intermittent fields like wind. First and foremost, spatial observations are not widely available for most variables (we had to rely on the help of bugs!), but definitely necessary when the spatial structure is concerned: Typical station networks are both too localized and too coarse for a reliable estimate of the spatial correlation structure; model-based analyses largely receive their fine-scale pattern from the underlying physical model and do not necessarily represent nature. One possible alternative is the use of much higher-resolved simulations (such as LES) as reference for NWP models – an interesting future application for SAD. A second difficulty in verifying fields like wind and temperature is the sheer variety of simultaneous processes, some of which are fixed in space or time, some of which are transient, all of which contribute to the overall pattern. Structural characteristics like scale and anisotropy can become much harder to interpret when mountains, land-sea breezes, diurnal cycles, pressure systems and convection all coincide in the same field. We were able to obtain useful information about a specific physical phenomenon by limiting the analysis to small spatial regions, times of day, weather situation and the divergent component of the flow.

Our final study capitalized further on the wavelets flexibility, to define a new location score – our counterpart to SAL's L – based on the phase of the complex wavelet coefficients. The benefit of applicability beyond precipitation is particularly relevant here because almost all previous location measures share this limitation. However, the difficulties discussed above must be considered as well. When only the location is of interest, interpolated station data may be sufficient as a reference, especially when it is enhanced with additional information from a highly resolved digital elevation model. However, the multitude of co-occurring processes makes it particularly hard to determine the origin of estimated errors and interpret them in terms of an overall "displacement". The possibility to separate the phase errors by scale and even spatial location partly remedies these issues.

5.2 Future research

No three year PhD project can hope to explore the full depth of its topic. This final section is therefore dedicated to some of the research directions that were either not pursued at all, or simply did not come to fruition in time.

Structural post-processing One natural direction for future research would be the development of a full post-processing algorithm based on the proof of concept given in Buschow & Friederichs (2021a). As stated in the conclusions of that paper (appendix C), the preliminary version of the algorithm mainly serves to further illustrate the ways in which the predicted structure was erroneous: One has to know the desired mean spectrum of the observations in order to correct the forecast. In an operational setting, the target spectrum is unknown, but could potentially be inferred from a weather situation dependent climatology or a rolling training window. Such an application might prove useful when the underlying forecast model systematically misrepresents the spatial structure. This is sure to be the case when the forecast model in question is relatively coarsely resolved compared to the observations (a regional or even global model). Whether or not our approach can deliver convincing results in such a setting, where the visual difference in spatial structure is large, remains to be seen. It may be necessary to develop a randomized procedure which generates an ensemble of post-processed fields to capture the unpredictable nature of small-scale variability. In any case, an in-depth study of the existing precipitation post-processing literature is needed in order to identify the current state of the art. An important starting point is Bellier et al. (2020), who develop a statistical down-scaling method which emphasizes the (directional) correlation structure of the rain fields. These authors also give an overview of existing approaches from the literature and define several desirable characteristics for precipitation dis-aggregation algorithms. Further relevant references include Scovell (2020), who used complex dual-tree wavelets for a stochastic downscaling application, and Nerini et al. (2017) who employed a windowed Fourier transform for similar purposes.

Uses outside of verification Besides verification against observations, there are other use-cases where our distance measures in wavelet space may be useful in the future. One straightforward extension of our work in Buschow & Friederichs (2020) would be the comparison between different versions of a model or members in an ensemble. When the variable of interest is prone to double penalties, it can be difficult to decide objectively, which changes have a strong impact and which configurations produce similar results. Our scores could, for example, be used to identify groups of members in a forecast ensemble and select a representative subsample for presentation or further processing.

A similar application would be analogue forecasting, where one or more predictions are generated by searching a data-base for historic examples of similar weather situations; when the fields are highly resolved and intermittent, sophisticated measures of similarity yield better results than classic, point-wise metrics. Keller et al. (2017) applied a neighborhood approach to search for precipitation analogues; structural distance measures may find different, perhaps complementary analogues.

Another closely related area is the study of recurrence statistics in climatological data-sets. Based on the theory of chaotic dynamical systems, Faranda et al. (2017)

propose to estimate the local attractor dimension and persistence in phase space from the statistics of extremely close recurrences. Their approach, which can be used to objectively define circulation regimes, identify recurrent patterns (Buschow & Friederichs, 2018) and potentially asses the predictability of a particular state, has since been applied to a variety of climate sub-systems. For precipitation data, the current recommendation (Messori & Faranda, 2021) is to use the point-wise differences between binary fields as the distance metric. Wavelet-transforms may produce a more useful projection of the high-dimensional phase space and lead to a more intuitive idea of precipitation regimes.

Other wavelet transforms Previous wavelet-based verification studies relied primarily on the classic two-dimensional DWT of Mallat (1989). Over the course of the present investigation, we have seen that the slightly more sophisticated DTCWT introduced in Kingsbury (1999) is better suited to the task of image analysis in virtually every way. Seeing that this approach was developed over two decades ago, it is a natural idea to investigate other, possibly more recent developments in the wavelet literature and consider their benefits for spatial verification tasks.

One interesting direction would be wavelet transforms that act in space *and* time. Given sufficiently frequent model output and observations, the temporal evolution of meteorological phenomena could be added to the verification procedure, thereby explicitly treating timing errors and including the movement of features in the structural analysis. A possible candidate for this would be the "spatio-temporal wavelet transform" (Kikuchi & Wang, 2010), which was recently used to compare modeled and observed convection in the tropical Atlantic by Brune et al. (2020).

One weakness of all wavelet-based verification tools, as well as most other spatial techniques, is the requirement for validation data on a regular grid. In many cases, the only widely available observations are point measurements at weather stations, thereby necessitating some form of interpolation, which introduces its own host of difficulties and possible errors. Such issues can potentially be circumvented using the so-called "second generation wavelets" of Sweldens (1995), which generalize the multi-resolution analysis to irregularly sampled data using the "lifting-scheme". Whether and how the structural information inferred from point-wise data in this manner can be compared to model data is an open question for future research. In principle, lifting schemes could also be an appropriate answer to issues with missing data (like gaps in radar images) and unknown boundary conditions, as noted by Sweldens (1995). Furthermore, Sweldens lists data on spherical surfaces as a natural application of second generation wavelets. In our studies, we have circumvented this issue by using appropriate map projections and small regional domains; when global data is to be verified, lifting may provide the necessary treatment of the grid geometry. A two-dimensional lifting scheme is described in Jansen et al. (2009); Shuman et al. (2013) give a relatively recent overview general signal-processing approaches on potentially irregular graphs. In a meteorological context, a two-dimensional lifting implementation of an MRA was used for spatial smoothing of highly resolved model output by Theis (2005). Lastly, a comprehensive open-source software package for wavelet-transforms on the sphere, originally developed for astronomical applications, is provided by the `s2let` python library (Leistedt et al., 2013).

Appendix A

Buschow et al. 2019

Geosci. Model Dev., 12, 3401–3418, 2019
https://doi.org/10.5194/gmd-12-3401-2019

Geoscientific
Model Development

Assessment of wavelet-based spatial verification by means of a stochastic precipitation model (wv_verif v0.1.0)

Sebastian Buschow, Jakiw Pidstrigach, and Petra Friederichs

Institute of Geosciences, University of Bonn, Bonn, Germany

Correspondence: Sebastian Buschow (sebastian.buschow@uni-bonn.de)

Received: 4 April 2019 – Discussion started: 24 April 2019
Revised: 27 June 2019 – Accepted: 5 July 2019 – Published: 5 August 2019

Abstract. The quality of precipitation forecasts is difficult to evaluate objectively because images with disjointed features surrounded by zero intensities cannot easily be compared pixel by pixel: any displacement between observed and predicted fields is punished twice, generally leading to better marks for coarser models. To answer the question of whether a highly resolved model truly delivers an improved representation of precipitation processes, alternative tools are thus needed. Wavelet transformations can be used to summarize high-dimensional data in a few numbers which characterize the field's texture. A comparison of the transformed fields judges models solely based on their ability to predict spatial structures. The fidelity of the forecast's overall pattern is thus investigated separately from potential errors in feature location. This study introduces several new wavelet-based structure scores for the verification of deterministic as well as ensemble predictions. Their properties are rigorously tested in an idealized setting: a recently developed stochastic model for precipitation extremes generates realistic pairs of synthetic observations and forecasts with prespecified spatial correlations. The wavelet scores are found to react sensitively to differences in structural properties, meaning that the objectively best forecast can be determined even in cases where this task is difficult to accomplish by naked eye. Random rain fields prove to be a useful test bed for any verification tool that aims for an assessment of structure.

1 Introduction

Typical precipitation fields are characterized by large empty areas, interspersed with patches of complicated structure. Forecasts of such intermittent patterns are difficult to verify because we cannot compare them to the observations in a grid-point-wise manner: if a given rain feature is forecast perfectly, but slightly displaced, point-wise verification will punish the error twice, once at the points where precipitation is missing and once at the points where it was erroneously placed. The correctly predicted structure is not rewarded in any way. Following the advent of high-resolution numerical weather predictions, this effect, known as *double penalty* (Ebert, 2008), has motivated the introduction of numerous new spatial verification tools.

In a comprehensive review of the field, Gilleland et al. (2009) identified four main strategies that deal with the double penalty problem and supply useful diagnostic information on the nature and gravity of forecast errors. The classification was updated to include an emerging fifth class in Dorninger et al. (2018). Proponents of the first strategy, the so-called neighbourhood approach, attempt to ameliorate the issue via successive application of spatial smoothing filters (Theis et al., 2005; Roberts and Lean, 2008). A second group of researchers including Keil and Craig (2009), Gilleland et al. (2010), and recently Han and Szunyogh (2018) explicitly measure and correct displacement errors by continuously deforming the forecast into the observed field. A third popular approach consists of automatically identifying discrete objects in each field and subsequently comparing the properties of these objects instead of the underlying fields. Examples from this category include the MODE technique of Davis et al. (2006) as well as the SAL technique by Wernli et al. (2008).

The fourth group of spatial verification strategies contains so-called scale-separation techniques, which employ some form of high- and low-pass filters to quantify errors on a hierarchy of scales. A classic example of this family is the wavelet-based intensity-scale score of Casati et al. (2004), which decomposes the difference field between observation and forecast via thresholding and an orthogonal wavelet transformation. The final class newly identified by Dorninger et al. (2018) contains the so-called *distance measures*, which exploit mathematical metrics between binary images developed for image processing applications. One example is Baddeley's delta metric, which was first employed as a verification tool in Gilleland (2011).

The basic idea of the method presented in this study, which can be classified as a scale-separation technique, is that errors, that neither relate to the marginal distribution nor to the location of individual features, should manifest themselves in the field's spatial covariance matrix. Direct estimates of all covariances would require unrealistically large ensemble data sets or restrictive distributional assumptions. Following a similar approach to scale-separated verification, Marzban and Sandgathe (2009), Scheuerer and Hamill (2015), and Ekström (2016) therefore base their verification on the fields' variograms. The variogram is directly related to the spatial auto-correlations (Bachmaier and Backes, 2011) but can be estimated from a single field under the assumption that pairwise differences between values at two grid points only depend on the distance between those locations (the so-called *intrinsic hypothesis* of Matheron, 1963). Similarly one could require stationarity of the spatial correlations themselves, in which case the desired information is contained within the field's Fourier transform. Both of these stationarity assumptions may be inadequate in realistic situations where the structure of the data varies systematically across the domain; for example, due to orographic forcing, the distribution of water bodies or persistent circulation features.

Weniger et al. (2017) have suggested an alternative approach based on wavelets. The key result in this context comes from the field of texture analysis, where Eckley et al. (2010) proved that the output of a two-dimensional discrete redundant wavelet transform (RDWT) is directly connected to the spatial covariances. The crucial advantage of their approach is that it merely requires the spatial variation of covariances to be *slow*, not zero – a property known as *local stationarity*. After some initial experiments by Weniger et al. (2017), this framework has successfully been applied to the ensemble verification of quantitative precipitation forecasts by Kapp et al. (2018). Their methodology consists of (1) performing the corrected RDWT, following Eckley et al. (2010), to obtain an unbiased estimate of the local wavelet spectra at all grid points, (2) averaging these spectra over space, (3) reducing the dimension of these average spectra via linear discriminant analysis, and (4) verifying the forecast via the logarithmic score.

In this study, we aim to expand on their pioneering work in several ways. Firstly, we argue that the aggregation method of simple spatial averaging is not the only sensible approach. An alternative is introduced which incidentally suggests a compact way of visualizing the results of the RDWT: instead of aggregating in the spatial domain, we first aggregate in the scale domain by calculating the dominant scale at each location. Secondly, we use both kinds of spatial aggregates to introduce a series of new wavelet-based scores. In contrast to Kapp et al. (2018), we consider both the ensemble case and the deterministic task of comparing individual fields while avoiding the need for further data reduction. We furthermore demonstrate how to obtain a well defined sign for the error, indicating whether forecast fields are scaled too small or too large. The experiments performed to study the properties of our scores constitute another main innovation: the recently developed stochastic rain model of Hewer (2018) allows us to set up a controlled yet realistic test bed, where the differences between synthetic forecasts and observations lie solely in the covariances and can be finely tuned at will. In contrast to similar tests performed by Marzban and Sandgathe (2009) and Scheuerer and Hamill (2015), our data are physically consistent and thus bear close resemblance to observed rain fields. Lastly, we consider the choice of mother wavelet in detail, using the rigorous wavelet-selection procedure of Goel and Vidakovic (1995). The sensitivity of all newly introduced scores to the wavelet choice is assessed as well.

The remainder of this paper is structured as follows. The stochastic model of Hewer (2018) is introduced in Sect. 2. Sections 3 and 4 respectively discuss the wavelet transformation and spatial aggregation in detail. The general sensitivity of the wavelet spectra to changes in correlation structure is experimentally tested in Sect. 5. Based on these results, we define several possible deterministic and probabilistic scores in Sect. 6. In a second set of experiments (Sect. 7), we simulate synthetic sets of observations and predictions and test our scores' ability to correctly determine the best forecast. A comprehensive discussion of all results is given in Sect. 8.

2 Data: stochastic rain fields

In order to test whether our methodology can indeed detect structural differences between rain fields, we need a reasonably large rain-like data set whose structure is, to some extent, known a priori. Faced with a similar task, Wernli et al. (2008), Ahijevych et al. (2009), and others have employed purely geometric test cases. While those experiments are educational, we would argue that the simple, regular texture of such data bears too little resemblance with reality to constitute a sensible test case for our purposes. As an alternative, Marzban and Sandgathe (2009) considered Gaussian random fields, which have the advantage that the texture is more interesting and can be changed continuously via the parameters of the correlation model. However, since precipitation is generally known to follow non-Gaussian distributions, the realism of this approach is arguably still lacking.

In this study, we generate a more realistic testing environment using the work of Hewer (2018), who developed a physically consistent stochastic model of precipitation fields based on the moisture budget:

$$P = \max\left(E - T - \boldsymbol{v} \cdot \nabla q - q \nabla \cdot \boldsymbol{v},\ \ 0 \right), \tag{1}$$

where P denotes precipitation, E is a constant evaporation rate (in practice set to zero without loss of generality), q is the absolute humidity, and $\boldsymbol{v} = (u, v)^T$ is the horizontal wind field. The threshold T specifies the percentage of the field with non-zero values, i.e. the *base rate*. The velocity and its divergence are represented via the two-dimensional Helmholtz decomposition, which reads

$$\boldsymbol{v} = \nabla \times \Psi + \nabla \chi \qquad \Rightarrow \qquad \nabla \cdot \boldsymbol{v} = \nabla^2 \chi\,,$$

where $\nabla \times \Psi := (-\partial_x \Psi, \partial_y \Psi)^T$ is the rotation of the streamfunction and χ is the velocity potential. The spatial process of P is thus completely determined by $(\Psi, \chi, q)^T$, which we model as a multivariate Gaussian random field with zero mean and covariance matrix

$$\begin{aligned} &\mathrm{Cov}\left((\Psi_{\boldsymbol{s}}, \chi_{\boldsymbol{s}}, q_{\boldsymbol{s}})^T, (\Psi_{\boldsymbol{t}}, \chi_{\boldsymbol{t}}, q_{\boldsymbol{t}})^T \right) = \\ &\qquad \boldsymbol{\Sigma}_{\Psi,\chi,q} \cdot M(\,||b(\boldsymbol{t} - \boldsymbol{s})||,\ \nu\,)\,. \end{aligned} \tag{2}$$

Here, $\boldsymbol{t}, \boldsymbol{s} \in \mathbb{R}^2$ are two locations within the 2-D domain and M is the Matérn covariance function. The parameter b governs the scale of the correlations and the smoothness parameter ν determines the differentiability of the paths. The matrix $\boldsymbol{\Sigma}_{\Psi,\chi,q}$ is set to unity for our experiments, meaning that the velocity components and humidity are uncorrelated. Preliminary tests have shown that these parameters have negligible effects on the structural properties of the resulting rain fields. The covariances needed to simulate a realization of P via Eq. (1), i.e.

$$\begin{aligned} \mathrm{Cov}\Bigg(&\Big[q_{\boldsymbol{s}},\ \nabla \cdot q_{\boldsymbol{s}},\ \nabla \chi_{\boldsymbol{s}} - \nabla \times \Psi_{\boldsymbol{s}},\ \nabla^2 \chi_{\boldsymbol{s}} \Big]^T \\ &\Big[q_{\boldsymbol{t}},\ \nabla \cdot q_{\boldsymbol{t}},\ \nabla \chi_{\boldsymbol{t}} - \nabla \times \Psi_{\boldsymbol{t}},\ \nabla^2 \chi_{\boldsymbol{t}} \Big]^T \Bigg), \end{aligned}$$

follow from Eq. (2) by taking the respective mean-square derivatives. In the special case where Ψ, χ, and q are uncorrelated, these three Gaussian fields, as well as the necessary first and second derivatives, can directly be simulated via the `RMcurlfree` model from the `R` package `RandomFields` (Schlather et al., 2013). While the underlying distributions of Ψ, χ and q are Gaussian, the precipitation process, consisting of non-linear combinations of the derived fields, can exhibit non-Gaussian behaviour. For further details, the reader is referred to Hewer et al. (2017), Hewer (2018), and references therein.

Figure 1 shows several realizations of P. Here, as in the rest of this study, we have normalized all fields to unit sum, thereby removing any differences in total intensity and allowing us to concentrate on structure alone. We recognize that the model produces realistic-looking rain fields, at least for moderately low smoothness (small ν) and large scales (small values of b). Two important restrictions imposed by Eq. (2) become apparent as well. Firstly, the model is isotropic, meaning that it cannot produce the elongated, linear structures which are typical of frontal precipitation fields. Secondly, covariances are stationary, implying the same texture across the entire domain. An anisotropic extension of this model is theoretically relatively straightforward (replacing the scalar parameter b by a rotation matrix), but the technical implementation remains a non-trivial problem. The search for a non-stationary version is an open research question in its own right.

3 The redundant discrete wavelet transform

The technical core of our methodology consists of projecting the fields onto a series of so-called daughter wavelets $\psi_{j,l,\boldsymbol{u}}(\boldsymbol{r}) : \mathbb{R}^2 \to \mathbb{R}$, which are all obtained from a mother wavelet $\psi(\boldsymbol{r})$ via scaling by $\boldsymbol{r} \to \boldsymbol{r}/j$, a shift by $\boldsymbol{r} \to \boldsymbol{r} - \boldsymbol{u}$ and rotation in the direction denoted by l. Such *wavelet transforms*, which generate a series of basis functions from a single mother ψ that is localized in space and frequency, allow for an efficient analysis of non-stationary signals on a hierarchy of scales and have attained great popularity in numerous applications. For a general introduction to the field, we recommend Vidakovic and Mueller (1994) and Addison (2017).

Before we can apply wavelets to our problem, we must choose a mother ψ and decide which values of $\{j, l, \boldsymbol{u}\}$ to allow. Starting with the latter decision and guided by our desire to capture the field's covariance structure, we follow Weniger et al. (2017) and Kapp et al. (2018) in choosing a redundant discrete wavelet transform (RDWT). In this framework, the shift $\boldsymbol{u}$ takes on all possible discrete values, meaning that the daughters are shifted to all locations on the discrete grid. The scale j is restricted to powers of 2 and the daughters have three orientations with $l = 1, 2, 3$ denoting the horizontal, vertical, and diagonal direction, respectively. The projection onto these daughter wavelets, for which efficient algorithms are implemented in the `R` package `wavethresh` (Nason, 2016), transforms a $2^J \times 2^J$ field into $3 \times J \times 2^J \times 2^J$ coefficients, one for each location, scale, and direction. Our decision in favour of the RDWT is motivated by a relevant result proven in Eckley et al. (2010). Let

$$X(\boldsymbol{r}) = \sum_{\text{all } j,l,\boldsymbol{u}} \underbrace{W_{j,l,\boldsymbol{u}}}_{\text{weight}} \cdot \underbrace{\psi_{j,l,\boldsymbol{u}}(\boldsymbol{r})}_{\text{daughter}} \cdot \underbrace{\xi_{j,l,\boldsymbol{u}}}_{\text{noise}} \tag{3}$$

be the so-called *two-dimensional locally stationary wavelet process* (henceforth LS2W). The random increments $\xi_{j,l,\boldsymbol{u}}$ are assumed to be Gaussian white noise. *Local stationarity* means that X's auto-correlation varies infinitely slowly in the limit of infinitely large domains or, equivalently, infinitely highly resolved versions of a unit-sized domain. This requirement is enforced by certain asymptotic regularity conditions on the weights $W_{j,l,\boldsymbol{u}}$. For all technical details the reader is referred to Eckley et al. (2010); Kapp et al. (2018) present

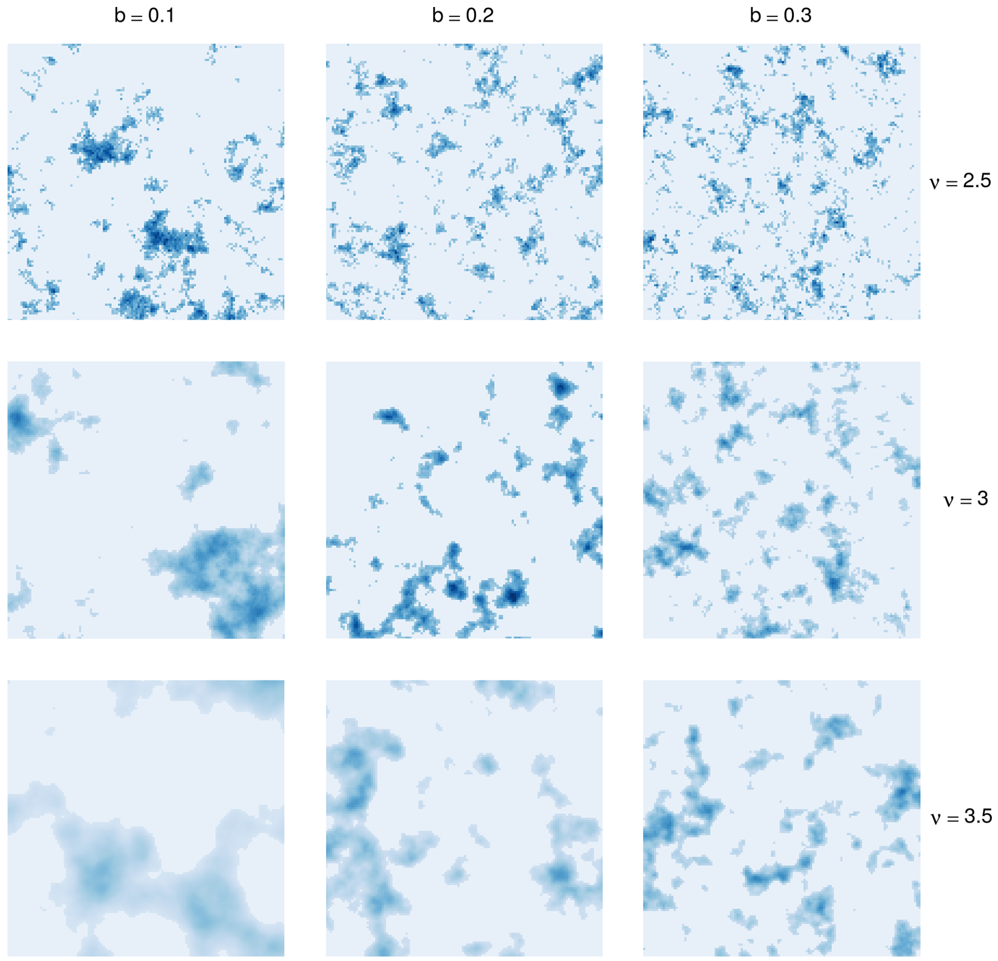


Figure 1. Example realizations of the stochastic rain model on a 128×128 grid for various choices of scale b and smoothness ν. The threshold T was chosen such that 20 % of the field has non-zero values.

a more condensed summary. The main result of Eckley et al. (2010) states that in the limit of an infinitely high spatial resolution, the autocovariances of X can directly be inferred from the squared weights $|W|^2$. In analogy to the Fourier transform, $|W|^2$ is referred to as the *local wavelet spectrum*. Eckley et al. (2010) have furthermore proven that the squared coefficients of X's RDWT constitute a biased estimator of this spectrum: due to the redundancy of the transform, the very large daughter wavelets all contain mostly the same information, leading to spectra which unduly over-emphasize the large scales. The bias is corrected via multiplication by a matrix $\mathbf{A}^{-1}$ which contains the correlations between the $\psi_{j,l,\boldsymbol{u}}$ and thus depends only on the choice of ψ and the size and resolution of the domain. Away from the asymptotic limit, this step occasionally introduces negative values to the spectra, which have no physical interpretation and pose some practical challenges in the subsequent steps. Preliminary investigations have shown that the abundance of this *negative energy* sharply decreases with the smoothness of the wavelet ψ and mostly averages out when mean spectra over the complete domain are considered (cf. Appendix, Fig. A3). Apart from the bias correction, the corrected local spectra also need to be smoothed spatially in order to obtain a consistent estimate. The complete procedure, including the computationally expensive calculation of $\mathbf{A}^{-1}$, is implemented in the `R` package `LS2W` (Eckley and Nason, 2011).

Having decided on a type of transformation, we must select a mother wavelet ψ. Our decision is restricted by the fact that the results of Eckley et al. (2010) have only been proven for the family of orthogonal Daubechies wavelets. These widely used functions, henceforth denoted D_N, have compact support in the spatial domain, increasing values of N indicate larger support sizes as well as greater smoothness. Smoother and hence more wave-like basis functions with better frequency localization are thus also more spread out in space. Figure 2 shows a few examples from this family. D_1 (panel a), the only family member that can be written in closed form, is widely known as the *Haar wavelet* (Haar, 1910) and has been applied in several previous verification studies (Casati et al., 2004; Weniger et al., 2017;

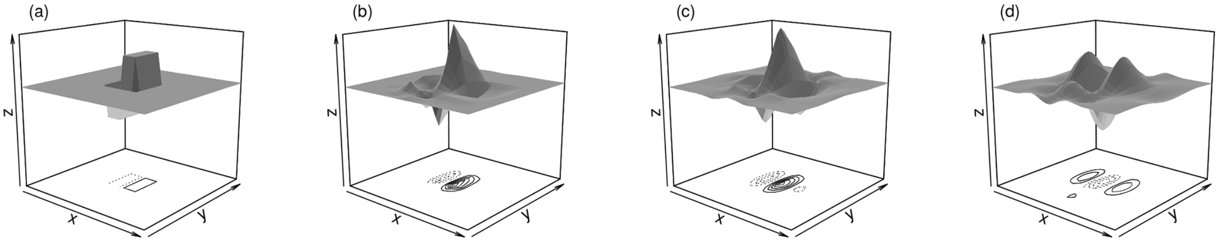


Figure 2. Two-dimensional Daubechies daughter wavelets D_1 and D_2 **(a, b)**, as well as least asymmetric D_4 **(c)** and D_8 **(d)**, vertical orientation.

Kapp et al., 2018). For $N > 3$, the constraints on smoothness and support length allow for multiple solutions, two of which are typically used: the *extremal phase* solutions are optimally concentrated near the origin of their support, while the *least asymmetric* versions have the greatest symmetry (Mallat, 1999). $D_{1,2,3}$ belongs in both sub-families; wherever a distinction is needed, we will label the two branches of the family as *ExP* and *LeA*, respectively. Among these available mother wavelets, we seek the basis that most closely resembles the data, thus justifying the model given in Eq. (3). To this end, we follow Goel and Vidakovic (1995) and rank wavelets by their ability to compress the original data: the sparser the representation (the more of the coefficients are negligibly small) in a given wavelet basis, the greater the similarity between basis functions and data. Relegating all details concerning this procedure to Appendix A, we merely note that the structure of the rain field, determined by the parameters b and ν, has substantially more impact on the efficiency of the compression than the choice of wavelet. Overall, the least asymmetric version of D_4 (shown in Fig. 2c) is most frequently selected as the best basis (28 % of cases), followed by D_1 and D_2 (21 % each). Unless otherwise noted, we will therefore employ LeA4 in all subsequent experiments. Considering the relatively small differences between wavelets, we hypothesize that the basis selection should have only minor effects on the behaviour of the resulting verification measures – a claim which is tested empirically in Sect. 7.

4 Wavelet spectra spatial aggregation

The previous step's redundant transform inflates the data by a factor of $3 \times J$, meaning that a radical dimension reduction is needed before verification can take place. Throughout this study, we will always begin this process by discarding the two largest scales, which are mostly determined by boundary conditions, and averaging over the three directions. The latter step is unproblematic, at least for our isotropic test cases. Next, the resulting fields must be spatially aggregated in a way that eliminates the double-penalty effect.

The straightforward approach to this task consists of simply averaging the wavelet spectra over all locations. The redundancy of the transform guarantees that this mean spectrum is invariant under shifts of the underlying field (Nason et al., 2000), thereby allowing us to circumvent double-penalty effects. Kapp et al. (2018) have already demonstrated that the spatial mean contains enough information to confidently distinguish between weather situations in a realistic setting. In particular, the difference between organized large-scale precipitation and scattered convection has a clear signature in these spectra – an observation that has recently been exploited by Brune et al. (2018), who defined a series of wavelet-based indices of convective organization using this approach. As mentioned above, we furthermore know that *negative energy*, introduced by the correction matrix $\mathbf{A}^{-1}$, mostly averages out in the spatial mean, provided that we choose a wavelet smoother than D_1 (cf. Appendix, Fig. A3).

In spite of these desirable properties, there are two main issues which motivate us to consider an alternative way of aggregation: if we normalize the mean spectrum to unit total energy, its individual values can be interpreted as the fraction of total *rain intensity* associated with a given scale and direction. It is easy to imagine cases where a very small fraction of the total precipitation area contains almost all of the total intensity and therefore dominates the mean spectrum. This is clearly at odds with the intuitive concept of *texture*. Furthermore, there is no obvious way of visualizing how individual parts of the domain contribute to the mean spectrum – if our visual assessment disagrees with the wavelet-based score, we can hardly look at all fields of coefficients at once in order to pinpoint the origin of the dispute. This second point leads us to introduce the *map of central scales* C: for every grid point (x, y) within the domain, we set $C_{x,y}$ to the centre of mass of the local wavelet spectrum. The resulting $2^J \times 2^J$ field of $C \in (1, J)$ is a straightforward visualization of the redundant wavelet transform, intuitively showing the dominant scale at each location. Since the centre of mass is only well defined for non-negative vectors, all negative values introduced by the bias correction via $\mathbf{A}^{-1}$ are set to zero before computing C.

To illustrate the concept, we have calculated the map of central scales for one of the test cases from the `SpatialVx` `R` package (Fig. 3, these data were originally studied by Ahijevych et al., 2009). Here, the original rain field was replaced by a logarithmic rain field, adding 2^{-3} to all grid points with zero rain in order to normalize the marginal distribu-

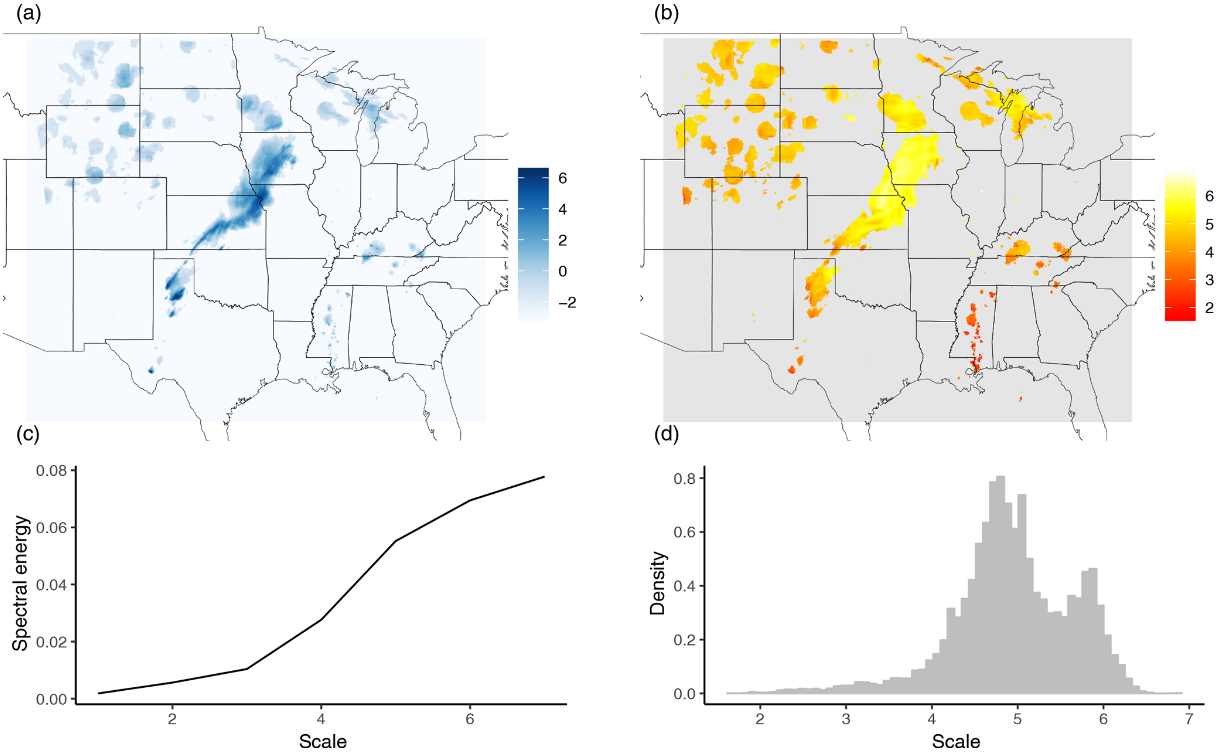


Figure 3. Logarithmic rain field **(a)** and corresponding map of central scales **(b)** from the stage II reanalysis on 13 May 2005. The field has been cut and padded with zeroes to 512 × 512, scales were calculated using the least asymmetric D_4 wavelet, only locations with non-zero rain are shown. Panels **(c)** and **(d)** show the corresponding mean spectrum and scale histogram, respectively.

tion (Casati et al., 2004) and reduce the impact of single extreme events. We see a clear distinction between the large frontal structure in the centre of the domain (scales 6–7), the medium-sized features in the upper-left quadrant (scale 4–5), and the very small objects on the lower right (scales ≤ 4). As an alternative to the spatial mean spectrum (Fig. 3c), we can base our scores on the histogram of C over all locations pooled together (Fig. 3d). Intuitively, this scale histogram summarizes which fraction of the total *area* is associated with features of various scales. We observe a clear bi-modal structure which nicely reflects the two dominant features on scales five and six.

5 Wavelet spectra sensitivity analysis

Before we design verification tools based on the mean wavelet spectra and histograms of central scales, it is instructive to study what these curves look like and how they react to changes in the model parameters b, ν, and T from Eq. (1). Can we correctly detect subtle differences in scale? What are the effects of smoothness and precipitation area? To answer these questions, we begin by simulating 100 realizations of our stochastic model on a 128×128 grid, first keeping the smoothness ν constant at 2.5 and varying the scale b between 0.1 and 0.5 (recall that large values of b indicate small-scaled features). For a second set of experiments, we simulate 100 fields with constant $b = 0.25$ and vary ν between 2.5 and 4. All of these fields are then normalized to unit sum (to eliminate differences in intensity), transformed and aggregated as described above.

Figure 4a shows the spatial mean spectra, averaged over all directions and realizations, as a function of the scale parameter b (on the y axis). As expected, an increase in b monotonically shifts the centre of these spectra towards smaller scales. Considering the experiment with variable ν (panel c), we find that an increase in smoothness results in a shift towards larger scales. This is in good agreement with the visual impression we get from the example realizations in Fig. 1. The corresponding scale histograms are shown in Fig. 4b and d. We observe that their centres, corresponding to the expectation values of the central scales, are shifted in the same directions (and to a similar extent) as the mean spectra.

In addition to the shift along the scale axis, we observe that the two model parameters have a secondary effect on the shapes of the curves: a decrease in b or ν goes along with flatter mean spectra – the energy is more evenly spread across scales. The histograms react similarly to b, larger scales coinciding with greater variance, while changes in ν have only a minor impact on the histogram's shape.

The final variable model parameter considered here is the threshold T, for which the expected reactions of our wavelet

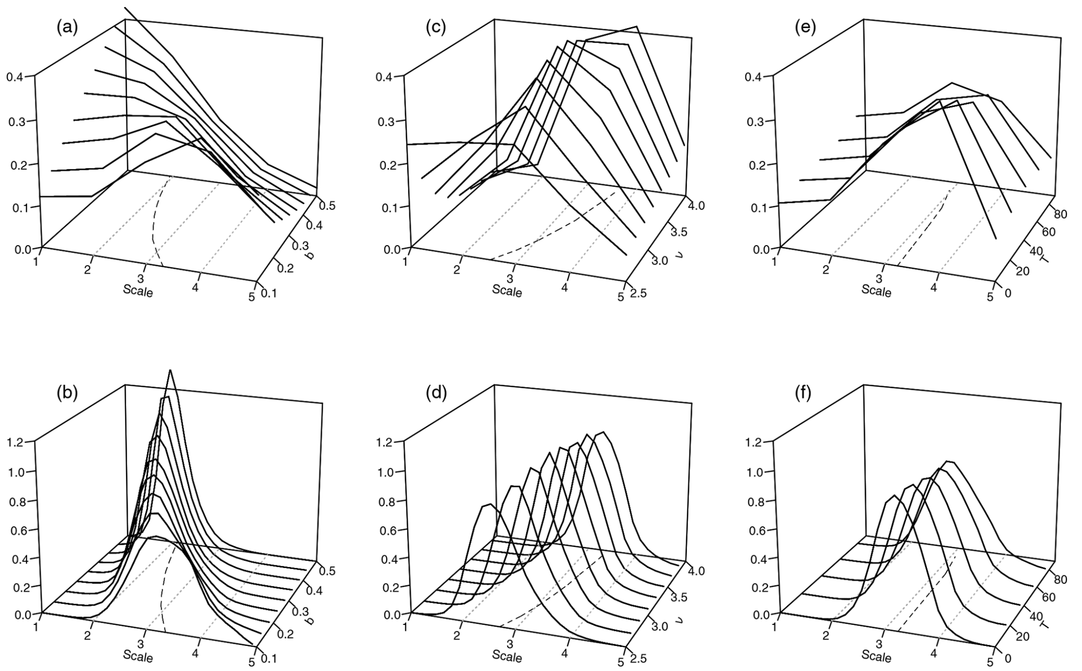


Figure 4. Mean spectra **(a, c, e)** and histograms of central scale **(b, d, f)**, as functions of the scale parameter b at $\nu = 2.5$ **(a, b)**, smoothness parameter ν at $b = 0.25$ **(c, d)**, and threshold T at $b = 0.1$ and $\nu = 2.5$ **(e, f)**. Dashed lines in the x–y-plane indicate the respective curve centres of mass; dotted grey lines (parallel to the y axis) were added for orientation.

characteristics are less clear: are fields with a larger fraction of precipitating area perceived to be scaled larger or smaller? To investigate this, we set $b = 0.1$ and $\nu = 2.5$ and vary the rain area between 10 % and 100 %. Figure 4e and f shows that the centres of the spectra and histograms hardly depend on T at all. The spread slightly increases with the threshold in both cases, but the changes are far more subtle than for the other two parameters.

In summary, we note that the two structural parameters ν and b have clearly visible effects on the mean spectra as well as the scale histograms. Metrics that compare the complete curves (as opposed to their centres alone) should be able to distinguish between errors in scale and smoothness since these characteristics have different effects on their location and spread. The effect of the threshold T is only moderate in comparison, but could potentially compensate errors in the other two parameters, which may occasionally lead to counterintuitive results.

6 Wavelet-based scores

Motivated by the previous section's results, we now introduce several possible scores, comparing the spectra and histograms of forecast and observed rain fields. Here, we consider the case of a single deterministic prediction, as well as ensemble forecasts.

6.1 Deterministic setting

From an observed field and a single deterministic forecast, we obtain the respective mean wavelet spectra and histograms of central scales as described above. If we naively compare these vectors in an element-wise way, we may fall victim to a new incarnation of the double-penalty problem since a small shift in one of the spectra (or histograms) will indeed be punished twice. Rubner et al. (2000) discuss this issue in great detail and suggest the *earth mover's distance* (henceforth EMD) as an alternative. The EMD between two non-negative vectors (histograms or spectra in our case) is calculated by numerically minimizing the cost of transforming one vector into the other, i.e. *moving the dirt from one arrangement of piles to another while doing the minimal amount of work*. Here, the locations of the piles corresponding to the histograms (spectra) are given by the centres of the bins (the scales of the spectrum), and the count (energy) determines the mass of the pile. For the simple one-dimensional case where the piles are regularly spaced, the EMD simplifies to the mean absolute difference between the two cumulative histograms (spectra) (Villani, 2003). This quantity is a true metric if the two vectors have the same norm, which is trivially true for the histograms. To achieve the same for the mean spectra, we normalize them to unit sum, thereby removing any bias in total intensity and concentrating solely on structure. Our first two deterministic wavelet-based structure scores are thus given by the EMD between the histograms of central scales (henceforth H_{emd}) and the normalized, spatially and directionally averaged, wavelet spectra (henceforth $\mathrm{Sp}_{\mathrm{emd}}$), respectively.

Being a metric, the EMD is positive and semi-definite and therefore yields no information on the direction of the error. We can obtain such a judgement by calculating, instead of the EMD, the difference between the respective centres of mass. For the histograms, this corresponds to the difference in expectation value. Rubner et al. (2000) have proven that the absolute value of this quantity is a lower bound of the EMD. Its sign indicates the direction in which the forecast spectrum or histogram is shifted, compared to the observations. We have thus obtained two additional scores, H_{cd} and $\mathrm{Sp}_{\mathrm{cd}}$, which are conceptually and computationally simpler than the EMD versions and allow us to decide whether the scales of the forecast fields are too large or too small.

6.2 Probabilistic setting

When predictions are made in the form of probability distributions (or samples from such a distribution), verification is typically performed using proper scoring rules (Gneiting and Raftery, 2007). Here, we treat scoring rules as cost functions to be minimized, meaning that low values indicate good forecasts. A function $\mathcal{S}$ that maps a probabilistic forecast and an observed event to the extended real line is then called a *proper* score when the predictive distribution F minimizes the expected value of $\mathcal{S}$ as long as the observations are drawn from F. In this case, there is no incentive to predict anything other than one's best knowledge of the truth. $\mathcal{S}$ is called *strictly proper* when F is the only forecast which attains that minimum. As mentioned above, Kapp et al. (2018) verified the spatial mean wavelet spectra via the logarithmic score, which necessitates a further dimension reduction step. In the interest of simplicity as well as consistency with our other scores, we employ the energy score (Gneiting and Raftery, 2007) instead, which is given by

$$\mathrm{En}(F,\boldsymbol{y}) = E_F|\boldsymbol{X}-\boldsymbol{y}| - 0.5E_F|\boldsymbol{X}-\boldsymbol{X}'|, \qquad (4)$$

where $\boldsymbol{y}$ is the observed vector, E_F denotes the expectation value under the multivariate distribution of the forecast F, and $\boldsymbol{X}$ and $\boldsymbol{X}'$ are independent random vectors with distribution F. Here, we substitute the observed mean spectrum for y and estimate F from the ensemble of predicted spectra. The resulting score, which we will denote as $\mathrm{Sp}_{\mathrm{en}}$, is proper in the sense that forecasters are encouraged to quote their true beliefs about the distribution of the spatial mean spectra.

The two previously introduced scores based on the histograms of central scales can directly be applied to the case of ensemble verification by estimating the forecast histogram from all ensemble members pooled together. In this setting where two distributions are compared directly, *proper divergences* (Thorarinsdottir et al., 2013) take the place of proper scores: a divergence, mapping predicted and observed distributions F and G to the real line, is called proper when its expected minimum lies at $F=G$. The square of H_{cd} corresponds to the mean value divergence, which is proper. H_{emd} is a special case of the Wasserstein distance, the propriety of which is only guaranteed in the limit of infinite sample sizes (Thorarinsdottir et al., 2013). Whether or not these divergences are useful verification tools in the probabilistic case will be tested empirically in Sect. 7.

All of our newly proposed wavelet-based texture scores are listed in Table 1.

6.3 Established alternatives

In order to benchmark the performance of our new scores, we compare them to potential non-wavelet alternatives from the literature. A first natural choice is the variogram score of Scheuerer and Hamill (2015), which is given by

$$V(F,\boldsymbol{y}) = \sum_{a,b=1}^{n} w_{a,b}\big(|y_a-y_b|^p - E_F[|X_a-X_b|^p]\big)^2, \qquad (5)$$

where $\boldsymbol{y}$, F, and X now correspond to the observed rain field, the distribution of the predicted rain fields, and a random field distributed according to the latter. a and b denote two grid points within the domain. The weights $w_{a,b}$ can be used to change the emphasis on pairs with small or large distances, while the exponent p governs the relative importance of single, extremely large differences. We include two versions of this score in our verification experiment: the naive choice $w_{a,b}=1$, $p=2$ (denoted V_{20} below) and the more robust configuration $w_{a,b}=|\boldsymbol{r}_a-\boldsymbol{r}_b|^{-1}$, $p=0.5$ ($V_{w,5}$ below), where $\boldsymbol{r}_a$ denotes the spatial location corresponding to the index a. Assuming stationarity of the data, we can efficiently calculate both of these scores by first aggregating the pairwise differences over all pairs with the same distance in space up to a pre-selected maximum distance. V_{20} then simplifies to the mean-square error between the two stationary variograms. The maximum distance is set to 20, which is a rough approximation of the range of the typical variograms of our test cases. Preliminary experiments have shown that this aggregation greatly improves the performances of $V_{w,5}$ and V_{20} in all of our experiments. It furthermore allows us to apply these scores to the case of deterministic forecasts.

As a second alternative verification tool, we include the S component of the well-known SAL (Wernli et al., 2008). This object-based structure score (1) identifies continuous rain objects in the observed and predicted rain field, (2) calculates the ratio between maximum and total precipitation in each object, (3) calculates averages over these ratios (weighted by the total precipitation of each object), and (4) compares these weighted averages of forecast and observation. The sign of S is chosen such that $S>0$ indicates forecasts which are scaled too large and/or too flat. In this study, we employ the original object identification algorithm by Wernli et al. (2008), setting the threshold to the maximum observed value divided by 15. We have checked that the sensitivity to this parameter is low

Table 1. Wavelet-based structure scores (top part) and established alternatives (bottom).

Abbreviation	Description	Probabilistic	Deterministic
$\mathrm{Sp}_{\mathrm{emd}}$	EMD of the mean spectra	no	yes
$\mathrm{Sp}_{\mathrm{cd}}$	distance in mean spectra's centre of mass	no	yes
H_{emd}	EMD of the scale histograms	yes	yes
H_{cd}	distance in the scale histograms' centre of mass	yes	yes
$\mathrm{Sp}_{\mathrm{en}}$	energy score of the predicted mean spectra	yes	no
RMSE	root-mean-square error between rain fields	no	yes
$V_{w,5}$	variogram score, $w_{a,b} = \lvert \boldsymbol{r}_a - \boldsymbol{r}_b \rvert^{-1}$, $p = 0.5$	yes	yes
V_{20}	variogram score, $w_{a,b} = 1$, $p = 2$	yes	yes
S	object-based structure score of Wernli et al. (2008)	yes	yes

Table 2. Varying parameters in Eq. (2) for the four groups of artificial ensemble forecasts.

Model	RL	SmL	RS	SmS
Smoothness ν	2.5	3	2.5	3
Scale b	0.1	0.1	0.2	0.2

in our test cases. For the purposes of ensemble verification, we employ a recently developed ensemble generalization of SAL (Radanovics et al., 2018). Here, the ratio between maximum and total predicted rain is averaged not only over rain objects, but also over the ensemble members.

Lastly, the naive root-mean-square error (RMSE) will be included in our deterministic verification experiment in order to confirm the necessity for more sophisticated methods of analysis.

7 Idealized verification experiments

For our first set of randomly drawn forecasts and observations from the model given by Eq. (1), we keep the threshold T constant such that 20 % of the fields have non-zero values and select four combinations of ν and b, listed in Table 2.

The resulting texture is rough and large scaled (RL), smooth and large scaled (SmL), rough and small scaled (RS), and smooth small scaled (SmS). One realization for each of those settings is depicted in Fig. 1. In the following sections, we interpret random samples of these models as observations and forecasts, thus allowing us to observe how frequently the truly best prediction (the one with the same parameters as the observation) is awarded the best score.

7.1 Ensemble setting

Beginning with the synthetic ensemble verification experiment, we draw 100 realizations each from RL and RS as our observations. For every observation (200 in total), we issue four ensemble predictions, consisting of 10 realizations from RL, SmL, RS, and SmS, respectively. Only one of these 10-member ensembles thus represents the correct correlation structure while the other three are wrong in either scale, smoothness, or both. Observation and ensembles are compared via the three wavelet scores H_{cd}, H_{emd}, and $\mathrm{Sp}_{\mathrm{en}}$ as well as the established alternatives for ensemble forecasts, i.e. S, V_{20}, and $V_{w,5}$.

Figure 5 shows the resulting score distributions. All scores are best when small, except for the two-sided S and H_{cd} where values near zero are optimal. Beginning with the case where the observations are drawn from RS (top row of Fig. 5), we observe that the four predictions are ranked quite similarly by all scores. Here, the correct forecast almost always receives the best mark, while SmL, which is most dissimilar from RS, fares worst. S and H_{cd} furthermore agree that all three false predictions are scaled too large. The task of determining the truly best forecast is substantially more complicated when the observations belong to RL (bottom row of Fig. 5): since SmS is both smoother and scaled smaller, the effects on the location of the spectra and histograms along the scale axis (cf. Fig. 4) compensate each other. These curves can therefore hardly be distinguished by their centres of mass alone. We recognize that RL and SmS consequently obtain similar values of H_{cd}, this score judging solely based on the centres. The other two wavelet scores achieve better discrimination, as does $V_{w,5}$. Concerning the signs of the error, we note that S and H_{cd} both consider RS too small and SmL too large. For SmS, H_{cd} is only slightly negative, indicating nearly correct scales. S is less affected by the compensating effect of increased smoothness and determines more clearly that SmS is smaller-scaled than RL. Its overall success rate is, however, not significantly better than that of H_{cd}.

Figure 6 summarizes the ability of the six tested probabilistic scores to correctly determine the best forecast ensemble. As discussed above, all scores are very successful at determining correct forecasts of RS. In the alternative setting (observations from RL), SmS is the most frequent wrong answer, receiving the smallest (absolute) values of V_{20}, S, H_{cd}, and H_{emd} in more than a quarter of cases. In contrast to the other scores, $\mathrm{Sp}_{\mathrm{en}}$ hardly ever erroneously prefers SmS over RL. Instead, SmL is wrongly selected most frequently, lead-

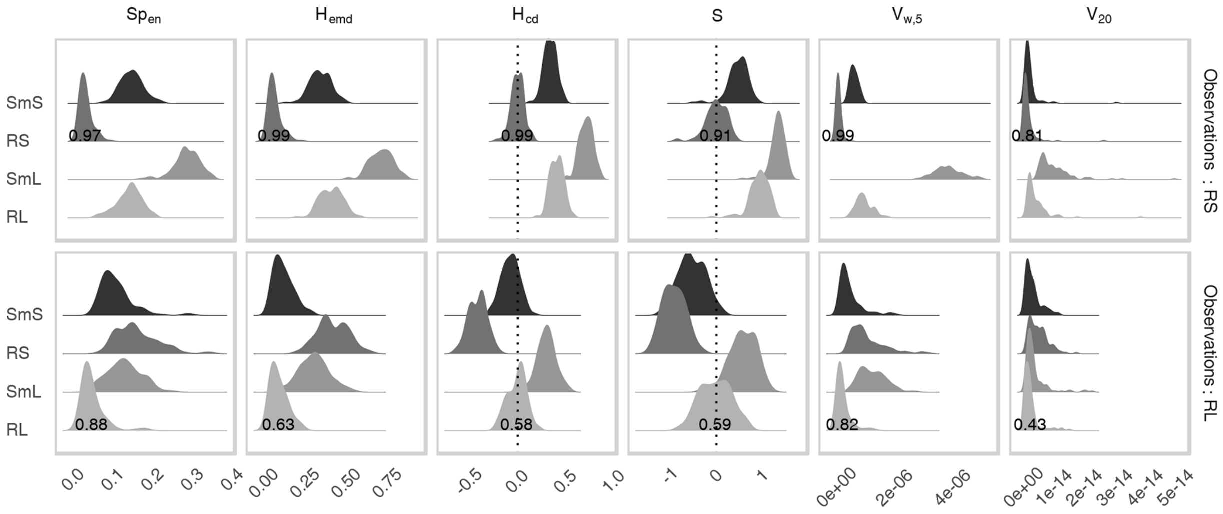


Figure 5. Distribution of all probabilistic scores for each of the forecast ensembles corresponding to the four models from Table 2. Top row: observations drawn from RS. Bottom row: observations drawn from RL. Numbers denote the fraction of cases in which the forecast from the correct distribution received the best score.

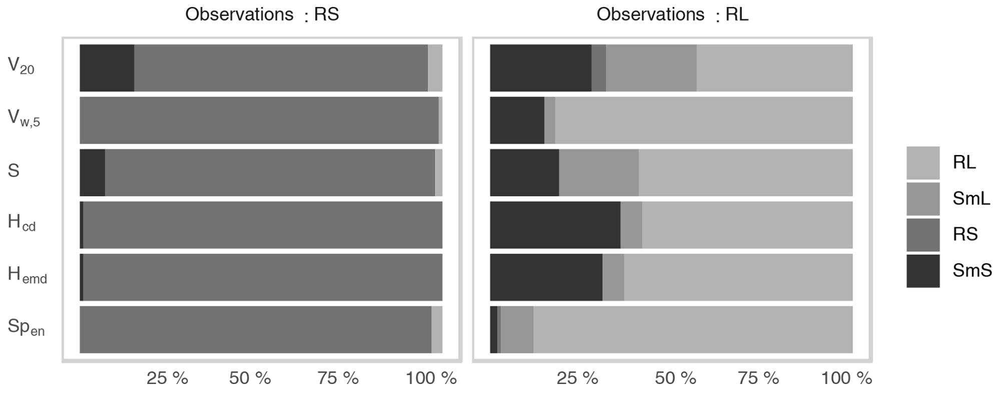


Figure 6. Percentage of cases where each of the four ensembles corresponding to the models in Table 2 was deemed the best forecast, separated by score and the model of the observation.

ing to the overall lowest error rate (12 %) in this part of the experiment.

7.2 Deterministic setting

Having investigated the behaviour of our probabilistic scores, we now consider the deterministic case: how successfully can we determine the truly best forecast, given only a single realization? The set-up for this experiment is the same as before, only the size of the forecast ensembles is reduced from 10 to 1. Since the resulting scores naturally have greater variances than before, we increase the number of observations to 1000 (500 each from RL and RS) in order to achieve similarly robust results. In addition to the four appropriate wavelet scores ($\mathrm{Sp}_{\mathrm{emd}}$, $\mathrm{Sp}_{\mathrm{cd}}$, H_{emd} and H_{cd}), we again calculate $V_{w,5}$ and V_{20} as well as the S component of the original SAL score. To ensure that the verification problem is sufficiently difficult, the root-mean-square error (RMSE) is included as a naive alternative as well.

Figure 7 reveals that correct forecasts are again easily identified by all of the wavelet-based scores when the observed fields belong to RS. As in the ensemble scenario, the main difficulty lies in the decision between SmS and RL in cases where the latter model generates the observations. The two EMD scores, which use the complete curves and not just their centres, clearly outperform the corresponding CD versions in this part of the experiment and detect RL correctly in the majority of cases. $V_{w,5}$ is similarly successful as the best wavelet-based score, faring marginally better than $\mathrm{Sp}_{\mathrm{emd}}$. The failure rates of V_{20} and S are again slightly higher. Unsurprisingly, the RMSE is completely unsuited to the task at hand, achieving less than 25 % correct verdicts overall. The inferiority to a random evaluation, which would, on average, be correct one-fourth of the time, is caused by

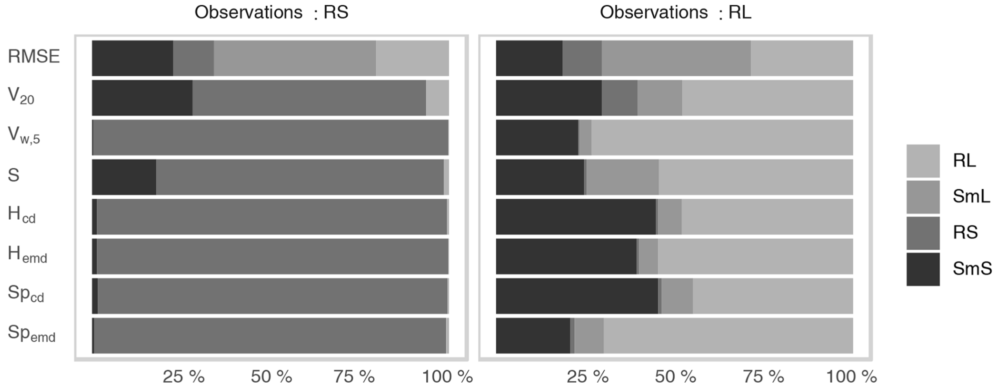


Figure 7. As Fig. 6, but for the deterministic verification experiment.

Table 3. Fraction of cases where the correct forecast received the best score for a range of extremal phase (ExP) and least asymmetric (LeA) Daubechies wavelets. LeA4 is the wavelet used for all other experiments in this study; ExP1 is the well-known Haar wavelet.

	Deterministic case				Ensemble case		
	$\mathrm{Sp}_{\mathrm{emd}}$	$\mathrm{Sp}_{\mathrm{cd}}$	H_{emd}	H_{cd}	$\mathrm{Sp}_{\mathrm{en}}$	H_{emd}	H_{cd}
ExP1	0.76	0.72	0.73	0.72	0.86	0.78	0.78
ExP2	0.83	0.73	0.8	0.77	0.92	0.82	0.83
ExP4	0.87	0.7	0.8	0.75	0.94	0.87	0.78
ExP6	0.87	0.7	0.82	0.73	0.94	0.86	0.74
LeA4	0.84	0.71	0.76	0.73	0.92	0.81	0.78
LeA6	0.86	0.69	0.76	0.69	0.94	0.88	0.8

Table 4. As Table 3, but without the bias-correction step.

	Deterministic case				Ensemble case		
	$\mathrm{Sp}_{\mathrm{emd}}$	$\mathrm{Sp}_{\mathrm{cd}}$	H_{emd}	H_{cd}	$\mathrm{Sp}_{\mathrm{en}}$	H_{emd}	H_{cd}
ExP1	0.66	0.6	0.63	0.63	0.68	0.76	0.74
ExP2	0.65	0.57	0.65	0.64	0.68	0.74	0.76
ExP4	0.65	0.56	0.61	0.6	0.68	0.72	0.72
ExP6	0.63	0.54	0.59	0.59	0.66	0.7	0.71
LeA4	0.63	0.56	0.64	0.63	0.68	0.7	0.68
LeA6	0.63	0.55	0.62	0.61	0.66	0.69	0.68

the fact that the model with the largest, smoothest features (SmL) has the least potential for double penalties and thus fares best in a point-wise comparison; in fact, RMSE orders the four models by their typical features size, irrespective of the distribution of the observation.

7.3 Wavelet choice and bias correction

One obvious question to ask is whether or not the choice of the mother wavelet has a significant impact on the success rates in the two experiments discussed above. To address this issue, we repeat both the deterministic and the ensemble verification processes for several Daubechies wavelets. Recalling the results of our objective wavelet selection (Sect. 3 and Appendix A), we expect no dramatic effects.

Table 3, listing the overall success rates for each tested wavelet, mostly confirms this expectation: in the deterministic case, $\mathrm{Sp}_{\mathrm{emd}}$ and H_{emd} are really only affected by the choice between the Haar wavelet, which performs worst, and any of its smoother cousins. The two centre-based scores ($\mathrm{Sp}_{\mathrm{cd}}$ and H_{cd}) show hardly any wavelet dependence at all. Sensitivities are overall slightly higher in the ensemble case. While D_1 again appears to be the worst choice, there are some differences between the other options, particularly for the two histogram scores. Generally speaking, the impacts of wavelet choice on our verification results are nonetheless rather limited, as long as the Haar wavelet is avoided.

To confirm that the bias correction following Eckley et al. (2010) is indeed a necessary part of our methodology, we repeat these experiments without applying the correction matrix $\mathbf{A}^{-1}$. Without discussing the details (Table 4), we merely note that the success rates decrease substantially (depending on score and wavelet), meaning that bias correction generally cannot be skipped.

7.4 Perturbed thresholds

Next, we consider the case where forecast and observations are subject to random perturbations which are not directly related to the underlying covariance model. One rather natural way of implementing this scenario consists of randomly perturbing the thresholds, i.e. the fractions of the domain covered by non-zero precipitation. In a realistic context, such random differences between forecast and observation could be associated with a displacement error which shifts unduly large or small parts of a precipitation field into the forecast domain.

Our experiments in Sect. 5 indicate that the wavelet-based scores should be relatively robust to small changes in the threshold T (cf. Fig. 4e and f). For the variogram scores, one might expect greater sensitivity since the presence of a fixed fraction of zero values greatly reduces the variance in

Table 5. Fraction of cases where the correct forecast received the best score. Top two rows: deterministic forecasts with and without perturbed threshold. Bottom: ensemble forecasts with and without perturbed thresholds.

		$\mathrm{Sp}_{\mathrm{en}}$	$\mathrm{Sp}_{\mathrm{emd}}$	$\mathrm{Sp}_{\mathrm{cd}}$	H_{emd}	H_{cd}	$V_{w,5}$	V_{20}	S	RMSE
det.	constant T	–	0.84	0.71	0.76	0.73	0.86	0.57	0.68	0.2
	random T	–	0.83	0.7	0.78	0.74	0.56	0.35	0.67	0.22
ens.	constant T	0.92	–	–	0.81	0.78	0.9	0.62	0.75	–
	random T	0.92	–	–	0.8	0.75	0.7	0.44	0.74	–

the pairwise distances from which the stationary variogram is estimated. To test these hypotheses, we again repeat the two verification experiments, this time randomly varying T such that the precipitation area, previously fixed at 20 %, is a uniform random variable between 15 % and 25 % of complete domain.

Looking at the resulting success rates (Table 5), we find our expectations largely confirmed: while variations in the precipitation coverage hardly influence our wavelet-based judgement, $V_{w,5}$ and V_{20} seem to strongly depend on this parameter, thus mostly losing their ability to determine the correct model. The performances of S and RMSE are only weakly influenced by variations in T.

8 Summary and discussion

The basic idea of this study is that the structure of precipitation fields can be isolated and subsequently compared using two-dimensional wavelet transforms. Building on the work of Eckley et al. (2010) and Kapp et al. (2018), we have argued that the corrected, smoothed version of the redundant discrete wavelet transform (RDWT) is an appropriate tool for this task since it is shift invariant and has a proven asymptotic connection with the correlation function of the underlying spatial process. This approach is theoretically more flexible than Fourier- or variogram-based methods which make some form of global stationarity assumption, while our method relies on the substantially weaker requirement of local stationarity.

Before wavelet-transformed forecasts and observations can be compared to one another, the spatial data must be aggregated in a way that avoids penalizing displacement errors twice. Besides the proven strategy (Kapp et al., 2018) of averaging the wavelet spectra over all locations, we have newly introduced the map of central scales as a potentially interesting alternative: by calculating the centre of mass for each local spectrum, we obtain a matrix of the same dimensions as the original field, each value quantifying the locally dominant scale. Aside from the possibility of compactly visualizing the output of the RDWT in a single image, the histogram of these scales can serve as an alternative basis for verification, emphasizing each scale based on the area in which it dominates, rather than the fraction of total rain intensity it represents.

In order to rigorously test the sensitivity of these aggregated wavelet transforms to changes in the structure of rain fields, a controlled but realistic test bed was needed. The stochastic precipitation model of Hewer (2018) constitutes a very convenient case study for our purposes: the construction based on the moisture budget and a Helmholtz-decomposed wind field allows for non-Gaussian behaviour and guarantees that the simulated data are more realistic than simple geometric patterns or Gaussian random fields. The model's structural properties can nonetheless be determined at will via the smoothness and scale parameter of the underlying Matérn fields, allowing us to simulate observations and forecasts with known error characteristics. In a realistic context, errors in scale correspond to misrepresentation of feature sizes (e.g. smoother representation of small-scale convective organization), while errors in smoothness correspond to forecast models with a resolution that is too course, which are incapable of reproducing fine structures.

In a first suite of experiments we found that the wavelet spectra do indeed react sensitively to changes in both of these parameters. In particular, errors in smoothness and scale have different signatures which can potentially be differentiated from one another. Encouraged by these results, we have defined several possible scores which compare mean spectra and scale histograms via the difference between their centres (H_{cd} and $\mathrm{Sp}_{\mathrm{cd}}$), their earth mover's distance (H_{emd} and $\mathrm{Sp}_{\mathrm{emd}}$), and the energy score ($\mathrm{Sp}_{\mathrm{en}}$). In our idealized verification experiments, the performance of the latter three scores, i.e. their ability to correctly determine the objectively best forecast, was on par with the best tested variogram score ($V_{w,5}$). The less robust V_{20}, as well as the SAL's structure component S and the simplistic RMSE, was clearly outperformed. H_{cd} and $\mathrm{Sp}_{\mathrm{cd}}$, while less proficient at finding the correct answer, do yield valuable auxiliary information in the form of the error's sign, answering the question of whether the predicted structure was too coarse or too fine. Keeping in mind that both spectra and histograms can have multi-modal structures in realistic non-stationary cases (compare Fig. 3d), a comparison based on centres alone is likely not sufficient and the EMD versions of these scores should be preferred. If a signed structure score is desired, we can simply multiply the respective EMD by the sign of the difference in centres.

All five wavelet scores were shown to be robust to small perturbations of the data, realized here as random changes to the fraction of non-zero rain. In these experiments, which essentially test the score's sensitivity to the sample climatology, the variograms largely lost their ability to determine the correct forecast. Interpreting this result, it is important to keep in mind that our wavelet scores were specifically designed to judge based on structure alone while the variogram-methodology of Scheuerer and Hamill (2015) allows for a more holistic assessment. Sensitivity to precipitation coverage is therefore not necessarily a disadvantage. If the goal is a pure assessment of structure, this dependence is undesirable.

The two free parameters of the variogram score, namely the exponent p and the choice of weights $w_{i,j}$, were found to have a significant impact on the resulting verification. We have also tested the sensitivity of the newly introduced wavelet scores to the choice of the mother wavelet. An objective wavelet-selection procedure following Goel and Vidakovic (1995) was performed and the verification experiments were repeated for a variety of possible choices. Summarizing both of these steps, we can conclude that the success of our wavelet-based verification depends only weakly on the choice of an appropriate mother wavelet. One somewhat surprising exception is the Haar wavelet, which was favoured by previous studies (cf. Weniger et al., 2017, and references therein) but turned out to be a suboptimal choice for our purposes.

Now that the merits of wavelet-based structure scores have been demonstrated in a controlled environment, further tests are needed to study their behaviour in real-world verification situations. One important open question concerns the use of direction information, which was neglected in the present study but may well be valuable in a more realistic scenario. It is furthermore worth noting that, in contrast to primarily rain-specific tools like SAL, our methodology can be applied to any variable of interest with no major changes besides the new selection of an appropriate mother wavelet. A simultaneous evaluation of, for example, wind components, humidity, and cloud-cover – using the exact same verification tool to assess structural agreement in each variable – is thus feasible and could answer interesting questions concerning the origins of specific systematic forecast deficiencies.

Code and data availability. All necessary `R` code for the simulation of the stochastic rain fields, as well as the wavelet-based forecast verification, is available from https://github.com/s6sebusc/wv_verif (last access: 31 July 2019). The specific version used in this paper was also archived at https://doi.org/10.5281/zenodo.3257510 (Buschow, 2019).

Appendix A: Entropy-based wavelet selection

To find the most appropriate wavelet, we calculate the entropy of the transform's squared coefficients (representing the energy of the transformed data) and select the wavelet with the smallest entropy. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be a vector with non-negative entries satisfying $\sum_i y_i = 1$. For our purposes, its entropy is defined as

$$s(\boldsymbol{y}) := -\sum_{i=1}^{n} y_i \log_2 y_i \quad \in \quad [0, \log_2(n)], \tag{A1}$$

where we set $0 \cdot \log_2(0) = 0$. Following Goel and Vidakovic (1995), the RDWT is replaced by its corresponding orthogonal decomposition, which is obtained by selecting every second of the finest-scale coefficients, every fourth on the second-finest scale, and so on. The number of data points is thus conserved under the transformation and we can compare the entropy of the transformed data to that of the original representation.

The outcome of this procedure depends on the structure of the data to be transformed, the smoothness of the wavelet, and the length of its support. To understand how these properties interact, we quantify smoothness via the number of vanishing moments: a wavelet ψ is said to have N vanishing moments if $\int x^q \psi(x) \mathrm{d}x = 0$ for $q = 0, \ldots, N-1$. This implies that polynomials of order $N-1$ have a very sparse representation in the wavelet basis corresponding to ψ. The theorem of Deny-Lions (Cohen, 2003) relates this property to a function's differentiability: loosely speaking, if f is N times differentiable, the error made when approximating f by polynomials of order $N-1$ is bounded by a constant times the energy of f's Nth derivative $f^{(N)}$. It follows that f is well represented by wavelets with N vanishing moments, as long as $f^{(N)}$ is not too large.

Besides more or less smooth regions within the rain fields (in our test cases governed by the parameter ν) and constant zero areas outside, the data we wish to transform also contain singularities at the edges of precipitating features. Here, $f^{(N)}$ is generally not small and wavelets with shorter support length are superior since fewer coefficients are affected by any given singularity. Heisenberg's uncertainty principle ensures that localization in space and approximation of polynomials (related to the localization in frequency) cannot both be optimal simultaneously: if a wavelet has N vanishing moments, then its support size (in one dimension) is at least $2N-1$. In proving this theorem, Daubechies (1988) introduced the D_N wavelets, which are optimal in the sense that they have N vanishing moments at the smallest possible support.

To illustrate the competing effects of support size and smoothness on the efficiency of the wavelet transformation, we simulate one-dimensional Gaussian random fields with Matérn covariances (same function M and parameters b and ν as in Eq. 2, but only one variable and one spatial dimension). Figure A1 neatly demonstrates the concepts discussed above: when the time series is uniformly smooth, the higher order wavelet D_4 delivers a far more efficient compression than D_1 (panels a, c, e). The situation changes when we truncate the data (b, d, f): while D_4 continues to be superior within the smooth regions, D_1, due to its shorter support, requires fewer coefficients to represent the regions of constant zero values. This trade-off between representing smooth internal structure and intermittency is precisely quantified by the entropy (defined in Eq. A1, values noted in the captions of Fig. A1), which measures the total degree of concentration on a small number of coefficients: while the D_4 does better in both cases, the relative and absolute improvement is worse in the cut off case, where we introduced artificial singularities.

Figure A2 shows the results of our entropy-based wavelet-selection procedure for the model given by Eq. (1). We observe that the model parameters have substantially more impact on the efficiency of the compression than the choice of wavelet. Fields with greater smoothness and larger scales (large values of ν and small values of b) are represented far more compactly than rough small-scale cases, irrespective of the chosen basis. The differences between wavelets, while small in comparison, reveal a systematic behaviour: increasing support length leads to monotonously worse compression and the least asymmetric wavelets tend to fit slightly better than their "extremal phase" counterparts. The Haar wavelet constitutes an exception to this pattern, its entropy being frequently larger than that of several of its smoother cousins.

Besides theoretical optimality motivated by Eq. (3), practical concerns can play an important role in the selection of an appropriate wavelet as well. Recalling that the bias-correction following Eckley et al. (2010) can introduce negative values to the spectra, which have no intuitive interpretation, we are interested in seeing whether the problem can be circumvented by selecting an appropriate mother wavelet. Fig. A3 shows the ratio between negative and positive energy in the mean spectra from the experiments discussed in Sect. 7.3. For D_1, this ratio is typically close to one-tenth. Such large quantities of negative energy are rare for D_2 and basically never occur in higher-order wavelets. This observation, while reassuring, does not alter our wavelet selection since the Haar wavelet was not favoured by the entropy-based approach either.

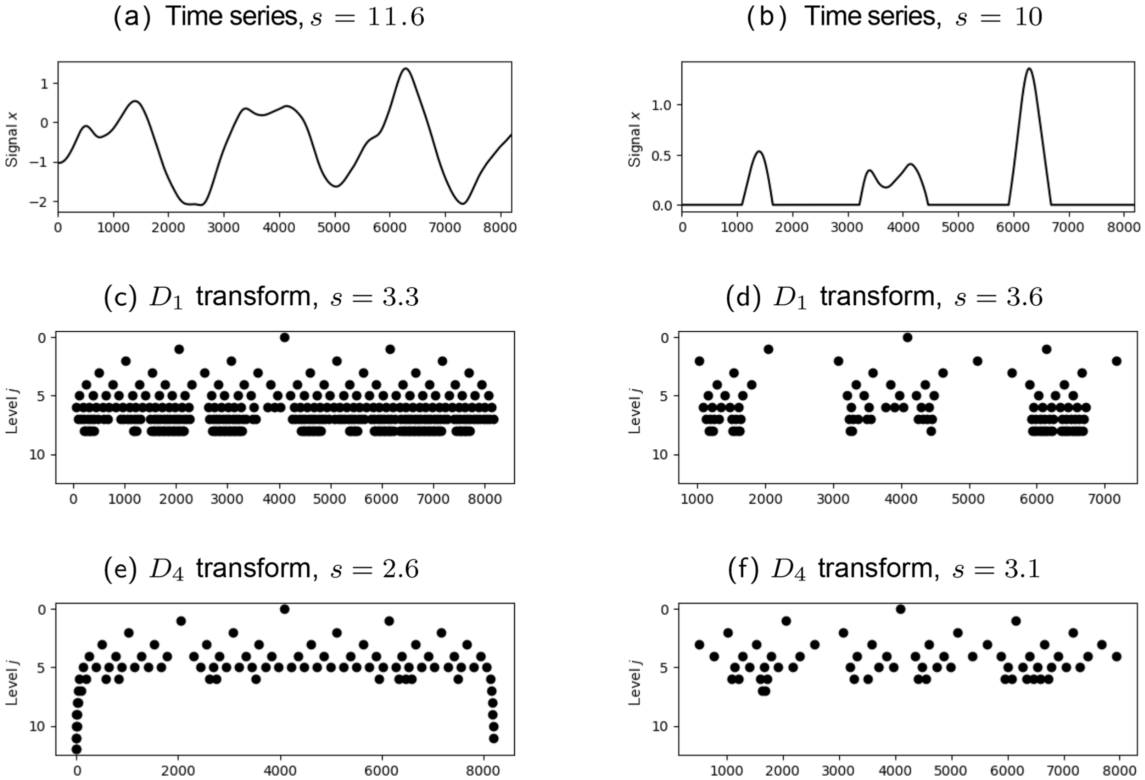


Figure A1. Realization of a one-dimensional Gaussian random vector with covariance $M(\nu = 3.5, b = 2)$ **(a)** and the corresponding values of the D_1 transform and least asymmetric D_4 transform **(c, e)** which are greater than 0.1. Panels **(b)**, **(d)**, and **(f)** are the corresponding plots for the cases where the vector is cut off at zero.

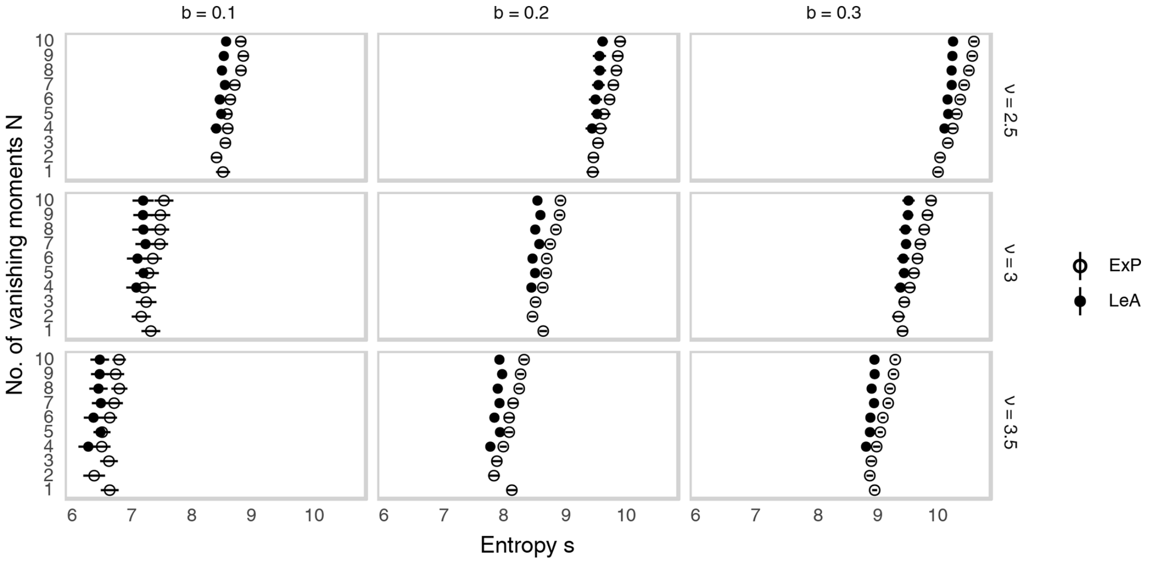


Figure A2. Entropy of the wavelet-transformed synthetic rain fields from Fig. 1 as a function of the wavelet's order N. Empty and filled dots correspond to the extremal phase version and least asymmetric version of D_N, respectively. Lines indicate 1 standard deviation, estimated from 10 realizations.

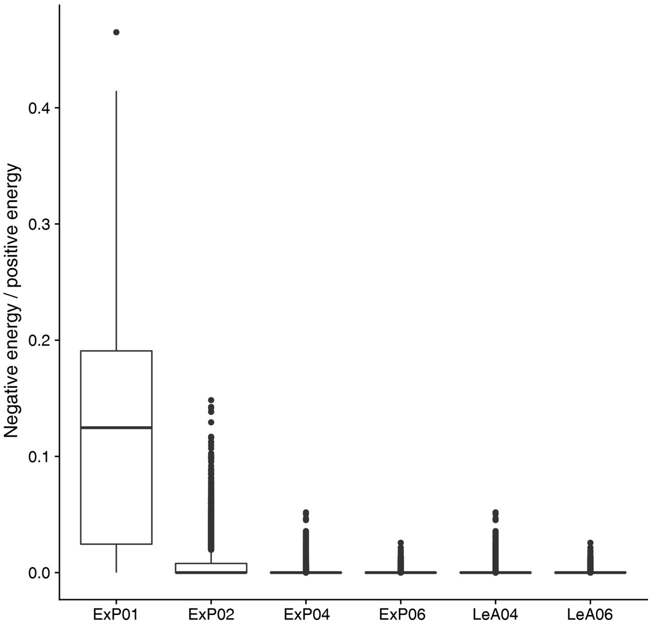


Figure A3. Ratio of negative to positive energy in the mean spectra (data set from Sect. 7.3, all models from Table 2, six selected mother wavelets as in Table 3).

Author contributions. SB and PF developed the basic concept and methodology for this work. SB designed and carried out the experiments used to test the methods. JP investigated and carried out the wavelet-selection procedure and led the writing and visualization in this part of the study. The rest of the writing and coding was led by SB with contributions from PF and JP. All authors contributed to the proof-reading and added valuable suggestions to the final draft.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We are very grateful to Rüdiger Hewer for providing the original program code, as well as invaluable guidance for the stochastic rain model. Further thanks go to Franka Nawrath for providing efficient code to calculate the variogram score, as well as helpful discussions concerning its implementation. We would also like to thank Michael Weniger for many suggestions concerning the use of wavelets for forecast verification. Finally, our thanks go to Joseph Bellier and one anonymous reviewer for their constructive criticism and thoughtful suggestions.

Financial support. This research has been supported by the DFG (grant no. FR 2976/2-1).

Review statement. This paper was edited by Simon Unterstrasser and reviewed by Joseph Bellier and one anonymous referee.

References

Addison, P. S.: The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance, CRC press, 2017.

Ahijevych, D., Gilleland, E., Brown, B. G., and Ebert, E. E.: Application of spatial verification methods to idealized and NWP-gridded precipitation forecasts, Weather Forecast., 24, 1485–1497, 2009.

Bachmaier, M. and Backes, M.: Variogram or semivariogram? Variance or semivariance? Allan variance or introducing a new term?, Math. Geosci., 43, 735–740, 2011.

Brune, S., Kapp, F., and Friederichs, P.: A wavelet-based analysis of convective organization in ICON large-eddy simulations, Q. J. Roy. Meteor. Soc., 144, 2812–2829, 2018.

Buschow, S.: wv_verif (Version 0.1.0), Zenodo, https://doi.org/10.5281/zenodo.3257511, 2019.

Casati, B., Ross, G., and Stephenson, D.: A new intensity-scale approach for the verification of spatial precipitation forecasts, Meteorol. Appl., 11, 141–154, 2004.

Cohen, A.: Numerical analysis of wavelet methods, vol. 32, Elsevier, 2003.

Daubechies, I.: Orthonormal bases of compactly supported wavelets, Commun. Pure Appl. Math., 41, 909–996, 1988.

Davis, C., Brown, B., and Bullock, R.: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas, Mon. Weather Rev., 134, 1772–1784, 2006.

Dorninger, M., Gilleland, E., Casati, B., Mittermaier, M. P., Ebert, E. E., Brown, B. G., and Wilson, L. J.: The Setup of the MesoVICT Project, B. Am. Meteorol. Soc" 99, 1887–1906, https://doi.org/10.1175/BAMS-D-17-0164.1, 2018.

Ebert, E. E.: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework, Meteorol. Appl., 15, 51–64, 2008.

Eckley, I. A., Nason, G. P., and Treloar, R. L.: Locally stationary wavelet fields with application to the modelling and analysis of image texture, J. Roy. Stat. Soc. C-Appl., 59, 595–616, 2010.

Eckely, I. A. and Nason, G. P.: LS2W: Implementing the Locally Stationary 2D Wavelet Process Approach in R, J. Stat. Softw., 43, 1–23, 2011.

Ekström, M.: Metrics to identify meaningful downscaling skill in WRF simulations of intense rainfall events, Environ. Model. Softw., 79, 267–284, 2016.

Gilleland, E.: Spatial forecast verification: Baddeley's delta metric applied to the ICP test cases, Weather Forecast., 26, 409–415, 2011.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E.: Intercomparison of spatial forecast verification methods, Weather Forecast., 24, 1416–1430, 2009.

Gilleland, E., Lindström, J., and Lindgren, F.: Analyzing the image warp forecast verification method on precipitation fields from the ICP, Weather Forecast., 25, 1249–1262, 2010.

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc., 102, 359–378, 2007.

Goel, P. K. and Vidakovic, B.: Wavelet transformations as diversity enhancers, Institute of Statistics & Decision Sciences, Duke University Durham, NC, 1995.

Haar, A.: Zur Theorie der orthogonalen Funktionensysteme, Mathematische Annalen, 69, 331–371, 1910.

Han, F. and Szunyogh, I.: A Technique for the Verification of Precipitation Forecasts and Its Application to a Problem of Predictability, Mon. Weather Rev., 146, 1303–1318, 2018.

Hewer, R.: Stochastisch-physikalische Modelle für Windfelder und Niederschlagsextreme, PhD thesis, University of Bonn, available at: http://hss.ulb.uni-bonn.de/2018/5122/5122.htm (last access: 31 July 2019), 2018.

Hewer, R., Friederichs, P., Hense, A., and Schlather, M.: A Matérn-Based Multivariate Gaussian Random Process for a Consistent Model of the Horizontal Wind Components and Related Variables, J. Atmos. Sci., 74, 3833–3845, 2017.

Kapp, F., Friederichs, P., Brune, S., and Weniger, M.: Spatial verification of high-resolution ensemble precipitation forecasts using local wavelet spectra, Meteorol. Z., 27, 467–480, 2018.

Keil, C. and Craig, G. C.: A displacement and amplitude score employing an optical flow technique, Weather Forecast., 24, 1297–1308, 2009.

Mallat, S.: A wavelet tour of signal processing, 3rd edition, Elsevier, Burlington MA, 294–296, 2009.

Marzban, C. and Sandgathe, S.: Verification with variograms, Weather Forecast., 24, 1102–1120, 2009.

Matheron, G.: Principles of geostatistics, Economic geology, 58, 1246–1266, 1963.

Nason, G.: wavethresh: Wavelets Statistics and Transforms, available at: https://CRAN.R-project.org/package=wavethresh (last access: 31 July 2019), r package version 4.6.8, 2016.

Nason, G. P., Von Sachs, R., and Kroisandt, G.: Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum, J. Roy. Stat. Soc. B, 62, 271–292, 2000.

Radanovics, S., Vidal, J.-P., and Sauquet, E.: Spatial verification of ensemble precipitation: an ensemble version of SAL, Weather Forecast., 33, 1001–1020, 2018.

Roberts, N. M. and Lean, H. W.: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events, Mon. Weather Rev., 136, 78–97, 2008.

Rubner, Y., Tomasi, C., and Guibas, L. J.: The earth mover's distance as a metric for image retrieval, Int. J. Comput. Vision, 40, 99–121, 2000.

Scheuerer, M. and Hamill, T. M.: Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities, Mon. Weather Rev., 143, 1321–1334, 2015.

Schlather, M., Menck, P., Singleton, R., Pfaff, B., and R Core Team: RandomFields: Simulation and Analysis of Random Fields, available at: https://CRAN.R-project.org/package=RandomFields (last access: 31 July 2019), r package version 2.0.66, 2013.

Theis, S., Hense, A., and Damrath, U.: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach, Meteorol. Appl., 12, 257–268, 2005.

Thorarinsdottir, T. L., Gneiting, T., and Gissibl, N.: Using proper divergence functions to evaluate climate models, SIAM/ASA Journal on Uncertainty Quantification, 1, 522–534, 2013.

Vidakovic, B. and Mueller, P.: Wavelets for kids, Instituto de Estadística, Universidad de Duke, 1994.

Villani, C.: Topics in Optimal Transportation, Graduate Studies in Mathematics, Volume 58, American Mathematical Society, Providence, Rhode Island, 1st Edn., p. 75, 2003.

Weniger, M., Kapp, F., and Friederichs, P.: Spatial verification using wavelet transforms: A review, Q. J. Roy. Meteor. Soc., 143, 120–136, 2017.

Wernli, H., Paulat, M., Hagen, M., and Frei, C.: SAL – A novel quality measure for the verification of quantitative precipitation forecasts, Mon. Weather Rev., 136, 4470–4487, 2008.

Appendix B

Buschow and Friederichs 2020

Adv. Stat. Clim. Meteorol. Oceanogr., 6, 13–30, 2020
https://doi.org/10.5194/ascmo-6-13-2020

ASCMO
Open Access

Using wavelets to verify the scale structure of precipitation forecasts

Sebastian Buschow and Petra Friederichs

Institute of Geosciences, University of Bonn, Auf dem Hügel 20, Bonn, Germany

Correspondence: Sebastian Buschow (sebastian.buschow@uni-bonn.de)

Received: 16 September 2019 – Revised: 6 January 2020 – Accepted: 23 January 2020 – Published: 6 March 2020

Abstract. Recently developed verification tools based on local wavelet spectra can isolate errors in the spatial structure of quantitative precipitation forecasts, thereby answering the question of whether the predicted rainfall variability is distributed correctly across a range of spatial scales. This study applies the wavelet-based structure scores to real numerical weather predictions and radar-derived observations for the first time. After tackling important practical concerns such as uncertain boundary conditions and missing data, the behaviour of the scores under realistic conditions is tested in selected case studies and analysed systematically across a large data set. Among the two tested wavelet scores, the approach based on the so-called map of central scales emerges as a particularly convenient and useful tool: summarizing the local spectrum at each pixel by its centre of mass results in a compact and informative visualization of the entire wavelet analysis. The histogram of these scales leads to a structure score which is straightforward to interpret and insensitive to free parameters like wavelet choice and boundary conditions. Its judgement is largely the same as that of the alternative approach (based on the spatial mean wavelet spectrum) and broadly consistent with other, established structural scores.

1 Introduction

The quantitative prediction of precipitation is a central task of modern weather forecasting. A demand for improved predictions of localized severe rainfall events, in particular, has been one of the main drivers behind the development of forecast models with increasingly fine resolutions (Baldauf et al., 2011; Seity et al., 2011), sophisticated parametrizations (Seifert and Beheng, 2006; Kuell and Bott, 2008) and assimilation of novel observation data (Stephan et al., 2008; Bick et al., 2016).

Whether or not the desired improvement has actually been achieved, however, is no trivial question. Since rain fields are inherently intermittent in space and time, a pixel-wise forecast verification can only reward the correct intensity, shape and structure of predicted rain patterns if their locations match exactly with the observed ones. Even a slight displacement between forecast and observation results in a double penalty, because the forecast is wrong in both the observed and the predicted location. The naive, grid-point-wise approach will generally favour coarse models over highly resolved ones and can neither assess the structure or intensity of displaced rain objects nor appropriately judge the severity of displacement errors. Recent years have seen the development of numerous so-called *spatial* verification techniques, which address the double penalty problem in a variety of ways (Gilleland et al., 2009; Dorninger et al., 2018). One strategy espoused by many of these techniques is to split the total forecast error into a number of (ideally orthogonal) components, thereby separating, for example, displacement from other kinds of errors. Following this idea, the present study uses a shift-invariant wavelet transform (Eckley et al., 2010) to isolate a single aspect of forecast performance, namely its structure. Our method, first introduced in Buschow et al. (2019), transforms a map of rain intensities into local wavelet spectra that measure the energy (variance) of the rain field for each combination of location and spatial scale. Under the assumption that auto-correlations vary only slowly in space, the connection between wavelet spectra and the spatial covariance function can be formalized via the theory of locally stationary wavelet processes (Eckley et al., 2010). In order to compare forecast and observation, we can either average the local spectra in space to obtain mean spectra, or calculate the dominant scale at each location and then evaluate the histograms of these central scales. Using a physics-based stochastic rain model (Hewer, 2018) as a controlled test bed, Buschow et al. (2019) have demonstrated that both approaches lead to double-penalty free verification procedures which can detect discrepancies between the observed and predicted correlation structure with great accuracy.

In the present study, we apply the wavelet-based structure scores of Buschow et al. (2019) to real numerical weather forecasts, focusing on the verification of deterministic predictions. Besides addressing some of the practical challenges associated with the non-idealized setting (boundary conditions, missing data, treatment of extremes), one main goal is to study which kinds of errors are typically evaluated by our method. Apart from the consideration of selected case studies, it is therefore instructive to compare the new approach to established alternatives from the rich literature of verification techniques.

Although the standard taxonomy of spatial verification techniques (Dorninger et al., 2018) classifies our method as a scale-separation approach, this class does not actually contain many useful objects of comparison. The most popular approach (Casati et al., 2004, ISS), while also relying on wavelets, studies the *scale of the error*, whereas our method assesses the *error of the scales*. The ISS therefore does not separate structure from displacement and is no direct "competitor" of our approach. Yano and Jakubiak (2016) employ a different type of wavelet transform to locate dominant features in space and scale before explicitly measuring their displacement error. Lastly Kapp et al. (2018), who developed the direct precursor to our method and employ the same wavelet transform, only consider ensemble forecasts and do not separate correlation structure from total variance. For our purposes, it is thus more helpful to group verification methods by the forecast attributes they aim to assess. In this way, we can identify the object-based structure error S of (Wernli et al., 2008) and the variogram-based scoring rules developed by Scheuerer and Hamill (2015) as two comparable pure structure scores.

To obtain robust results on the merits and interrelationship of the object-, variogram- and wavelet-based structure verification, we consider a large set of highly resolved forecasts from the COSMO-DE ensemble prediction system (COSMO-DE-EPS). The hourly adjusted radar product RADOLAN, as well as the regional reanalysis COSMO-REA2 (Wahl et al., 2017), serve as our reference fields. Although we verify each member of COSMO-DE-EPS individually, the ensemble nature of this data set is nonetheless very useful for our purposes. Besides giving us a great number of individual predictions (20 forecasts on 127 selected days), we can exploit the fact that each ensemble prediction consists of 20 realizations from a distribution which changes from case to case to set up idealized experiments: presented with a single member from one of the 127 ensembles, can our scores find the other 19 fields based on their similar correlation structure alone?

The remainder of this paper begins, in Sect. 2, with an overview of all relevant data sets. Section 3 details all steps related to the wavelet transform and its spatial aggregation. To get the first overview of the results of this transform, we analyse the climatology of observed and predicted spectra in Sect. 4. The wavelet-based structure scores of Buschow et al. (2019) are introduced and applied to two selected case studies in Sect. 5. Section 6 reviews the alternative scores from the literature before the verification of the full COSMO-DE-EPS data set in Sect. 7. Here, we study the relationship between all structure scores (Sect. 7.1), assess their discriminatory abilities (Sect. 7.2) and test the sensitivity of our wavelet scores to the free parameters of the method (Sect. 7.3). The paper concludes with a discussion and outlook in Sect. 8.

2 Data

As mentioned in the introduction, this study relies on COSMO-DE-EPS forecasts and COSMO-REA2 reanalysis data (Wahl et al., 2017, henceforth REA2), both of which were previously considered by Kapp et al. (2018). The COSMO-DE ensemble prediction system (Peralta et al., 2012), which has been operational at DWD since May 2012, is based on the non-hydrostatic regional NWP-model COSMO (Baldauf et al., 2011), run at a convection-permitting resolution of 2.8 km in a domain covering Germany and parts of all neighbouring countries (dashed lines in Fig. 1). The 20 ensemble members are generated by combining four boundary conditions with five slightly perturbed physics parametrizations.

The regional reanalysis REA2 is based on a similar version of COSMO, albeit run on a slightly larger domain (white mask in Fig. 1) and at finer resolution of 2 km. As in Kapp et al. (2018), the slight difference in grid is resolved via simple nearest neighbour interpolation to the coarser grid. We have checked that the choice of interpolation scheme has very little impact on the results of our verification procedure. The reanalysis contains information from conventional observations, assimilated in a continuous nudging scheme, as well as radar observations which were included via latent heat nudging. The latter point in particular makes REA2 an attractive validation data set for our purposes since it encompasses direct measurements of rainfall while avoiding systematic discrepancies with the model due to measurement errors or spatial interpolation schemes.

Highly resolved regional reanalyses, while clearly convenient, are not available in most parts of the world and may also contain the same biases as the numerical models verified against them. It is thus of great interest to know whether our methodology can also be applied to direct observational data. In this study, we therefore use DWD's hourly RADOLAN-RW (Winterrath et al., 2018) product as our

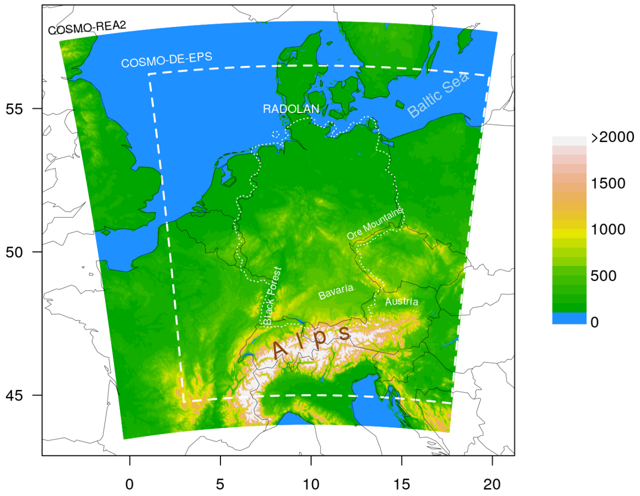


Figure 1. Domain and model orography of COSMO-REA2 in metres. Dashed lines delineate the COSMO-DE-EPS domain, and the dotted line corresponds to the maximum extent of the RADOLAN-RW data set used in this study.

main validation data set. Rain gauge adjusted radar products such as RADOLAN are more widely available and additionally allow us to verify both model and reanalysis against more direct observation data which is completely independent from any dynamical model. Kapp et al. (2018) did not use radar data in order to avoid issues with missing data. This study will explore how big such effects actually are. As for REA2, we bridge the slight difference in nominal resolution (RADOLAN being available at 1 km × 1 km) via nearest neighbour interpolation to the COSMO-DE-EPS grid. Due to the adjustment with rain gauge data, the RADOLAN-RW product is cropped to roughly the German national borders (dotted line in Fig. 1). For the purposes of verification, values outside of the RADOLAN domain, as well as the occasional missing values within, are set to zero. To ensure a fair comparison, the same pixels are set to zero in the forecast and reanalysis fields as well.

Forecasts of hourly rain sums were provided by DWD for the complete year 2011. Since our focus is on an evaluation of the rain field's texture, it stands to reason that the total rain area has to reach some minimum extent since very small rain objects leave us with too few data to confidently estimate the spatial correlations. In this study, we therefore select only cases where at least 5 % of the pixels in the RADOLAN-field have non-zero rain. We furthermore consider only the afternoon hours (16:00–19:00 UTC) in order to ensure comparable lead times. For each day which meets our criteria, we select the hour with the greatest total rain area. This selection procedure leaves us with 127 cases for which the ensemble issues a total of 2540 individual predictions.

In order to roughly classify the 127 case studies according to the processes which generate precipitation, we have manually checked the corresponding DWD analysis maps (freely available from http://www1.wetter3.de/, last access: February 2020) and the registered lightning events (observed by the community project http://www.lightningmaps.org, last access: February 2020). For each day, we note the occurrence of cold fronts, warm fronts, other fronts (quasi-stationary and occlusion fronts), convergence lines and deep moist convection (observed lightning being a proxy for the latter) in the domain. The auxiliary data set is summarized in Fig. 2. We observe that the majority of notable afternoon precipitation episodes in 2011 was associated with lightning (indicating convective processes), often in combination with occlusion or quasi-stationary fronts. The considered time span is furthermore long enough to contain several examples of both purely frontal and purely convective events.

3 Estimation of local wavelet spectra

3.1 Redundant discrete wavelet transforms and local stationarity

Our first objective is to extract the structural properties of observed and predicted fields in a shift-invariant manner. This is achieved by projecting the data, given as a matrix M of dimension $n_x \times n_y$, onto an overcomplete set of basis functions of the form $\psi_{j,d,u}(r) = s_j^{-1/2} \psi_d \left(\frac{r-u}{s_j} \right)$. These so-called *daughter wavelets* are obtained from their mother wavelet $\psi(r)$ via a shift u, scaling s_j and change in orientation, here denoted by the index d. The redundant discrete wavelet transform (RDWT) is defined by scales which are whole powers of two ($s_j = 2^j,\ j \in \{1, 2, \ldots, J\}$), includes three directions ($d = 1$: vertical, $d = 2$: horizontal, $d = 3$: diagonal) and allows shifts to all locations on the grid of the data. The redundancy introduced in this manner ensures that this transformation is shift invariant in the sense that a shift of the input field merely leads to a shift of the coefficient fields. Without this property, the outcome of the verification would depend on the absolute location of rain features within the domain. One basic requirement of the transformation is that the dimensions of M are exactly $n_x = n_y = 2^J$ – we discuss solutions to this boundary problem in some detail in Sect. 3.3.

At this point, we face two natural questions: how are the wavelet coefficients related to the structure of the underlying field, i.e., its spatial covariance matrix, and how should we deal with the great redundancy of the transformed field? Both of these issues can be resolved by assuming that our data are generated by a locally stationary two-dimensional wavelet process (henceforth LS2W). This two-dimensional stochastic process introduced by Eckley et al. (2010) is defined as

$$X(r) = \sum_{j=1}^{J} \sum_{d=1}^{3} \sum_{\text{all } u} W_{j,d,u} \psi_{j,d,u}(r) \xi_{j,d,u}, \qquad (1)$$

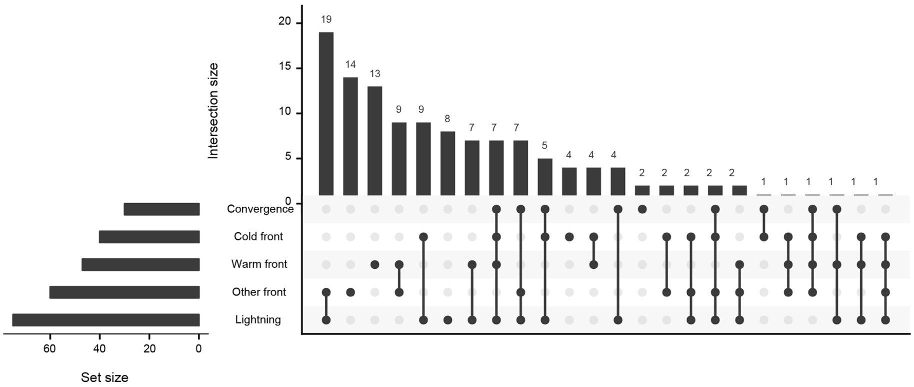


Figure 2. Frequency of weather events and their combinations during the 127 d considered. Data visualized using the UpSetR R package (Conway et al., 2017).

where the $W_{j,d,u}$ represent fixed weights associated with each daughter wavelet and $\xi_{j,d,u}$ is a random white-noise increment. We assume that the spatial covariance of X varies only slowly with r. This requirement of *local stationarity* is weaker than global stationarity and can be formalized as constraints on the regularity of $W_{j,d,u}$ (Eckley et al., 2010). If local stationarity holds, it can be shown that the spatial autocovariances of X in the limit of an infinitely large domain are completely determined by, and can be inferred from, the set of all $|W_{j,d,u}|^2$. Moreover, the squared wavelet coefficient corresponding to $\psi_{j,d,u}(r)$ is a biased estimator of $|W_{j,d,u}|^2$. The bias, which mostly consists of an over-emphasis on the very large scales, can be removed by multiplication with a wavelet-specific matrix A_ψ^{-1}. In analogy to the Fourier spectrum, the $3 \times J$ bias-corrected squared coefficients at each grid point are called the *local wavelet spectrum*. Since any practical application falls outside the realm of asymptotic limits, the bias correction is only approximate, occasionally overshoots its target and introduces negative values to the local spectra. We will set such values, which have no useful interpretation as “energy”, to zero before proceeding with our verification.

The need for a bias correction limits our choice of mother wavelet ψ to the Daubechies family (Daubechies, 1992) for which Eckley et al. (2010) derived the corresponding matrices A_ψ^{-1}. We refer to the compactly supported Daubechies wavelets as D_n. Intuitively, large values of the index $n \in \mathbb{N}$ correspond to smooth functions with good localization in frequency, whereas small n means good localization in space, i.e., a small support size.

The support sizes of the first four Daubechies daughter wavelets are listed in Table 1. A daughter with support size greater than 2^J is no longer unambiguously localized since it “wraps around” the domain more than once (some grid points are sampled multiple times due to the cyclic convolutions of the transform). To avoid this effect, we truncate the local spectra at the largest scale that fits inside the domain. In order to avoid spreading the information from these untrustworthy daughters to the rest of the spectrum (and incidentally spreading information from the uncertain boundaries), scales that are too large are removed prior to bias correction.

For the model given by Eq. (1) to be appropriate, we select the D_n which is most similar to the data using the wavelet selection procedure of Goel and Vidakovic (1995). A few details concerning this step are given in Appendix A. For the present data set, D_2 emerges as the overall winner and is used for the rest of this investigation. Consequently, the largest used scale is $j = 7$ (see Table 1). The three directional versions of D_2 are shown in Fig. 3. Observing their complicated structure, we recognize that the location within the support of $\psi_{j,d,u}$ to which the corresponding spectral value should be assigned is not obvious. As a heuristic solution, we simply select the centre of mass of $\psi^2_{j,d,u}$. Features in the resulting local spectra are thus located close to the corresponding features in the input image.

Concluding this section, we note that our spectrum is not a consistent estimator of $|W_{j,d,u}|^2$ (it has non-vanishing variance in the limit of infinite domain sizes), which necessitates a spatial smoothing of the wavelet coefficients (Eckley et al., 2010). Unless noted otherwise we will omit this step from our present investigation for several reasons: firstly, smoothing introduces a number of additional free parameters which are undesirable for a verification procedure. Secondly, information from the uncertain boundary regions (introduced by expanding the field to $2^J \times 2^J$) is spread across the domain. Lastly, some smoothing algorithms can incur significant additional computational costs. Asymptotic inconsistency is therefore accepted as the cost of a more streamlined verification procedure.

3.2 Logarithmic transformation

Before applying the RDWT to our observed and predicted rain fields, we set all values below 0.1 mm to zero, 0.1 mm being the smallest non-zero value registered by RADOLAN. This step is generally advisable as it removes extremely low-

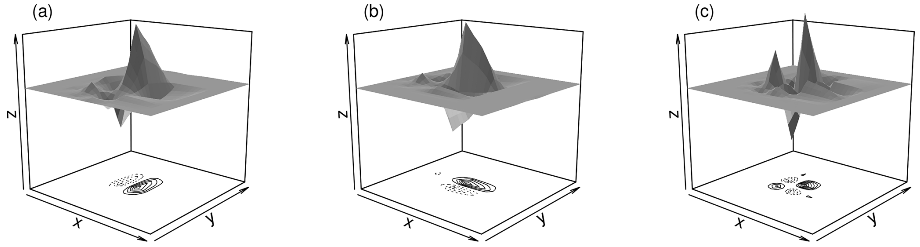


Figure 3. Vertical **(a)**, horizontal **(b)** and diagonal **(c)** daughter wavelet for D_2.

Table 1. Side length of the daughter wavelets' support as a function of the scale j for the first 10 Daubechies wavelets. For each mother wavelet, the star marks the largest daughter wavelet with support size smaller than 2^9.

$j=$	1	2	3	4	5	6	7	8	9	10
D_1	2	4	8	16	32	64	128	256*	512	1024
D_2	4	10	22	46	94	190	382*	766	1534	3070
D_3	6	16	36	76	156	316*	636	1276	2556	5116
D_4	8	22	50	106	218	442*	890	1786	3578	7162

intensity model noise which cannot be interpreted as an actual forecast of precipitation. Next, we replace the original rain fields by their binary logarithm. Casati et al. (2004) argue that this procedure corresponds to an approximate "normalization" of the data. Schleiss et al. (2014), who studied the non-stationary structure of rain fields, concur that this type of variance stabilization facilitates structural analysis.

Thinking visually, the log-transform can be interpreted as a change in colour scale: very few meteorological publications visualize precipitation on a linear scale since it frequently over-emphasizes small, intense showers while rendering the boundary between rain and no rain invisible. In fact, only 5 of the 46 figures depicting rain fields in publications cited in this paper or Buschow et al. (2019) have linear colour scales. The typical step-wise alternatives have many bins near zero and few bins at large values. It is easy to imagine situations where a human assessor will disagree with algorithmically calculated scores if the scores are based on the original data (*linear colour scale*) while the human is looking at transformed data. The conflict is resolved by basing both judgements on the logarithm of the fields: a logarithmic colour scale achieves a similar effect as the step-wise alternatives mentioned above and can easily be used as the input for our algorithm. This step furthermore dampens the potential impact of strongly localized extreme events on our evaluation: without such precaution, a single high-intensity rain object could overshadow the rest of the field, shifting the overall distribution of power to very small scales.

It should be noted that the logarithm introduces one additional free parameter, namely the new value assigned to pixels with zero rain. For this study, it will be set to $\log_2(0.1) \approx -3$, i.e., the logarithm of the smallest considered non-zero intensity. We have checked that moderate changes to this parameter hardly impact the local wavelet spectra.

3.3 Boundary conditions and missing data

Before our wavelet transformation can be applied, the input field needs to undergo a transformation $\mathbb{R}^{n_x \times n_y} \to \mathbb{R}^{2^J \times 2^J}$, which (i) continues the input realistically at the domain edge while (ii) altering the values within the original domain as little as possible. Ideally, this procedure should (iii) be mathematically simple and leave few degrees of freedom. It is furthermore desirable that (iv) the appropriateness of the boundary condition does not depend strongly on the data itself. After the wavelet transform, the original domain is cut out of the fields of wavelet coefficients.

Regarding requirements (ii–iv), the reflective boundary conditions employed by Brune et al. (2018) are a very attractive option: by simply mirroring the domain at each side until the result is larger than $2^J \times 2^J$ and then cutting out the desired square, the fields can be extended to arbitrary dimensions without altering the original data. This transformation is furthermore inexpensive and has no free parameters and the structure outside of the original domain is completely determined by the structure within. We therefore generally recommend the use of reflective boundaries, *as long as the domain boundary is a rectangle*. In the present case, however, the effective domain edge is given by the irregularly shaped RADOLAN region (see Fig. 1), making the mirroring procedure impractical. To ensure a fair comparison of forecasts, reanalysis and observations, we resort to zero boundaries, meaning that all pixels for which no RADOLAN data are available are set to zero.

We note that a large fraction of the RADOLAN-fields used contain further missing data due to failure of individual radars, thus creating even longer and more complicated boundaries. Any rain object which touches these boundaries generates an artificially sharp edge which might, in general, affect the resulting wavelet spectra in unexpected ways. The importance of such effects is tested empirically in Sects. 4 and 7.

3.4 Aggregation of local wavelet spectra

The redundant wavelet transform results in $3 \times J$ spectral values at each grid point. In this study, we will follow Buschow et al. (2019) and average the spectra over the three directions, leaving us with one value per scale (some reasons for discarding the directional information are given in Sect. 8). Before the structure information contained in the local wavelet spectra can be used for analysis and verification, further data reduction is required.

The straightforward approach consists of simply averaging the local spectra over the complete domain. Kapp et al. (2018) first demonstrated that the mean spectra are a solid basis for forecast verification. This strategy generally leaves open which feature in the underlying rain field corresponds to which energy component – the localization potential of the wavelets is under-utilized. Buschow et al. (2019) therefore suggested the *map of central scales* as an alternative aggregation of the local wavelet spectra: instead of averaging in space, each local spectrum is summarized by its centre of mass. The resulting array of z_C has the same dimensions as the original field; the value at each pixel denotes the dominant scale at that location. The authors cited above showed that this form of visualization nicely separates small-scale from large-scale features. The histogram of central scales can replace the spatial mean spectrum as the basis of wavelet-based verification.

We note that the greater their distance to the next rain pixel, the larger the scales on which areas without rain will appear. The addition of a tiny non-zero intensity to such a region can completely alter the local central scales. The spatial mean spectra are naturally insensitive to regions with zero intensity; for the scale histograms we simply remove them from the analysis.

Algorithm 1 Wavelet analysis of rain fields

```
Input: rain field $R$, list of pixels not missing from RADOLAN $L$
Output: mean spectrum, map of central scales $z_C$,
histogram of central scales

1: set values $R < 0.1mm \leftarrow R = 0mm$
2: set $R \notin L \leftarrow 0mm$
3: pad $R$ with zeroes up to $2^9 \times 2^9$
4: set $R \leftarrow \log_2(R' + 0.1mm)$
5: apply $D_2$ transform, select scales 1-7
6: apply bias correction with $\mathbf{A}^{-1}$, set negative spectral values to
   zero
7: average local spectra over the three directions
8: average local spectra over all pixels $\in L$, normalize to unit sum
   $\rightarrow$ mean spectrum
9: get centre of mass for each local spectrum $\rightarrow$
   map of central scales $z_C$
10: get normalized histogram of $z_C$ at pixels $\in L$ with $R > 0 \rightarrow$
    histogram of central scales
```

4 Climatology of wavelet spectra

For a first overview of the spatial structure in our data, we apply the complete wavelet analysis (summarized in Algorithm 1) to each of the 127×22 rain fields. The resulting mean spectra and scale histograms are then averaged over days related to different weather situations (Fig. 4). We observe that purely convective cases, where thunderstorms occurred without direct connection to a frontal structure, are clearly recognized as small in scale, with energy peaking at scale five (panel a) and the most frequent central scale being near four. The reverse situation, i.e., fronts without significant thunderstorm activity, is characterized by a shift of energy towards larger scales (energy concentrated at scale seven, most centres near scale six). The forecast ensemble and REA2 agree closely on this regime behaviour; the relatively tight spread encompasses the observed spectra in nearly all cases. The fact that almost no variability resides on scales 1 and 2 is hardly surprising since the effective resolution of the COSMO model, below which all processes are unrealistically damped, is at 4 to 5 grid boxes (Bierdel et al., 2012).

For the purely frontal cases, as well as the overall climatology, precipitation in RADOLAN lives on systematically smaller scales than in the two model-based data sets, with histograms shifted by about 0.5, reduced energy at scale seven and increased energy below scale 5. Interestingly, this discrepancy is not evident for the purely convective cases where the curves corresponding to RADOLAN are even closer to the centre of the ensemble range than REA2.

To assess the impact of the imperfect, padded boundary conditions on the climatology of these wavelet spectra, we have repeated the analysis for REA2 without setting pixels missing from RADOLAN to zero (neglecting the second step of Algorithm 1). As one might expect due to the possibility for overall larger features, the resulting curves (dotted lines in Fig. 4) are slightly shifted toward large scales. The effect is, however, small compared to both the spread of the ensemble and the difference between ensemble mean, RADOLAN and REA2.

Besides the climatologies of the spatially aggregated wavelet spectra, we are also interested in their average distribution across the domain. The map of central scales allows us to investigate this behaviour in a straightforward manner by simply averaging the locally dominant scales at each pixel over all instances with rain. To ensure that the results are reasonably robust, we only consider grid points with at least three full weeks of non-zero data.

The resulting pattern of average central scales for the reanalysis is shown in Fig. 5a. For this calculation no RADOLAN mask was applied, thus enabling us to study the variability across the complete COSMO-DE domain. We observe that the distribution of predominantly small and large scales is closely tied to the orography: the Alps, Ore Mountains, Black Forest and central German highlands are all as-

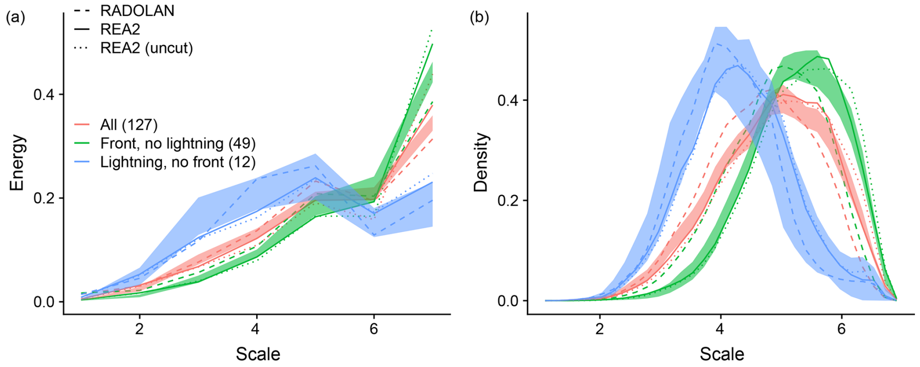


Figure 4. Normalized spatial mean spectra **(a)** and histograms of central scales **(b)**, averaged over cases with fronts and no convection (green), convection and no fronts (blue), and all cases (red). Areas indicate the range of these mean curves over the 20 ensemble members. Solid and dashed lines correspond to REA2 and RADOLAN, respectively. The dotted line represents the REA2 spectra obtained without masking the fields with the available RADOLAN data.

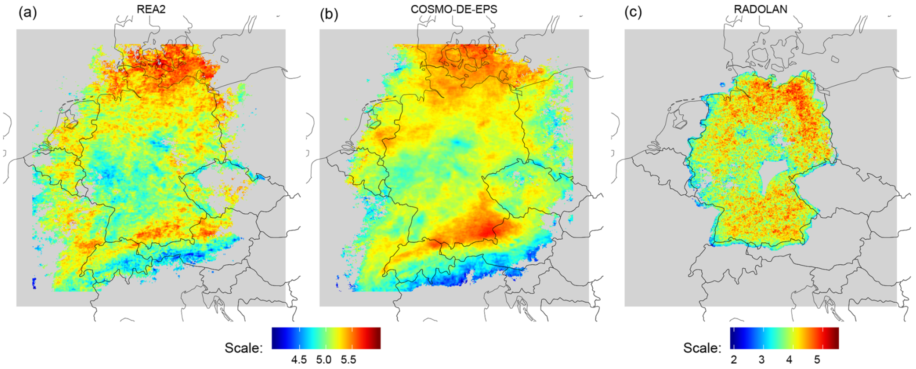


Figure 5. Map of central scales, averaged over all instants with non-zero precipitation for COSMO-REA2 **(a)**, COSMO-DE-EPS (**b**, averaged over all 20 members) and RADOLAN (**c**, individual colour bar). Pixels with fewer than 21 d with precipitation were discarded. The RADOLAN mask was *not* applied to REA2 and COSMO-DE-EPS.

sociated with decreased central scales. The Baltic Sea, northern German flatlands and Alpine foothills in Bavaria and Austria, on the other hand, tend to experience larger precipitation features.

The corresponding climatological map for the forecasts, averaged here over all ensemble members, is very similar to the reanalysis albeit with slightly larger scales in the southern half of the domain. The picture for RADOLAN, on the other hand, looks completely different (Fig. 5c; note the separate colour scale). Most notably, the overall scales are decreased by roughly 1. Due to the limited area – both the Alps and the Baltic sea are outside the domain – and sharp edges caused by missing data, very little of the structure described above can be recognized.

For a direct and fair comparison of models and observation, we repeat the calculation of the climatological maps of central scales for REA2 and COSMO-DE-EPS, this time including only pixels for which RADOLAN data are not missing. Noting furthermore that the differences in scale vary mainly in the meridional direction, we average these maps over all longitudes; the results are shown in Fig. 6. In this visualization, we find that the overall pattern of larger scales in southern and northern Germany and smaller scales near the centre is present in all three data sets after all. The RADOLAN profile is qualitatively similar to the others, but shifted down by nearly one scale.

Figure 6 furthermore allows us to assess the differences between groups of ensemble members. Anticipating the results, we have coloured ensemble members according to their physics setting. We find that members with the first physics setting, i.e., an increased entrainment rate (Theis et al., 2014), produce more small-scale variability than the others. Conversely, members with the fifth parameter setting, i.e., increased turbulent length scale, favour large-scale variability. No clear-cut pattern emerges when we sort the ensemble members by their boundary condition (not shown).

Throughout northern and central Germany, the reanalysis lies near the centre of the ensemble spread. In the South, however, all ensemble members produce systematically larger features than REA2. Since the slight discrepancy in internal resolution is constant across the domain, this discrepancy is likely the result of continuous data assimilation.

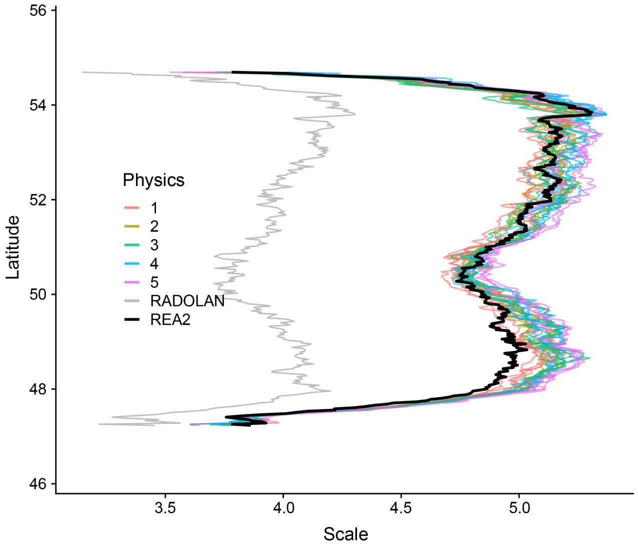


Figure 6. Map of central scales, averaged over all instants with non-zero precipitation and all longitudes. Ensemble members with the same physics setting have the same colour, and RADOLAN and REA2 are black and grey, respectively. Only pixels with available RADOLAN observations and at least 21 d of non-zero rain were included.

5 Wavelet-based scores

5.1 Scores based on the mean spectra and scale histograms

Following Buschow et al. (2019), we compare the scale histograms of two rain fields, i.e., forecast and observation, via the earth mover's distance (henceforth EMD): the count in each histogram bin constitutes a pile of earth located at the bin's centre. The EMD is given by the minimum work (dirt moved times distance travelled) required to transport the predicted arrangement of piles into the observed one. We prefer this type of comparison over an element-wise difference because it treats shifts between neighbouring scales appropriately: a displacement from one bin to the next increases the total work and thus the EMD only slightly. A discrepancy by several scales, which would lead to the same element-wise difference between the histograms, is punished more strongly. For further details about the merits of the EMD, the reader is referred to Rubner et al. (2000). The EMD between the two scale histograms (henceforth HEMD) constitutes our first wavelet-based score.

The second score, SEMD is analogously given by the EMD between the two normalized and spatially and directionally averaged spectra. Here, the locations of the dirt piles are given by the scales $j \in \{1, \ldots, J\}$, the spectral energy corresponds to the amount of dirt. The normalization of the spectra eliminates differences in total intensity and guarantees that the EMD is a true metric, meaning that only perfectly predicted spectra achieve perfect scores.

As mentioned in Buschow et al. (2019), we can obtain a sign associated with the EMD by calculating the distance between the centres of the two curves, i.e., the difference in expectation value for HEMD and the difference in central scale for SEMD. When desired, the sign of these differences can be attached to SEMD and HEMD in order to assess the directions of the forecast errors (too large or too small).

5.2 Case study: 19 June 2011

To get a first impression of the kinds of errors which determine the outcome of our wavelet-based verification, we consider a case study for which the quality of the ensemble members was deemed below average by both of our scores. On 19 June 2011, a secondary depression near the end of its life cycle made landfall on the German North Sea coast and traversed northern Germany during the afternoon hours. Between 15:00 and 16:00 UTC, RADOLAN observed a large-scale rain band near the cyclone's centre in eastern Germany and a large number of smaller, relatively intense, features across the rest of the domain (Fig. 7a). The forecast considered in the example (member five, Fig. 7c) features a single, substantially rounder, larger and smoother field in the east and only a few scattered objects with very low intensity besides. This discrepancy is clearly reflected by a surplus of large-scale variability in both the mean spectra (panel b) and the scale histograms (panel e). The resulting earth mover's distances amount to approximately one full scale in both cases. Here, we have visualized the corresponding transports as river plots (coloured lines between the histograms). Considering the maps of central scales (panels d and f), we find that the features in the images are classified just as expected with the large rain band living near scale 5 in RADOLAN and scale 6 in the forecast, while the smaller features lie closer to scales 3 and 4.

5.3 Case study: 26 February 2011

Our second case study similarly features a depression crossing northern Germany. In contrast to the previous example, the dominant weather phenomena are associated not with the cyclone itself, but with its frontal system enclosing a very narrow warm sector which crosses western Germany during the afternoon of 26 February 2011 (Fig. 8). The resulting rain field, as observed by RADOLAN (Fig. 9), consists of two narrow rain bands, one with medium intensity associated with the cold front in the west and one with very low intensities related to the warm front in the east. Neither the reanalysis nor ensemble member 6 exhibit a separation between the precipitation fields of the two fronts, both showing a single broad rain field across south-western Germany instead. Member 1, on the other hand, produces two narrow rain bands, albeit with increased width and length as well as slightly wrong locations compared to RADOLAN.

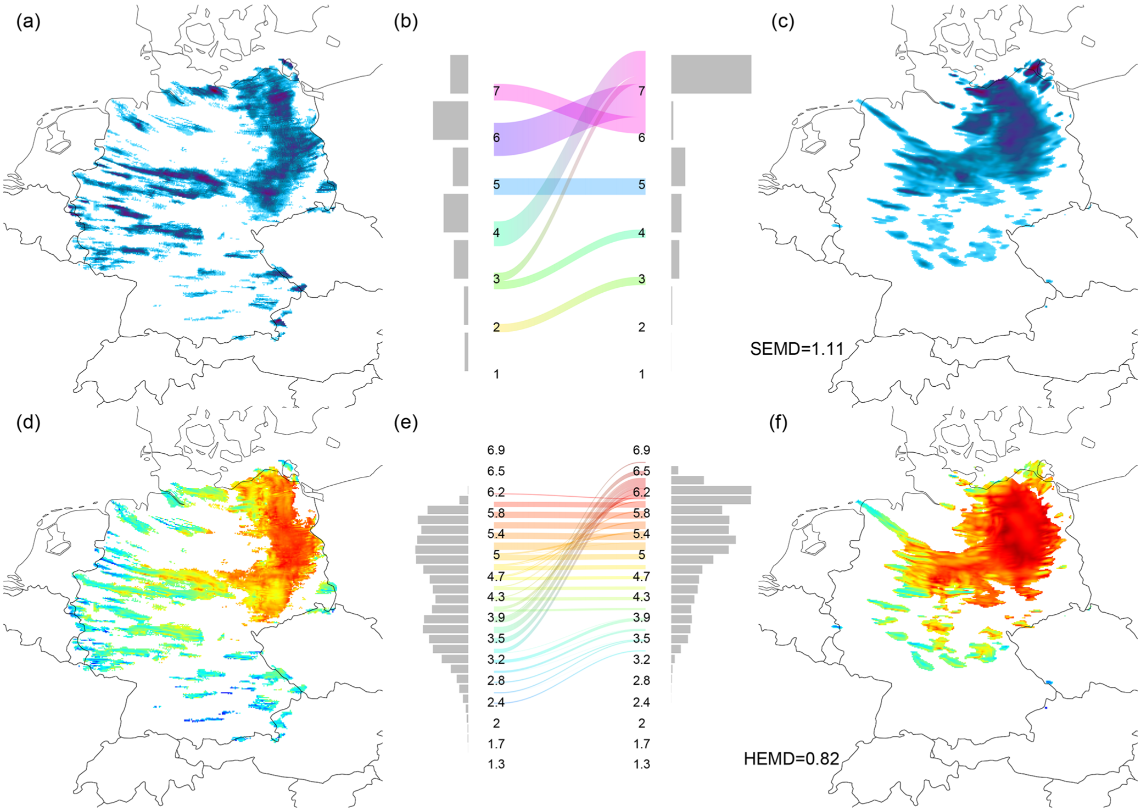


Figure 7. Wavelet-based verification for 19 June 2011 at 16:00 UTC: observed field (RADOLAN, **a**); observed spectrum, EMD components and forecast spectrum (**b**); forecast field (Member 5, **c**). Bottom row: observed map of central scales (**d**); corresponding histogram, EMD components, forecast scale histogram (**e**); forecast map of scales (**f**).

In terms of the overall structure, the first ensemble member is arguably superior to member 6 and REA2. A point-wise verification measure like the root mean square error does not reward the correctly simulated separation into two rain bands. The map of central scales (bottom row of Fig. 9), on the other hand, adequately registers two disjoint rain bands as smaller than the unified pattern. Consequently, member 1 receives a substantially better score (HEMD ≈ 0.5) than member 6 or REA (both close to HEMD $= 1$).

6 Non-wavelet scores

To investigate which properties of a forecast are punished or rewarded by our wavelet-based verification, one natural approach is to compare the scores presented above to alternative verification methods which also focus on the field's structure.

Our first candidate is the structure component of SAL (Wernli et al., 2008, S). For the calculation of S, which is implemented in the SpatialVx R package (Gilleland, 2018), observed and predicted rain field are decomposed into discrete objects. Here, we use the standard algorithm of the R package, which first smooths the data with a simple disc kernel, then discards all pixels below a given threshold $R_{\min}$ and groups continuous regions of non-zero pixels into separate objects. For each object (i), the ratio between total and maximal precipitation is calculated as

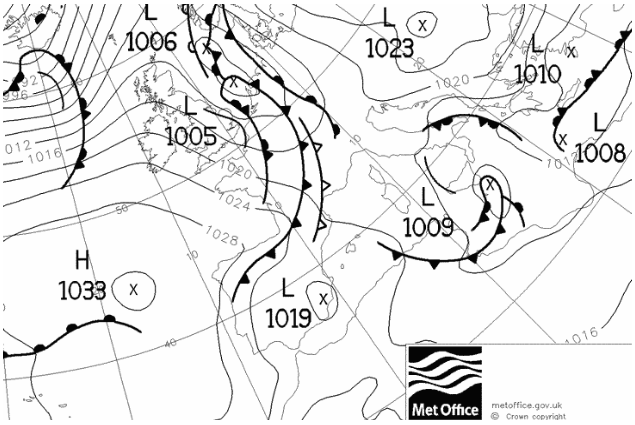


Figure 8. UK Met Office surface pressure chart for 26 February 2011 18:00 UTC (cropped). Contains public sector information licensed under the Open Government Licence v1.0.

$$V_{(i)} = R_{\mathrm{tot},(i)} / R_{\max,(i)}\,, \tag{2}$$

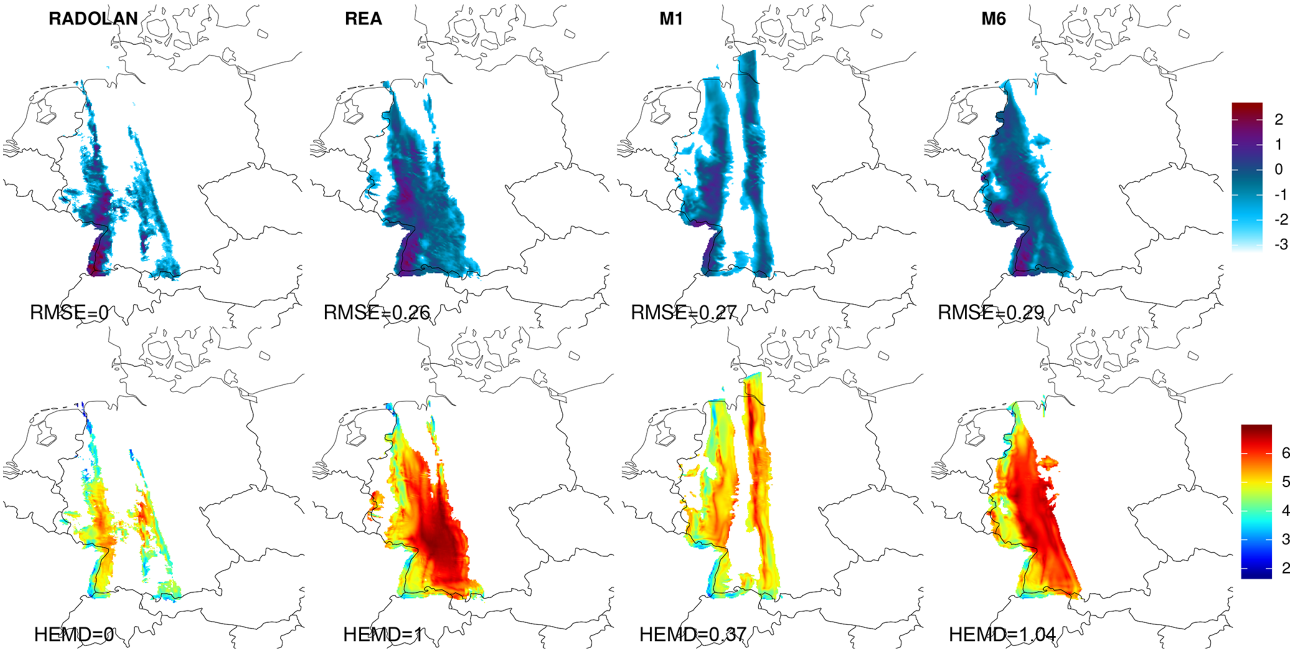


Figure 9. Logarithmic rain fields for 26 February 2011 at 19:00 UTC (top row) and corresponding maps of central scales (bottom). From left to right: RADOLAN, REA and COSMO-DE-EPS ensemble members 1 and 6. All fields were cropped to the extent of the available RADOLAN data.

where $R_{\mathrm{tot},(i)}$ and $R_{\mathrm{max},(i)}$ refer to the total and maximum intensity of the object, respectively. This "peakedness" is averaged over all objects in both fields separately, weighted by $R_{\mathrm{tot},(i)}$. S is then given by the relative difference in (weighted) mean peakedness of forecast and observation. The sign is chosen such that $S > 0$ indicates forecasts with features that are not peaked enough, i.e., too large and/or too flat.

The key parameter of this procedure is the threshold R_{min}, which can, depending on the data, have a strong impact on the outcome of the verification (Weniger and Friederichs, 2016). Radanovics et al. (2018) point out that such effects can be minimized as long as thresholds below the respective minimum positive values of the fields are avoided. This property is met by choosing individual thresholds for forecast and observation, truncating each field at $1/15$ of the 95 %-quantile of non-zero values. This approach greatly decreases the computational cost of the procedure since the object decomposition has to be repeated only once per field, not once per combination of observation and forecast. We have checked that the results hardly differ from those obtained with a common threshold.

Our second object of comparison is the weighted p-variogram score of Scheuerer and Hamill (2015). Originally designed for ensemble verification of multivariate quantities, Buschow et al. (2019) adapted this score to a deterministic setting. Assuming stationarity of the data, the score simplifies to the mean squared difference between observed and predicted empirical p variogram, weighted by the inverse distance d^{-1} between pairs of points, i.e.,

$$\mathrm{VGS} = \sum_{\text{all } 0<d<d_{\mathrm{max}}} d^{-1} \left(\sum_{|r_i - r_j| = d} \left| R_{\mathrm{obs}}(r_i) - R_{\mathrm{obs}}(r_j) \right|^p - \sum_{|r_i - r_j| = d} \left| R_{\mathrm{for}}(r_i) - R_{\mathrm{for}}(r_j) \right|^p \right)^2 . \quad (3)$$

Here, $R_{\mathrm{obs/for}}(r_i)$ denotes the observed or predicted rain value at a given location r_i. In contrast to SEMD, HEMD and S, scores of this form depend explicitly on the variance of the two fields: for $p = 2$, i.e., the classic variogram, the expected squared differences between distant points converges exactly to the variance; changes in this parameter shift the curves up and down. Since we wish to isolate structure from intensity errors, we set $p = 2$ and standardize all fields to unit variance before calculating VGS. This guarantees that all curves converge to the same value; their remaining differences are due to discrepancies in correlation structure. Noting that the inverse distance weighting limits the impact of very distant pairs, we set $d_{\mathrm{max}} = 50\,px$.

In order to check how strongly VGS and the other supposed structure scores depend on intensity errors, we include SAL's amplitude component A, given as the relative difference in total rain, in our experiments as well. All wavelet and non-wavelet scores used in this study are listed in Table 2, the optimal score in each case is zero. The wavelet and variogram transformations are applied to the logarithmic rain fields for the reasons detailed in Sect. 3.2. This transformation is not appropriate for S and A because the resulting negative values lead to unexpected behaviour of the score definitions. These scores are therefore based on the untransformed rain fields for which they were originally developed.

7 Verification of COSMO-DE-EPS in 2011

To study the behaviour of our structure verification in aggregate, we apply the wavelet analysis of Algorithm 1 to all 127×22 fields in our data set to obtain the mean spectra and scale histograms on which SEMD and HEMD are based. Similarly, we calculate the total precipitation (basis for A), the average structure function V (Eq. 2, basis for S) and the weighted stationary variogram (basis for VGS). Every field is then compared to every other field, giving us approximately four million realizations of each score listed in Table 2. Different subsets of this large data set are then used to address the following questions:

1. How are these scores related to each other?

2. Can the structure scores discriminate good forecasts from bad ones?

3. How sensitive are the wavelet scores to the choice of mother wavelet, the log-transform, the boundary conditions and the choice of reference data?

The following sections address each of these questions in turn.

7.1 Comparison between scores

For a first overview of the verification results, we consider the distributions of all scores (absolute values) for the 20 forecasts issued on each of the 127 d, verified against RADOLAN. In Fig. 10, we have first separated the resulting distributions by weather situation: days where precipitation was generated by *a single type of weather phenomenon* (warm front, cold front etc.) are shown in individual box plots, and all other days are grouped into the class “multiple”.

It appears that, at least qualitatively, HEMD, SEMD and VGS are in fair agreement: purely convective days and pure cold fronts (of which our data set contains eight and four cases, respectively; see Fig. 2) were forecast best (lowest scores), followed by warm fronts and other fronts. S agrees in the convective cases, but sees no clear differences between the front types. The two pure convergence-line cases received the unanimously worst scores, but the small sample size prohibits any general conclusions from this observation. The amplitude score A, which does not measure structural properties, shows no great variation across weather situation, the only exception being the four cold-front cases, the total amplitude of which was predicted unusually well.

To quantify how close the agreement between the different scores actually is, we calculate their correlation matrix, shown in Fig. 11a. Unsurprisingly, the strongest connection is found between the two wavelet scores (0.85), both of which also have a notable connection to the variogram score (0.64 and 0.68). The object-based S is slightly less similar to the other structure scores and shows the closest relationship with the amplitude error A. SEMD, HEMD and VGS, on the other hand, are only weakly linked to A.

To get a broader overview of these interrelations in cases where forecast and observation may be very dissimilar, we have also calculated the same correlations over all possible pairs of forecast and observation date (Fig. 11b). Across this data set, which includes some exceedingly bad predictions, the similarity between all four structure scores increases slightly, and SEMD and HEMD become nearly identical. The connection to the amplitude error A mostly vanishes.

In the next step, we include the sign of S and endow SEMD and HEMD with the signs of the corresponding centre differences as described in Sect. 5.1. These scores now measure not only the severity of the structural error, but also the direction, i.e., too small or too large. In accordance with the classic SAL definition, the signs are chosen such that positive values indicate a forecast with too much large-scale variability. The joint distributions of the three signed scores are shown in Fig. 12. Here, we have again included all 127×127 combinations of days in order to probe a broad range of good and bad forecasts. HEMD and SEMD agree on the sign of the error in 93 % of cases, and the sign of S matches roughly 85 % of the time. As a result, the correlations rise to $\mathrm{cor}(S, \mathrm{HEMD}) = 0.87$ and $\mathrm{cor}(S, \mathrm{SEMD}) = 0.85$, respectively. The bivariate histograms furthermore show that extreme disagreements, which would appear in the upper left and lower right quadrants of the histograms, are rare. The functional relationship of these scores follow a sigmoid-type function.

7.2 Discrimination

The previous section has shown that structure scores based on wavelets, variograms and object properties pass similar, but by no means identical judgement of forecast quality. A natural question is which (if any) of these assessments is correct in the sense that the best forecast receives the best score. In a realistic setting, this question cannot be answered because the objectively best forecast is unknown. As a surrogate, we can consider the ensemble forecast issued for each day as the “correct” prediction and compare it to the 126 forecasts issued for the other days: if the prediction system were perfect and weather patterns never repeated, a sharp verification tool should give the best scores to matching days.

The leftmost bars in Fig. 13 show the median rank of those supposedly best forecasts, verified against RADOLAN. Since there are 20 forecasts per day, the ideal rank is 10. Although such perfect scores are not observed, matching days

Table 2. All scores used in Sect. 7. $J_{\max}$ and $J_{\min}$ refer to the largest and smallest considered scale of the wavelet decomposition. In this study, $J_{\max} - J_{\min} = 7 - 1 = 6$. The optimal value of each score is zero.

Abb.	Description	Range	Signed	log(rain)
HEMD	EMD between histograms of central scale	$[0, J_{\max} - J_{\min}]$	(yes)	yes
SEMD	EMD between dir. averaged mean spectra	$[0, J_{\max} - J_{\min}]$	(yes)	yes
VGS	Weighted stationary variogram score, $p = 2$	$[0, \infty)$	no	yes
S	Relative difference in average feature "peakedness"	$[-2, 2]$	yes	no
A	Relative difference in total rain intensity	$[-2, 2]$	yes	no

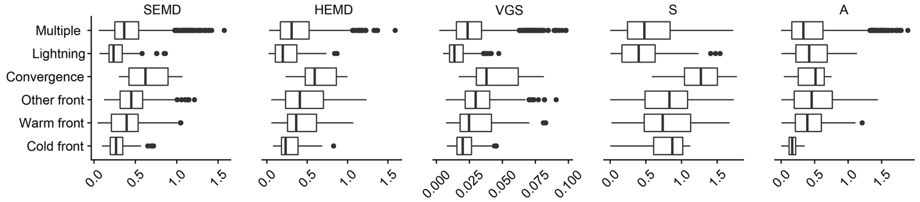


Figure 10. Distribution of absolute values for all scores (matching forecast and observation dates), separated by weather event.

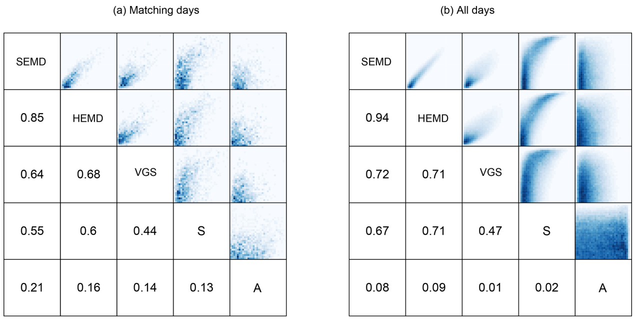


Figure 11. Lower triangle: correlations between the absolute values of all scores, calculated over **(a)** the 20×127 pairs belonging to matching days and **(b)** all $20 \times 127 \times 127$ combinations of forecast and observation. The upper triangles show bi-variate histograms for all combinations of scores.

are nonetheless typically among the 25 % best forecasts, with SEMD issuing the lowest median rank and S the highest. When we use REA2 as the reference instead of RADOLAN, the ranks of all scores improve by about 100 – all structure scores clearly indicate that the COSMO-DE-EPS predictions are structurally more similar to the reanalysis than the observations.

To focus on the discriminatory abilities of our scores, we can take the quality of the predictions out of the equation by selecting a member of the forecast ensemble as the "observation" against which all other forecasts are verified. Ideally, the 20 ensemble members constitute independent realizations from a single distribution which changes from day to day. When forecast and observation share neither physics setting nor boundary conditions (centre of Fig. 13), the rankings for matching days improve with respect to all four scores. In a perfect world, the matching forecasts would rank at number six (since there are 12 unrelated ensemble members). In reality, the ranks are between 326 for VGS and 424 for S. Switching from an unrelated member to an "observation" which shares the forecast's physics settings (of which there are four, making the perfect rank two) only marginally lowers the ranks.

As a final experiment, we select an observation which has the same boundary conditions as the prediction. Visual inspection of example forecast ensembles shows that these members are often extremely similar to one another. As a result, SEMD, HEMD and VGS consider only a handful of other predictions superior to those that share both the boundaries and the date of the observation (rightmost bars in Fig. 13). S, on the other hand, still prefers over 160 other forecasts over the "correct" ones, indicating weaker discriminatory ability.

7.3 Sensitivity of the wavelet scores

Concluding this statistical analysis of our wavelet-based scores, we consider their sensitivity to the free parameters of the method. To this end, the complete verification procedure is repeated three times: once with the Haar wavelet instead of D_2, once without the logarithmic transformation and once without setting pixels missing from RADOLAN to zero. The resulting joint distributions of original and altered scores are shown in Fig. 14. Here, we have again included all pairs of observation and forecast days in the bi-variate histograms (colours).

Recalling the outcome of the wavelet selection (Sect. A), as well as the results reported in Buschow et al. (2019), we expect the impact of the chosen mother wavelet to be weak. Figure 14a clearly confirms this expectation: SEMD experiences only minor changes, and the scores remain correlated

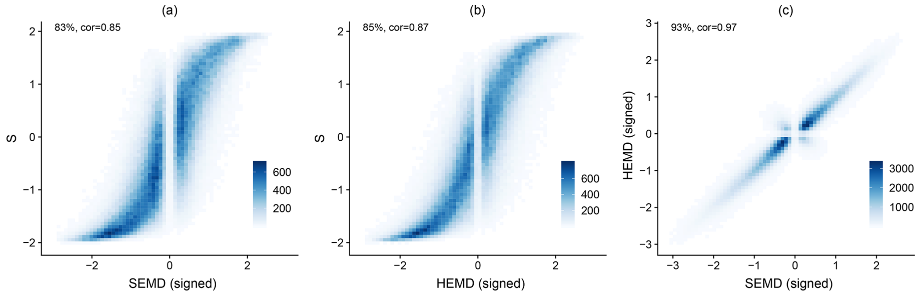


Figure 12. Bi-variate histograms of SEMD and S **(a)**, HEMD and S **(b)**, and SEMD and HEMD **(c)**. The two wavelet scores have been endowed with the sign of the corresponding difference in centre. Percentages indicate the fraction of cases where the two scores have the same sign; cor denotes the correlation.

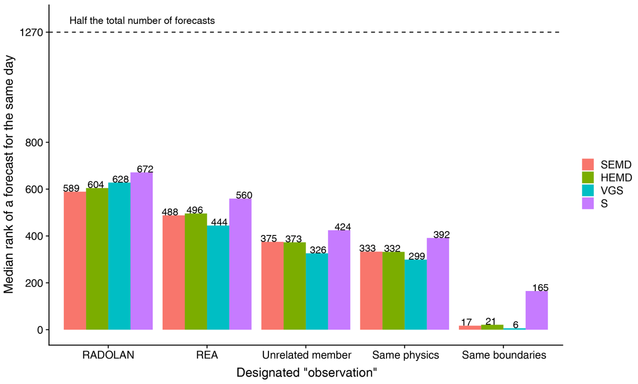


Figure 13. Median rank of the score obtained by the 20 ensemble members belonging to the same day as the observation among the set of all 2540 forecasts. From left to right, the designated "observations" are RADOLAN, REA2, an ensemble member which shares neither boundary conditions nor physics settings with the forecast, an ensemble member which shares the physics settings, and an ensemble member which shares the boundary conditions.

at 0.96; HEMD is even less sensitive (cor $= 0.98$). We furthermore observe no outliers, indicating that the verdict never changes abruptly as a result of switching from one wavelet to another.

Based on the discussion in Sect. 3.2, we expect the logarithmic transform to have a greater influence on the result of the verification. For SEMD, our expectation is confirmed (cor ≤ 0.85, wide distribution), and HEMD is notably less affected by the change in "colour scale".

The experiment without the RADOLAN mask (panel c) constitutes an ideal test for the impact of the wavelet-transform's boundary conditions: originally all values beyond the long and complicated edge of the available RADOLAN data were simply set to zero; now we replace them with the actually available model output, i.e., perfect boundary conditions. The resulting difference in scores is comparable in magnitude to that of the logarithmic transform, but the distribution is different. While the overall correlation over all cases is high, the range of occurring differences is broader, meaning that individual fields with prominent features near or beyond the border can experience a strong shift in the verification result. HEMD is again less sensitive than SEMD and produces fewer outliers.

In a final step, we consider the impact of the chosen validation data (Fig. 14d). As one might expect based on the results of previous sections, the change from RADOLAN to REA2 as "observation" can result in completely different verification results, the sensitivity of both scores being similar in this instance.

All correlations discussed so far decrease monotonically when only matching pairs of forecast and observation date, i.e., reasonably good forecasts, are considered (black dots in Fig. 14). The qualitative results remain unchanged; HEMD is the less sensitive score and the mother wavelet has the least impact, while logarithm and boundary condition are more important. The strongest decrease in correlations occurs for the choice of validation data, meaning that, in our data set, the ranking of individual forecasts for matching days changes almost completely depending on the chosen observations. We note, however, that none of the effects discussed in this section has a strong systematic component – the expected scores (white dots in Fig. 14) are nearly unchanged in all four sensitivity experiments.

8 Summary and discussion

This study has applied the wavelet-based pure structure verification of Buschow et al. (2019) to the systematic evaluation of numerical weather predictions against radar observations, as well as a regional reanalysis.

In the first step, we have studied the climatological properties of the local wavelet spectra. Similar analyses of the predicted average spatial structure were carried out by Willeit et al. (2015) and Wong and Skamarock (2016) using Fourier transforms. Aggregation of these mean spectral properties according to the weather situation has confirmed the findings of Brune et al. (2018), who report that wavelet spectra are very well suited to differentiate between rain fields with different degrees of spatial organization. We furthermore find

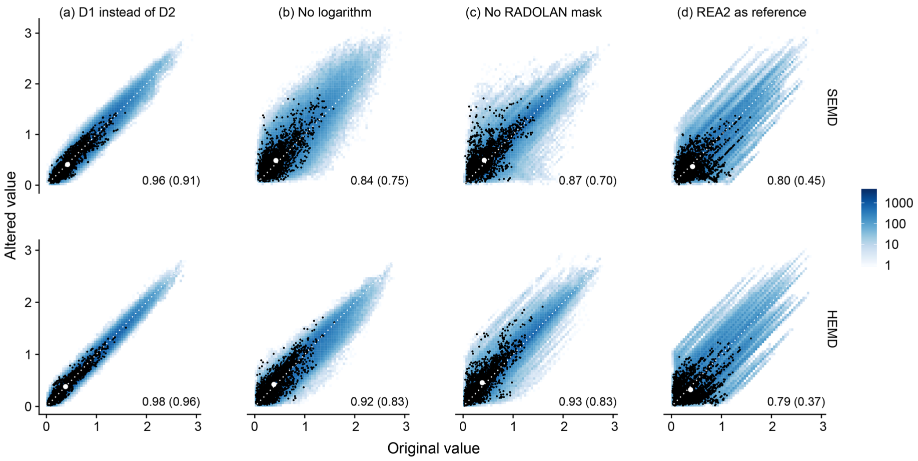


Figure 14. Bivariate histograms of the original wavelet-based scores (on the x axis) against their altered versions (y axis), including all combinations of forecast and observation date. For **(a)**–**(c)**, RADOLAN is the reference; **(d)** compares scores against RADOLAN to scores against REA2. Numbers indicate the correlation over all scores, and the number in brackets is the correlation obtained for matching days only (marked by black dots). The white dot represents the mean original and altered values for matching days.

that forecasts and reanalysis, which are based on similar configurations of the same NWP model, have very nearly the same average structure. RADOLAN, on the other hand, is systematically shifted towards smaller scales in most situations. For purely convective rain fields, however, the forecast ensemble is more similar to RADOLAN than to REA2. The latter observation indicates that the discrepancy in scale is not exclusively due to the slight difference in native resolution (1 km for RADOLAN, 2 km for REA2 and 2.8 km for COSMO-DE-EPS) since the grid spacing also differs between forecast and reanalysis and does not depend on the weather situation. By masking the forecasts with the available radar measurements, missing data have been ruled out as a possible explanation as well. We therefore conclude that, irrespective of boundary conditions, physics settings and data-assimilation scheme, the COSMO model tends to produce frontal and other large-scale precipitation patterns which are too large and too smooth.

An evaluation of the temporal mean map of central scales has shown that the discrepancy is mostly constant in space. This step furthermore revealed that the variation in average structure across the ensemble is mostly determined by the physics parametrization. A systematic discrepancy between predictions and reanalysis was furthermore detected over southern Germany. Since the difference in model resolution is constant in space, this observation indicates that the model has an internal tendency to under-represent small-scale variability in this region. Overall this type of climatological analysis has proven to be a useful first evaluation of the average model performance. The natural possibility to localize errors in space constitutes an advantage over the Fourier approach of Willeit et al. (2015) and Wong and Skamarock (2016).

Our second set of results concerns the typical behaviour of the two wavelet-based structure scores SEMD and HEMD. Buschow et al. (2019) report that these scores, as well as the object-based S and the variogram score VGS, can discriminate between good and bad predictions of spatial structure in a controlled environment. Exploiting the fact that each individual forecast ensemble essentially contains 20 draws from an ever-changing probability distribution, we have demonstrated that many of the results previously obtained with synthetic rain fields can be transferred to the real world: all four scores are reasonably good at distinguishing matching forecasts from non-matching ones, S being the worst at this exercise and VGS marginally better than the two wavelet alternatives. Interpreting this experiment, is important to realize that discrimination is not the only desirable property for the scores under consideration, since we also wish to isolate information on the field's structure from all other kinds of errors.

To learn more about the kinds of forecast errors punished by our structure scores, we have considered two selected case studies. Here, HEMD was found to be particularly easy to interpret since we can plot the map of central scales on which it is based. In this manner we found that the score can, for example, reward the correctly predicted split precipitation field in a nearly but not completely occluded frontal system, or punish the lack of small-scale rain features surrounding a secondary depression.

A statistical analysis across the complete data set revealed that, in realistic forecast situations, HEMD and SEMD are usually in very close agreement with each other. The wavelets furthermore typically find the same sign of the error as the object-based S. The moderate correlation between S and the wavelet scores is likely due to low-intensity areas which are removed during the object identification procedure required for SAL, but may have a big impact on the average wavelet spectra. The variogram-based VGS is, on average, more similar to the wavelets. Here, the remaining differences are probably related to the fact that the incarnation of VGS, recommended by Scheuerer and Hamill (2015) and employed in this study, down-weights long-distance correlations while the wavelet spectra treat all scales equitably. It is worth noting that the overall performance of the variogram score is surprisingly good, despite the questionable assumption of spatial stationarity.

Based on the discussion above, we can overall recommend HEMD as a useful tool for purely structural verification of quantitative precipitation forecasts. Its verdict is very similar to that of SEMD, but less sensitive to the choice of the mother wavelet and boundary conditions, and easier to interpret thanks to the underlying map of central scales. We have demonstrated that our score can provide useful additional information on a very specific aspect of forecast performance and should be used in conjunction with other techniques which isolate errors in feature location, intensity and total area.

Another property, which has so far been left out of the analysis, is the orientation and anisotropy of the rain fields. Since several important weather phenomena such as fronts and squall lines have very characteristic anisotropic shapes, these are clearly relevant aspects of forecast quality to which all scores tested in this study are insensitive. We have intentionally removed the directional information from our wavelet spectra because the underlying transformation is invariant under shifts, but not under rotations. Consequently, the perceived degree of anisotropy, as well as the difference in the orientation of two fields, depends on the orientation itself – one could rotate observation and forecast simultaneously in the exact same way and receive a changed verification result. To avoid this problem, future studies will explore the use of different wavelet transforms which have the necessary redundancy in both location and orientation. A second important direction for future research is the application to the problem of wind verification, which faces many of the same issues as precipitation and has recently received much attention in the spatial verification community (Dorninger et al., 2018).

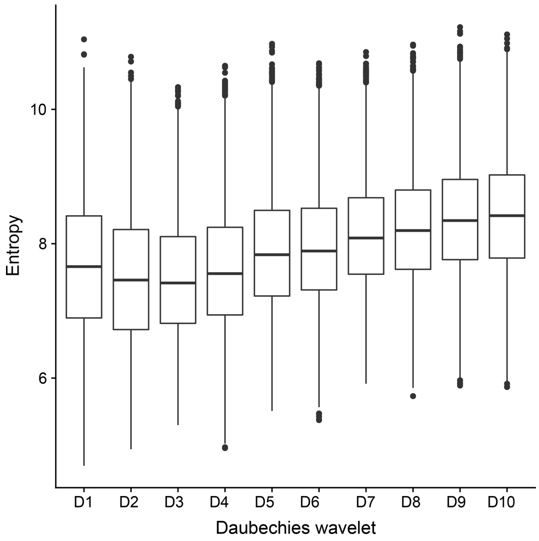


Figure A1. Entropy of the transforms for the first 10 Daubechies wavelets (specifically the "extremal phase" versions). Points denote the median, lines the interquartile range over all forecasts and observations from our data set.

Appendix A: Wavelet selection

In order to objectively select the most appropriate mother wavelet, we follow Goel and Vidakovic (1995), who demonstrate that the similarity between data and basis function can be optimized by minimizing the entropy of the mother wavelet's corresponding orthogonal transform. In a nutshell, wavelets with many vanishing moments and large support areas are good at representing smooth internal structures while shorter wavelets can handle discontinuities better. For a more detailed discussion of this approach and its appropriateness to our application, we refer to Buschow et al. (2019). Applying the same method to synthetic rain fields with tunable smoothness and scale, these authors found that the differences between the Daubechies wavelets are only moderate compared to the difference between parameter settings – the wavelet spectra are determined mostly by the structure of the field, not the shape of the basis function.

Figure A1, summarizing the entropies for all rain fields from our data set, largely confirms this result. While the optimum lies between one and four vanishing moments, the differences between these wavelets of short to intermediate smoothness are marginal compared to the sample variability across the different fields. Faced with the choice between D_2 and D_3, which have very nearly identical results, we select D_2 because it has a shorter support, thereby allowing us to utilize the first seven scales (see Table 1).

Data availability. The RADKLIM data set is available from the DWD servers (https://doi.org/10.5676/DWD/RADKLIM_RW_V2017.002, Winterrath et al., 2018). The regional reanalysis COSMO-REA2 is available from the "Hans-Ertel-Zentrum" website: ftp://ftp.meteo.uni-bonn.de/pub/reana/COSMO-REA2/ (Wahl et al., 2017). COSMO-DE-EPS forecasts are available from DWD upon request.

Author contributions. SB and PF developed the idea for this study and designed the methodology. Data handling, analysis and visualization were carried out by SB, who also lead the writing. Both authors contributed to the discussion and interpretation of the results.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors are grateful to Sebastian Brune for his help in obtaining the RADOLAN data and Velibor Pejcic for valuable advice concerning their interpretation.

Financial support. This research has been supported by the Deutsche Forschungsgemeinschaft (grant no. FR 2976/2-1).

Review statement. This paper was edited by Chris Forest and reviewed by three anonymous referees.

References

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., and Reinhardt, T.: Operational convective-scale numerical weather prediction with the COSMO model: description and sensitivities, Mon. Weather Rev., 139, 3887–3905, 2011.

Bick, T., Simmer, C., Trömel, S., Wapler, K., Hendricks Franssen, H.-J., Stephan, K., Blahak, U., Schraff, C., Reich, H., Zeng, Y., and Potthast, R.: Assimilation of 3D radar reflectivities with an ensemble Kalman filter on the convective scale, Q. J. Roy. Meteorol. Soc., 142, 1490–1504, 2016.

Bierdel, L., Friederichs, P., and Bentzien, S.: Spatial kinetic energy spectra in the convection-permitting limited-area NWP model COSMO-DE, Meteorol. Z., 21, 245–258, https://doi.org/10.1127/0941-2948/2012/0319, 2012.

Brune, S., Kapp, F., and Friederichs, P.: A wavelet-based analysis of convective organization in ICON large-eddy simulations, Q. J. Roy. Meteorol. Soc., 144, 2812–2829, 2018.

Buschow, S., Pidstrigach, J., and Friederichs, P.: Assessment of wavelet-based spatial verification by means of a stochastic precipitation model (wv_verif v0.1.0), Geosci. Model Dev., 12, 3401–3418, https://doi.org/10.5194/gmd-12-3401-2019, 2019.

Casati, B., Ross, G., and Stephenson, D.: A new intensity-scale approach for the verification of spatial precipitation forecasts, Meteor. Appl., 11, 141–154, 2004.

Conway, J. R., Lex, A., and Gehlenborg, N.: UpSetR: an R package for the visualization of intersecting sets and their properties, Bioinformatics, 33, 2938–2940, 2017.

Daubechies, I.: Ten lectures on wavelets, vol. 61, Siam, 1992.

Dorninger, M., Gilleland, E., Casati, B., Mittermaier, M. P., Ebert, E. E., Brown, B. G., and Wilson, L. J.: The setup of the MesoVICT Project, B. Am. Meteorol. Soc., 99, 1887–1906, 2018.

Eckley, I. A., Nason, G. P., and Treloar, R. L.: Locally stationary wavelet fields with application to the modelling and analysis of image texture, J. Roy. Stat. Soc. C, 59, 595–616, 2010.

Gilleland, E.: SpatialVx: Spatial Forecast Verification, available at: https://CRAN.R-project.org/package=SpatialVx (last access: February 2020), r package version 0.6-3, 2018.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E.: Intercomparison of spatial forecast verification methods, Weather Forecast., 24, 1416–1430, 2009.

Goel, P. K. and Vidakovic, B.: Wavelet transformations as diversity enhancers, Institute of Statistics & Decision Sciences, Duke University Durham, NC, 1995.

Hewer, R.: Stochastisch-physikalische Modelle für Windfelder und Niederschlagsextreme, Ph.D. thesis, University of Bonn, 2018.

Kapp, F., Friederichs, P., Brune, S., and Weniger, M.: Spatial verification of high-resolution ensemble precipitation forecasts using local wavelet spectra, Meteorol. Z., 27, 467–480, 2018.

Kuell, V. and Bott, A.: A hybrid convection scheme for use in nonhydrostatic numerical weather prediction models, Meteorol. Z., 17, 775–783, 2008.

Peralta, C., Ben Bouallègue, Z., Theis, S., Gebhardt, C., and Buchhold, M.: Accounting for initial condition uncertainties in COSMO-DE-EPS, J. Geophys. Res.-Atmos., 117, D07108, https://doi.org/10.1029/2011JD016581, 2012.

Radanovics, S., Vidal, J.-P., and Sauquet, E.: Spatial verification of ensemble precipitation: an ensemble version of SAL, Weather Forecast., 33, 1001–1020, 2018.

Rubner, Y., Tomasi, C., and Guibas, L. J.: The earth mover's distance as a metric for image retrieval, Int. J. Comput. Vision, 40, 99–121, 2000.

Scheuerer, M. and Hamill, T. M.: Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities, Mon. Weather Rev., 143, 1321–1334, 2015.

Schleiss, M., Chamoun, S., and Berne, A.: Nonstationarity in Intermittent Rainfall: The "Dry Drift", J. Hydrometeorol., 15, 1189–1204, https://doi.org/10.1175/JHM-D-13-095.1, 2014.

Seifert, A. and Beheng, K. D.: A two-moment cloud microphysics parameterization for mixed-phase clouds. Part 1: Model description, Meteorol. Atmos. Phys., 92, 45–66, 2006.

Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., and Masson, V.: The AROME-France convective-scale operational model, Mon. Weather Rev., 139, 976–991, 2011.

Stephan, K., Klink, S., and Schraff, C.: Assimilation of radar-derived rain rates into the convective-scale model COSMO-DE at DWD, Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, Appl. Meteorol. Phys. Oceanogr., 134, 1315–1326, 2008.

Theis, S., Gebhardt, C., and Bouallegue, Z. B.: Beschreibung des COSMO-DE-EPS und seiner Ausgabe in die Datenbanken des DWD, Deutscher Wetterdienst, 2014.

Wahl, S., Bollmeyer, C., Crewell, S., Figura, C., Friederichs, P., Hense, A., Keller, J. D., and Ohlwein, C.: A novel convective-scale regional reanalysis COSMO-REA2: Improving the representation of precipitation, Meteorol. Z., 26, 345–361, https://doi.org/10.1127/metz/2017/0824, 2017 (data available at: ftp://ftp.meteo.uni-bonn.de/pub/reana/COSMO-REA2/, last access: March 2020).

Weniger, M. and Friederichs, P.: Using the SAL technique for spatial verification of cloud processes: A sensitivity analysis, J. Appl. Meteorol. Climatol., 55, 2091–2108, 2016.

Wernli, H., Paulat, M., Hagen, M., and Frei, C.: SAL – A novel quality measure for the verification of quantitative precipitation forecasts, Mon. Weather Rev., 136, 4470–4487, 2008.

Willeit, M., Amorati, R., Montani, A., Pavan, V., and Tesini, M. S.: Comparison of spectral characteristics of precipitation from radar estimates and COSMO-model predicted fields, Meteorol. Atmos. Phys., 127, 191–203, 2015.

Winterrath, T., Brendel, C., Mario, H., Junghänel, T., Klameth, A., Walawender, E., Weigl, E., and Becker, A.: RADKLIM Version 2017.002: Reprocessed gauge-adjusted radar data, one-hour precipitation sums (RW), https://doi.org/10.5676/DWD/RADKLIM_RW_V2017.002, 2018.

Wong, M. and Skamarock, W. C.: Spectral characteristics of convective-scale precipitation observations and forecasts, Mon. Weather Rev., 144, 4183–4196, 2016.

Yano, J.-I. and Jakubiak, B.: Wavelet-based verification of the quantitative precipitation forecast, Dynam. Atmos. Oceans, 74, 14–29, 2016.

Appendix C

Buschow and Friederichs 2021a

Received: 28 July 2020 | Revised: 13 November 2020 | Accepted: 11 December 2020

DOI: 10.1002/qj.3964

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

SAD: Verifying the scale, anisotropy and direction of precipitation forecasts

Sebastian Buschow | **Petra Friederichs**

Institute of Geosciences, University of Bonn, Germany

Correspondence
S. Buschow, Institute of Geosciences, University of Bonn, Bonn 53113, Germany.
Email: sebastian.buschow@uni-bonn.de

Funding information
German Research Foundation (DFG), Grant/Award Number: FR 2976/2-1

Abstract

One important attribute of meteorological forecasts is their representation of spatial structures. While several existing verification methods explicitly measure a structure error, they mostly produce a single value with no simple interpretation. Extending a recently developed wavelet-based verification method, this study separately evaluates the predicted spatial scale, orientation and degree of anisotropy. The scale component has been rigorously tested in previous work and is known to assess the quality of a forecast similar to other, established methods. However, directional aspects of spatial structure are less frequently considered in the verification literature. Since important weather phenomena related to fronts, coastlines and orography have distinctly anisotropic signatures, their representation in meteorological models is clearly of interest. The ability of the new wavelet approach to accurately evaluate directional properties is demonstrated using idealized and realistic test cases from the MesoVICT project. A comparison of precipitation forecasts from several forecasting systems reveals that errors in scale and direction can occur independently and should be treated as separate aspects of forecast quality. In a final step, we use the inverse wavelet transform to define a simple post-processing algorithm that corrects the structural errors. The procedure improves visual similarity with the observations, as well as the objective scores.

KEYWORDS

MesoVICT, precipitation forecasts, structure error, verification, wavelets

1 | INTRODUCTION

The errors of modern weather forecasts can take many different forms. While everyday users may only notice that their weather app failed to predict rainfall at a specific point in space and time, such a mistake could have several possible meanings. Perhaps the coherence of a frontal precipitation band was underestimated. The likelihood of convective initiation in a certain region may have been misjudged. Alternatively, the simulated pattern was perfectly adequate but its spatial location was wrong. In some cases, the precipitating process is so small and short-lived that no present-day weather model could be expected to foresee its exact timing and placement.

In order to obtain useful diagnostic information on the merits of highly resolved simulations, many forecast verification tools aim to separate the various types of error from one another. Most prominently, spatial displacements tend to mask all other kinds of error in a point-wise evaluation. To tackle this issue, a multitude of so-called "spatial" verification techniques have been developed throughout the last two decades. A first intercomparison of these methods was undertaken within the intercomparison project (ICP; Gilleland *et al.*, 2009), which classified the various approaches and attempted to elucidate their differences and similarities using a set of standardized test cases. The Mesoscale Verification Intercomparison over Complex Terrain (MesoVICT; Dorninger *et al.*, 2018), launched in 2014, constitutes the second phase of the ICP and focuses on the effects of uneven terrain and uncertain observations and considers forecasts of both precipitation and wind.

This study participates in MesoVICT by using both the realistic test cases of Dorninger *et al.* (2018) and the recently presented geometric tests of Gilleland *et al.* (2020). Our focus lies on isolating and understanding errors in the predicted spatial structure of quantitative precipitation forecasts. Using a two-dimensional wavelet transform, we want to separately determine whether the predicted structure was (a) too small or too large, (b) too directed or too round, and (c) oriented along the correct angle.

Several popular methods from the rich spatial verification literature have previously been used to determine a "structure" error. Using the field deformation technique of Keil and Craig (2007), Han and Szunyogh (2016) approximately corrected the forecast's location and intensity and referred to the residual error as "structural". While straightforward and intuitive, this kind of approach yields no further information on how exactly the pattern was mis-forecast. Furthermore, it should be noted that any field deformation approach which allows for a divergent optical flow will be sensitive, simultaneously, to errors in both the spatial scale and anisotropy and therefore cannot truly separate structure from location.

A more intuitive notion of structural disagreement can be obtained using object-based methods that decompose the fields into features and measure their individual properties. Such techniques are typically adapted to the special case of precipitation forecasts where well-defined discrete objects are known to exist. The popular SAL method of Wernli *et al.* (2008) defines its structure component S via the ratio between total and maximum precipitation in each object. The resulting score is related to the size and number of objects as well as the tail behaviour of the marginal distribution; directional aspects are neglected by S. Interestingly, the relative placements of the individual objects are also not included in S. It is clear that a number of small features in close proximity to one another can form a large-scale structure, perhaps driven by a single meteorological process. Such fractured pattern may result from the driving process itself, the complex terrain in which it occurs, or, in case of observational data, the measurement technique. SAL does not consider this as an element of structure, but instead includes the relative placement of the object in the location component (Wernli *et al.*, 2008). A further related drawback of this otherwise useful technique is its potential sensitivity to the details of the object identification algorithm (Weniger and Friederichs, 2016).

The popular Method for Object-based Diagnostic Evaluation (MODE; Davis *et al.*, 2006) provides a wide framework in which numerous structural properties such as feature size, aspect ratio, aspect angle and even curvature can be evaluated. Like SAL, it is mostly adapted to precipitation, can be sensitive to the object-defining algorithm and does not simply allow multiple features to form an organized super-structure on larger scales. More generally, these techniques are inherently *single-scaled*. If objects are detected by smoothing the field with a kernel of size σ and thresholding at a value T, then the pair (σ, T) defines a spatial scale: larger values of T isolate smaller intense regions, larger values of σ lead to the union of increasingly distant features into single, larger objects. If we calculate, for example, the average aspect ratio of the objects detected in this manner (following Davis *et al.*, 2006), the result is characteristic of the anisotropy on the scale defined by (σ, T). Re-arranging the objects in space (larger-scale variability) or re-arranging the pixels within an object (smaller-scale variability) leave the result of the analysis unchanged, since the aspect ratio depends only on an object's shape, not its position or internal composition.

Avoiding such pitfalls of the object-based methods, several authors have based their structural verification on indirect estimates of the spatial correlation structure. Marzban and Sandgathe (2009), Scheuerer and Hamill (2015) and Ekström (2016) rely on empirical variograms for this purpose. The latter study in particular achieves an intuitive notion of the predicted and observed spatial scale by estimating the variogram's range. Without defining a verification score, Willeit *et al.* (2015) study the climatological structure of forecast precipitation fields using Fourier transforms. Wong and Skamarock (2016) extended this work using directional information from the 2D Fourier spectra.

A similar direction was pursued in Buschow *et al.* (2019). Building on the work of Kapp *et al.* (2018), they used a two-dimensional redundant discrete wavelet transform (RDWT) to analyze the scale on which spatial variability occurs, both globally and at each grid point. The resulting scores effectively isolate scale errors in idealized (Buschow *et al.*, 2019) as well as realistic situations (Buschow and Friederichs, 2020). A first analysis of anisotropy, using the same kind of wavelet transform, was undertaken by Brune *et al.* (2018) who included a measure of anisotropy as one component of their wavelet-based convective organization index (WOI). Following the same approach, Brune *et al.* (2020) furthermore attempted to infer the local direction in precipitation fields from the relative contributions of horizontal and vertical features to the total variability. However, as discussed in Buschow and Friederichs (2020), the classic discrete wavelet transform has inherent shortcomings in its representation of directional structures which make it especially unsuitable to the task of forecast verification. With this transform, one could rotate forecast and observation by the same angle and receive a very different score.

In the context of image processing, the same issues were long ago recognized by Kingsbury (1999), who addressed the problem using complex-valued wavelets. Their so-called dual-tree complex wavelet transform (DTCWT) forms the new basis for our verification method. A first meteorological application of this technique was recently presented by Scovell (2020) who used it to incorporate anisotropy into a stochastic noise generator for precipitation nowcasting. Nerini *et al.* (2017) pursued a similar route in their application of a localized Fourier transform to the task of reproducing non-stationary, non-isotropic rainfall variability.

After introducing the relevant datasets in Section 2, we explain in Section 3 why the original RDWT is unsuited to analyze directions and how the DTCWT solves the problem. The next step (Section 4) is to extend the idea of a central scale (Buschow *et al.*, 2019) to include anisotropy and direction. This leads to the definition of new wavelet-based structure scores in Section 5. Experiments with geometric test patterns (Section 6) and the realistic MesoVICT forecasts (Section 7) demonstrate that the new wavelets allow for the same kind of sensitive scale-verification as their predecessors. In addition, they yield valuable information on the forecast's degree of anisotropy and predominant orientation. Section 8 uses the inverse DTCWT to define a simple algorithm for correcting the structural errors detected by our approach. The algorithm is tested on individual forecasts as well as the dataset as a whole. We discuss the outcomes of all experiments in Section 9.

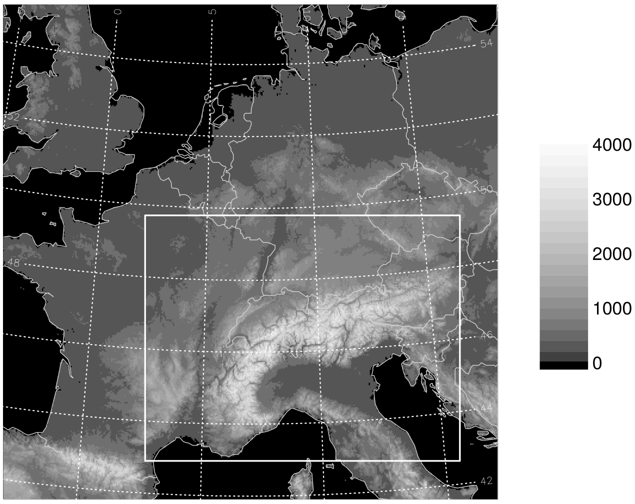


FIGURE 1 VERA orography in metres and minimum common domain for the MesoVICT dataset (white rectangle)

TABLE 1 MesoVICT cases, reproduced from Dorninger *et al.* (2018)

Case	Date	Weather event
1	20–22 June 2007	Strong convective activity north of the Alps followed by a cold front
2	18–21 July 2007	Strong convective events across an airmass boundary impinging the Alps from northwest
3	25–29 September 2007	Accelerating cold front north of the Alps and cyclogenesis in the Gulf of Genoa with heavy rains south of the Alps (Venice flood)
4	6–8 August 2007	Squall line ahead of cold front crossing the Alps causing widespread thunderstorms
5	18 September 2007	Cold front crossing the Alps causing severe thunderstorm in Slovenia
6	8–10 July 2007	Subtropical air mass advected into the Alpine region causing widespread thunderstorms

2 | DATA

In contrast to the largely homogeneous Great Plains considered in the ICP, the study area of the MesoVICT project focuses on a small, mountainous region surrounding the European Alps (Figure 1). Six case-studies on interesting weather situations in summer and autumn 2007 were selected (Table 1, reproduced from Dorninger *et al.* (2018)). Gridded analysis data of precipitation and wind are provided by the Vienna Enhanced Resolution Analysis (VERA; Bica *et al.*, 2007) which incorporates station observations and topographic information but no data from numerical models. While a variety of forecast datasets are in principle available within MesoVICT, we focus on four deterministic models which cover the entire region for all cases: the Swiss COSMO (initialized at 0000 UTC), CMH from the Canadian weather service (initialized at 0600 UTC) as well as BOLAM007 and MOLO0225 from the Institute for Environmental Protection and Research

TABLE 2 Summary of all used datasets

Abbreviation	Organization	Lead times	Δx	Citations
CMH	Environment Canada	+1 hr … +24 hr	2.5 km	McTaggart-Cowan (2009)
COSMO	MeteoSwiss	+6 hr … +24 hr	2.2 km	Ament and Arpagaus (2009)
BOLAM007	ISPRA	+12 hr … +35 hr	0.07°	Mariani and Casaioli (2018)
MOLO0225	ISPRA	+12 hr … +35 hr	0.0225°	Mariani and Casaioli (2018)
VERA	University of Vienna	—	8 km	Bica *et al.* (2007)
RADKLIM	DWD	—	1 km	Winterrath *et al.* (2018)

(ISPRA, both initialized at 1200 UTC, first 12 hr discarded). To avoid obvious artifacts of model spin-up, we consider only time steps from 0700 UTC to 2300 UTC. Further details and the references for each model are given in Table 2.

In Section 7.2, we move beyond the domain of the MesoVICT project and validate VERA, BOLAM007 and MOLO0225 against the gauge-adjusted radar climatology RADKLIM of Winterrath *et al.* (2018). The dates and other characteristics of the datasets remain the same, but the domain for this experiment is defined by the German national borders. Missing pixels in RADKLIM (outside Germany or due to radar failures) are set to zero in all fields to ensure comparability.

Hourly rain sums for all forecasts have been interpolated to the VERA grid at a common resolution of approximately 8 km. After cropping the data to the core regions where all datasets have non-missing values, we obtain 133×88 grid points for the Alpine domain and 99×116 for Germany. These regions are symmetrically extended to 256×256 because our implementation of the wavelet transform requires the input dimensions to be whole powers of two. Buschow and Friederichs (2020) discuss several possible ways of handling these boundary conditions and conclude that, in theory, reflective boundaries are the most elegant and appropriate solution. However, this approach may no longer be viable because we are interested not only in the scale but also in the direction and anisotropy detected by the wavelets. It is easy to imagine situations where the latter two properties are distorted, when we reflect the input image at the edges. To avoid such effects, we pad the fields with zeros instead. Following Kapp *et al.* (2018), we linearly decrease the original values to zero across ten pixels along each side in order to smooth out potential artificial edges.

Rain values below 0.1 mm are set to 0 mm, then all values x are replaced by $\log_2(x + 0.1)$ before the wavelet transform is applied. Buschow and Friederichs (2020) discuss the rationale behind this step in detail. Simply, typical plots of rain fields use logarithmic or similar colour scales in order to visualize both local extreme events and extended areas of moderate intensities. Similar ideas apply to our automatic analysis of spatial structure. The results of the wavelet transform are generally easier to understand if they are based on the same data transformation as the plots used for visual inspection.

3 | THE DUAL-TREE COMPLEX WAVELET TRANSFORM

This section introduces the basics of discrete wavelet transforms in a very concise manner. To readers who are completely new to wavelets, we recommend Torrence and Compo (1998) and Weniger and Friederichs (2016) for an introduction in a meterological context, as well the general textbook of Daubechies (1992).

Classic wavelet transforms start by selecting a function $\psi(r)$, $r \in \mathbb{R}^n$, which is localized in both space and frequency and integrates to zero over its domain of definition. From this so-called mother wavelet, a set of daughter wavelets are derived via shifting and re-scaling, i.e., $\psi_{s,u}(r) = s^{-1/2}\psi((r-u)/s)$. In a multidimensional space $\mathbb{R}^n$ with $n > 1$, the daughters can furthermore have various spatial orientations, which we will denote by the index d. A signal (a time series for $n = 1$, an image for $n = 2$) is projected onto the $\psi_{s,u(,d)}$, thereby decomposing it into components with specific scales and (in 2D) directions. This so-called *wavelet transform* is similar to the well-known Fourier transform except that the basis functions are also localized in time (or space in 2D), which allows for the correct treatment of non-stationary signals.

The bottom row of Figure 2 shows the three directional daughter wavelets associated with the two-dimensional discrete wavelet transform (DWT). Their mother is the least asymmetric Daubechies wavelet with six vanishing moments. By shifting several scaled versions of these directed, localized wave-forms across an image, one can localize and study features of various spatial scales and orientations. Looking at the diagonal daughter (45° in the bottom row of Figure 2), it becomes obvious that all attempts at deriving direction and anisotropy from these basis functions are flawed: while the vertical and horizontal daughter wavelets are rotated versions of one another,

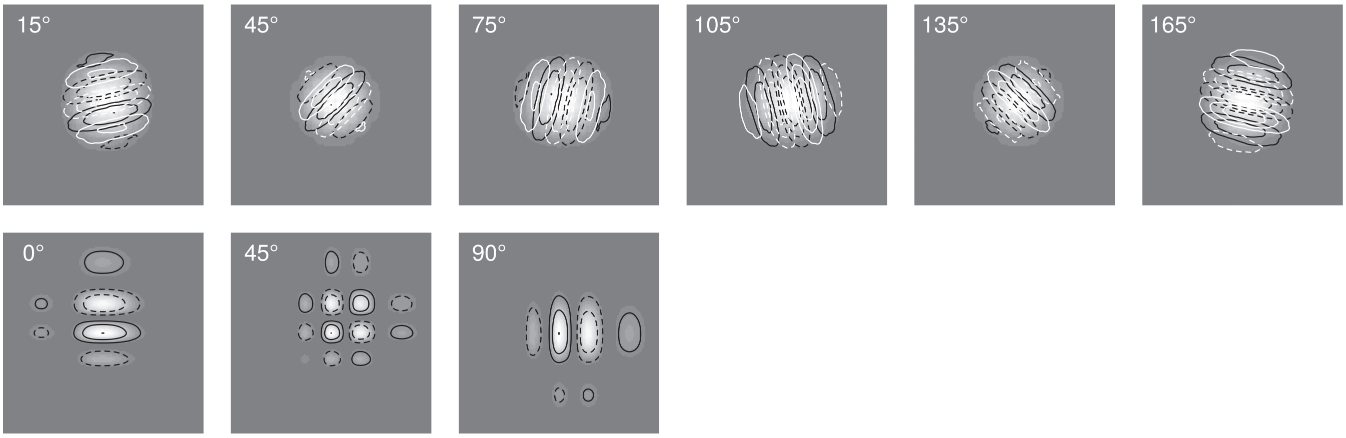


FIGURE 2 Directed daughter wavelets of the DTCWT (top row) and the regular DWT ("extremal phase" Daubechies wavelet number 6, bottom). Solid and dashed lines indicate positive and negative values, respectively, the background shading indicates the absolute amplitudes squared. White contours correspond to the imagery part of the DTCWT daughters

the daughter for 45° clearly lives on a smaller scale and cannot distinguish between the two diagonals. Using this wavelet transform, it is thus impossible to decide whether an edge is oriented at +45° or −45°. The degree to which one of the three directions dominates over the others, that is, the estimated anisotropy of a given feature, furthermore depends on the scale and the orientation of that feature.

To understand the origin of (as well as the solution to) these undesirable effects, we must briefly discuss the algorithms by which wavelet transforms are implemented. In principle, one could convolve the signal with each scaled and oriented daughter wavelet individually. This procedure is used in continuous wavelet transforms (CWTs), which allow arbitrary scales s and involve a high degree of redundancy at high computational costs. A far more efficient algorithm was introduced by Mallat (1989), paving the way for innumerable modern wavelet methods: instead of defining a continuous function ψ, the mother wavelet is represented by a finite set of filter coefficients $g_{1,\dots,n}$. Next, the so-called father wavelet $\phi(x)$ is defined by the filter coefficients $h_k = (-1)^k g_{n-1-k}$. The father wavelet is thus a reversed version of the mother, where the sign of every second coefficient has been flipped. Loosely speaking, mom performs a differentiation (high-pass) while dad is an averaging (low-pass) filter.

In one dimension, the so-called discrete wavelet transform (DWT) is then implemented by (1) convolving the signal with g and discarding every second value from the result to obtain the wavelet coefficients at the finest scale, (2) convolving the signal by h (again discarding half of the values) to obtain the input for the next coarser scale, and (3) repeating (1) and (2) until only a single value remains. By dropping every second value, we effectively shift the smallest-scaled daughter wavelet ($s = 2^0$) to every second location in the domain, the next larger one ($s = 2^1$) to every fourth, and so on, thereby removing the redundancy. The scales of this so-called "decimated" transform are no longer continuous but whole powers of two, and the larger wavelets are shifted to fewer locations. As a result, the daughter wavelets of this transformation form an orthogonal basis. The algorithm can be adapted to obtain values at all possible locations by simply not discarding any values and instead inserting zeros between the filter coefficients h_k and g_k after each level, resulting in a redundant discrete wavelet transform (henceforth RDWT).

The efficient DWT algorithm has a straightforward extension to higher dimensions which wasalso introduced by Mallat (1989) (shown in Figure 3). Given a two-dimensional matrix of input values, convolve the rows with h and then the columns with g to obtain the vertical daughter coefficients. The horizontal daughter coefficients result from applying g to the rows and h to the columns; the diagonal daughter is the product of applying g to both rows and columns. Application of h in both directions gives the input for the next coarser scale. This procedure, which can be implemented with decimation or redundancy just as in the 1D case, generates the three directional daughters seen in Figure 2. That explains the reduced scale of the diagonal wavelets (being the product of two high-pass filters), as well as the absence of a fourth filter for the other diagonal.

Recognizing the shortcomings of the classic DWT, Kingsbury (1999) introduced the so-called dual-tree complex wavelet transform (henceforth DTCWT). Instead of a single real-valued mother, they defined a complex-valued $\psi = \psi_r + i\psi_i$ with corresponding filters h_i, h_r, g_i, g_r. The two mother wavelets ψ_r and ψ_i are each other's Hilbert transform, meaning that they are 90° out of phase with each other. In two dimensions, the complex transform can be implemented by performing four regular DWT s (as in Figure 3) with all possible combinations of $h_{i,r}, g_{i,r}$

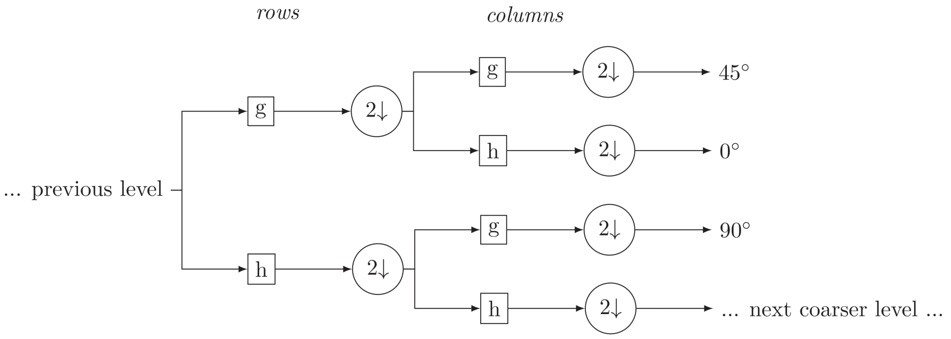


FIGURE 3 One level of the two-dimensional discrete wavelet transform (dwt). g and h denote applications of the high- or low-pass filter, respectively. "2↓" signifies a down-sampling step

applied to the rows and columns. The twelve resulting sets of coefficients are then re-combined into six complex directional daughter wavelet coefficients. A set of corresponding daughter wavelets is shown in the top row of Figure 2. Each of these functions represents one distinct direction, and the two diagonals are no longer ambiguous. Here, we have furthermore applied the method of Kingsbury (2006) to obtain optimized diagonal wavelets with very nearly the same scale as their sisters. The two issues preventing us from utilizing the directional information of the wavelet transform are thus resolved. The absolute values, shown as background shading in Figure 2, reveal a further advantage of this transform. While real and imaginary parts both constitute wave forms, the Hilbert property means that the modulus monotonously decreases from the centre of the support. Image features can thus unambiguously be located within the support of each daughter wavelet – a task which is less straightforward for the Daubechies wavelets. As a final benefit, Selesnick *et al.* (2005) report that the complex nature of the coefficients greatly reduces the shift dependence of the transform. While for the regular DWT, we must always rely on the computationally more expensive redundant transform, we can obtain robust information on the global structure of a field from the decimated DTCWT as well. We demonstrate the effective equivalence of the two transforms in Appendix B.

Regardless of the merits of the decimated DTCWT, it does not deliver fully localized information because the large-scale coefficients are only available on increasingly coarse grids due to the downsampling (cf. Figure 3). If we are interested in local characteristics at every location, a fully redundant transform is needed. In this case, an over-emphasis on very large scales, caused by their great redundancy (large overlapping areas), must be avoided. Here, we follow Kapp *et al.* (2018), Brune *et al.* (2018) and Buschow *et al.* (2019) and rely on the theory of locally stationary wavelet processes (Eckley *et al.*, 2010) to remove this large-scale bias. In a nutshell, it can be shown that the squared local wavelet coefficients have a well-defined relationship with the spatial covariances if we multiply them by a bias-correction matrix which depends on the domain size and choice of mother wavelet. This step mostly reduces the values of large-scale coefficients and re-distributes their energy to smaller scales. The theory was extended to the redundant DTCWT by Nelson *et al.* (2018). Following Buschow *et al.* (2019), any negative "energy" values introduced by the bias correction are set to zero. Based on the discussion in Buschow and Friederichs (2020), we furthermore discard the three largest scales due to their ambiguous localization (basis functions being larger than the entire domain).

4 | ANALYZING SCALE AND DIRECTION

In order to compactly summarize the output of the wavelet transform, Buschow *et al.* (2019) studied the central scales of the wavelet spectra. Let e_j be the (bias-corrected) squared wavelet coefficient for scale $j \in \{1, \dots, J\}$, averaged over all directions. Now consider the e_j as point-masses, located along a line at the coordinates $z_j = j$. The central scale of the wavelet spectrum is then defined as the centre of mass z_c of that arrangement. A plot of these central scales for each pixel of an input image compactly visualizes the result of the wavelet analysis by showing the dominant scale at each location. Buschow *et al.* (2019) demonstrated how this map of central scales can also serve as the basis for spatial verification.

We now extend the idea to the case of the directional wavelet spectra produced by the DTCWT. Noting that the energy of the 15° daughter wavelet should be next to those for 45° and 165°, both being a 30° rotation away (Figure 2), the natural geometry in which to arrange the $6 \times J$ coefficients is a prism with hexagonal base. Figure 4 schematically shows this arrangement. The energies for the six directions are placed along the vertices of a regular hexagon, parallel to the x–y plane. The various scales j correspond to different values of the z coordinate. Indexing

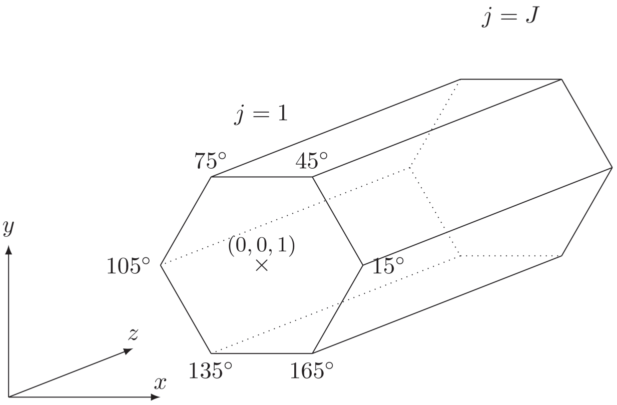


FIGURE 4 Geometry used to define the central scale, radius and angle of the DTCWT spectra

the directions by d, the coordinates for the value $e_{j,d}$ are

$$\begin{aligned} x_{j,d} &= a\cos\{60(d-1)\pi/180\},\\ y_{j,d} &= a\sin\{60(d-1)\pi/180\},\\ z_{j,d} &= j\ , \end{aligned} \tag{1}$$

where a denotes the arbitrarily fixed circumradius of the hexagon. Calculating the centre of mass in this geometry leads to the same central scale z_c as before. The other two central components x_c and y_c contain information on the preferred direction and degree of anisotropy. We can easily separate these two properties by transforming from the x–y plane to polar ρ, θ coordinates. The central radius $\rho_c = \sqrt{x_c^2 + y_c^2}$ then measures the total degree of anisotropy, averaged over all scales. From the central angle $\theta_c = \arctan 2(y_c, x_c)$, we can derive the angle in image space as $\varphi_c = 15° + \theta_c/2$. Note that a is merely a multiplicative factor dertermining the scale of ρ. A more detailed example of these ideas is discussed in Appendix A.

5 | DEFINITION OF SCORES

Buschow *et al.* (2019) introduced the structure score *semd*, which is given by the earth mover's distance (EMD; Rubner *et al.*, 2000) between two direction-averaged spatial mean spectra: the energy e_j of scale j is considered a point mass located at the position $z = j$ along the real line. *semd* measures the minimum cost of transporting all energy from one spectrum to another. Both spectra are normalized to unit sum, making the EMD a true metric. The EMD was preferred over other metrics because it appropriately measures shifts in the spectra as well as differences in their shape.

Based on the ideas from Section 4, the extension of *semd* to the case of directed spectra is very straightforward. Simply place the energies $e_{j,d}$ corresponding to the scales j and directions d at the corresponding vertices of the hexagon (Figure 4) and solve the transport problem to obtain the EMD. We will refer to this directed version of *semd* as *semd*$_d$. The radius a in Equation 1 i.e., the ratio between width and length of the prism *within which the centre resides* (cf. Figure 4), governs the relative contributions of errors in scale, direction and anisotropy to the total value of the score. For the purposes of this paper, we will set $a = (J-1)/2$ corresponding to equal weights for both components ($e_{1,15°}$ is equally far away from $e_{1,105°}$ and $e_{J,15°}$). A more in-depth explanation of this score, including the mathematical definition of the EMD, is given in Appendix A. In addition to this summary score, we introduce three helpful auxiliary quantities:

$$d\rho = \rho_c^{(for)} - \rho_c^{(obs)}, \tag{2}$$

$$d\varphi = \begin{cases} \varphi^{(for)} - \varphi^{(obs)} - 180° & \text{for } \varphi^{(for)} - \varphi^{(obs)} > 90°,\\ \varphi^{(for)} - \varphi^{(obs)} + 180° & \text{for } \varphi^{(for)} - \varphi^{(obs)} < -90°,\\ \varphi^{(for)} - \varphi^{(obs)} & \text{otherwise,} \end{cases} \tag{3}$$

$$dz = z_c^{(for)} - z_c^{(obs)}\ , \tag{4}$$

where $^{(for)}$ and $^{(obs)}$ denote quantities related to forecast and observation, respectively. The difference in central scales dz was studied in Buschow *et al.* (2019) (under the name Sp$_{cd}$). These authors note that dz is a lower bound on *semd* (Rubner *et al.*, 2000) which gives a rough estimate of the scale error and, crucially, determines its sign (too small or too large in scale). Analogously, we have now defined the signed anisotropy error $d\rho$ and the angular error $d\varphi$. It is important to note that $d\varphi$ is only relevant if the predicted and the observed field are both reasonably anisotropic – a circle can be rotated by any angle without actually changing. While this quantity is thus useful and intuitive for individual comparisons, it cannot simply be aggregated over many cases. We therefore define the combined anisotropy and direction error

$$dxy = \sqrt{\left(x_c^{(for)} - x_c^{(obs)}\right)^2 + \left(y_c^{(for)} - y_c^{(obs)}\right)^2}\ . \tag{5}$$

In Section 7, we show that *semd*$_d$ can usually be explained as a linear combination of dxy and $|dz|$, the difference in the shape of the spectra playing only a minor role.

Concluding this section, we note that the complete verification procedure has been implemented in the `sad` R-package (Buschow, 2020).

6 | GEOMETRIC TEST CASES

As a first test of our structural forecast verification based on scale, anisotropy and direction, we consider the

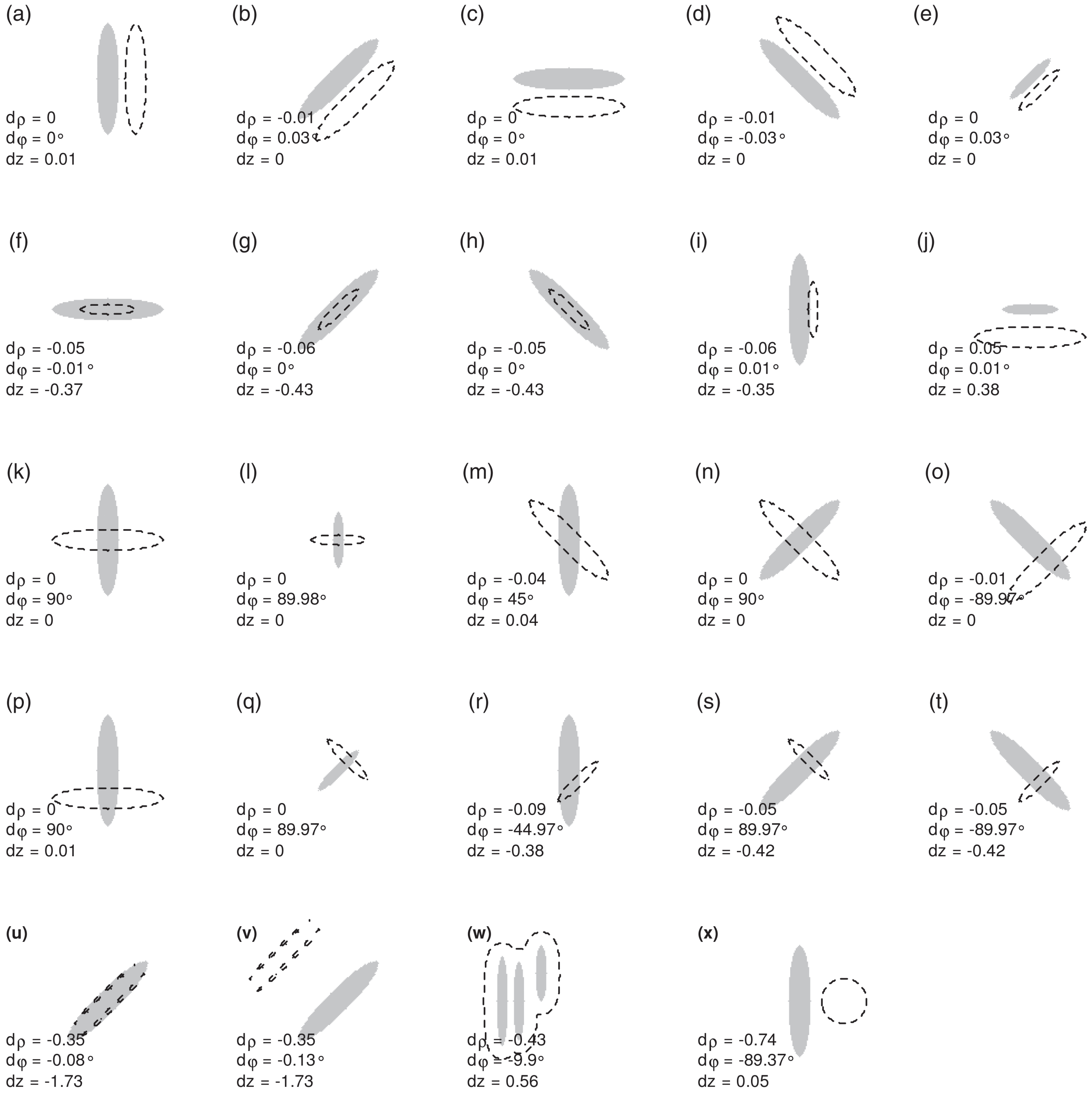


FIGURE 5 Comparisons between elliptical MesoVICT test cases suggested by Gilleland *et al.* In each panel, $d\rho$, $d\varphi$ and dz are anisotropy, angle and scale of the "forecast" (dashed contour) minus the "observations" (grey area)

geometric test cases proposed by Gilleland *et al.* (2020). These authors present a set of 50 binary images and suggest 55 pairwise comparisons between them. Here, we will discuss only the 23 comparisons between the elliptical test images because these are most relevant and interesting for our purposes.

Figure 5 shows all of the elliptical comparisons and the resulting values of $d\rho$, $d\varphi$ and dz. We have calculated the scores in this figure based on the decimated version of the DTCWT in order to test its remaining dependence on location and orientation. In Figure 5a–e, forecast and observation differ only in their location. As expected, all of our scores are close to zero with only very minor variations due to the remaining shift variance.

The situation in Figure 5f–j is more interesting. Besides possible displacement errors, the predicted area is now also too small (f–i) or too large (j). Intuitively, we expect $d\rho = d\varphi = 0$ and an identical non-zero scale error dz in all of these cases. While $d\varphi$ is indeed almost exactly zero, $d\rho$ indicates that the small ellipse appears slightly less anisotropic than the larger one. The values of the scale error differ by less than one tenth of a whole scale and have the expected sign. Figure 5k–t introduce errors in the orientation. In all of those tests, $d\varphi$ detects the rotation with nearly perfect precision (errors <0.05°), irrespective of the presence of additional errors in location or scale.

In the last three test cases (Figure 5u–w), more complicated patterns are compared. In (u) and (v), the observation is a regular ellipse, while the forecast consists of very small oblong shapes along the ellipses' boundaries. As expected, the orientation is found to be very nearly correct, whereas the scale of the forecast is deemed far too small. As for the slightly too small ellipses in previous comparisons, $d\rho$ indicates that the predicted pattern is too isotropic. In the final suggested test (Figure 5w), the observation consists of three small ellipses; the forecast shows a single large feature enveloping the three. As expected, the scale error dz is strongly positive. The anisotropy error is significantly smaller than zero which is in good agreement with our subjective judgement as well. A slight clockwise rotation, indicated by $d\varphi$, also seems reasonable since the three small ellipses are vertically oriented, while the combined pattern extends somewhat more along the diagonal. Since none of the suggested comparisons feature an obvious error in anisotropy, we have added an extra case Figure 5x which compares a circle to an ellipse. As expected, the strongly negative value of $d\rho$ correctly detects the discrepancy. Note that the apparent rotation by -90° is meaningless because one of the two images has a low value of ρ.

In summary, the geometric tests show that our structure verification overall works as intended. The angular component in particular yields almost perfect results and is very robust to changes in location, scale, overall orientation and anisotropy of the fields to be compared. We have seen that the remaining shift-dependence of the decimated DTCWT plays only a minor role for our purposes (see also Appendix B) with discrepancies on the order of ~0.05 in all three components (compare, for example, Figures 5(r) to (s), (u) to (v), and (i) to (j)). Conversely, this also means that forecast errors smaller than 0.05 can generally be regarded as negligible.

7 | REALISTIC TEST CASES

7.1 | Verification against VERA

For a first impression of our verification technique in a realistic situation, we consider a single time step from the second MesoVICT case. Here, we focus on only two competing forecasts. Figure 6a–c show the hourly rain intensity analyzed by VERA and predicted by CMH and COSMO. At this time, precipitation was mainly induced by a quasi-stationary airmass boundary extending roughly from the German–French border to the southwesterly corner of the domain. VERA shows a relatively linear rain feature along the Rhine and a number of more amorphous cells throughout France and Switzerland. CMH overestimates the rain area slightly and the total intensity strongly by producing a nearly round rain field in the north and numerous very small convective cells across the rest of the domain. COSMO, on the other hand, simulates a single linear feature along the airmass border. According to the maps of z_c (Figure 6d–f), the spatial scales are well represented by COSMO while the structure of CMH is overall slightly too small ($dz \approx -0.3$). As expected, CMH is slightly too isotropic, whereas COSMO appears far more directed than the observations ($d\rho \approx 0.46$). In addition, a rotational error of about 14° is assigned to COSMO, which is also in good agreement with our visual impression. The slightly worse scale and much better anisotropy add up to substantially better overall rating for CMH ($semd_d \approx 0.2$) than COSMO ($semd_d \approx 0.41$).

To get an overview of the complete MesoVICT dataset, we apply the decimated DTCWT to all fields and calculate the central components ρ_c, φ_c and z_c. Figure 7 displays the distributions of the evaluated central statistics for each hourly field as well as the the total rain area and total intensity, separated by case and model. Starting with the two simple, non-wavelet quantities, we observe that all models are able to simulate approximately correct rain totals, at least as far as the average over all cases is concerned (case 5 being an exception where all models frequently predict too little rain). The rain area, on the other hand, is systematically underestimated, especially by MOLO0225. One possible interpretation would be that this model simulates variability on smaller scales than those analyzed by VERA. We can partly confirm this hypothesis with the help of z_c which shows that MOLO0225 , as well as COSMO and CMH , operate on smaller scales than VERA. BOLAM007, with its nominal resolution of approximately 7 km, produces similar, in some cases even larger, scales than VERA. The order of the five datasets, BOLAM007 being largest, followed by VERA, COSMO, CMH and the very fine-scaled MOLO0225, is consistent across all six cases.

Next, we are interested in the directional structure. Looking at the distributions of ρ_c in Figure 7, we note that the degree of anisotropy of each model depends on the weather situation. Cases 3 (featuring a Genoa cyclone), 4 (dominated by large convective cells across the Alpine region) and 6 (no organizing frontal structure) are less directed than the remaining three cases, in which cold fronts and an airmass boundary dominate the weather

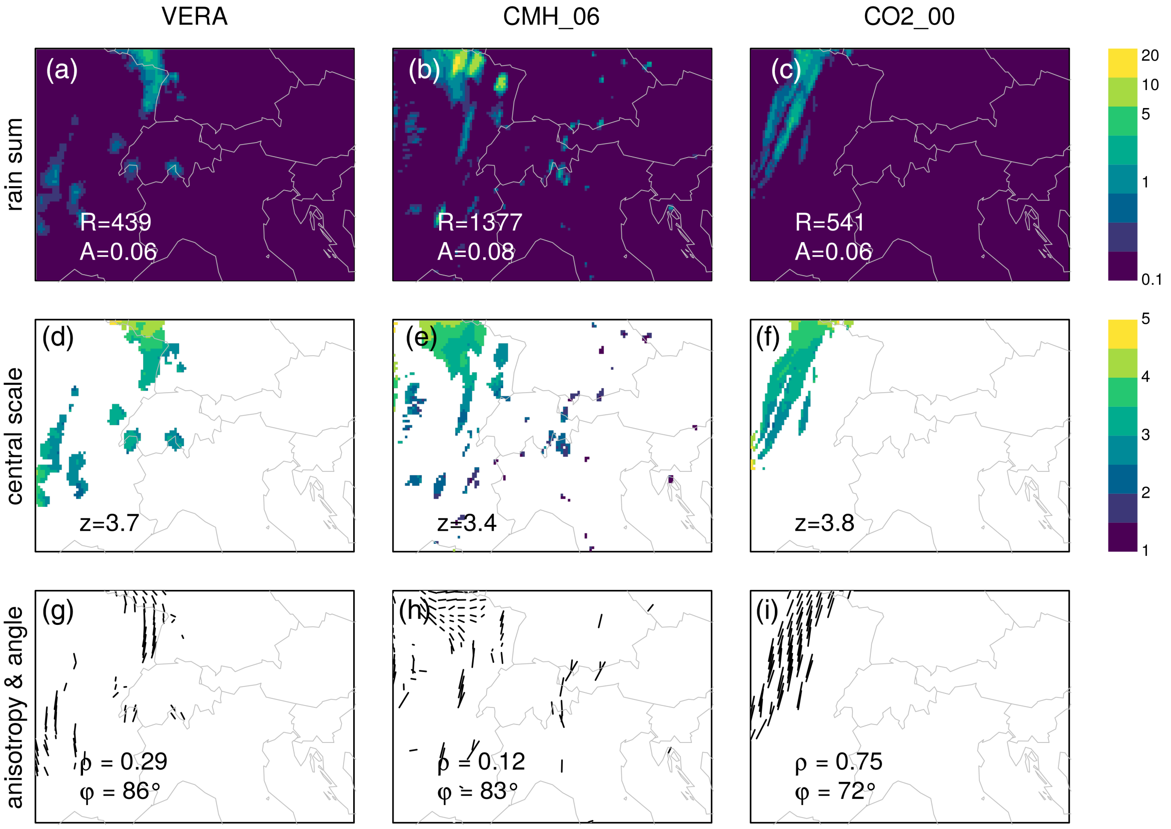


FIGURE 6 Wavelet-based analysis of observed and predicted precipitation on 20 August 2007, 1100 UTC. (a–c) Hourly rain accumulation with total rain sum in mm (R) and area fraction (A), (d–f) map of central scales with mean central scale (z) and (g–i) anisotropy (length of the arrows) and angle (direction of the arrows) with mean anisotropy (ρ) and mean angle (φ) [Colour figure can be viewed at wileyonlinelibrary.com]

patterns. In comparison to VERA, CMH reproduces the average distribution of ρ_c very well. COSMO has the largest positive bias in ρ_c (is far too anisotropic), followed by BOLAM007 and MOLO0225.

The angles φ_c are not included in Figure 7 because box-plots can be misleading for a circular quantity. Figure 8 therefore shows the corresponding histograms instead. As expected, the two strongly directed cases 1 and 2 have a clear preferential direction around 45° and 60°, respectively, corresponding to the alignment of the airmass boundaries present in these cases. All models reproduce the analyzed direction reasonably well in the first case; BOLAM007 and MOLO0225 exhibit slight rotation errors in case 2. The anisotropic cases 3 and 4 feature a wide variety of directions, which are not particularly well matched by any of the models – recall that errors in the orientation are not meaningful when ρ_c is small. Case 5, which only encompasses 24 hr, has well-defined directions related to the cold front crossing the domain. All models represent the 45° orientation of this feature reasonably well. VERA's secondary peak at 90° is caused by a large rain area being cut off at the domain's eastern edge during the final time-steps of the day. Despite its relatively low anisotropy, case 6 also exhibits a well defined peak around 45°, which is present in VERA and all four forecast models. This phenomenon is likely related to the shape of the western flank of the Alps, where many of the precipitation events during this case-study were triggered.

So far, we have only assessed the modelled and observed statistics of spatial structures in each of the six cases. Figure 9 shows the corresponding distributions of the structure scores from Section 5 with respect to VERA, calculated at each time step and separated by case. COSMO and BOLAM007's systematic overestimation of ρ_c is reflected in increased values of the combined anisotropy / direction score dxy. Cases 2 and 5 are deemed particularly bad, while errors in the other, overall more isotropic, cases are less severe. In total, the representation of directional structures in CMH and MOLO0225 is notably better than in the other two models. The opposite result emerges for the representation of spatial scales where CMH and MOLO0225 are the worst candidates with strongly negative values of dz. Despite its nominally finer resolution, COSMO is only slightly too small in scale; the low-resolution BOLAM007 fares best.

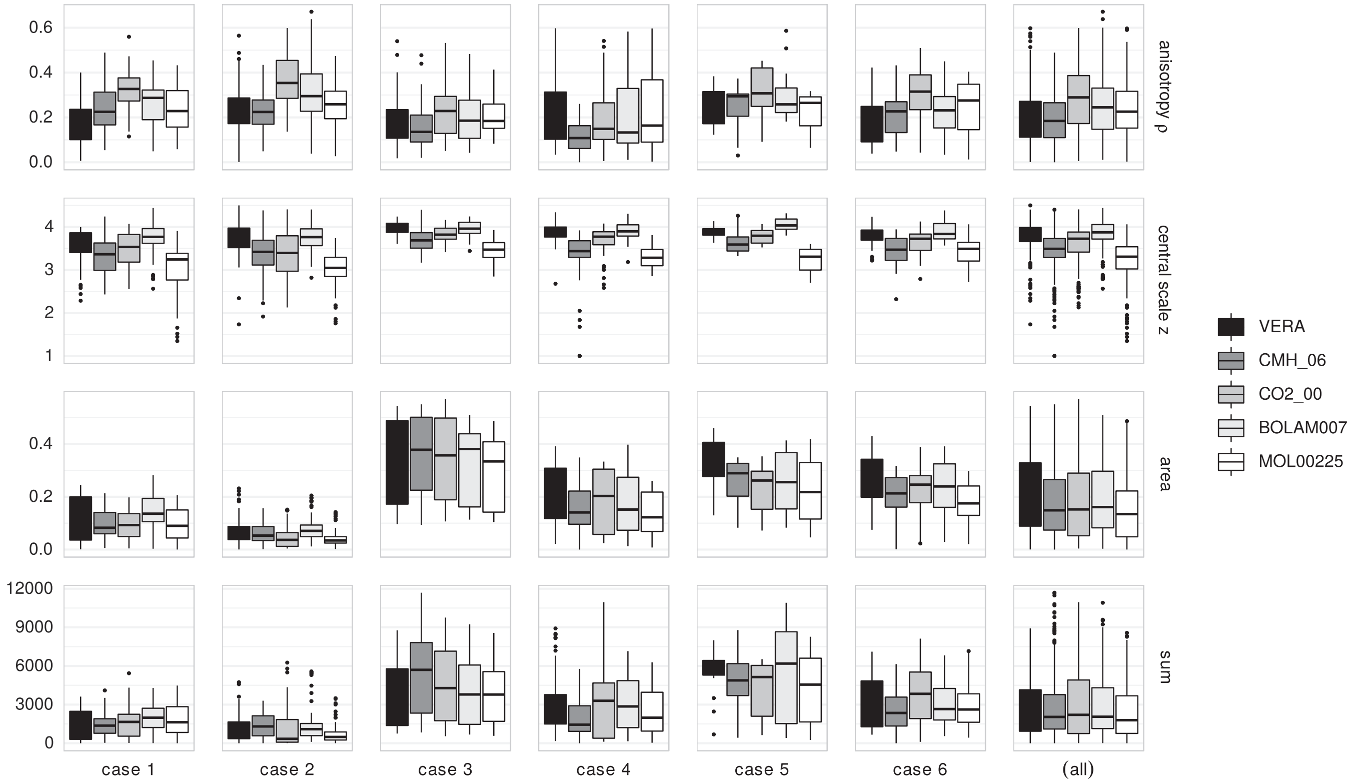


FIGURE 7 Degree of (row 1) anisotropy, (row 2) central scale, (row 3) fraction of the domain with non-zero rain and (row 4) total rain sum in mm for each of the six MesoVICT test cases and for all cases together

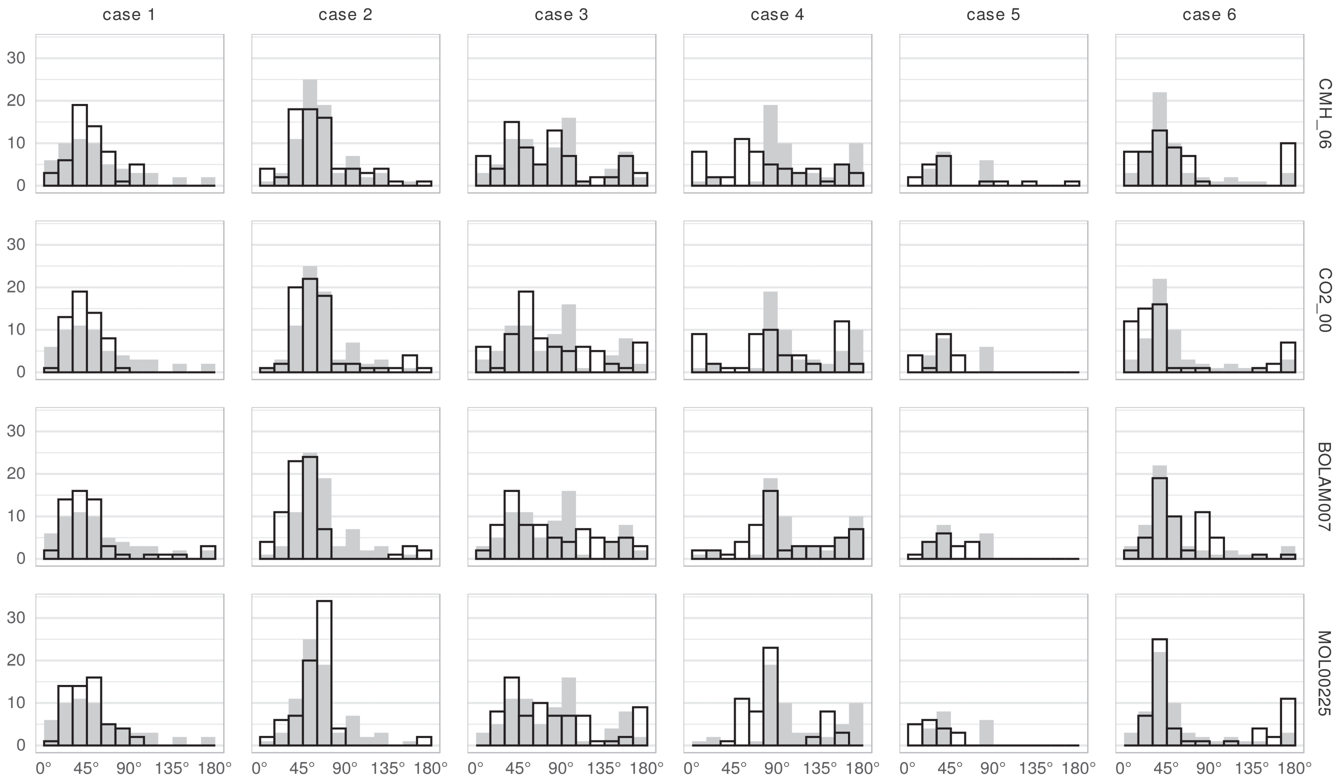


FIGURE 8 Histograms of dominant directions φ_c for all data forecasts (open bars) and VERA (grey bars), separated by case

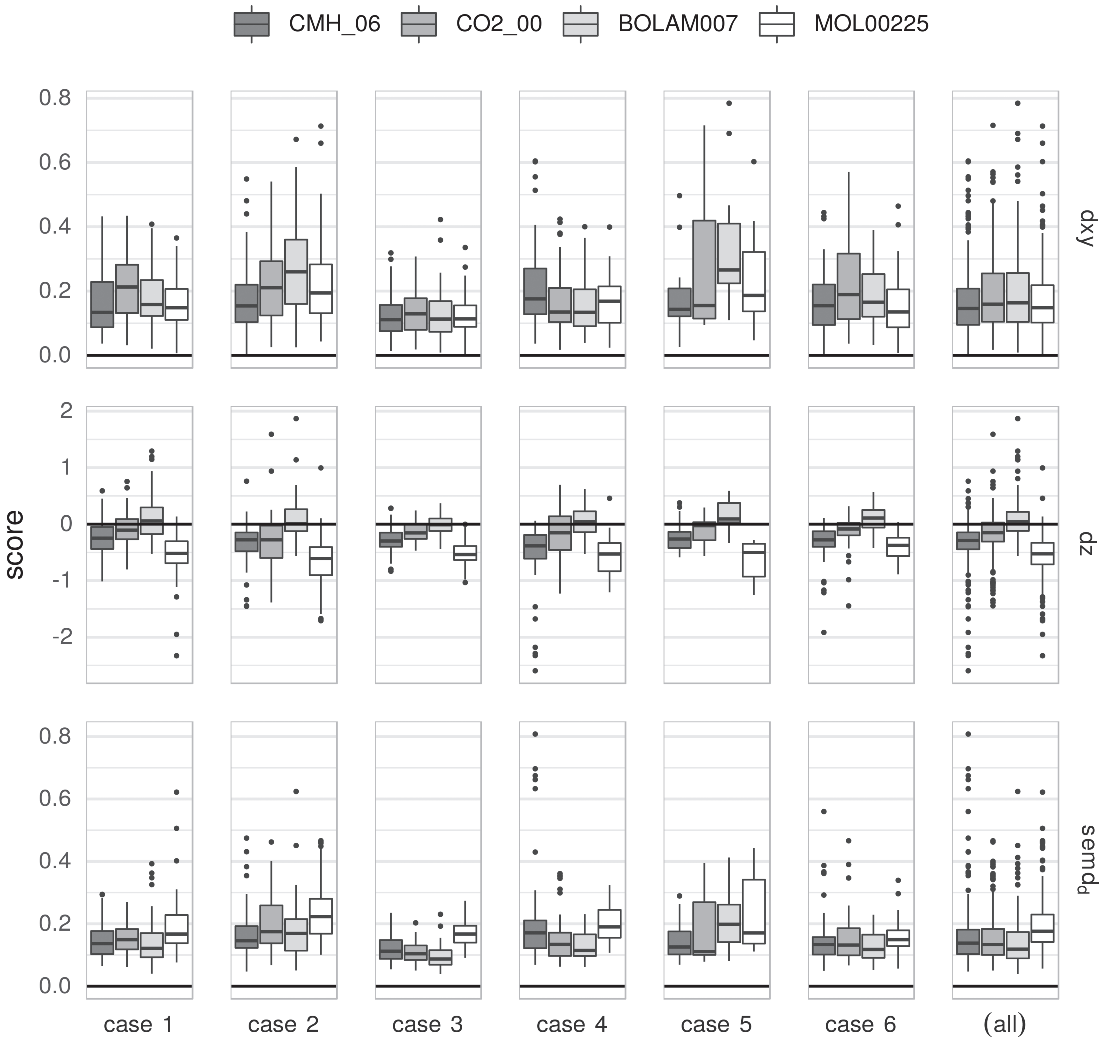


FIGURE 9 Combined (row 1) anisotropy/direction error, (row 2) difference in scale and (row 3) complete spectral EMD between VERA and the forecast models for the six cases. Positive values of *dz* indicate that the forecast is too large in scale

TABLE 3 Coefficient of determination R^2 for linear regressions of the scores on the left of ∼ against those on the right for each model individually and for all models

	CMH	COSMO	BOLAM007	MOLO0225	All
$\|dz\|{\sim}dxy$	0.23	0.04	0.11	0.05	0.05
$semd_d \sim \|dz\|$	0.82	0.58	0.51	0.69	0.69
$semd_d \sim dxy$	0.61	0.57	0.79	0.48	0.48
$semd_d \sim \|dz\|{+}dxy$	0.98	0.96	0.98	0.97	0.97

$semd_d$, shown in row 3 of Figure 9, combines both kinds of structural errors into a single score that takes into account the complete distribution of energy across directions and scales. We find that the large scale-error makes MOLO0225 the overall loser in each individual case and in total, despite its good representation of directions. For CMH and COSMO, the two kinds of error tend to average out, leading to nearly identical scores in aggregate. BOLAM007 emerges as the overall winner, largely due to having similar spatial scales as VERA.

To get a rough idea of the interrelations between our structure scores, we perform a series of linear regressions between them and consider the degree of determination measured by R^2. Table 3 confirms that the scale and directional errors are largely independent of one another. The relative contributions of *dxy* and *dz* to the overall score $semd_d$ differ from model to model, scale dominating for CMH and MOLO0225, direction for BOLAM007 and both contributing equally for COSMO. In all cases, as well as in total, $semd_d$ can almost entirely be explained as a linear combination of the shift in scale and the shift in direction/anisotropy ($R^2 > 0.95$).

7.2 | Verification against RADKLIM

Up to this point, we have assumed VERA as a flawless representation of the true precipitation fields and interpreted all discrepancies as forecast errors. However, the fact that the coarsest model achieved the best overall rating raises some suspicions. Can VERA appropriately represent the scale structure of precipitation fields? To address this question, we shift our attention from the Alpine MesoVICT domain to Germany, where the radar-based RADKLIM dataset offers spatial observations at 1 km nominal resolution. CMH and COSMO are not available in this domain, so the other three datasets are now treated as competing forecasts to be verified against RADKLIM.

Before looking at the scores, it is again instructive to get a visual first impression from an example case. Figure 10 shows an instance of scattered convective cells across Germany during MesoVICT case 6. As expected, VERA and RADKLIM agree very well on the placement and approximate shape of the individual cells. Accordingly, the degree of anisotropy and the overall direction φ are nearly identical. However, RADKLIM reveals a much finer-scaled texture and distributes more precipitation across a smaller total area, leading to a scale error of $dz = 0.5$ for VERA. BOLAM007 and MOLO0225 both under-forecast the overall rain intensity and area and produce patterns slightly too isotropic. Variability in MOLO0225 occurs on the same small scales as in RADKLIM, and BOLAM007 is again far more similar to VERA.

The distribution of scores shown in Figure 11 reveals that our example case was in fact representative of an overall trend: while the scale errors of MOLO0225 are centred around zero, both BOLAM007 and VERA exhibit a bias towards larger scales ($dz \approx 0.3$). In terms of directional structure, VERA is by far the most similar to RADKLIM, followed by MOLO0225 and BOLAM007. With respect to the summary score $semd_d$, VERA and MOLO0225 are thus tied for first place, both performing substantially better than BOLAM007. It is worth noting that the distributions of the scores for the two forecast models have substantially heavier tails than for VERA. These outliers represent complete mis-forecasts of spatial structure, which naturally do not occur in observational datasets like VERA.

8 | CORRECTING STRUCTURAL ERRORS

Errors related to the marginal distribution of a forecast can generally be corrected if the desired distribution is known. Such a calibration procedure may be desirable to improve the forecast or to remove marginal errors before applying further verification methods. Most spatial verification techniques do not suggest a simple way of correcting the errors they detect; the wavelet approach is an exception to this rule. As detailed in Section 3, a wavelet transform is essentially just a change of basis, which can be reversed. Similar to the well-known Fourier case, the discrete wavelet transform allows for analysis *and* synthesis. To correct the errors in the spatial mean wavelet spectrum, we can therefore (1) transform forecasts and observations, (2) multiply the forecast values at each location, scale and direction by the corresponding ratio between total observed and predicted energy and (3) reverse the transform to obtain a corrected version of the forecast. The spatial distribution of the energy of the resulting image is that of the prediction, but its distribution over scale and direction corresponds to that of the observations. The complete procedure is given by Algorithm 1. The logarithmic transform in step 2 and the limitation to scales $\leq J$ ensure that the correction is consistent with our verification. By restoring the original mean and variance of the log-transformed field in step 10, we concentrate on the spatial structure without attempting to correct the margins as well.

Algorithm 1 is applied to all forecasts in the Alpine MesoVICT dataset; the reference in each case is the VERA analysis. Figure 12 shows four examples which illustrate the effects of our structural correction. In the first case Figure 12a–c, a forecast by the MOLO0225 model was deemed too small and too anisotropic. The algorithm smooths the field, rounds the linear pattern and visibly reduces small-scale variability. The result has near-perfect scale and direction properties, while maintaining the same arrangement of features as the original.

The second example Figure 12d–f depicts a situation in which the BOLAM007 model predicted a single large-scale rain band over the Alps, whereas VERA sees a number of smaller, disjointed cells. As expected, the correction converts the continuous rain area of the forecast into several smaller objects, thereby increasing the visual similarity with VERA.

While the previous two predictions were too anisotropic, Figure 12g–i show a forecast from the CMH model that was both too small-scaled and too round. After correction, much of the small-scale variability has disappeared, while the elongated shape in the centre of the domain has been rendered more coherent and linear.

As our final example, we have included a complete mis-forecast from the CMH model, which fails to simulate the front seen in VERA and produces scattered small-scaled precipitation across the Alps. This extreme example, which is clearly related to model spin-up

Algorithm 1. Correction of structure errors

```
Input: forecast $F$, reference $R$, largest scale $J$, minimum value $R_{\min}$
Output: corrected forecast $F'$
 1: Set values $<R_{\min}$ to zero
 2: Set $F \rightarrow \log_2(F + R_{\min})$, $R \rightarrow \log_2(R + R_{\min})$
 3: Standardize $F$ and $R$ to zero mean, unit variance
 4: Forward transform $dtF = \mathrm{dtcwt}(F)$, $dtR = \mathrm{dtcwt}(R)$
 5: for all scales $j = 1, \dots, J$, directions $d = 1, \dots, 6$ do
 6:     Calculate sum over all grid points $i$: $e_F = \sum_i |dtF_{i,j,d}|^2$, $e_R = \sum_i |dtR_{i,j,d}|^2$
 7:     set all $dtF_{i,j,d} \rightarrow dtF_{i,j,d} \cdot e_R / e_F$
 8: end for
 9: Inverse transform $F' = \mathrm{dtcwt}^{-1}(dtF)$
10: Reset mean and variance of $F'$ to the values before step 3
11: Set $F' \rightarrow 2^{F'} - R_{\min}$
12: Set values $< R_{\min}$ to zero
```

in CMH, serves to demonstrate the limitations of the algorithm. The global adjustment of the wavelet spectra cannot possibly create a cold front in which no precipitation has been simulated. Instead, the two largest cells at the western domain edge are united into a smooth, elongated feature; most of the remaining small-scale variability is removed.

Having seen that the correction algorithm produces realistic-looking fields while greatly improving the visual similarity between forecast and observation, we now quantify its influence on the verification results. As expected, Figure 13 shows that both directionality and scale, measured by *dxy* and *dz*, are greatly improved. The fact that these scores are not *exactly* zero is due to the (necessary) truncation step (Algorithm 1, step 12). In addition, the inverse wavelet transform used here (following Kingsbury 2006) is not perfect due to the special treatment of the diagonal directions. The improvement of the scores is nonetheless immense, indicating that these effects play no great role – the algorithm works as intended.

While the improvement in the wavelet scores is thus almost guaranteed by design, it is interesting to see whether beneficial effects on the structural forecast skill are observed by other verification methods as well. To this end, we apply the object-based structure score S of Wernli *et al.* (2008) (using 1/15 of the observed and predicted 90% quantiles as thresholds) and the variogram score *vgs* of Scheuerer and Hamill (2015). Following Buschow and Friederichs (2020), we use the stationary, isotropic, inverse-distance-weighted version of *vgs* with $p = 2$ and scale each field by its standard deviation to concentrate on verifying the correlation structure. The bottom panels in Figure 13 confirm that both the object-based S and the variogram score *vgs* measure a significant improvement after the wavelet-based structure correction. In particular, S originally also detects the substantial scale errors of MOLO0225 and CMH; after our adjustment, these models are deemed as good as COSMO and BOLAM007, both of which see a modest improvement in S as well.

9 | DISCUSSION

The central goal of this study is to present a verification technique that evaluates the predicted spatial structure in terms of scale, anisotropy and direction. This level of detailed structural analysis is enabled by the complex dual-tree wavelet transform of Kingsbury (1999), which comprises six directional filters on a range of spatial scales. Using data from the MesoVICT project, we have demonstrated that the DTCWT can indeed replace the classic discrete wavelet transform in an analysis of spatial scales. All previous results concerning the usefulness of such an analysis for spatial forecast verification (Kapp *et al.*, 2018; Buschow *et al.*, 2019; Buschow and Friederichs, 2020) or the quantification of convective organization (Brune *et al.*, 2018; 2020) remain

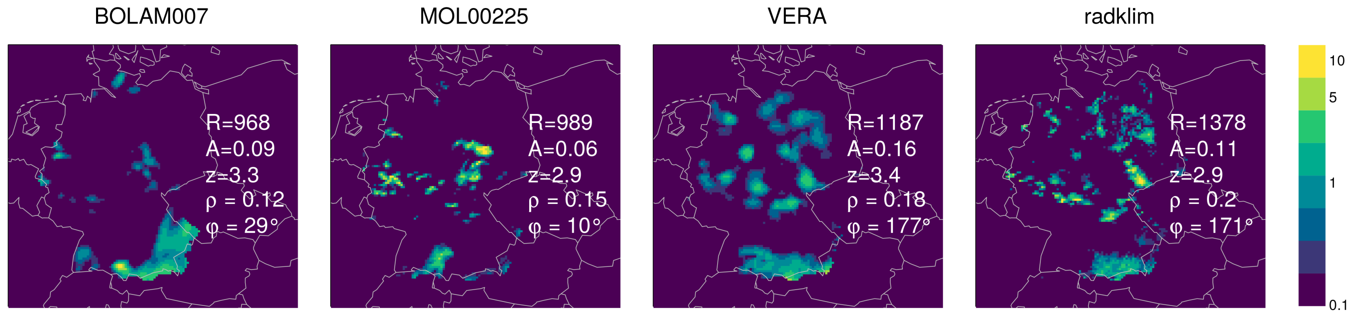


FIGURE 10 Predicted, analyzed and observed rain fields and structural characteristics on 9 July 2007 at 1900 UTC [Colour figure can be viewed at wileyonlinelibrary.com]

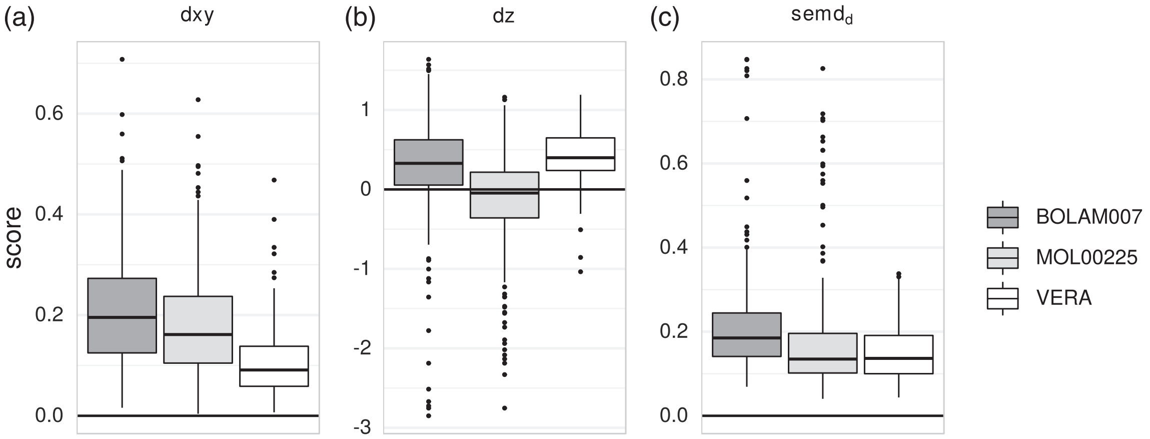


FIGURE 11 (a) Anisotropy, (b) scale and (c) total structure error of BOLAM007, MOLO0225 and VERA, verified against RADKLIM in the Germany domain. Only cases with at least 100 non-zero rain pixels in the RADKLIM image were included

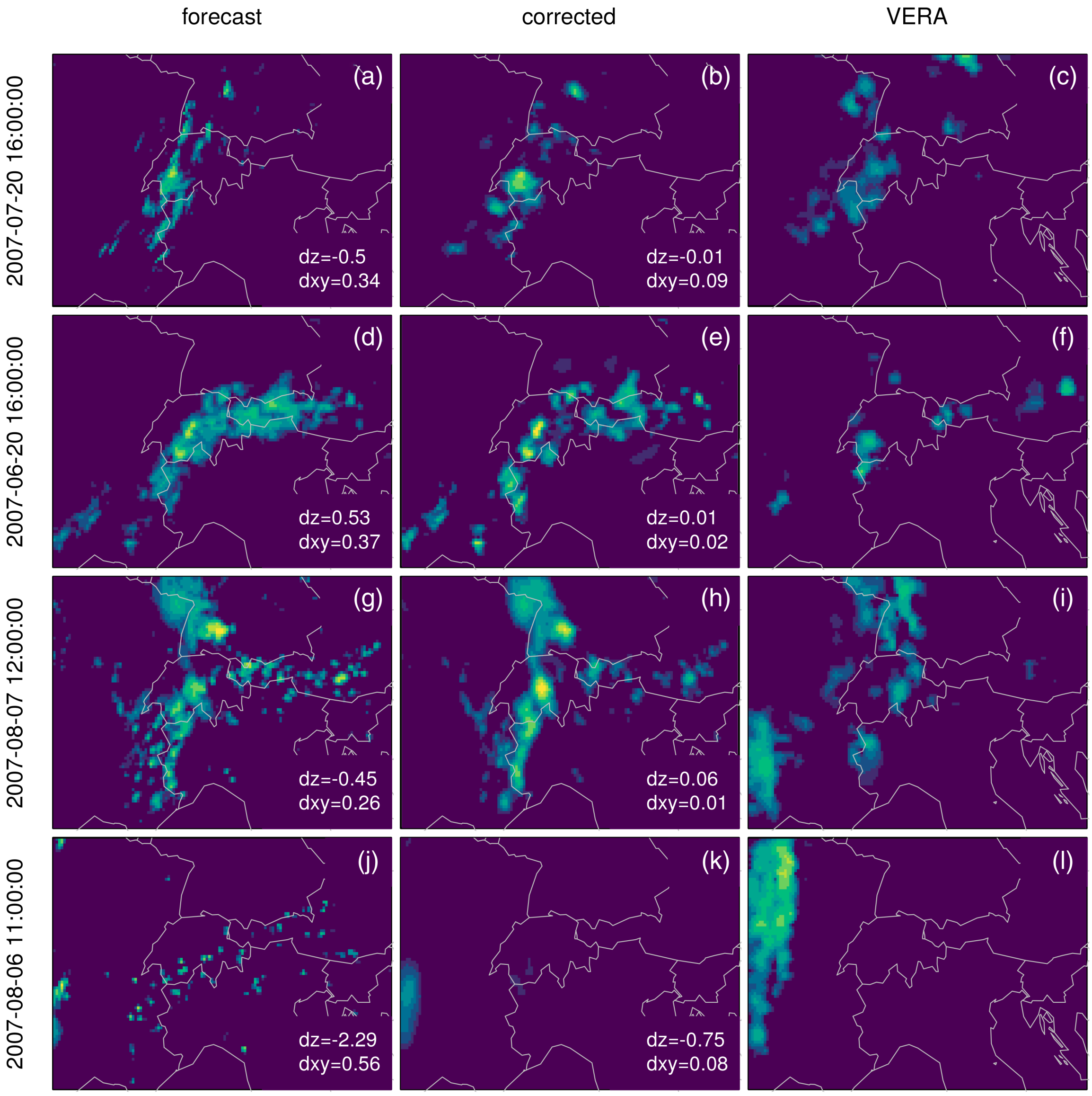


FIGURE 12 (a, d, g, j) Original forecasts, (b, e“h, k) corrected versions, and (c, f, i, l) the corresponding VERA analysis for four cases [Colour figure can be viewed at wileyonlinelibrary.com]

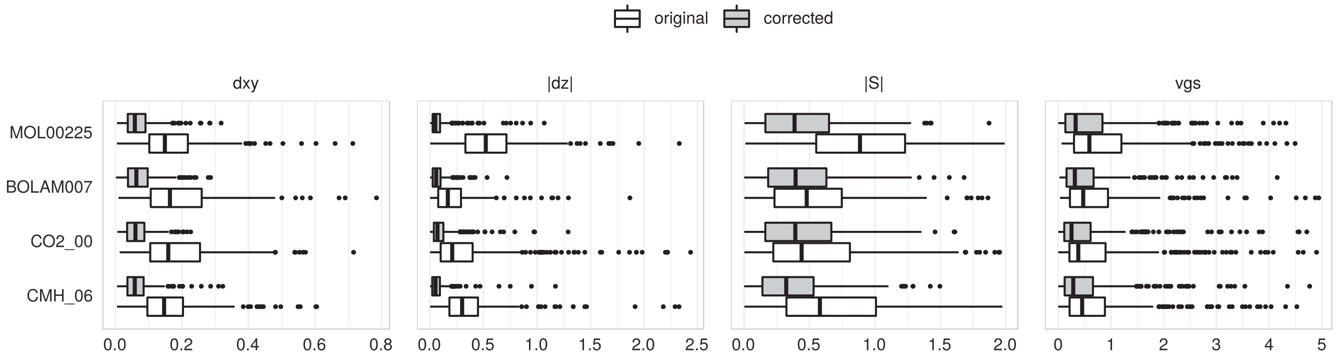


FIGURE 13 Distributions of structure scores for the original and corrected predictions of each model

valid for the dual-tree version. When only global characteristics are of interest, we can even replace the computationally expensive redundant transform by the extremely efficient decimated version. This potentially enables the use of our methods in contexts where time constraints, very large spatial domain sizes ($n_x > 1024$) or the sheer number of fields would make the redundant transform impractical. Incidentally, the decimated transform also allows for an efficient implementation of the global WOI of Brune *et al.* (2018).

However, the key innovation of the dual-tree approach lies in the analysis of the directional structure, which is impossible with the three directional filters of the classic DWT. Building on the idea of a central scale, we have introduced two further components to the centre of the wavelet spectrum, namely the degree of anisotropy ρ_c and the angle φ_c. The geometric MesoVICT test cases demonstrate that ρ_c adequately distinguishes between elongated and round patterns; the analysis of directions using φ_c is nearly flawless for simple geometric shapes.

For the purposes of verification, we advocate the use of a combined anisotropy–direction score since errors in φ_c are not meaningful when the anisotropy is low. This score, denoted here as *dxy* was used together with the signed difference in central scale *dz* and the combined structure score $semd_d$ to verify precipitation forecasts from four competing models within the MesoVICT framework. Perhaps the most important lesson from this experiment is that scale and directionality represent two independent aspects of forecast quality. Compared to VERA, CMH and COSMO achieve nearly the same average values of $semd_d$, but the composition of the structure score is very different: while COSMO simulates structures that are systematically too linear, leading to higher values of *dxy*, CMH produces excessive small-scale variability early in the day ($dz < 0$). MOLO0225 is tied for the lowest *dxy* with CMH but simulates far smaller structures than any other model. It may be interesting to note that the scale of the precipitation fields is not entirely determined by the model's nominal spatial resolution, which is nearly the same for CMH, COSMO and MOLO0225. On the other hand, BOLAM007 has a much coarser grid spacing of approximately 7 km and, somewhat expectedly, produces nearly the same distribution of spatial scales as VERA, which gives it the best overall scores in this comparison.

Should we thus conclude that the coarsest model delivers the most realistic representation of spatial structure? Doubting this, we have compared VERA, BOLAM007 and MOLO0225 to the radar-based RADKLIM dataset, which represents a realistic spatial observation of rainfall. Here MOLO0225, the overall loser in the previous comparison, is rewarded for simulating the same fine spatial scales as seen in the observations; both VERA and BOLAM007 are deemed too large. MOLO0225's representation of directionality is good but not as good as VERA's, again leaving two "forecasts" tied in terms of the overall score $semd_d$. This is a good example of independence between directional and scale-related aspects of the spatial structure: even if a comparatively coarsely resolved validation dataset hinders our analysis of spatial scales, we can nonetheless study the direction and directedness.

The degree of anisotropy and the distribution of dominant directions in a realistic precipitation field are closely tied to meteorological processes like organized or unorganized convection, moving airmass boundaries and pressure systems, as well as the interaction of these processes with the local topography. Our results demonstrate that the wavelet-based approach can meaningfully verify these directional aspects of spatial structure, and thereby indirectly the simulation of the underlying processes. Together with the analysis of scales and simple statistics such as the total rain area and accumulation, we can obtain a very detailed, objective picture of the spatial pattern in various observed and simulated fields.

An advantage of the wavelet-based approach is the existence of an inverse transform which allows us to correct the detected structural errors by a rather straightforward algorithm. The resulting post-processed fields combine the spatial placement of the predicted rain field with the global structure of the observations. We have shown that this procedure produces realistic-looking results with greatly improved wavelet and non-wavelet structure scores. The correction procedure has three main benefits. First, it enhances our intuitive understanding of the wavelet-based verification by showing us what an improved version of the forecast would have been. Second, the errors detected by the wavelet scores can be removed from the forecast before other scores are applied. In this manner, one can eliminate structure errors before verifying other aspects like the location of the predicted objects. Third, if a forecasting system exhibits strong systematic biases in its spatial structure (as was the case for MOLO0225 here), a correction to the observed climatological spectra could actually improve the value of the forecasts. How exactly such a structural post-processing could be implemented and whether it has any real-world utility is a question for future research.

ACKNOWLEDGEMENTS

Sebastian Buschow was funded by the DFG under grant no. FR 2976/2-1. We are grateful to Barbara Casati and one anonymous reviewer for their helpful criticism and encouraging remarks. Further thanks go to Eric Gilleland at NCAR and Stefano Mariani at ISPRA for providing the MesoVICT datasets. We also thank Alex Gibberd for his advice on the implementation of the bias correction for the dual-tree model and Rich Wareham for providing an open source Python implementation of the DTCWT. Open access funding enabled and organized by Projekt DEAL.

ORCID

Sebastian Buschow https://orcid.org/0000-0003-4750-361X

REFERENCES

Ament, F. and Arpagaus, M. (2009). dphase_cosmoch2: COSMO model forecasts (2.2 km) run by MeteoSwiss for the MAP D-PHASE project. Deutscher Wetterdienst, Offenbach, Germany. https://gisc.dwd.de/wisportal/showMetadata.jsp?xml=de.dkrz.wdcc.iso2150323; accessed 22 December 2020.

Bica, B., Steinacker, R., Lotteraner, C. and Suklitsch, M. (2007) A new concept for high resolution temperature analysis over complex terrain. *Theoretical and Applied Climatology*, 90, 173–183. https://doi.org/10.1007/s00704-006-0280-2

Brune, S., Kapp, F. and Friederichs, P. (2018) A wavelet-based analysis of convective organization in ICON large-eddy simulations. *Quarterly Journal of the Royal Meteorological Society*, 144, 2812–2829. https://doi.org/10.1002/qj.3409

Brune, S., Buschow, S. and Friederichs, P. (2020) Observations and high-resolution simulations of convective precipitation organization over the tropical Atlantic. *Quarterly Journal of the Royal Meteorological Society*. https://doi.org/10.1002/qj.3751

Buschow, S. (2020) SAD: Verify the Scale, Anisotropy and Direction of weather forecasts. R package version 0.1.3. https://CRAN.R-project.org/package=sad

Buschow, S. and Friederichs, P. (2020) Using wavelets to verify the scale structure of precipitation forecasts. *Advances in Statistical Climatology, Meteorology and Oceanography*, 6(1), 13–30. https://doi.org/10.5194/ascmo-6-13-2020

Buschow, S., Pidstrigach, J. and Friederichs, P. (2019) Assessment of wavelet-based spatial verification by means of a stochastic precipitation model (wv_verif v0.1.0). *Geoscientific Model Development*, 12(8), 3401–3418. https://doi.org/10.5194/gmd-12-3401-2019

Daubechies, I. (1992) *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM.

Davis, C., Brown, B. and Bullock, R. (2006) Object-based verification of precipitation forecasts. Part I: methodology and application to mesoscale rain areas. *Monthly Weather Review*, 134(7), 1772–1784. https://doi.org/10.1175/MWR3145.1

Dorninger, M., Gilleland, E., Casati, B., Mittermaier, M.P., Ebert, E.E., Brown, B.G. and Wilson, L.J. (2018) The setup of the MesoVICT project. *Bulletin of the American Meteorological Society*, 99(9), 1887–1906. https://doi.org/10.1175/BAMS-D-17-0164.1

Eckley, I.A., Nason, G.P. and Treloar, R.L. (2010) Locally stationary wavelet fields with application to the modelling and analysis of image texture: modelling and analysis of image texture. *Journal of the Royal Statistical Society, Series C*, 59, 596–616. https://doi.org/10.1111/j.1467-9876.2009.00721.x

Ekström, M. (2016) Metrics to identify meaningful downscaling skill in WRF simulations of intense rainfall events. *Environmental Modelling & Software*, 79, 267–284. https://doi.org/10.1016/j.envsoft.2016.01.012

Gilleland, E., Ahijevych, D., Barbara G. Brown, Casati, B. and Ebert, E.E. (2009) Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, 24(5), 1416–1430. https://doi.org/10.1175/2009WAF2222269.1

Gilleland, E., Skok, G., Brown, B.G., Casati, B., Dorninger, M., Mittermaier, M.P., Roberts, N. and Wilson, L.J. (2020) A novel set of geometric verification test fields with application to distance measures. *Monthly Weather Review*, 148(4), 1653–1673

Han, F. and Szunyogh, I. (2016) A morphing-based technique for the verification of precipitation forecasts. *Monthly Weather Review*, 144(1), 295–313. https://doi.org/10.1175/MWR-D-15-0172.1

Kapp, F., Friederichs, P., Brune, S. and Weniger, M. (2018) Spatial verification of high-resolution ensemble precipitation forecasts using local wavelet spectra. *Meteorologische Zeitschrift*, 27(6), 467–480. ISSN 0941-2948

Keil, C. and Craig, G.C. (2007) A displacement-based error measure applied in a regional ensemble forecasting system. *Monthly Weather Review*, 135(9), 3248–3259. https://doi.org/10.1175/MWR3457.1

Kingsbury, N. (1999) Image processing with complex wavelets. *Philosophical Transactions of the Royal Society of London, Series A*, 357(1760), 2543–2560. https://doi.org/10.1098/rsta.1999.0447

Kingsbury, N. (2006). *Rotation-invariant local feature matching with complex wavelets, in 14th European Signal Processing Conference, Florence, Italy*. Piscataway, NJ: IEEE.

Mallat, S.G. (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7), 674–693. http://dx.doi.org/10.1109/34.192463.

Mariani, S. and Casaioli, M. (2018) Effects of model domain extent and horizontal grid size on contiguous rain area (CRA) analysis: a MesoVICT study. *Meteorologische Zeitschrift*, 27(6), 481–502. https://doi.org/10.1127/metz/2018/0897. ISSN 0941-2948

Marzban, C. and Sandgathe, S. (2009) Verification with variograms. *Weather and Forecasting*, 24(4), 1102–1120. https://doi.org/10.1175/2009WAF2222122.1

McTaggart-Cowan, R. (2009). Metadata for dphase_cmcgemh: regional GEM model high resolution forecast run by CMC for the MAP D-PHASE project. World Data Center for Climate (WDCC) at DKRZ. https://doi.org/10.1594/WDCC/dphase_cmcgemh

Nelson, J.D.B., Gibberd, A.J., Nafornita, C. and Kingsbury, N. (2018) The locally stationary dual-tree complex wavelet model. *Statistics and Computing*, 28(6), 1139–1154. https://doi.org/10.1007/s11222-017-9784-0

Nerini, D., Besic, N., Sideris, I., Germann, U. and Foresti, L. (2017) A non-stationary stochastic ensemble generator for radar rainfall fields based on the short-space Fourier transform. *Hydrology and Earth System Sciences*, 21(6), 2777–2797. https://doi.org/10.5194/hess-21-2777-2017

Rubner, Y., Tomasi, C. and Guibas, L.J. (2000) The Earth Mover's Distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2), 99–121

Scheuerer, M. and Hamill, T.M. (2015) Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4), 1321–1334. https://doi.org/10.1175/MWR-D-14-00269.1

Scovell, R.W. (2020) Applications of directional wavelets, universal multifractals and anisotropic scaling in ensemble nowcasting: a review of methods with case studies. *Quarterly Journal of the Royal Meteorological Society*. https://doi.org/10.1002/qj.3780

Selesnick, I.W., Baraniuk, R.G. and Kingsbury, N.C. (2005) The dual-tree complex wavelet transform. *IEEE Signal Processing Magazine*, 22(6), 123–151. https://doi.org/10.1109/MSP.2005.1550194

Torrence, C. and Compo, G.P. (1998) A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1), 61–78

Urbanek, S. and Rubner, Y. (2012) *emdist: Earth Mover's Distance*. https://CRAN.R-project.org/package=emdist. R package version 0.3-1.

Weniger, M. and Friederichs, P. (2016) Using the SAL technique for spatial verification of cloud processes: a sensitivity analysis. *Journal of Applied Meteorology and Climatology*, 55(9), 2091–2108. https://doi.org/10.1175/JAMC-D-15-0311.1

Wernli, H., Paulat, M., Hagen, M. and Frei, C. (2008) SAL – A novel quality measure for the verification of quantitative precipitation forecasts. *Monthly Weather Review*, 136(11), 4470–4487. https://doi.org/10.1175/2008MWR2415.1

Willeit, M., Amorati, R., Montani, A., Pavan, V. and Tesini, M.S. (2015) Comparison of spectral characteristics of precipitation from radar estimates and COSMO-model predicted fields. *Meteorology and Atmospheric Physics*, 127(2), 191–203. https://doi.org/10.1007/s00703-014-0359-8

Winterrath, T., Brendel, C., Mario, H., Junghänel, T., Klameth, A., Walawender, E., Weigl, E. and Becker, A. (2018). RADKLIM version 2017.002: reprocessed gauge-adjusted radar data, one-hour precipitation sums (RW).

Wong, M. and Skamarock, W.C. (2016) Spectral characteristics of convective-scale precipitation observations and forecasts. *Monthly Weather Review*, 144(11), 4183–4196. https://doi.org/10.1175/MWR-D-16-0183.1

How to cite this article: Buschow S, Friederichs P. SAD: Verifying the scale, anisotropy and direction of precipitation forecasts. *QJR Meteorol. Soc*. 2021;1–21. https://doi.org/10.1002/qj.3964

APPENDIX A. HEXAGONAL ARRANGEMENT AND THE EMD

To further illustrate the concepts introduced in Section 4, we consider the wavelet spectra corresponding to the rain fields shown in Figure 12l,j. In Figure A1a, b, we have listed the energy values corresponding to each combination of scale j and direction d. For our calculations, each of these energies $e_{j,d}$ is treated as a point mass located at the coordinates $x_{j,d}, y_{j,d}, z_{j,d}$ (cf. Equation 1). The point masses are visualized as spheres of different volumes in Figure A1c, d. The frontal structure from Figure 12l leads to a concentration of mass at scale five and the directions around 90°. Conversely, the small, isotropic pattern of Figure 12j is reflected by a more even distribution across all directions and the three smallest scales.

ρ_c, θ_c and z_c are the barycentre of this arrangement of point masses, represented in cylindrical coordinates. Scores like $d\rho$ and dz are simply given by the difference between the central coordinates of two spectra. These simple scores are useful because they are easy to interpret, but they neglect some information on the full distribution of energy (Buschow *et al.*, 2019). To define a summary score that includes all information from the mean spectrum, we therefore use the Earth Mover's Distance (EMD) which measures the minimum total cost of transforming one arrangement of point masses into another (Rubner *et al.*, 2000). Let $m = 1, \ldots, 6J$ be an index enumerating all combinations of scale and direction (j, d). One spectrum is transformed into another by successively transferring amounts of "mass" (in our case spectral energy) $f_{m \rightarrow n} > 0$ from locations m in the first spectrum to locations n in the second spectrum. Recalling that our "masses" are normalized such that $\sum_i e_i^{(k)} = 1$ ($k = 1, 2$ denoting the first and second spectrum), we seek a set of mass transfers which satisfy

$$\sum_m f_{m \to n} = e_n^{(2)} \quad \text{(the result of the transport is spectrum 2)} ,$$

$$\sum_n f_{m \to n} = e_m^{(1)} \quad \text{(all mass from spectrum 1 is transported somewhere)} .$$

Denoting the Euclidean distance between locations m and n by

$$d_{m,n} = \sqrt{(x_m - x_n)^2 + (y_m - y_n)^2 + (z_m - z_n)^2} ,$$

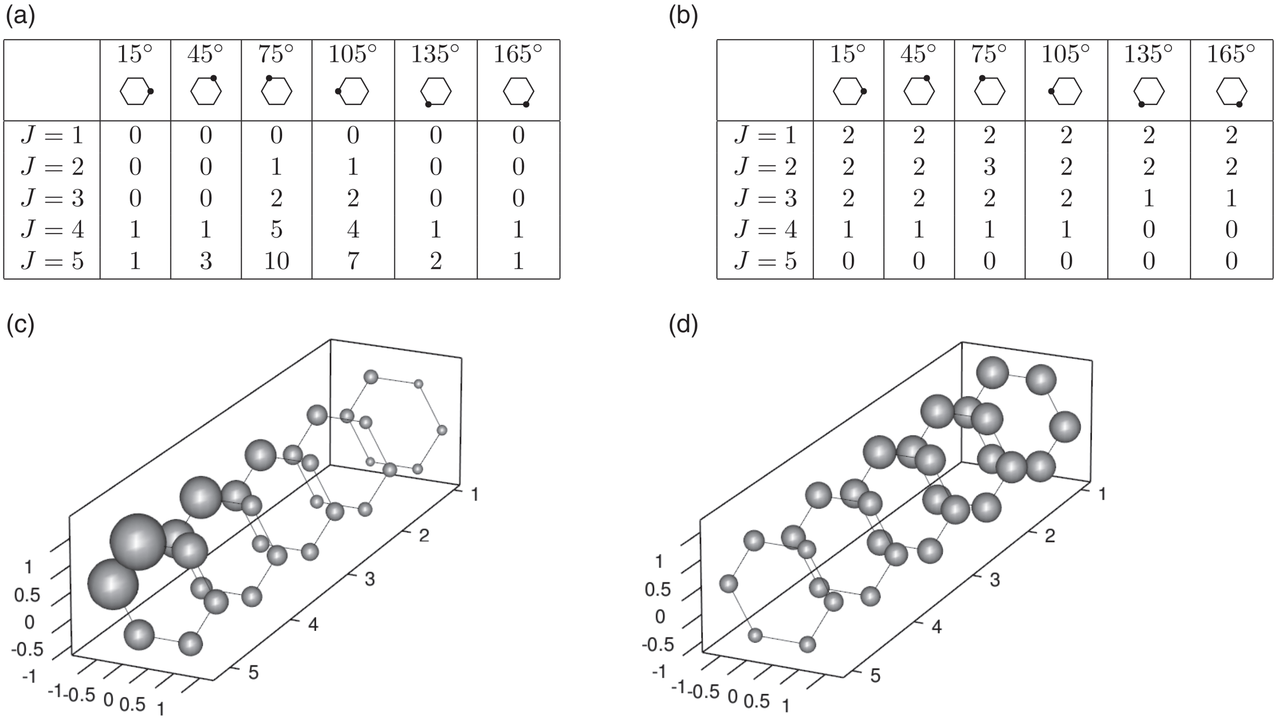
we can write the total work of a particular transport scheme as $\sum_{m,n} d_{m,n} f_{m \to n}$. The EMD is given by the minimum work needed, i.e.,

$$semd_d = \min \left(\sum_{m,n=1}^{6 \cdot J} d_{m,n} f_{m \to n} \right) , \qquad \text{(A1)}$$

where the minimum is taken over all possible sets of transports $(f_{m \to n})_{m,n}$ that satisfy the requirements above. The solution of the optimization problem is found numerically via the `emdist` R-package (Urbanek and Rubner, 2012).

APPENDIX B. COMPARISON BETWEEN WAVELET TRANSFORMS

To quantify the impact of our choice of wavelet transform on the resulting structure analysis, we transform each field in the original MesoVICT dataset (Alpine domain, VERA, CMH, COSMO, BOLAM007, MOLO0225) three times: once with the decimated DTCWT which was used throughout Sections 6 and 7, once with the redundant version (used to produce Figure 6) and a third time with the sixth "Extremal Phase" Daubechies wavelet (redundant version with bias correction). Figure B1a shows the central scales z_c resulting from the spatially averaged, bias-corrected wavelet spectra. Apart from a slight linear offset, the agreement between the three analyses is close to perfect ($R^2 \approx 0.99$), and we observe no surprising outliers and no nonlinear effects. This confirms our claim that the DTCWT analyses scales in nearly exactly the same way as the usual DWT used by Buschow and Friederichs (2020). For the anisotropy ρ_c, we only compare the decimated DTCWT to its undecimated version since the DWT is not expected to agree with the dual-tree results here. Figure B1b shows almost no systematic bias; the correlation is again very high ($R^2 \approx 0.95$). We conclude that the global scale and anisotropy can be inferred from the decimated DTCWT just as well as from the undecimated version without incurring any significant, systematic double penalty.


(a)

	15°	45°	75°	105°	135°	165°
$J = 1$	0	0	0	0	0	0
$J = 2$	0	0	1	1	0	0
$J = 3$	0	0	2	2	0	0
$J = 4$	1	1	5	4	1	1
$J = 5$	1	3	10	7	2	1

(b)

	15°	45°	75°	105°	135°	165°
$J = 1$	2	2	2	2	2	2
$J = 2$	2	2	3	2	2	2
$J = 3$	2	2	2	2	1	1
$J = 4$	1	1	1	1	0	0
$J = 5$	0	0	0	0	0	0

FIGURE A1 Mean wavelet spectra for the fields in Figure 12l,j. (a"b") Energy values, re-scaled and rounded to integers between 0 and 10. (c, d) Representation in a hexagonal arrangement (Figure 4); the volume of the spheres is proportional to the energy listed in the table

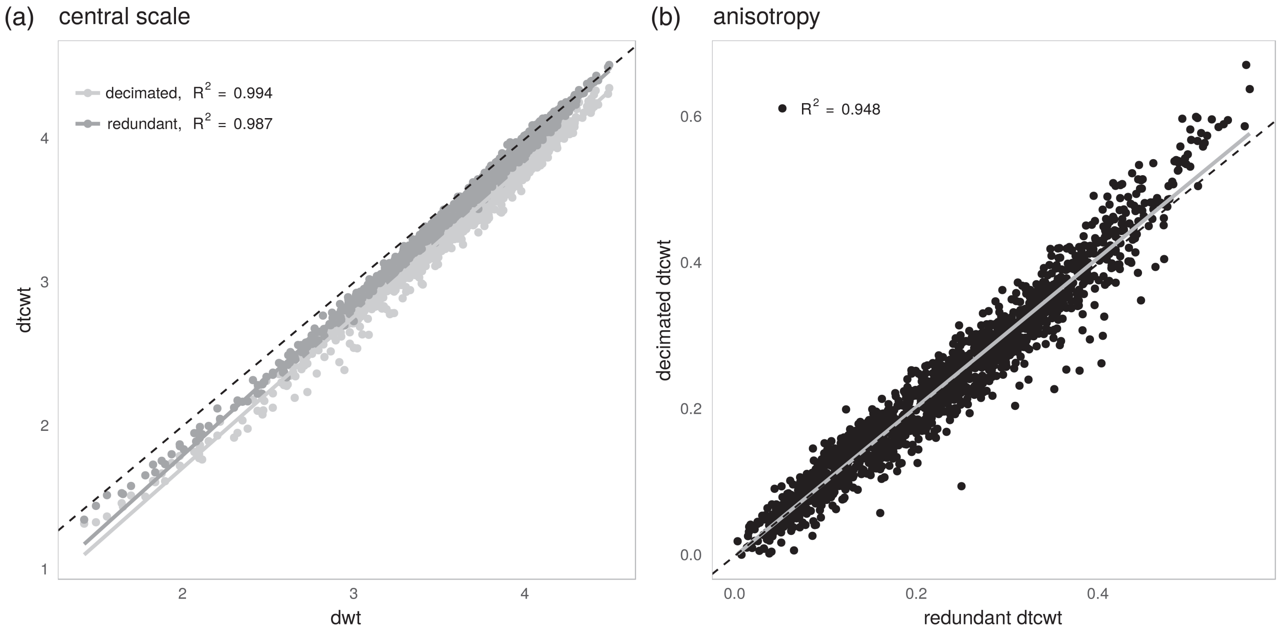


FIGURE B1 Dependence of structural characteristics on the wavelet transform. (a) central scale z_c of the DWT (DB6) against that of the decimated and redundant DTCWT. (b) Anisotropy ρ_c of the redundant and decimated DTCWT. The solid line indicates linear regression, and the dashed line marks the unit diagonal

Appendix D

Buschow and Friederichs 2021b

Verification of Near Surface Wind Patterns in Germany using Clear Air Radar Echoes

Sebastian Buschow[1] and Petra Friederichs[1]

[1]Institute of Geosciences, University of Bonn, Bonn, Germany

Correspondence: Sebastian Buschow (sebastian.buschow@uni-bonn.de)

Abstract. The verification of high-resolution meteorological models requires highly resolved validation data and appropriate tools of analysis. While much progress has been made in the case of precipitation, wind fields have received less attention, largely due to a lack of spatial measurements. Clear-sky radar echoes could be an unexpected part of the solution by affording us an indirect look at horizontal wind patterns: Regions of horizontal convergence attract non-meteorological scatterers such as insects; their concentration visualizes the structure of the convergence field. Using a two-dimensional wavelet transform, this study demonstrates how divergences and reflectivities can be quantitatively compared in terms of their spatial scale, (horizontal) anisotropy and direction. A long-term validation of the highly resolved regional reanalysis COSMO-REA2 against the German radar composite RADOLAN shows surprisingly close agreement. Despite theoretically predicted problems with simulations in or near the 'grey-zone' of turbulence, COSMO-REA2 is shown to produce a realistic diurnal cycle of the spatial scales larger than 8km. In agreement with the literature, the orientation of the patterns in both data-sets closely follows the mean wind direction. Conversely, an analysis of the horizontal anisotropy reveals that the model has an unrealistic tendency towards highly linear, roll like patterns early in the day.

1 Introduction

Modern numerical weather models at horizontal resolutions on the order of $1-10\,km$ are generally believed to be useful, but their added value compared to coarser models is not easy to quantify. On the one hand, the precise placement of very small features continues to be largely unpredictable. In a gridpoint-by-gridpoint comparison, highly resolved models are punished twice for slight location errors in features which coarser models do not attempt to simulate at all. On the other hand, a single error value summarizing the realism of a highly complex meteorological field is not very informative. To address these issues, a large variety of so-called spatial verification techniques has been developed in recent years. A first systematic survey of the field was undertaken in the spatial forecast verification Inter-Comparison Project (Gilleland et al., 2009, ICP). At this point, almost all efforts were focused on the verification of precipitation forecasts, for several reasons: Firstly, the improved representation of convective precipitation was a main incentive for the development of mesoscale weather models. Secondly, the intermittent nature of rain fields makes the aforementioned double-penalty problem particularly obvious. Lastly, radar (and to a lesser degree, satellite) observations readily provide high-resolution spatial observations of precipitation.

The second phase of the ICP project (Dorninger et al., 2018, MesoVICT) has highlighted the need for a spatial verification of other meteorological variables, particularly wind: Wind fields at kilometer resolutions can produce highly complex patterns with potential impacts on convective initiation, wind energy, air quality and aviation safety. The task of verifying spatial wind forecasts poses practical, methodological and theoretical challenges.

From a practical point of view, we face a lack of spatial observations: Model analyses (e.g. used for wind verification by Zschenderlein et al. (2019)) conveniently provide highly resolved, gap-free data but the realism of the underlying model would have to be verified against some other data beforehand. Interpolated station data (for example the VERA analysis used within MesoVICT) are generally too coarsely resolved to represent structures on the scale of single kilometers, denser station networks such as the WegenerNet data-set used by Schlager et al. (2019) are rare. Bousquet et al. (2008) and Beck et al. (2014) use Multi-Doppler wind retrievals from the French national radar network to verify wind predictions from the AROME model. This approach is very appealing but limited to cases with precipitation. In addition, Doppler-derived wind composites are not yet widely available.

Skinner et al. (2016) present a very interesting alternative using single-Doppler azimuthal wind shear as a proxy for low-level rotation. Their study also highlights some of the main methodological challenges related to wind verification: Most spatial verification techniques were developed for scalar quantities which can be decomposed into discrete objects via thresholding. How should such techniques be adapted to vector fields where non-zero variability is present at every location and the existence of well defined objects is not guaranteed? Skinner et al. (2016), who are interested in tornado forming mesocyclones, chose to focus on the rotational component of the wind field by verifying only the horizontal vorticity. Model and observations are subjected to several spatial filters and then thresholded at manually selected values before the object based MODE technique (Davis et al., 2009) and the image-morphing DAS of Keil and Craig (2009) are applied. Their approach is justified because well-defined objects, i.e., tornadic supercells, clearly exist in the specific case study under consideration. Bousquet et al. (2008) find a similar answer to the vector-problem by verifying horizontal divergences against the corresponding values from the French Multi-Doppler network. Besides point-wise measures, these authors apply a simple scale-separation approach based on a Haar wavelet decomposition of the wind fields. Other recent attempts at spatial wind verification include Zschenderlein et al. (2019) who apply the object-based SAL technique (Wernli et al., 2008) to tresholded predictions of gusts (i.e. absolute wind speed), and Skok and Hladnik (2018) who sort wind vectors into classes based on their speed and direction and use the popular fractions skill score (Roberts and Lean, 2008, FSS) to find the scales on which the predicted classes agree with the observations.

In this study, we take a similar route as Skinner et al. (2016) but instead of the rotational component we focus on the horizontal divergence of the near surface wind field. Under the right environmental conditions, the spatial pattern of this divergence field can be observed in widely available radar reflectivity data: On warm, rain-free days, convergent boundary layer circulations attract swarms of insects which are drawn in and actively attempt to resist the vertical motion of updraughts (Wilson et al., 1994). The resulting increased concentration of biological scatterers within the radar beam reflects the pattern of convergence and divergence. Numerous studies including Weckwerth et al. (1997, 1999); Thurston et al. (2016); Banghoff et al. (2020) have used this kind of data to study the dominant patterns of boundary layer organization. Atkinson and Wu Zhang

(1996) identified mesoscale shallow convection, organized in the form of cells or horizontal rolls, as the most prominent of those patterns. Numerous studies have used radar data to observe these phenomena (see references in Banghoff et al. (2020)); Banghoff et al. (2020) also present a first long-time climatology using ten years of reflectivities and Doppler velocities from a single radar station in Oklahoma. They manually classified radar images from over 1000 days into cells, rolls and unorganized patterns, reporting organized features on 92 % of summer days without rain. Santellanes et al. (2021) exploited this data-set to study the environmental conditions that favor the different modes of organization.

In the present investigation, we aim to study a similarly large data-base of reflectivities from the German RADOLAN-RX composite and compare it to divergence structures from the regional reanalysis COSMO-REA2 (Wahl et al., 2017), covering the time-span from 2007 to 2013. We limit our analysis to small environments around each radar station and consider both the entire COSMO-REA2 time-series (for an overall model climatology) and the subset where clear air radar echoes are available (for verification).

For a fair, quantitative validation of the model, the spatial patterns must be analyzed objectively. Here, we rely on the wavelet-based SAD verification methodology of Buschow and Friederichs (2021) which applies a series of directed filters to objectively determine the dominant spatial Scale, Anisotropy and Direction in an image. A closely related approach was used to define a wavelet-based index of convective organization in radar and satellite images by Brune et al. (2021).

To what extent a model at $\mathcal{O}(1km)$ horizontal resolution can be expected to realistically represent boundary layer circulations in the so-called 'Grey-Zone' regime (Wyngaard, 2004) between parametrized and resolved turbulence is a difficult question which poses further theoretical challenges to the verification process. Section 2 therefore briefly summarizes some of the relevant theoretical and experimental results from the literature. Data and methodology are described in sections 3 and 4. Section 5 presents the results of our analysis, including the model-based climatology of divergence structures and its validation against RADOLAN. Some discussion of our findings is given in section 6, section 7 examines what conclusions can be drawn and identifies avenues for future research.

2 Theory and modelling of mesoscale shallow convection

Zhou et al. (2014) have demonstrated how occurrence and basic properties of shallow convective circulation in the atmospheric boundary layer can be understood in analogy to Rayleigh Bénard thermal instability. In the classic framework, the circulation regime of a fluid between two heated plates is determined by the Rayleigh number

$$\mathrm{Ra} = \frac{g\alpha}{k\nu} \cdot \beta d^4\,, \tag{1}$$

where d is the distance between the plates, $\beta = dT/dz$ is the temperature gradient, and the coefficients g, α, k, ν denote gravitational acceleration, thermal expansion coefficient, thermal conductivity and kinematic viscosity, respectively. Eddies of wavelength λ start to grow when Ra exceeds a critical value $\mathrm{Ra}_c(\lambda)$. The qualitative sketch in figure 1 shows that this marginal stability curve has a global minimum near $\lambda = 2d$. For Ra<$\mathrm{Ra}_c(2d)$, the flow is laminar and heat is exchanged via conduction. When Ra is increased to $\mathrm{Ra}_c(2d)$, convective cells are initiated with a wavelength of roughly twice the depth of the fluid. Zhou

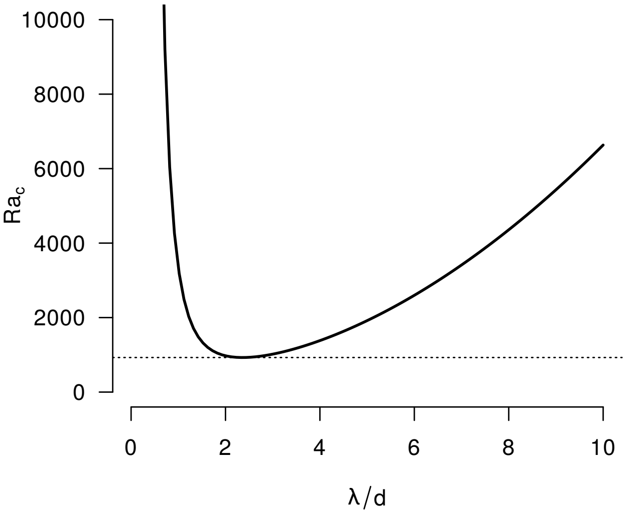


Figure 1. Marginal stability curve of Rayleigh-Bénard convection for the classic rigid-rigid boundary conditions. For any given wavelength λ (relative to the fluid depth d), Eddies grow if the Rayleigh number lies above the curve and decay otherwise.

et al. (2014) argue that an analogous stability curve applies to the atmospheric boundary layer. In this case, Ra is replaced by a turbulent Rayleigh number of similar form as Eq. 1 wherein the depth d is replaced by the boundary layer height H. On a sunny day, the earth's surface is heated and the vertical temperature gradient, as well as the height of the boundary layer increase. The theory predicts that, once a critical Ra is crossed, the initial wavelength of the circulation should be near $\lambda = 2H \approx 3km$; both smaller and larger eddies begin to develop later.

The simulation of this process is challenging because a model with grid-spacing δ can never resolve eddies with $\lambda < 2\delta$. In large eddy simulations with $\delta << 2H$, convection will correctly be initiated at the natural critical Ra$_c$ with a wavelength of $\sim 2H$. Current NWP models, on the other hand, have $\delta \gtrsim 2H$. In this case, the first eddies to form as Ra increases have $\lambda \approx \delta$ and initiate at a grid-spacing dependent value Ra$_c(\delta)$. For global or regional models with $\delta \gtrsim 10\,km$, the critical value is so large that such circulations will never form under realistic conditions. Modern mesoscale models, however, operate at $\delta = \mathcal{O}(1km)$ and Ra$_c(\delta)$ becomes attainable. The result is a potentially unrealistic model circulation, the scale and initiation time of which depends on δ. This is one example of the so-called *Terra Incognita* or Grey-Zone of turbulence (Wyngaard, 2004; Honnert et al., 2020), where the highest energy vortices are too large to be adequately represented by the boundary layer parametrization but too small to be explicitly resolved by the dynamical core of the model. Ching et al. (2014) observed this phenomenon in nested WRF simulations, Poll et al. (2017) detected it in TerrSysMP, the atmospheric component of which is COMSO. Using LES runs of the same models as a reference, both of these studies found that simulations with grid spacing on the kilometer scale initiate turbulence too late and too energetically. In the present study, we will investigate how frequently such small-scaled circulations occur in the climatology of COSMO-REA2 and how they compare to radar observations.

3 Data

3.1 COSMO-REA2

For a systematic investigation of low-level divergence structures, we ideally need a long, homogeneous time series of high resolution model data. The regional reanalysis COSMO-REA2 is uniquely suited for our need as it provides seven years (2007-2013) of hourly output from the mesoscale model COSMO (Baldauf et al., 2011) at a horizontal resolution of $0.018°$ or roughly 2 km. The model was run with 50 vertical levels over a domain covering Germany and the neighbouring countries. For a full description of the used physics parametrizations, we refer to Wahl et al. (2017) and references therein. For our purposes, it is important to note that boundary layer fluxes are handled by a level-2.5 TKE-closure, shallow convection is parametrized via a modified Tiedtke mass-flux scheme while deep moist convection is left to the dynamic core. The data assimilation uses a continuous nudging scheme to relax the prognostic temperature, wind speed, pressure and relative humidity towards observations from stations, radiosondes, aircraft, ships and buoys. In addition, rain rates from radar observations are assimilated via latent heat nudging (Stephan et al., 2008, LHN). Thus, on clear air days, the only source of mesoscale information (LHN) is inactive, meaning that while data assimilation can help create realistic environmental conditions, the fine-scale structure of the fields is a product of the dynamics and physics of the model. Horizontal divergences were calculated from the hourly 10 m wind vector fields as a simple finite difference approximation.

3.2 RADOLAN RX

RADOLAN (Radar online adjustment, 'RADar OnLine ANeichung') RX is the operational radar reflectivity composite of the 16 C-band radars operated by the German weather service. The output has a spatio-temporal resolution of $1\,km \times 1\,km \times 5\,min$ and covers Germany and parts of its neighbours. The underlying radar scans are performed at an orography following elevation angle ($\sim 1°$) with an azimuthal resolution of $1°$ and a range resolution of 250 m. Due to the beam geometry, the true native resolution of the reflectivity composite, as well as the height for which it is representative, depends heavily on the distance to the radar station. Pejcic et al. (2020) show that the beams reach typical boundary layer heights of $1 - 1.5$ km at about 100 km from the radar location. Therefore, relevant clear-air echoes caused by insects that cannot survive at low temperatures are expected to be found only in the immediate vicinity of the radars.

To get an idea of the type of data we rely on for our model validation, it is instructive to consider an example case. Figure 2 (a) displays the RADOLAN RX composite at noon on 2009-07-29. Aside from a few showers over the North Sea, no appreciable precipitation was observed in Germany on this warm summer day. Temperatures reached values in the high twenties and a high pressure system centred near the German-Polish border generated weak south-easterly flow. Despite the absence of rain, most radars in the composite are surrounded by a disk of low but non-zero reflectivities (-10 to 5 dBZ). While the size, shape and mean intensity of the disks varies, a consistent fine-scaled cellular pattern can be observed throughout central, northern and eastern Germany. Moreover the regions of increased reflectivity are coherently organized in a line-like fashion along a SW/NE direction. Figure 2 (b), showing the corresponding wind and divergence field from COSMO-REA2, reveals that the orientation of the reflectivity lines is broadly consistent with the overall direction of low level flow. Furthermore, the divergence field is

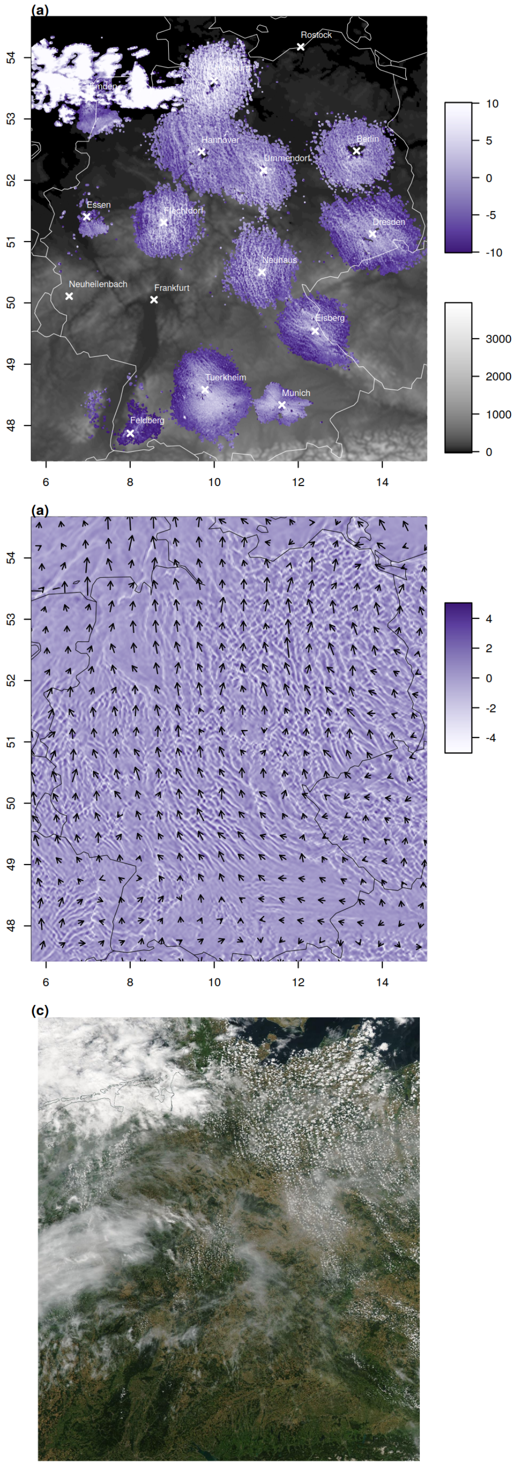


Figure 2. RADOLAN RX reflectivity in dBZ (a), COSMO-REA2 10 m divergence (b) and AQUA MODIS satellite image (c) on 2009-07-29 12:00 UTC.

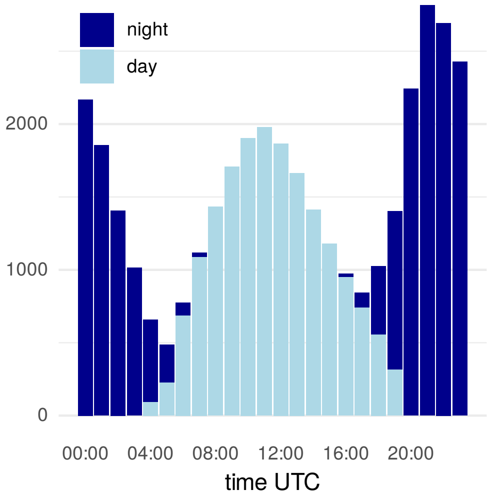


Figure 3. Number of complete clear air radar echoes at the twelve selected radars, separately for night and day as defined by sunrise and sunset.

characterized by small scaled patterns of cells and lines with alternating convergence and divergence, the size and orientation of which roughly resembles the radar pattern. Throughout eastern Germany, where the divergences are strongest, the satellite image in panel (c) shows the typical chains of Cumulus clouds often associated with mesoscale shallow convection (Atkinson and Wu Zhang, 1996). A visual comparison of the reflectivities around, for example, the Berlin radar with the simulated divergences and the clouds in that region leads us to hypothesize that the boundary layer processes in COSMO-REA2 are not entirely unrealistic.

3.3 Data availability

As mentioned above, clear-air echoes typically only occur in a small environment around each radar. We therefore limit our study to circular regions with $64\,km$ radius, centred at the 16 radar station which were active throughout the COSMO-REA2 period. While simulated divergences are readily available at every such grid point for each hour between 2007 and 2013, the availability of clear-sky echoes depends on many factors including local topography, technical details of the radars, radar processing at DWD and the life-cycle of the biological scatterers. We consider an individual radar image incomplete if less than $50\,\%$ of pixels within our $64\,km$ radius around the radar are above $-10\,dBZ$ (visual analysis of many example images has shown that no significant signals exist between roughly $-10\,dBZ$ and the smallest stored value of $-32.5\,dBZ$). From the remaining data, we must filter out rainy episodes, defined here somewhat arbitrarily as cases where at least 100 pixels exceed $+10\,dBZ$. We will refer to all remaining images as *complete*.

Table 1 shows that such complete clear air echoes are overall rare (well below $5\,\%$ of all hourly images) and their frequency varies considerably between radars. For this study, we neglect the four radar stations with the fewest data, thereby removing

Table 1. Number of hourly incomplete, rainy, nighttime and complete daytime hourly radar images per station. The top four radars are excluded from further analysis.

	incomplete	rain	night	day
Frankfurt	54841	6335	41	104
Emden	56064	5065	60	132
Essen	54229	6889	58	145
Rostock	54627	6059	295	340
Hamburg	53556	6866	351	548
Munich	51315	9131	181	694
Feldberg	52143	7419	855	904
Ummendorf	52806	6847	666	1002
Neuhaus	50355	8357	1527	1082
Berlin	52075	6935	1011	1300
Flechtdorf	49117	9033	1830	1341
Hannover	49199	8846	1478	1798
Eisberg	48088	9154	2100	1979
Tuerkheim	45576	10672	3044	2029
Neuheilenbach	47286	8731	3107	2197
Dresden	45787	9462	3122	2950

Table 2. Number of complete hourly non-rainy daytime radar echoes at the twelve selected radar stations.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2007	12	1	34	191	446	797	293	147	81	90	11	1
2008	3	1	24	13	212	808	1333	124	87	38	6	11
2009	16	12	38	26	264	209	1379	892	406	66	0	34
2010	36	73	52	45	74	541	1684	171	84	5	3	24
2011	2	31	13	145	210	716	741	190	139	59	2	7
2012	10	27	44	22	211	289	750	301	91	16	3	2
2013	53	18	53	46	65	318	1740	509	93	7	2	1

two urban (Essen, Frankfurt) and two coastal locations (Emden, Rostock). The twelve remaining radars give us roughly 20 thousand individual hourly images for comparison with COSMO-REA2. When studying the diurnal cycles below, we will furthermore include radar data at the full 5 min resolution which gives us over 200 thousand images.

In table 2, we see that the vast majority of clear sky echoes occurs during summer, particularly June and July, with considerable variability between the years. The preference for the warm season is expected since both insect activity and boundary layer height are increased by higher temperatures. Consequently, the daytime frequency of available data follows a diurnal cycle as well (figure 3). In addition, there is a large second population of night time cases. The sudden increase in clear air echoes at dusk, as well as their absence in winter, hints at migrating swarms of insects as a likely explanation (Drake and Reynolds, 2012). We exclude these data because (1) the weaker nighttime convergences are less likely to influence the pattern of the insect cloud and (2) migrating swarms tend to inhabit thin layers near an atmospheric inversion which only partly intersect the radar beam (cf. p.237 f. in Drake and Reynolds (2012)).

4 Methods

4.1 Wavelet analysis

The idea of this study is to compare the correlation structures of the radar reflectivities and divergence fields, summarized in terms of scale, anisotropy and direction. To extract these properties from divergence and reflectivity images, we use the SAD methodology of Buschow and Friederichs (2021): The image to be analyzed is convolved with a series of localized 2D wave-forms with varying scale and orientation. The analyzing filters are the so-called *daughter wavelets* which are generated by shifting, scaling and rotating a single, carefully designed wave function, the *mother wavelet*. The square of one wavelet coefficient, i.e., the result convolving the image with one of the daughters, represents the amount of variance present at a particular location for a particular combination of spatial scale and orientation. The dual-tree complex wavelet transform (Selesnick et al., 2005) used in this study provides daughter wavelets with six orientations and up to J scales for an image of size $2^J \times 2^J$. Following Buschow and Friederichs (2021), the largest three scales are removed because their support is larger than the image, rendering their interpretation ambiguous. After spatial averaging, a radar image with 128×128 pixels is thus summarized by 4×6 values, the so-called wavelet spectrum. To extract the scale, anisotropy and direction from this spectrum, we treat the $J \times 6$ values as point-masses arranged in a 3D space such that the six directions for one scale are at the vertices of a hexagon in the $x-y$-plane and the hexagons for the J scales are located at $z = 1, \ldots, J$. The centre of mass of these point masses has three components in cylindrical coordinates:

- The central scale $z \in [1, J]$ measures the dominant spatial scale of the image. If all variance was at spatial scale j, then $z = j$; if all scales contain equal variance, then $z = (J-1)/2$.
- The radius $\rho \in [0, 1]$ describes the anisotropy. If all directions have equal variance, then the centre of mass is in the middle of the hexagon and $\rho = 0$; if all energy is concentrated in one direction, then $\rho = 1$.
- From the angular coordinate, we can determine the dominant orientation angle $\varphi \in [0°, 180°]$. Note that φ is only meaningful if the anisotropy ρ is non-zero.

For a detailed description of the calculation of these properties, as well as the details of the wavelet transform itself, we refer to Buschow and Friederichs (2021) and references therein. The software for this analysis is freely available in the open source `dualtrees` R-package (Buschow et al., 2020).

The central scale z is a dimensionless quantity which cannot be analytically transformed into an equivalent Fourier wavelength. Since the actual physical size of the patterns is of some interest in the present study, we derive an empirical relationship based on test images with fixed wavelength in appendix A. We find that, in the range of $1.5 < z < 2.5$, the relationship is well described by a linear fit with

$$\lambda \approx z \cdot 9\,km - 5.4\,km \tag{2}$$

It is important to note that this relationship is only approximately valid for the specific wavelets, scales and wave-like test images used in the present study. This equivalent wavelength is furthermore not identical to the spacing between wave-crests used as the measure of horizontal scale by Banghoff et al. (2020) because our λ includes also the scale perpendicular to the orientation of the features.

To make the distribution of angles φ interpretable, we compute the angles of intersection between φ and the model wind direction (averaged over the regions around each radar). A relative angle $\Delta\varphi = 0°$ thus means that the patterns align with the wind direction whereas $\Delta\varphi = 90°$ indicates an orthogonal orientation.

4.2 Boundary conditions and pre-processing

The wavelet analysis described above requires data on a regular grid, ideally of size $2^n \times 2^n$ to ensure fast computation times, discontinuities at the boundaries must be avoided. This is only a minor factor for intermittent fields like rain but very important for data with non-zero values along each border. To achieve periodic boundaries, we cut out a $128\,km \times 128\,km$ region (128 and 64 pixels for RADOLAN and COSMO-REA2, respectively) around each radar location and apply a circular Tukey window to smoothly reduce the field to zero (for divergences) or -10 dBz (for reflectivities) towards each side. A rectangular boundary would introduce spurious horizontal and vertical directions to the wavelet spectra.

For the reflectivity data, further pre-processing steps are required. Firstly, some radar images contain erroneous isolated pixels with unusual intensities which would artificially reduce the analyzed spatial scales. Following Lagrange et al. (2018), we therefore compare each pixel to the average over its eight nearest neighbours. If the difference is greater than 10 dBZ, the pixel value is replaced by the neighbourhood average. Secondly, the reflectivities around each radar often contain gaps of very small reflectivities (< -10 dBZ), caused for example by buildings, mountains or water bodies without scattering insects. These arbitrarily shaped holes introduce an artificial pattern which is unrelated to the wind field and needs to be removed. Here, we fill in the gaps with a simple algorithm which iteratively replaces values below -10 dBZ with an average over the neighbouring non-missing pixels. The details of the gap-filling algorithm, as well as a demonstration of its effectiveness are given in appendix B.

Lastly, a comparison between the wavelet spectra of two images would normally require that both data sets be given on the same grid. In our case, we can avoid re-gridding either field since the spatial resolutions differ by a factor of two. The second scale in RADOLAN thus corresponds to the first scale of COSMO-REA2. We can therefore simply remove the smallest scale from the radar image to make the spectra comparable. We have checked that the results are virtually identical when the radar

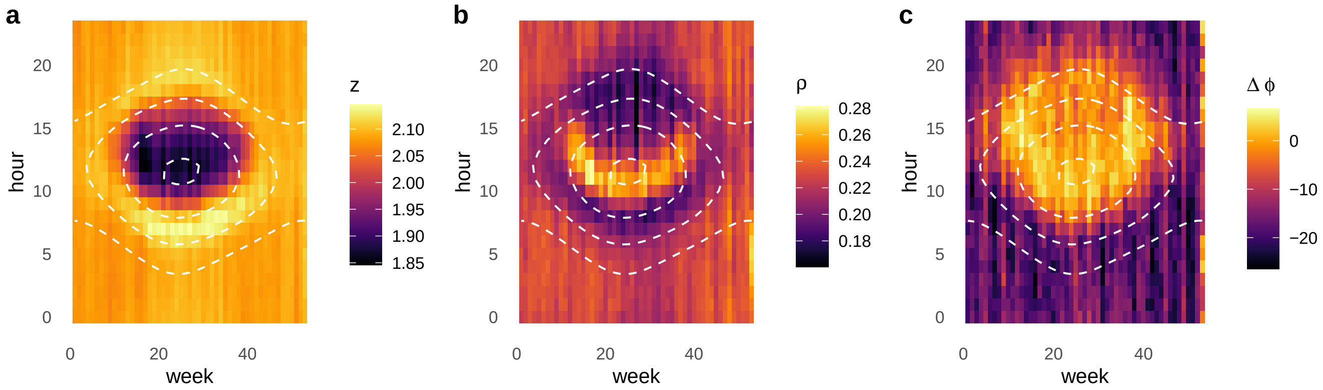


Figure 4. Average central scale (a), anisotropy (b) and angle relative to the mean wind (c), calculated from COSMO-REA2 (2007-2013) in the environment of the selected radar stations. White contours mark the sun's elevation angle at $0^\circ, 20^\circ, 40^\circ, 60^\circ$.

images are bilinearly re-mapped instead. The largest daughter wavelet that fits into our domain is $j = 4$ for RADOLAN and $j = 3$ for COSMO-REA2, giving us three comparable scales with six directions each.

5 Results

5.1 Climatology of divergence structures in COSMO-REA2

Based on section 2, we can expect that small-scaled, cellular circulations will form on warm sunny days, favored by high pressure and low wind speeds. Following the diurnal cycle of the boundary layer depth, these circulations start out small and become larger over the course of the day. According to Poll et al. (2017), Banghoff et al. (2020) and references therein, we furthermore expect to see a more linear mode of organization on windier days. The orientation of these roll-like structures will generally follow the mean wind direction (Weckwerth et al., 1997). Both cells and rolls should leave an imprint on the scale and anisotropy and direction of the horizontal divergence fields. We therefore cut out square regions of 64×64 pixels around the twelve selected radar stations (table 1) and apply the wavelet analysis described above for all hourly time-steps from 2007 to 2013.

As a first overview, we average the scale z, anisotropy ρ and direction relative to the mean wind $\Delta\varphi$ over the hours of the day and weeks of the year. Figure 4 shows that all three simulated variables undergo pronounced diurnal and annual cycles. During nighttime, the average central scales of the divergence fields remain close to $z \approx 2$ (about 13 km) with no strong variations between seasons. After the solar elevation exceeds roughly 40°, z approaches a clear minimum around noon before increasing again during the afternoon. This region of small values is surrounded by a ring of increased scales a few hours after sunrise and around sunset. These largest average scales coincide with a similar ring of unusually low anisotropy (figure 4 b). ρ reaches a maximum during the early hours of the small-scale phase before decreasing during the afternoon. Concerning the orientation

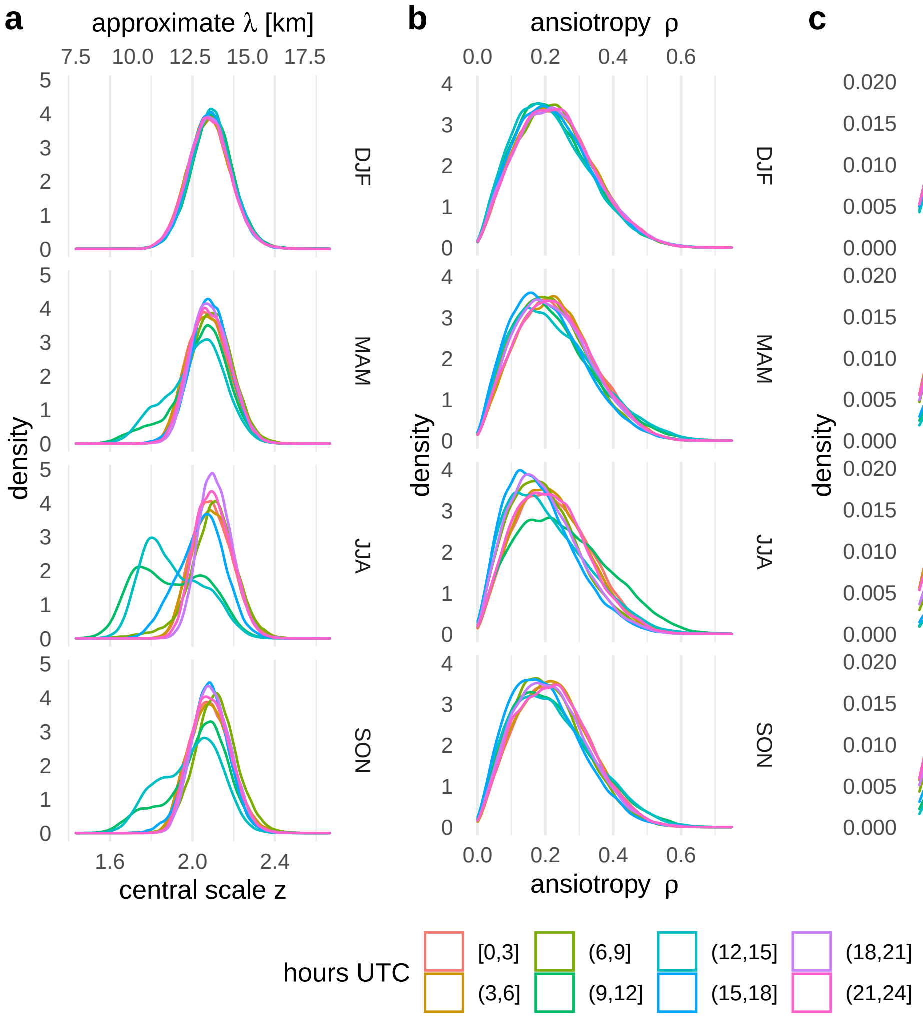


Figure 5. Estimated probability densities (kernel estimates) for the scale z (a, converted into an approximate wavelength λ via equation 2), anisotropy ρ (b) and relative angle $\Delta\varphi$ (c) for different seasons and times of day.

of the divergence field (panel c), we observe that the small-scale pattern is typically aligned with the mean wind direction while the larger scaled nighttime patterns are not.

250 As expected, the simulated small-scaled circulations thus impress their diurnal life-cycle on the mean spatial structure of the divergence field. To see how prominent these features are, compared to the overall variability, we now consider probability densities of the three structural quantities, separated by season and time of day (figure 5).

For the spatial scales in panel (a), we find that the prominent minimum around noon is indeed a common occurrence in all seasons except winter, indicated by bi-modal distributions between 9 and 15 UTC. During summer in particular, the smaller-

255 scaled mode, centred near $z \approx 1.75$ or $\lambda \approx 10\,km$, is more likely than $z > 2$. Two modes can be seen with similar clarity in

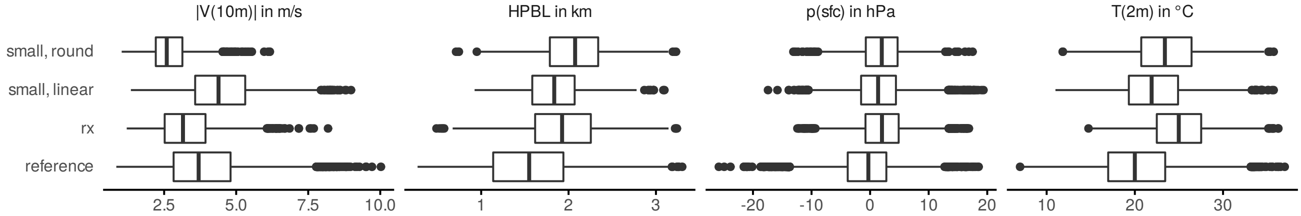


Figure 6. REA2 wind speed, boundary layer height, surface pressure anomaly and 2 m temperature during summer (JJA) between 11 UTC and 13 UTC, averaged around the selected radar locations. “small and round” cases have $z < 1.86, \rho < 0.12$, “small and linear” is $z < 1.86, \rho > 0.32$. The boxplot labeled “rx” contains all instances where at least one clear air radar echo is available.

the distribution of orientations (figure 5 c): During winter or nighttime, orientations along the wind direction are rare, most angles are closer to $-75°$. In the other three seasons, $\Delta\varphi \approx 0$ is by far the most likely value during daytime. The signal in the anisotropy (figure 5 b), on the other hand, is far weaker: A clearly increased likelihood for anisotropic features is only evident in summer between 9 and 12 UTC and the change in the distribution is far less pronounced than for z. While the formation of exceptionally small structures, oriented along the mean wind, is thus a common occurrence, the increased linearity around noon seen in figure 4 b can only occasionally be observed.

Next, we are interested in the typical weather situation associated with the occurrence of these small and / or linear patterns. To this end, we focus on the three hours around noon during the summer season and search for cases where both ρ and z are in the bottom 5 % of their climatological distribution (“small and round” mode). For the “small and linear” mode, we select those cases where z is in the bottom 5 % whereas ρ is in the top 5 % of its distribution. At these time-steps, as well as the remaining “reference” cases, we compute spatial averages around the selected radar stations for several relevant variables from COSMO-REA2.

Figure 6 shows that boundary layer height, 2 m temperature and surface pressure see a moderate increase during time-steps with small and linear patterns and a stronger increase if the pattern is small and round. In the latter cases, the median temperature is close to $25°C$ and the boundary layer rarely falls below 2 km. Simultaneously, the average 10 m wind speed is strongly reduced. Conversely the small and linear mode is associated with a significantly increased wind speed. Hence the boundary layer circulation in COSMO-REA2 qualitatively resembles Rayleigh Bénard convection.

In preparation for the quantitative comparison with radar data, figure 6 also includes the environmental conditions for days where at least one clear-sky RADOLAN image is available. We find that the radar echoes occur mostly on very warm days with moderately increased boundary layer depth and decreased wind speeds. This is consistent with the assumption of insects as the primary origin of these echoes. The observations thus mostly sample cases where small-scale circulations are likely to occur.

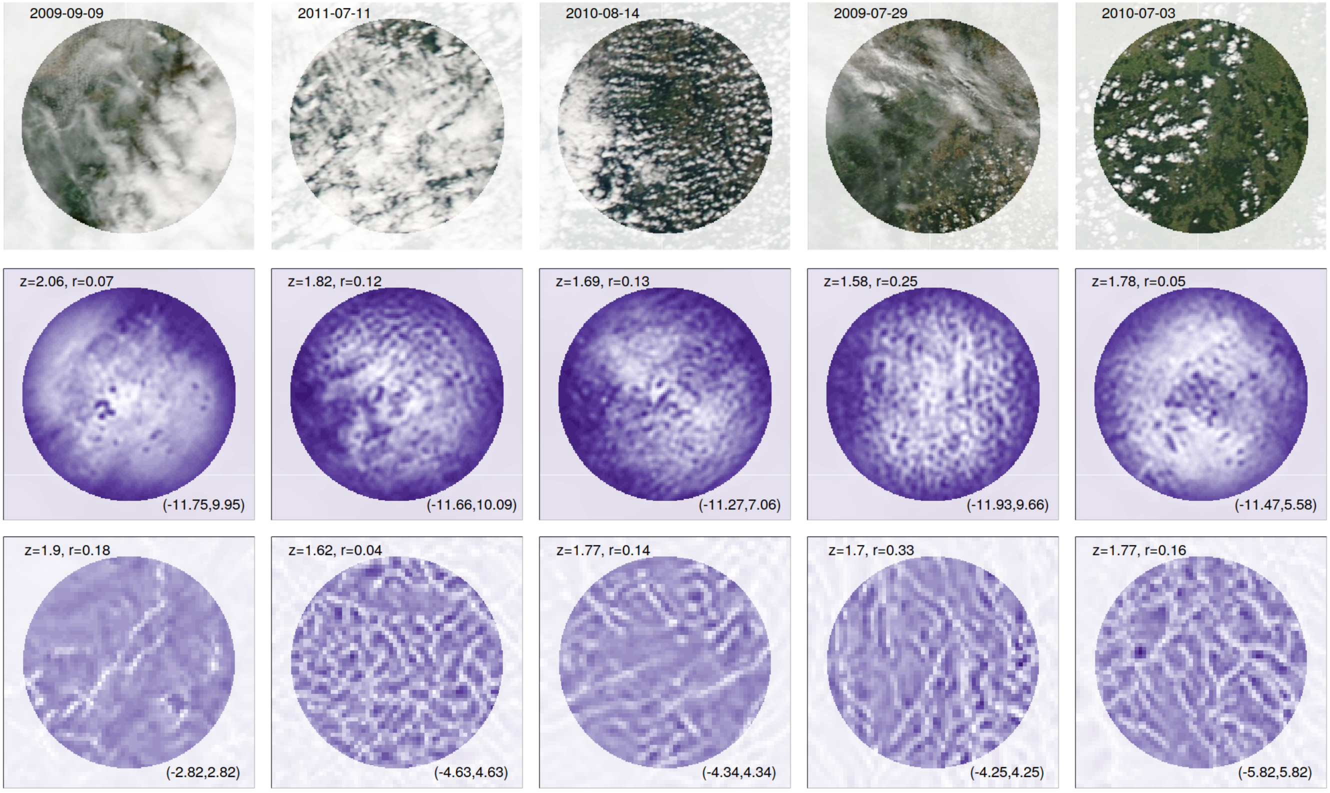


Figure 7. Randomly selected examples from the set of available, non-rainy 12 UTC radar images at Flechtdorf. Top row: Aqua MODIS snapshots (wvs.earthdata.nasa.gov, timing only approximately matches 12 UTC). Middle: RADOLAN RX reflectivity. Bottom: COSMO-REA2 10 m divergence. Light colors indicate high reflectivity and convergence, respectively. Numbers in the top left corner indicate the analyzed scale and anisotropy, the range of reflectivity / divergence values is given in the bottom right.

5.2 Verification against radar reflectivities

In this section, we attempt to assess the realism of our model-based climatology using the clear-sky radar reflectivity data from RADOLAN. Besides cases with too many missing or rainy pixels, we also exclude all nighttime images. The remaining data is subjected to the wavelet analysis as described in section 4.2.

Before analysing the statistics of the entire dataset, we briefly consider a few individual examples. Figure 7 shows five randomly selected cases from the Flechtdorf radar station. The 12 UTC time step was chosen so that a visible satellite image from MODIS is available at approximately the same time. For consistency with the wavelet-based analysis, we have removed the smallest-scaled features from the RADOLAN images by transforming to wavelet-space, setting the coefficients at level 1 to zero and transforming back.

The first two examples (leftmost columns) feature a closed cloud-cover; model and observation agree on a relatively large structure on 2009-09-09 and small, isotropic cells on 2011-07-11. The remaining three cases are all relatively small in scale

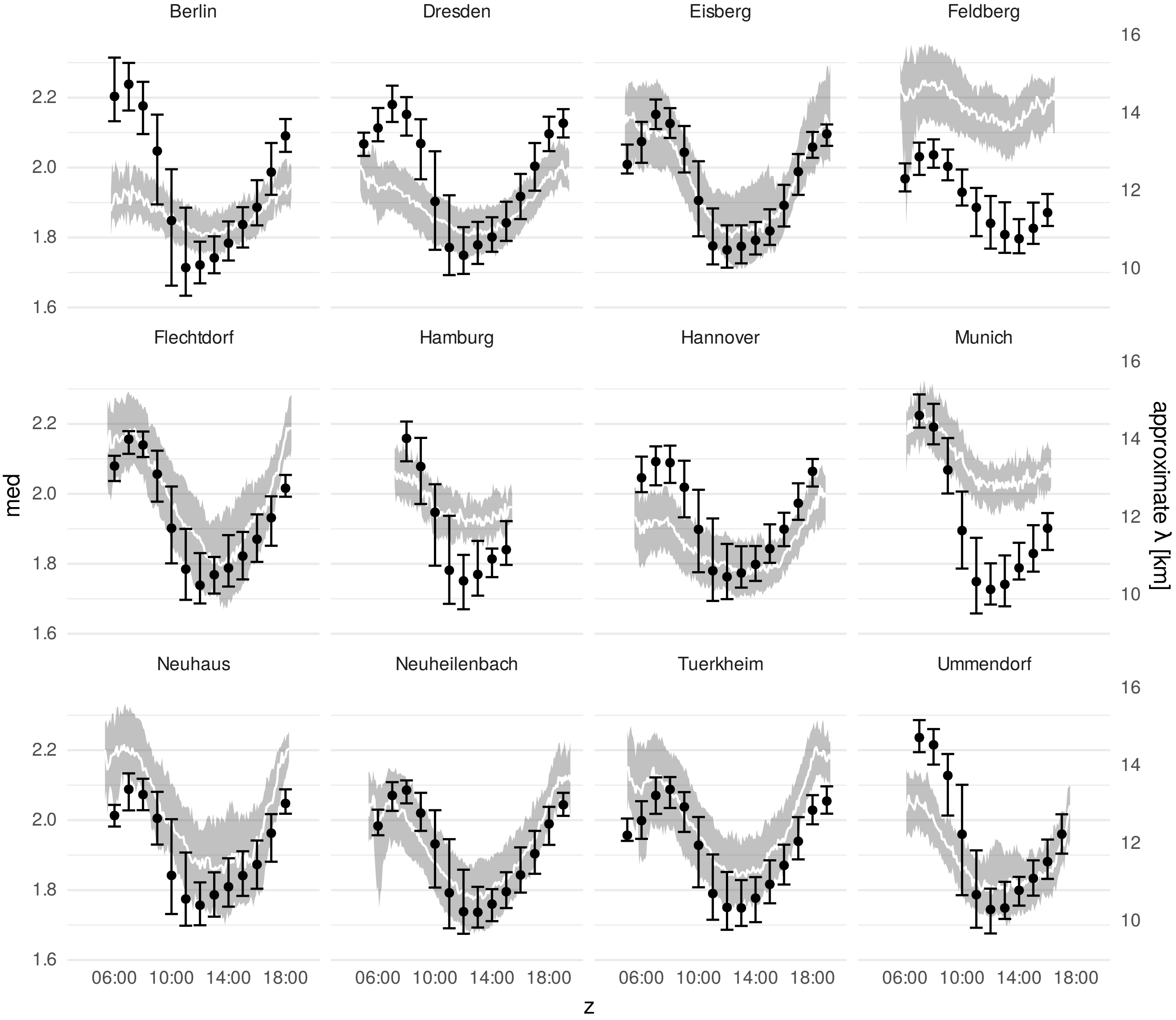


Figure 8. Diurnal cycles of spatial scales from 5 min radar data (areas and lines) and hourly COSMO-REA2 10 m divergence (points and error bars). Grey area and error bars indicate inter-quartile range, white line and black dots mark the median. Only cases with complete (see section 3.3) clear air echoes are included.

with both data-sets agreeing that 2009-07-29, i.e., our example from figure 2, has the smallest and most anisotropic structure. Overall, the decent visual similarity between COSMO-REA2 and RADOLAN is reflected in small to moderate differences in ρ and z.

Figure 8 shows a quantitative comparison of the modelled and observed diurnal cycles of central scales. In addition to the hourly data for which corresponding COSMO-REA2 divergences are available, we have included all other 5 min time-steps with complete clear-air echoes as well. The results can be separated into two main groups: At the rural radar stations in Eisberg, Flechtdorf, Neuhaus, Neuheilenbach, Türkheim and Ummendorf, the agreement betweeen model and observations is surprisingly good. COSMO-REA2 reproduces not only the correct evolution of the diurnal cycle but also similar spatial scales with a large overlap in the inter-quartile ranges. In contrast, the observed spatial patterns at the three largest German cities of Berlin, Hamburg and Munich, differ significantly from the modelled values, as well as from the other stations. Hannover and Dresden have more data than the other urban locations (cf. table 1) and show better agreement with the model. Here, the observed cycle is flatter but resembles its modelled counterpart in the afternoon. The unusual behaviour of the Feldberg/Schwarzwald station is likely the result of its mountainous surrounding which causes both additional ground clutter and changes to the local circulation, neither of which is resolved by the 2 km model orography. It is however worth noting that, despite the offset, both data sets agree that the smallest-scaled patterns occur later in the day than at other stations.

Good agreement between model and observations can be seen in the distribution of the angle φ as well. In figure 9, we have pooled all radars together and consider only full hours where the model wind direction is known. Cases with small observed anisotropy ($\rho \leq 0.1$), i.e., ambiguous orientation, were removed as well. We find that, between 10 and 17 UTC, both sets of images are usually oriented within $\pm 15^\circ$ of the mean model wind direction; the distributions of RADOLAN and COSMO-REA2 match almost perfectly. Before and after this interval, which coincides with the small-scale phase of the diurnal cycle, a wider variety of orientations is possible.

While the scale and orientation are thus in reasonably good agreement, the same can not be said for the anisotropy. Figure 10 shows that the observations are almost universally more isotropic than the model fields. The pattern of increasing linearity towards a maximum before noon seen in figure 4 b is clearly present in this sample of the model data. The observations, on the other hand, hardly contain this pattern at all with only a very weak maximum at 11 UTC and nearly constant values during the afternoon.

Aside from the climatological distribution and diurnal cycle, we are interested in the model's ability to represent the day-to-day variability of the spatial divergence patterns. For z and ρ, we can eliminate the overall bias and diurnal behaviour by subtracting the long-time mean for every daytime hour from the respective time series. To avoid residual effects of the annual cycle, we limit this analysis to the summer season. Timing errors within each day are furthermore removed by taking the daily minimum of z and maximum of ρ. Figure 11 a reveals that the remaining scale anomalies in COSMO-REA2 and RADOLAN are slightly correlated with many remaining errors below 0.1 and almost all below 0.2 (outer lines). As expected, the correlation is even lower for ρ (figure 11 b) and the typical errors are relatively large even after the bias has been removed.

6 Discussion

The results of section 5.1 and 5.2 raise several intertwined questions: What level of realism can be expected of the reanalysed small-scale structure? To what extent can the RADOLAN data-set be used to validate the simulation? How appropriate was the

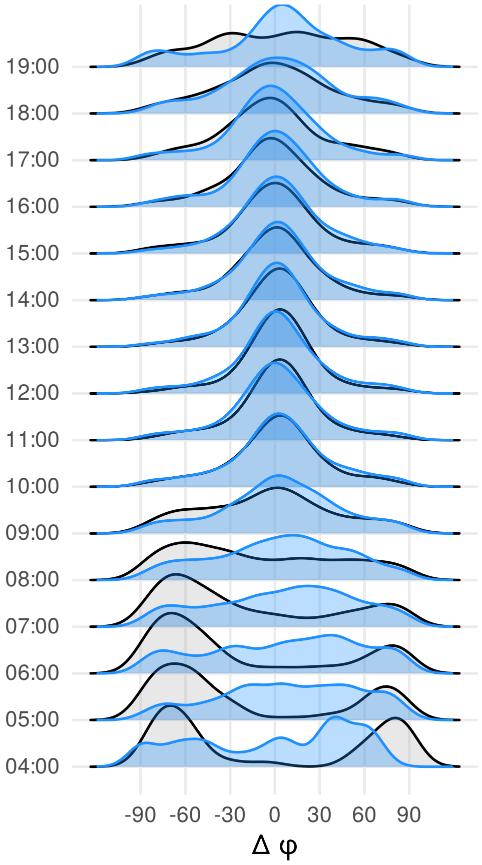


Figure 9. Distribution of orientations relative to the COSMO-REA2 mean wind throughout the day. COSMO-REA2 shown in black, RADOLAN in blue. Only complete, on-rainy daytime cases with $\rho(\text{RADOLAN}) > 0.1$ are included.

wavelet-based analysis for the task at hand? Concerning the trustworthiness of COSMO-REA2, it must be remembered that the local divergence patterns are primarily the product of the model dynamics and parametrized turbulence, not the data assimilation. The environmental conditions which drive the formation of a particular mode of small-scale organization, however, can be expected to have good accuracy due to the continuous input of wind speed, humidity and pressure from weather stations. It is therefore not surprising that the model can accurately represent diurnal and annual cycles and differentiate between days with organized and unorganized situations. Consequently, the model climatology as described in section 5.1 qualitatively agrees with our expectations from the literature. Whether or not the simulated small-scale structure can itself be trusted is questionable in light of the theory discussed in section 2. Our comparison with RADOLAN clear-air data suggests that, despite the proximity to the Grey-Zone, the modelled structures are not overall unrealistic. In interpreting this result, we must recall that the difference in native resolution between RADOLAN and COSMO-REA2 was handled by deleting the smallest scale from RADOLAN. We have thereby filtered out any variability below the model's effective resolution. Figure 8 therefore does *not*

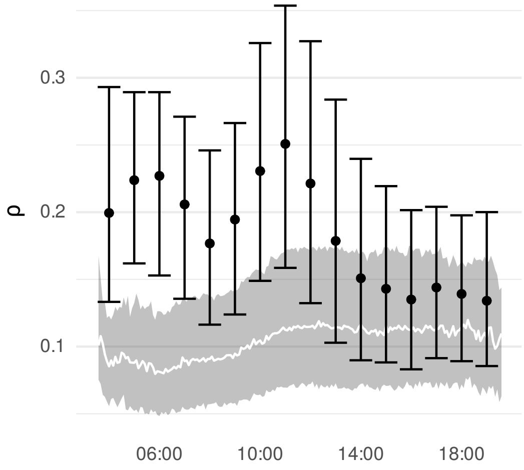


Figure 10. As figure 8, but for ρ and without separation by radars.

indicate that the mesoscale model successfully simulates the spatial scales present in the real atmosphere. We can merely see that the *remaining* variability (upwards of $\lambda \approx 8\,km$), which both data-sets *can* represent, matches the observed diurnal cycle decently, especially at the rural stations.

As predicted by Zhou et al. (2014), the wavelengths of the simulated eddies are near the smallest scale resolved by the model. We note, however, that the underlying resolution of RADOLAN is $1°$ in azimuth- and $250\,m$ in range-direction. Inside our $64\,km$ radius, and particularly close to the radar, the internal resolution of the measurements is considerably finer than the used $1\,km \times 1\,km$ grid. There is thus no obvious technical reason why, after filtering, RADOLAN should have to exhibit increased variability on the same scale as the model. We have experimentally re-calculated the central scales of the radar images including the previously removed smallest scales and found a slight shift in the cycle towards earlier hours. Conversely, if we remove the second smallest scale as well, a shift in the opposite direction emerges. This supports our interpretation that the model simulates the patterns seen in the observations with an approximately correct diurnal cycle, *on the scales we included*; smaller-scaled variability, which would initiate earlier in the day, is resolved by neither COSMO nor RADOLAN. It should furthermore be noted that we make no direct statements about the intensity (variance) of the circulations. Such information cannot easily be inferred because the absolute radar reflectivities depend on the technical details of the radar, applied pre-processing and the unknown overall concentration of biological scatterers.

The greater disagreement at the urban radar locations has two main explanations. On the one hand, it is likely that buildings and unrelated radio signals introduce excessive noise into the images, overshadowing the natural signal. This explanation is supported by the lack of complete images at the Essen and Frankfurt stations, both of which are located in highly urbanized regions (Frankfurt is the city with the most skyscrapers in Germany). On the other hand, the urban landscape itself can influence

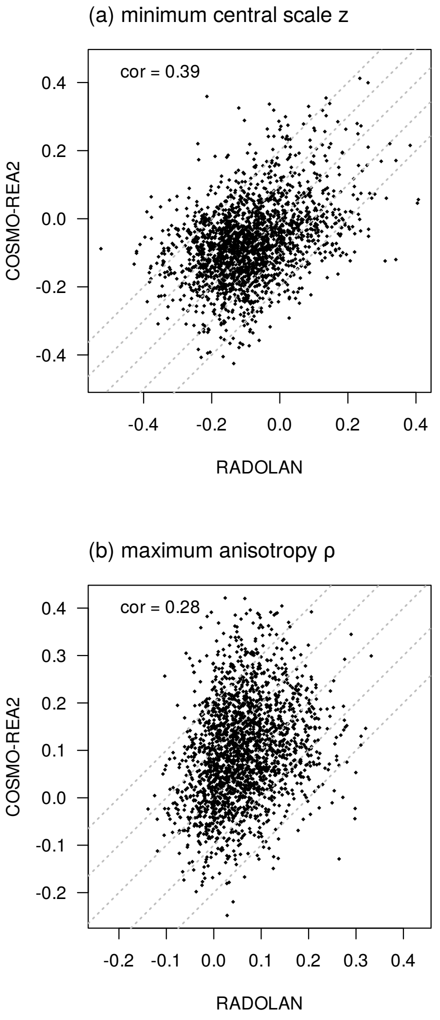


Figure 11. Scatter-plot of daytime minimum scale (a) and daytime maximum anisotropy (b) anomalies during daytime in summer (JJA) from RADOLAN (x-axis) and COSMO-REA2 (y-axis). Anomalies were calculated by subtracting the respective mean values from every hour of the day. Dashed lines mark errors $dz, d\rho = \{-0.2, -0.1, 0, 0.1, 0.2\}$.

the near-surface circulation in ways which are not resolved by the model. The similar effects of small-scale orography likely explain the special behaviour at the Feldberg/Schwarzwald station.

Aside from spatial scales, the anisotropy of the divergence pattern, i.e., the difference between linear and cellular organization, is of interest. Here, the model's tendency towards more linear patterns earlier in the day could not be confirmed observationally. On the one hand, it is plausible that the lack of finer-scale variability leads to the simulation of unnaturally regular stripes. On the other hand, gaps and noise have a larger impact on the anisotropy than the scale (cf. appendix B), making these results somewhat less robust.

Lastly, it should again be emphasized that our clear air data-set provides no information on nighttime and winter and is biased towards cases with high temperatures where small-scale circulations are likely to occur. Our validation is therefore mostly conditional on the occurrence of these phenomena; whether or not the model correctly differentiates between days with and without organized shallow convection could only partly be judged (cf. figure 11).

7 Conclusions and outlook

The main goal of this study was to explore the use of clear-sky radar data for the evaluation of simulated low-level divergence structures. A wavelet-based verification methodology, developed and extensively tested for precipitation data, was used to summarize the spatial patterns in terms of scale, anisotropy and direction. We have demonstrated that model-based divergences and radar reflectivities are comparable at this level of abstraction. Our investigation of the German radar network has shown that usable clear sky echoes are rare overall and almost non-existent in winter. This supports the assumption that such daytime echoes are caused by small insects, the life cycle and habitats of which may also explain the substantial differences between radars as well as strong year to year variations. The relatively long time-span from 2007 to 2013 nonetheless resulted in a robust data set of over 20.000 individual images, mostly during summer, where the modelled patterns could be verified against spatial observations. At most radar locations, both data sets show a very similar diurnal cycle in the spatial scales and orientations with a strong preference for small-scaled ($\lambda \approx 10\,km$) features around noon. The orientation during the small-scaled phase of the cycle is almost always within 15° of the mean wind. The fact that this observation holds for both data sets also implicitly confirms that the model adequately represents the mean wind direction. COSMO-REA2 furthermore simulated a trend towards increasingly linear features at the start of the small-scaled phase which could not be found in the observations. As discussed above, a more complete set of observations might be able to clarify whether this indicates deficiencies of the model or the observations or (likely) both.

Based on the overall decent agreement with the radar observations, we may put some trust in the model's behaviour at the unobserved parts of the time series as well. If COSMO-REA2 is thus to be believed, mesoscale shallow convection, favored by high pressure (clear skies) and temperatures, as well as weak winds, is a common occurrence in Germany in all seasons except winter; during JJA, the small-scale mode is more likely than the larger-scaled configuration. Its onset a few hours after sunrise is characterized by a transition phase with larger scaled, isotropic divergence patterns, the orientation of which switches from $\sim 70^\circ$ to $\sim 0^\circ$ with respect to the mean wind direction. While most patterns are isotropic, i.e., cellular in nature, there is also a weaker signal of linear organization. This more roll-like mode is most often simulated during JJA between 9 and 12 UTC and preferably occurs when winds are unusually strong and the boundary layer is shallower than in the cellular cases. These simulated features are qualitatively consistent with the theory, as well as previous observations of mesoscale shallow convection.

Concerning future prospects, it must be emphasized that we have relied on only the most widely available kind of radar observations. Modern dual-polarization Doppler radars produce a wealth of further information, which would for example allow us to confidently separate insect-related echoes from unhelpful noise and clear up the nature of the night-time echoes (Zrnic and Ryzhkov, 1998; Melnikov et al., 2015). Additionally, parameters like mean wind speed and direction, and even the boundary layer height (Banghoff et al., 01 Aug. 2018) could be inferred directly from the radar instead of relying on the model (Banghoff et al., 2020). Lastly, we re-iterate that small scales below $\sim 8\,km$ were filtered out in this study in order to fairly evaluate the mesoscale model. Depending on their frequency, weather radars can observe much finer details of the turbulent boundary layer. A similar strategy to ours could therefore also provide useful information for the objective validation of realistic large eddy simulations as in Thurston et al. (2016); Poll et al. (2017); Bauer et al. (2020); Ito et al. (2020); Pantillon et al. (2020).

Code and data availability. Software for the dual-tree wavelet transformation is available in the `dualtrees` R-package (Buschow et al., 2020). In addition, the specific version (0.1.4) used for this manuscript has been permanently archived at https://doi.org/10.5281/zenodo.5027277 (Buschow, 2021a). COSMO-REA2 is currently available from the website of the Hans Ertel centre (reanalysis.meteo.uni-bonn.de). RADOLAN is available via the DWD OpenData portal (opendata.dwd.de). The cropped reflectivity and divergence fields around the used radar station have been archived at https://doi.org/10.5281/zenodo.5036447, together with all auxiliary data and software needed to fully reproduce the figures in this manuscript from scratch (Buschow, 2021b).

Appendix A: Empirical relationship between scale and wavelength

To approximately translate the central scale into an equivalent Fourier wavelength λ, we apply the exact method described in section 4.2 to synthetic test images of pure sine-waves, given by

$$f(x,y) = \sin\left(2\pi(k_x x + k_y y)\right) + \epsilon$$

where ϵ is a Gaussian white noise term with zero mean and variance 0.04. Figure A1 shows that the relationship between z and λ is nearly linear for this idealized signal. For $z < 1.5$ and $z > 2.5$, the curve becomes non-linear because most variance is outside the range of scales covered by our wavelet transform. The linear fit yields $\lambda/\Delta_x = 4.464 \cdot z - 2.765$. Since we are merely interested in a rough approximation with round numbers, we simplify the result for $\Delta_x \approx 2\,km$ to obtain equation 2.

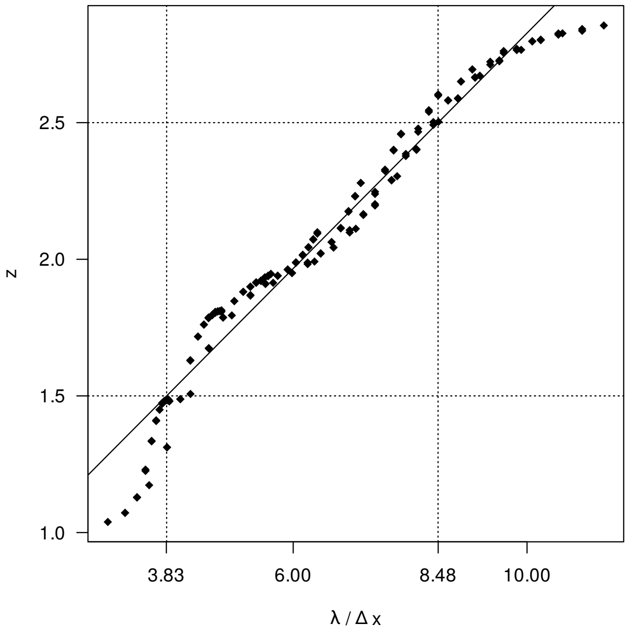


Figure A1. Wavelength $\lambda = 1/\sqrt{k_x^2 + k_y^2}$ against central scale z.

Appendix B: Filling the gaps in the radar images

For this study, we are not interested in the radar reflectivities themselves, or even their full spatial correlation function, but only the estimates structural characteristics ρ, φ, z. To mitigate the effects of holes, i.e., regions with $Z \leq -10\,dBz$, in the radar images, we implement a simple iterative algorithm to smoothly fill in the gaps: (1) Find missing points with at least one non-missing neighbour, (2) replace values of those points with an average over the up to eight adjacent non-missing values and (3) repeat from (1) until all gaps are filled. The result is similar to inverse distance interpolation but (at least in our implementations of the two algorithms) considerably faster. To test the success of our approach, we select 300 nearly complete (less than 3 % missing data) clear-air radar echoes from our data-set and artificially add the gaps form 300 other randomly selected incomplete images. In figure B1, we compare ρ, φ, z, estimated with and without the gap-filling algorithm. As expected, the impact of the gaps is massive but our algorithm mostly mitigates the effects. We have repeated the experiment with inverse distance interpolation (not shown) and found no substantial improvement over the iterative procedure.

Author contributions. SB had the idea for this work, both authors jointly developed the original methodology. Writing and coding was led by SB, with suggestions and additions from PF. Both authors contributed to the final draft and proof-reading.

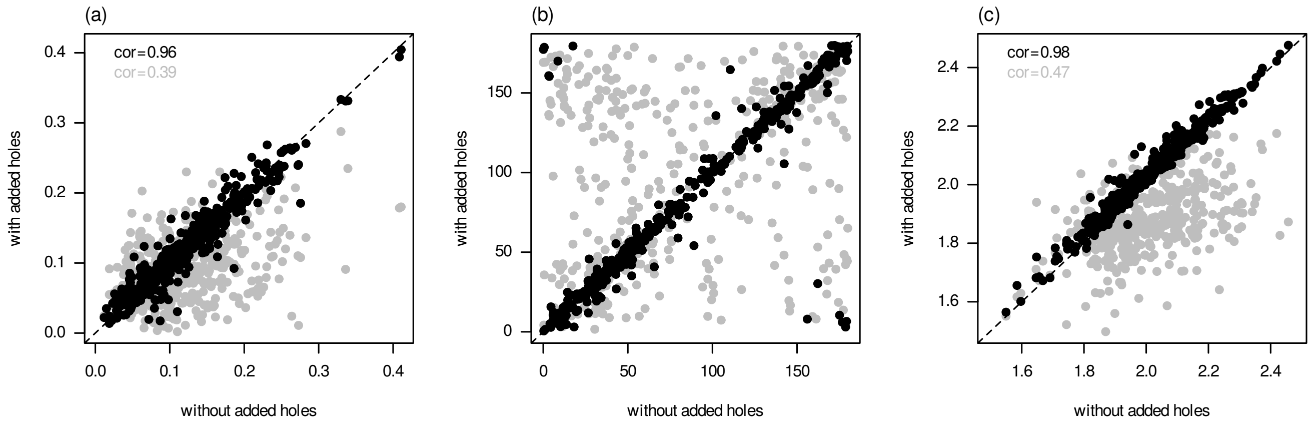


Figure B1. Anisotropy ρ (a), angle φ (b) and scale z (c) estimated from nearly complete images (x-axis) and images with added holes (y-axis). Black dots show the results of the iterative gap-filling algorithms; values obtained without gap filling are shown in grey.

Competing interests. The authors declare that they have no conflict of interest.

References

Atkinson, B. W. and Wu Zhang, J.: Mesoscale shallow convection in the atmosphere, Reviews of Geophysics, 34, 403–431, https://doi.org/10.1029/96RG02623, 1996.

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., and Reinhardt, T.: Operational Convective-Scale Numerical Weather Prediction with the COSMO Model: Description and Sensitivities, Monthly Weather Review, 139, 3887–3905, https://doi.org/10.1175/MWR-D-10-05013.1, 2011.

Banghoff, J. R., Stensrud, D. J., and Kumjian, M. R.: Convective Boundary Layer Depth Estimation from S-Band Dual-Polarization Radar, Journal of Atmospheric and Oceanic Technology, 35, 1723 – 1733, https://doi.org/10.1175/JTECH-D-17-0210.1, 01 Aug. 2018.

Banghoff, J. R., Sorber, J. D., Stensrud, D. J., Young, G. S., and Kumjian, M. R.: A 10-Year Warm-Season Climatology of Horizontal Convective Rolls and Cellular Convection in Central Oklahoma, Monthly Weather Review, 148, 21–42, https://doi.org/10.1175/MWR-D-19-0136.1, 2020.

Bauer, H.-S., Muppa, S. K., Wulfmeyer, V., Behrendt, A., Warrach-Sagi, K., and Späth, F.: Multi-nested WRF simulations for studying planetary boundary layer processes on the turbulence-permitting scale in a realistic mesoscale environment, Tellus A: Dynamic Meteorology and Oceanography, 72, 1–28, https://doi.org/10.1080/16000870.2020.1761740, 2020.

Beck, J., Nuret, M., and Bousquet, O.: Model Wind Field Forecast Verification Using Multiple-Doppler Syntheses from a National Radar Network, Weather and Forecasting, 29, 331–348, https://doi.org/10.1175/WAF-D-13-00068.1, 2014.

Bousquet, O., Montmerle, T., and Tabary, P.: Using operationally synthesized multiple-Doppler winds for high resolution horizontal wind forecast verification: OPERATIONAL DOPPLER RADAR NETWORKS, Geophysical Research Letters, 35, https://doi.org/10.1029/2008GL033975, 2008.

Brune, S., Buschow, S., and Friederichs, P.: The Local Wavelet-based Organization Index – Quantification, Localization and Classification of Convective Organization from Radar and Satellite Data, Quarterly Journal of the Royal Meteorological Society, 2021.

Buschow, S.: dualtrees: gmd-2021-128, https://doi.org/10.5281/zenodo.5027277, 2021a.

Buschow, S.: Code and data for Buschow and Friederichs (2021) "Verification of Near Surface Wind Patterns in Germany using Clear Air Radar Echoes", https://doi.org/10.5281/zenodo.5036447, 2021b.

Buschow, S. and Friederichs, P.: SAD: Verifying the scale, anisotropy and direction of precipitation forecasts, Quarterly Journal of the Royal Meteorological Society, 2021.

Buschow, S., Kingsbury, N., and Wareham, R.: dualtrees: Decimated and Undecimated 2D Complex Dual-Tree Wavelet Transform, https://CRAN.R-project.org/package=dualtrees, r package version 0.1.4, 2020.

Ching, J., Rotunno, R., LeMone, M., Martilli, A., Kosovic, B., Jimenez, P. A., and Dudhia, J.: Convectively Induced Secondary Circulations in Fine-Grid Mesoscale Numerical Weather Prediction Models, Monthly Weather Review, 142, 3284–3302, https://doi.org/10.1175/MWR-D-13-00318.1, 2014.

Davis, C. A., Brown, B. G., Bullock, R., and Halley-Gotway, J.: The Method for Object-Based Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program, Weather and Forecasting, 24, 1252–1267, https://doi.org/10.1175/2009WAF2222241.1, 2009.

Dorninger, M., Gilleland, E., Casati, B., Mittermaier, M. P., Ebert, E. E., Brown, B. G., and Wilson, L. J.: The Setup of the MesoVICT Project, Bulletin of the American Meteorological Society, 99, 1887–1906, https://doi.org/10.1175/BAMS-D-17-0164.1, 2018.

Drake, V. A. and Reynolds, D. R.: Radar entomology: observing insect flight and migration, Cabi, 2012.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E.: Intercomparison of Spatial Forecast Verification Methods, Weather and Forecasting, 24, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1, 2009.

Honnert, R., Efstathiou, G. A., Beare, R. J., Ito, J., Lock, A., Neggers, R., Plant, R. S., Shin, H. H., Tomassini, L., and Zhou, B.: The Atmospheric Boundary Layer and the “Gray Zone” of Turbulence: A Critical Review, Journal of Geophysical Research: Atmospheres, 125, https://doi.org/10.1029/2019JD030317, 2020.

Ito, J., Niino, H., and Yoshino, K.: Large Eddy Simulation on Horizontal Convective Rolls that Caused an Aircraft Accident during its Landing at Narita Airport, Geophysical Research Letters, 47, https://doi.org/10.1029/2020GL086999, 2020.

Keil, C. and Craig, G. C.: A Displacement and Amplitude Score Employing an Optical Flow Technique, Weather and Forecasting, 24, 1297–1308, https://doi.org/10.1175/2009WAF2222247.1, 2009.

Lagrange, M., Andrieu, H., Emmanuel, I., Busquets, G., and Loubrié, S.: Classification of rainfall radar images using the scattering transform, Journal of Hydrology, 556, 972–979, https://doi.org/10.1016/j.jhydrol.2016.06.063, 2018.

Melnikov, V. M., Istok, M. J., and Westbrook, J. K.: Asymmetric radar echo patterns from insects, Journal of Atmospheric and Oceanic Technology, 32, 659–674, 2015.

Pantillon, F., Adler, B., Corsmeier, U., Knippertz, P., Wieser, A., and Hansen, A.: Formation of Wind Gusts in an Extratropical Cyclone in Light of Doppler Lidar Observations and Large-Eddy Simulations, Monthly Weather Review, 148, 353–375, https://doi.org/10.1175/MWR-D-19-0241.1, 2020.

Pejcic, V., Saavedra Garfias, P., Mühlbauer, K., Trömel, S., and Simmer, C.: Comparison between precipitation estimates of ground-based weather radar composites and GPM’s DPR rainfall product over Germany, Meteorologische Zeitschrift, p. 94062, https://doi.org/10.1127/metz/2020/1039, 2020.

Poll, S., Shrestha, P., and Simmer, C.: Modelling convectively induced secondary circulations in the *terra incognita* with TerrSysMP: Modelling CISCs in the Terra Incognita with TerrSysMP, Quarterly Journal of the Royal Meteorological Society, 143, 2352–2361, https://doi.org/10.1002/qj.3088, 2017.

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

Santellanes, S. R., Young, G. S., Stensrud, D. J., Kumjian, M. R., and Pan, Y.: Environmental Conditions Associated with Horizontal Convective Rolls, Cellular Convection, and No Organized Circulations, Monthly Weather Review, 2021.

Schlager, C., Kirchengast, G., Fuchsberger, J., Kann, A., and Truhetz, H.: A spatial evaluation of high-resolution wind fields from empirical and dynamical modeling in hilly and mountainous terrain, Geoscientific Model Development, 12, 2855–2873, https://doi.org/10.5194/gmd-12-2855-2019, 2019.

Selesnick, I., Baraniuk, R., and Kingsbury, N.: The dual-tree complex wavelet transform, IEEE Signal Processing Magazine, 22, 123–151, https://doi.org/10.1109/MSP.2005.1550194, 2005.

Skinner, P. S., Wicker, L. J., Wheatley, D. M., and Knopfmeier, K. H.: Application of Two Spatial Verification Methods to Ensemble Forecasts of Low-Level Rotation, Weather and Forecasting, 31, 713–735, https://doi.org/10.1175/WAF-D-15-0129.1, 2016.

Skok, G. and Hladnik, V.: Verification of Gridded Wind Forecasts in Complex Alpine Terrain: A New Wind Verification Methodology Based on the Neighborhood Approach, Monthly Weather Review, 146, 63–75, https://doi.org/10.1175/MWR-D-16-0471.1, 2018.

Stephan, K., Klink, S., and Schraff, C.: Assimilation of radar-derived rain rates into the convective-scale model COSMO-DE at DWD, Quarterly Journal of the Royal Meteorological Society, 134, 1315–1326, https://doi.org/10.1002/qj.269, 2008.

Thurston, W., Fawcett, R. J. B., Tory, K. J., and Kepert, J. D.: Simulating boundary-layer rolls with a numerical weather prediction model: Simulating Boundary-Layer Rolls with a NWP model, Quarterly Journal of the Royal Meteorological Society, 142, 211–223, https://doi.org/10.1002/qj.2646, 2016.

Wahl, S., Bollmeyer, C., Crewell, S., Figura, C., Friederichs, P., Hense, A., Keller, J. D., and Ohlwein, C.: A novel convective-scale regional reanalysis COSMO-REA2: Improving the representation of precipitation, Meteorologische Zeitschrift, 26, 345–361, https://doi.org/10.1127/metz/2017/0824, 2017.

Weckwerth, T. M., Wilson, J. W., Wakimoto, R. M., and Crook, N. A.: Horizontal Convective Rolls: Determining the Environmental Conditions Supporting their Existence and Characteristics, MONTHLY WEATHER REVIEW, 125, 22, 1997.

Weckwerth, T. M., Horst, T. W., and Wilson, J. W.: An Observational Study of the Evolution of Horizontal Convective Rolls, MONTHLY WEATHER REVIEW, 127, 20, 1999.

Wernli, H., Paulat, M., Hagen, M., and Frei, C.: SAL—A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts, Monthly Weather Review, 136, 4470–4487, https://doi.org/10.1175/2008MWR2415.1, 2008.

Wilson, J. W., Weckwerth, T. M., Vivekanandan, J., Wakimoto, R. M., and Russell, R. W.: Boundary layer clear-air radar echoes: Origin of echoes and accuracy of derived winds, Journal of Atmospheric and Oceanic Technology, 11, 1184–1206, 1994.

Wyngaard, J. C.: Toward Numerical Modeling in the “Terra Incognita”, JOURNAL OF THE ATMOSPHERIC SCIENCES, 61, 11, 2004.

Zhou, B., Simon, J. S., and Chow, F. K.: The Convective Boundary Layer in the Terra Incognita, Journal of the Atmospheric Sciences, 71, 2545–2563, https://doi.org/10.1175/JAS-D-13-0356.1, 2014.

Zrnic, D. S. and Ryzhkov, A. V.: Observations of insects and birds with a polarimetric radar, IEEE Transactions on Geoscience and Remote Sensing, 36, 661–668, 1998.

Zschenderlein, P., Pardowitz, T., and Ulbrich, U.: Application of an object-based verification method to ensemble forecasts of 10-m wind gusts during winter storms, Meteorologische Zeitschrift, p. 90341, https://doi.org/10.1127/metz/2019/0880, 2019.

Appendix E

Unpublished location score manuscript

Measuring Displacement Errors with Complex Wavelets

Sebastian Buschow

July 27, 2021

When highly-resolved precipitation forecasts are verified against observations, displacement errors tend to overshadow all other aspects of forecast quality. The appropriate treatment and explicit measurement of such errors remains a challenging task. This study explores a new verification approach which uses the phase of complex wavelet coefficients to quantify spatially varying displacements. Idealized and realistic test cases from the MesoVICT project demonstrate that our approach yields helpful results in a variety of situations where popular alternatives may struggle. Potential benefits of very high spatial resolutions can be identified even when the observational data-set is coarsely resolved itself. The new approach can furthermore be applied not only to precipitation but also variables such as wind speed and potential temperature, thereby overcoming a limitation of many established location scores.

1. Introduction

Location errors are the main reason why simulated meteorological fields like precipitation cannot be directly compared to observations in a grid-point wise manner. If a forecast mis-places a rain field, large differences are seen both at the predicted and the true, observed location of the feature; possible similarities between the two images are not rewarded and the magnitude of the displacement is not quantified. Due to this *double-penalty* effect, quality measures like the point-wise RMSE prefer forecasts with large, smooth structures to finer-scaled, arguably more realistic models. The need for informative, objective evaluation of high-resolution forecasting systems lead to the development of numerous new "spatial" verification techniques, which were surveyed in the Spatial Forecast Verification Methods Intercomparison Project [Gilleland et al., 2009, ICP] and its successor project MesoVICT (Mesoscale Verification Intercomparison over Complex Terrain, Dorninger et al. [2018]). Besides facilitating the development of new methods and providing standardized test cases for their comparison, these projects classified nearly all existing approaches into five classes. The common thread is that each class uses a specific abstract representation of the underlying fields: *Neighbourhood* methods apply smoothing filters to essentially compare the average rainfall characteristics around each location. Similarly, *scale separation* methods split forecast and observation into individual frequency components via band-pass filters. *Binary distance measures* instead rely on the computation of distance maps which measure the distance from each pixel to the next rainy pixel in one of the images. *Field deformation* methods consider the optimization problem of transforming one image into the other via an optical flow field. The fifth and perhaps most popular group of methods decomposes the fields into discrete *objects* and compares their properties.

Many of these methods aim to isolate individual aspects of forecast quality in order to specify the nature of a forecast's error and, ideally, hint at the reasons for the shortcoming. Such scores would be useful even in the absence of double penalties, since the realism of a complex simulated field can hardly be described by a single number. A prime example is the object-based SAL method of Wernli et al. [2008], which identifies errors in Structure, Amplitude and Location. Weniger and Friederichs [2016] pointed out that SAL can be highly sensitive to the specifics of the object identification procedure and may not be appropriate for variables other than precipitation. Motivated by this, we developed a new structure verification method based on scale-separation instead of object decomposition [Kapp et al., 2018, Buschow et al., 2019, Buschow and Friederichs, 2020]. In its final form, presented in Buschow and Friederichs [2021], this approach compares the spatial Scale, Anisotropy and Direction. The scale-component of SAD is similar to SAL's S, albeit more specific in its interpretation (predicted structure too small or too large) and more effective in detecting the correct correlation structure [Buschow et al., 2019]. The other two components measure how strongly a field is directed and what its preferred orientation is; two aspects which are neglected by S.

Taking further inspiration from SAL, this study aims to define a location-score based on the same scale-decomposition used in SAD. On the one hand, a verification that considers only the correlation structure (and perhaps the marginal distribution) is incomplete since forecasts with no location-information are useless for many applications. On the other hand, the complex wavelet transform, which forms the technical core of SAD, affords us the unique opportunity to extract location information from the phases of the complex coefficients. As a further motivation, our literature survey in section 2 below indicates that SAL's location score is often used but rarely useful. In addition, wavelet transforms require no strong assumptions on the structure of the underlying fields, such as the existence of discrete objects and meaningful thresholds. Our new methodology is therefore not limited to intermittent, precipitation-like data but can be applied to any meteorological field of interest.

The remainder of this paper is structured as follows: In Section 2, we review some of the most popular location scores from the literature, including SAL. Section 3 briefly introduces the SAD structure verification and the wavelet transform on which it is based. We define the new, phase-based location score in section 4 and demonstrate its behaviour in a series of idealized tests. Realistic test cases from the MesoVICT project are introduced in section 5 and verified using the new and old location scores in section 6. Here we focus mainly on precipitation; potential temperatures and wind speed are included as a proof of concept as well. Section 7 summarizes the outcomes of our study and discusses the merits, as well as limitations of all tested scores.

2. Established displacement measures

Perhaps the most widely used pure location score is the L-component of the object-based Structure, Amplitude and Location score [Wernli et al., 2008, SAL]. Denoting the centre of mass in the observed and forecast field by $\mathbf{r}^{(\text{obs})}$ and $\mathbf{r}^{(\text{forc})}$, respectively, one half of L is defined as

$$L_1 = \frac{|\mathbf{r}^{(\text{obs})} - \mathbf{r}^{(\text{forc})}|}{L_{max}}, \tag{1}$$

where L_{max} is the longest distance between two grid-points within the domain. For the other half of L, rain fields are decomposed into discrete objects. In this study, we use the standard object identification procedure advocated in Wernli et al. [2009] and implemented in the SpaialVx R-library: (1) convert precipitation into binary fields by thresholding at 1/15 times the 95-th percentile of non-zero values in the respective field, (2) smooth the binary mask with a disk-kernel and (3) group continuous non-zero regions into individual objects. We compute the centres $\mathbf{r}_{1,\ldots,N}$ and precipitation totals $R_{1,\ldots,N}$ of all N objects in one of the fields and define the scatter around the overall centre $\mathbf{r}$ as $\Delta r = \sum_{i=1}^{N} R_i |\mathbf{r}_i - \mathbf{r}| / \sum_{i=1}^{N} R_i$. The second half of L is then given by

$$L_2 = 2 \cdot \frac{|\Delta r^{(\text{obs})} - \Delta r^{(\text{forc})}|}{L_{max}}. \tag{2}$$

The overall location score $L = L_1 + L_2$ is therefore in the interval $[0, 2]$ and consists of equal contributions from the overall centre of mass and the scattering around that centre. Considering the continued popularity and wide-spread use of SAL, it is worth pointing out that L has repeatedly failed to produce useful information on forecast performance. Table 2 in the appendix summarizes the results of L in 20 verification studies. Only three of these authors obtain any interpretable information from the location component [Hanley et al., 2013, Navascués, 2013, Davolio et al., 2017]; the others either fail to mention it entirely [Früh et al., 2007, Zimmer et al., 2008, Zacharov, 2013], or explicitly state that L remained uninformative (e.g. Wittmann et al. [2010], Lindstedt et al. [2015], Kann et al. [2015], Maurer et al. [2017]). While this list is not exhaustive, it nonetheless demonstrates that (1) there is considerable interest in a pure location score and (2) SAL's location component is frequently uninformative. An obvious explanation for this shortcoming is that the co-occurrence of multiple different location errors in one forecast may be handled incorrectly. L_1 is invariant under any rearrangement of the fields that leaves the centre of mass unchanged. The behaviour of L_2, which is supposed to compensate for such effects, is not obvious when the number, placement and intensity of objects can differ between the two fields in a variety of ways.

Field deformation is another widely cited approach to the explicit measurement of location errors. By computing an optimal vector field which transforms one image into the other (the so-called *optical flow*), these techniques account for varying displacements in different parts of an image. Since the flow field is generally not divergence free, such scores register not only displacements but also errors in the spatial structure and, in the case of precipitation, the rate of occurrence. While field deformation methods are frequently mentioned in lists of popular spatial verification methods, they are comparatively rarely used. Of the three deformation approaches included in the original ICP [Gilleland et al., 2010, Marzban and Sandgathe, 2010, Keil and Craig, 2009], only the Displacement and Amplitude Score (DAS) of Keil and Craig [2007, 2009] appears to have seen use in multiple later studies. Their “pyramid matching” algorithm performs an exhaustive search for the best shift of each individual pixel for a series of coarse-grained versions of the two fields and combines the resulting displacement vectors into an optical flow field (for a complete description see the papers cited above). While this pragmatic approach is not guaranteed to find the true optimum, it will always give a result after a predetermined number of steps (unlike other methods which may even fail to converge) and can be computed at reasonably low cost. Compared to SAL, DAS is far less widely used, likely because it is less easy to understand and implement. In fact, most subsequent applications are either co-authored by one of the original authors [Tafferner et al., 2008, Craig et al., 2012, Lange and Craig, 2014] or acknowledge them for providing the code and / or assisting with the implementation [Nan et al., 2010, Skinner et al., 2016, Han and Szunyogh, 2016]. For this study, we have developed an implementation of DAS based on the `imager` `R`-library [Barthelme, 2018]. We will refer to the vector magnitude of the flow which transforms the forecast into the observation, averaged over all locations with observed precipitation, as D_{KC}.

A third way of quantifying location errors is given by binary distance measures. The basis for these scores is the so-called distance map $d(\mathbf{r}, X)$, which measures the distance from an arbitrary location $\mathbf{r}$ to the nearest element of the set X of grid-points where the binary field under consideration has the value 1. For a recent review and comparison of distance measures used in forecast verification, we refer to Gilleland et al. [2020]. In this study, we use Baddeley's delta metric [Gilleland, 2011] as an example from this class. Denoting by A and B the sets of grid-points where predicted and observed precipitation exceed 0.1 mm, it is defined as

$$BD = \left[\frac{1}{N} \sum_{i=1}^{N} |w(d(\mathbf{r}_i, A)) - w(d(\mathbf{r}_i, B))|^p \right]^{1/p} . \tag{3}$$

Here, we use the default SpatialVx implementation where $p = 2$ and the weight function is just the identity $w(d) = d$. Thus, for each pixel in the domain, we compare the distance to the next rainy pixel in forecast and observation. BD rewards overlap and can measure displacement errors but also reacts to difference in the general shape and spatial distribution of the rain areas.

3. Wavelet based structure verification (SAD)

Wavelets were among the first proposed solutions to the double penalty problem in forecast verification. The basic concept is to represent an image (in our case a meteorological field) as a superposition of functions $\psi_{j,d,l}$ which are limited to a specific *scale* j, direction d and location l.[1] Classically, a suitable set of these so-called *daughter wavelets* $\psi_{j,d,l}$ is obtained by applying a re-scaling, rotation and shift to a single *mother wavelet* ψ. To qualify as a wavelet, ψ must

[1] For the sake of simpler notation, we have implicitly assumed that the locations l can be counted by a scalar index; general wavelet transforms can allow arbitrary locations in $\mathbb{R}^2$.

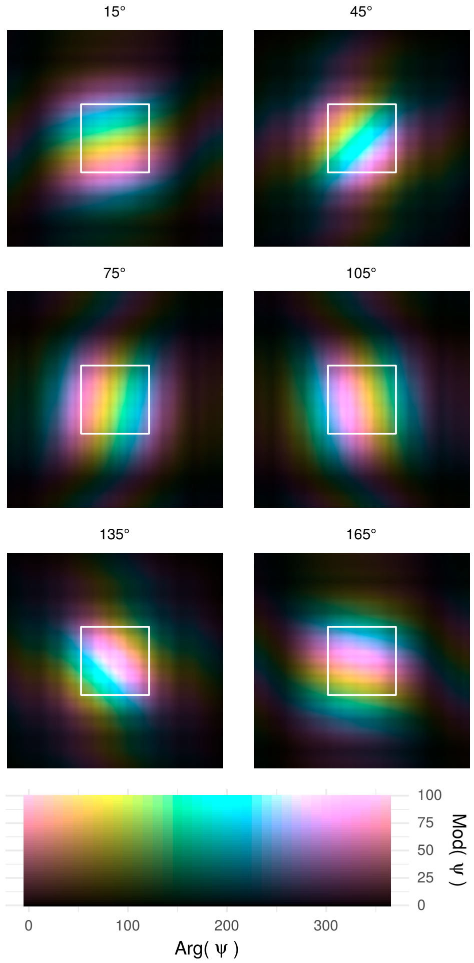


Figure 1: Complex daughter wavelets at scale $j = 6$ on a 200×200 domain in hcl-colorspace: Phase mapped to the hue, chroma and luminance correspond to the modulus. Images are cropped to a 200×200 region around the maximum amplitude, white boxes show an area of $2^6 \times 2^6$ pixels.

integrate to zero (localization in space) and its Fourier transform must decay sufficiently quickly (localization in frequency). The expansion coefficient for a specific daughter is defined as the scalar product with the image $I(x,y)$, i.e.,

$$c_{j,d,l} = \int_{\mathbb{R}} \int_{\mathbb{R}} I(x,y)\psi_{j,d,l}(x,y)dxdy\,. \tag{4}$$

Almost all forecast verification approaches based on wavelets rely on the multi-resolution analysis (MRA) algorithm of Mallat [1989]. The MRA is a wavelet transform that allows only scales which are whole powers of two, i.e., $\psi_j(x,y) = 2^{-j/2}\psi(2^{-j}x, 2^{-j}y)$, and shifts the daughter wavelets at scale j in increments of 2^j. In two dimensions, the transform is not implemented as an explicit convolution (as written in equation 4) but by a series of discrete high- and low-pass filters applied recursively to the rows and columns of the image. This separable construction leads to an orthogonal decomposition with three directions, namely horizontal (high-pass on the rows, low-pass on the columns), vertical (vice-versa) and diagonal (high-pass on rows and columns). The popular ISS verification method [Casati et al., 2004] uses an MRA to split the overall MSE between two binary images into its components on the various spatial scales. The double penalty effect is thereby limited to the small-scale side of the decomposition while skill on larger scales can still be rewarded.

In Buschow and Friederichs [2021], we pursued a different approach and used wavelets to isolate information on the spatial correlation structure of the images while ignoring location errors entirely. In principle, this could be achieved by summing up the squared MRA-coefficients over all locations l to obtain a *wavelet spectrum*. In analogy to the Fourier spectrum, information on the correlation structure could consequently be inferred following Eckley et al. [2010]. However, Mallat's original MRA is ill-suited to this task for two main reasons: (1) the distribution of energy across scales and directions changes abruptly when the input image undergoes a small shift and (2) the diagonally oriented daughter wavelets are ambivalent in their orientation ($\pm 45°$) and smaller in scale than their sisters. Both issues are resolved by switching to the so-called *dual-tree complex wavelet transform* [Kingsbury, 1999, dtcwt]: The real-valued mother ψ is replaced by a complex-valued function $\psi_r + i\psi_i$ where the real and imaginary part are out of phase by 90°. In two-dimensions, this is realized by a suitable set of four separate MRAs, the coefficients of which are re-combined into six uniquely oriented, complex daughter wavelets. For the details of the algorithm, we refer to the helpful tutorial-paper by Selesnick et al. [2005]. Figure 1 shows the six different orientations at scale $j = 6$. As in the original MRA, wavelets at this scale are shifted in increments of 2^6, i.e., the region inside the white box is represented by one complex coefficient per direction. We note that the support of the wavelet (image region where $|\psi_{j,d,l}| > 0$) is larger than the box it represents, which raises the question of boundary conditions, even if the input image were a square of size $2^J \times 2^J$. In this paper we will avoid this issue by (1) reflecting the fields at the edges[2] and (2) removing all coefficients which are either larger than the input image, centred outside of the input image, or touching the outer boundary. In accordance with Buschow and Friederichs [2021], the largest three scales are thereby removed entirely from the analysis.

[2] Here, reflection is preferred over padding because the former is appropriate for both precipitation and other variables.

The resulting grids at the largest scales can also be seen in figure 12 at the end of section 6.

The dtcwt's quadruple redundancy (two complex numbers for each pixel in the input image) gives it near-perfect invariance under shifts: We can sum up the squared coefficients $c_{j,d,l}^2$ over all locations to obtain a $J \times 6$ wavelet spectrum that changes only mildly when the underlying $2^J \times 2^J$ image is shifted. From this spectrum, we obtain the central Scale $z \in [1, J]$, degree of Anisotropy $\rho \in [0, 1]$ and preferred Direction $\varphi \in [0, \pi]$ by treating the $c_{j,d,l}^2$ as point-masses located along the edges of a hexagonal prism and computing the centre of mass. These structural characteristics can then be compared between forecast and observation. For a detailed explanation of the SAD verification method, we refer to Buschow and Friederichs [2021]. In this paper, we will use it only briefly to summarize the observed spatial structure of our test cases. Our goal in the next section is to derive a location score, which is exactly complementary to the SAD structure scores.

4. A wavelet based location score

The dtcwt achieves near-perfect shift invariance thanks to its complex basis functions: Shifts of the input image are encoded in the phase of the complex coefficients while the amplitudes, averaged over all locations, are almost invariant. To get an intuition for this behaviour, consider again figure 1 and suppose that the image $I(x, y)$ to be transformed consists of a single non-zero pixel located somewhere inside the white box. At scale $j = 6$, the entire box is represented by a single complex coefficient for each direction, namely the scalar product between I and the daughter wavelet ψ centered on the box. Since I is zero everywhere except for one pixel, the resulting coefficient is simply the value of the daughter wavelet *at that pixel*. In other words, the complex number shown at any point inside the box in figure 1 is exactly the value of the $c_{j,d,l}$ we would obtain if I was 1 at that point and zero elsewhere. When we move the non-zero pixel around, the absolute value of the coefficient (the luminance and chroma in figure 1) remains nearly constant, but the phase Φ (the hue in our plot) changes. The basic idea of our location score is to use this change in phase to estimate the displacement between two images. This approach is particularly promising because the relationship between phase and displacement is approximately linear. For the Fourier transform of a shifted signal $x(\cdot - \tau)$, it is easy to show that

$$\mathcal{F}\{x(\cdot - \tau)\}(\omega) = e^{-2\pi i \omega \tau} \mathcal{F}\{x\}(\omega)\,, \tag{5}$$

meaning that a time-shift by τ results in a frequency-dependent phase shift by $-2\pi\omega\tau$. Since the real and imaginary part of the dtcwt wavelets have the same 90° offset as the Fourier basis, their local phase behaviour is similar and should thus be close to linear. We derive a location score as follows:

1. perform the dtcwt of forecast and observation

2. at every location, scale and direction compute the phase difference
$\Delta\Phi_{j,d,l} = \min(\ |\Phi_{j,d,l}(obs) - \Phi_{j,d,l}(for)|,\ 2\pi - |\Phi_{j,d,l}(obs) - \Phi_{j,d,l}(for)|\ )/\pi$

3. take a weighted average of $\Delta\Phi$ over all locations and directions to obtain a scale-dependent

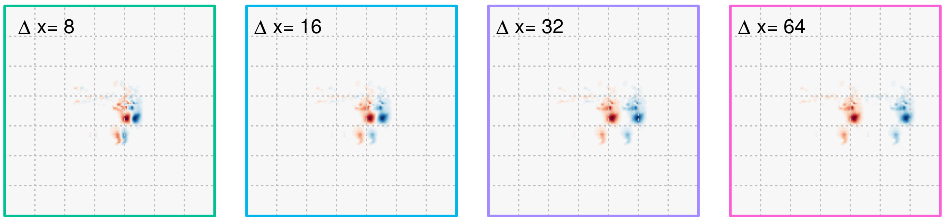


Figure 2: Difference between shifted and original test image used in figure 3

location error

$$\Delta\Phi(j) = \sum_{\text{all } l} \sum_{d=1}^{6} \frac{w_{j,d,l}}{\sum_{d,l} w_{j,d,l}} \Delta\Phi_{j,d,l} \tag{6}$$

The division by π gives us a score between 0 and 1, where $\Delta\Phi(j) = 0$ indicates that the coefficients for scale j are perfectly in phase and $\Delta\Phi(j) = 1$ is the largest possible phase shift of 180°. Intuitively, the worst possible location score should be assigned to a forecast which contains no information on the location of the observed features at all. In this case, the predicted phase $\Phi(for)$ is a uniform random variable on the unit circle. Due to the rotational symmetry of the problem, we can set $\Phi(obs) = 0^\circ$ without loss of generality and find for the expected phase error

$$E\left[\, \min(|\Phi(obs) - \Phi(for)|, 2\pi - |\Phi(obs) - \Phi(for)|)\, \right] = E\left[\, |\Phi(for)|\, \right] = 0.5\pi\,. \tag{7}$$

To ensure that this remains the worst case for our verification score, we will therefore consider $\Delta\Phi > 0.5$ equally bad as $\Delta\Phi = 0.5$. This is also the value that materializes when the intensity in at least one of the images is zero, since the phase Φ is computed with finite precision as the arc-tangent of the ratio between two very small numbers. It is therefore clear that the spatial average of phase differences must be weighted by the amplitudes of the coefficients: Without weighting, regions of correct negative forecasts would contribute $\Delta\Phi = 0.5$, i.e., the worst possible score! To prevent this, we weight the phase differences by the total observed and forecast energy, i.e.,

$$w_{j,d,l} = c^2_{j,d,l}(obs) + c^2_{j,d,l}(for)\,. \tag{8}$$

The resulting score is thus symmetrical with respect to exchanging forecast and observation, ignores featureless regions and focuses on the most important part of the two fields.

For a first impression of the phase-based location score, we perform a simple experiment and compare a rain field from the MesoVICT data set to shifted versions of itself (figure 2). Figure 3 a shows the resulting weighted average phase differences $\Delta\Phi(j)$. As we expected due to the analogy to Fourier, the phase shift is indeed initially linear, with a slope depending on the scale j. We observe that $\Delta\Phi$ reaches 0.5, i.e., 90°, at shifts around 2^{j-1} (see dashed, colored vertical lines) and then oscillates around that limit value. The oscillation, caused by random re-alignment of image features at large displacements[3], is a further reason to treat $\Delta\Phi > 0.5$ and $\Delta\Phi = 0.5$

[3] large compared to the daughter wavelet

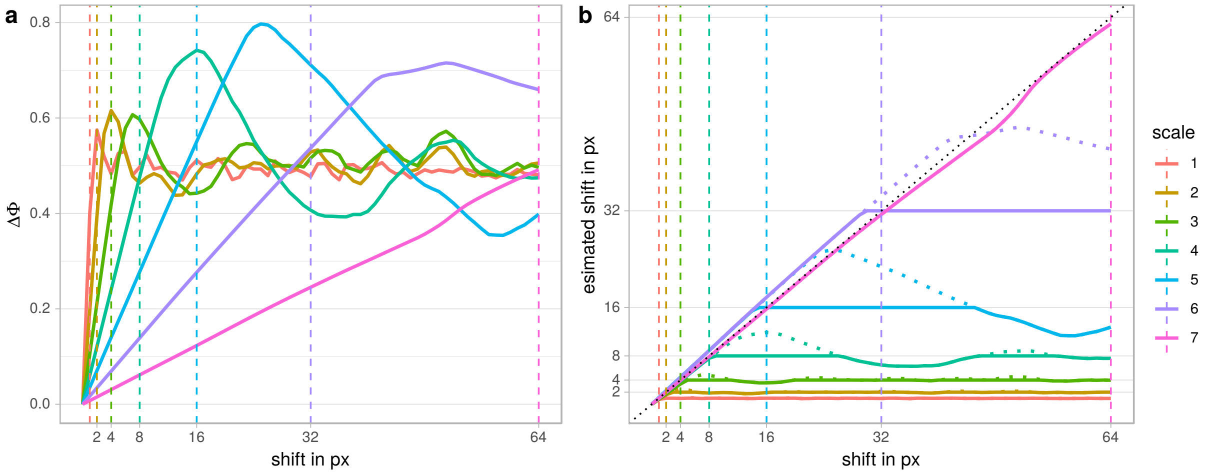


Figure 3: Scale-dependent phase-shift $\Delta\Phi(j)$ (a) and estimated shift in image space $2^j\Delta\Phi(j)$ (b) as a function of the true shift applied to the input image. For the solid lines in (b), $\Delta\Phi$ is cut off at 0.5.

equally.

We can use the observation that $\Delta\Phi(j) = 0.5$ is attained after $0.5 \cdot 2^j$ pixels to obtain a simple quantitative estimate of the displacement:

$$\Delta x(j) \begin{cases} \approx 2^j \cdot \Delta\Phi_j & \text{for } \Delta\Phi_j < 0.5 \\ > 2^{j-1} & \text{otherwise} \end{cases} \tag{9}$$

Figure 3 b shows that this estimator works quite well for our example. All scales agree approximately on the correct result until the shift exceeds 2^j at which point the corresponding estimate saturates. At very large displacements, we notice that the slope is not perfect, especially for $j = 6, 7$, but the deviations remain small. We confirmed that equation 9 is typically a good approximation across a large number of similar experiments (not shown).

Whenever a single summary measure of the overall displacement errors is needed, we will use the maximum estimated displacement, henceforth called

$$d_\Phi = \max_j(\Delta x(j)) \,. \tag{10}$$

For a single rigid displacement, d_Φ is our best estimate of the true error. In this simple case, the displacement error is generally easy to estimate, for example as the distance in centroids (equation 1). Due to their localized nature, the wavelets should, however, also be able to correctly handle more difficult situations with multiple different displacements. d_Φ then represents the most severe location error in the forecast. To demonstrate this capability, we consider some of the geometric test images from the MesoVICT project [Gilleland et al., 2020], shown in figure 4.

- The true displacement of $\Delta x = 40$ is almost correctly identified in (a-c). All scales smaller than $j = 7$ are saturated, the result depends only weakly on the position and orientation of the features.

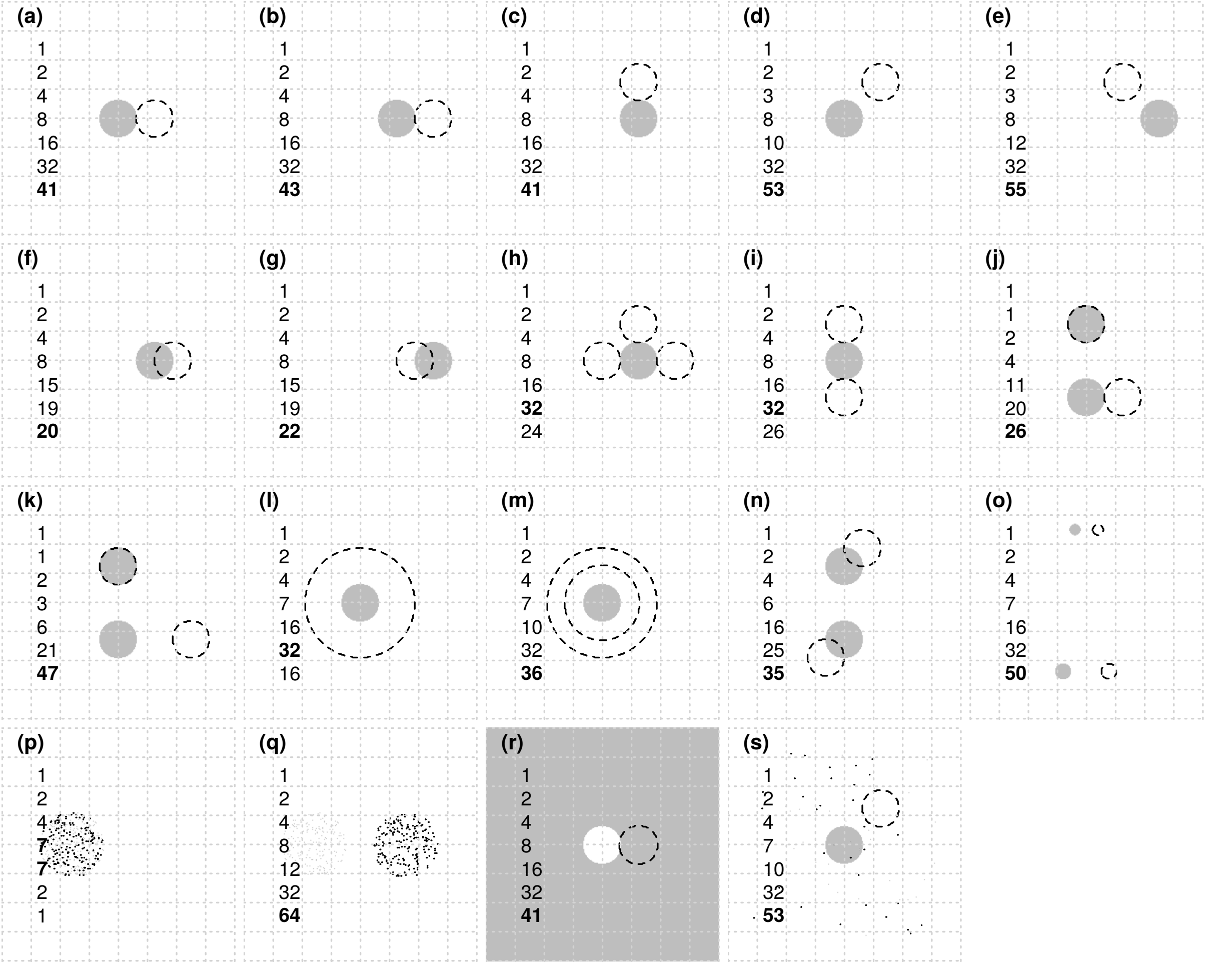


Figure 4: Estimated location errors for the circular test images. Grey areas mark the observation, dashed contours the forecasts. From top to bottom, the numbers indicate $\Delta x(j = 1, ..., 7)$, the final estimate is marked in bold. The distance between two grid lines is 32.

- (d) and (e) are correctly identified as worse than (a-c). (g) and (h) are recognized as better, the two largest scales approximately agree on the result, the estimated values are reasonably close to the correct solutions (57 in d and e, 20 in g, h).

- The addition of further features around the observation in (h) and (i) is considered an improvement over (a-c). These are two examples where the biggest error does not reside on the largest scale: With respect to $j = 7$, the placement of the features is decent; on smaller scales it is just as bad as in (a-c).

- Similarly, the additional hit in (j) leads to an overall improved score over the otherwise identical case (a).

- (k) is deemed worse than (a) on the largest scale but better on the small scales.

- (l) looks like a decent forecast on the largest scale while (m) is bad across almost all scales.

- The displacement in (n) is recognized but the two largest scales don't agree on a value. The shift indicated by $j = 6$ is close to the correct answer, at $j = 7$ each of the four daughters likely sees part of the unrelated feature and interprets it as a miss.

- (o) is among the overall worst forecasts.

- The correctly placed region of scattered pixels in (p) receives nearly perfect scores, the shifted region in (q) is maximally bad.

- The scores are nearly invariant under inversions of the image (compare r to b) and addition of noise (compare s to d).

Overall, these results confirm that sensible estimates of displacement errors can be extracted from the wavelet phase in scenarios which are more complex than a single, image-wide shift.

One aspect which is hardly tested by the MesoVICT cases is the direction dependence of the scores. It is easy to see that $\Delta\Phi$ will depend on the direction of the displacement errors if the features under consideration have a preferred orientation: Suppose the observed object is an elongated rectangle. If the forecast is displaced parallel to the longer edge, the two fields remain in phase along the overlapping section of their long edges, leading to a small overall phase error. A displacement perpendicular to the longer side will bring a much bigger portion of the image out of phase, thereby increasing $\Delta\Phi$ more quickly. Consequently, we cannot always recover the "true" shift as well as in figure 3 because the score effectively rewards forecasts with longer overlapping edges. In a real verification setting, where forecast errors are not the result of a simple image translation, this property is not necessarily a flaw.

Lastly, the localization of the daughter wavelets also allows us to display spatially varying displacement errors on a map. In our example case below (figure 7), we will use this to draw contours around the regions with the largest contributions to the overall $\Delta\Phi$. At the end of section 6, we also discuss the option of averaging $\Delta\Phi$ over time instead of space to obtain a map of scale-dependent mean displacements.

#	dates	time steps	weather events
1	20-22 Jun 2007	24	widespread convection
2	18-21 Jul 2007	48	airmass boundary
3	25-29 Sep 2007	48	cut-off low, Genoa cyclone
4	6-8 Aug 2007	24	airmass boundary, convergence line
5	18 Sep 2007	0	cold front
6	8-10 Jul 2007	24	widespread convection

Table 1: Dates of the MesoVICT test cases with a list of dominant weather events and the number of time steps used in this study.

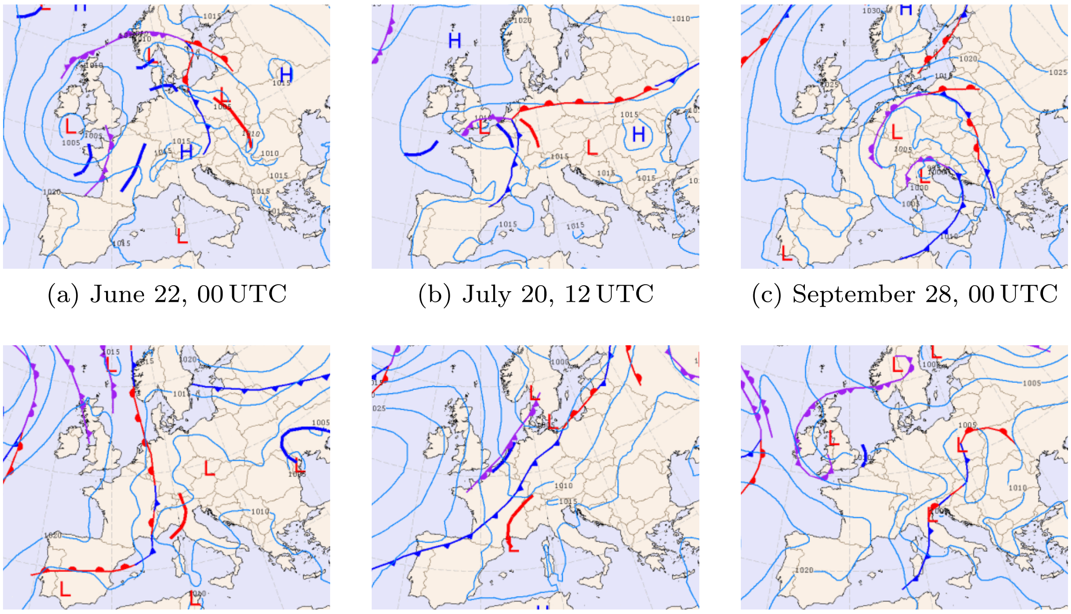
5. Data

The MesoVICT project relies on the Vienna Enhanced Resolution Analysis [Bica et al., 2007, VERA] as observational data against which all forecast models are verified. This model independent data set enhances interpolated station observations with thermal and dynamical fingerprints to produce maps of meteorological variables. In this study, we focus mainly on hourly precipitation sums, which are inferred from station observations alone. In addition, we explore the use of the novel score for absolute wind speed, potential and equivalent potential temperatures. For the latter two variables, the fingerprint method was applied, thereby introducing information from a finer-scaled orography beyond the resolution of the station network [Dorninger et al., 2018]. All data are interpolated to a regular 8 km grid covering central Europe (see maps in figure 7).

When the analysis domain is small compared to the typical features to be verified, displacement errors are hard to diagnose accurately because patterns are quickly displaced into or out of the domain. In the interest of avoiding such effects, as well as streamlining the experiment as a whole, we focus on two forecast models which cover the entire VERA domain: The hydrostatic BOLAM (BOlogna Limited Area Model), run at 0.07° resolution and the non-hydrostatic, convection permitting MOLOCH (MOdello LOCale) with 0.0225° grid-spacing which receives its boundary conditions from BOLAM. Re-forecasts with the 2015 operational version of this model chain were performed at ISPRA for the MesoVICT project [Mariani and Casaioli, 2018]. Both models were initialized at 12 UTC each day and run for 84 h (MOLOCH) and 108 h (BOLAM). The first twelve hour of each run were discarded as model spin-up time.

Table 1 lists the dates of the six MesoVICT test cases. With the exception of number five, all cases cover multiple days, thereby giving us the opportunity to compare forecasts from the same model with different lead-times. For the purpose of testing a new verification measure, this is convenient as it allows us to probe a wide range of error magnitudes and gives us a clear a priori expectation for which forecasts should, on average, be better than others. To take full advantage of this idea, we select those time-steps for which three different forecasts from each of the two models are available. This leaves us with the last day of cases 1, 4 and 6 and the last two days of cases 2 and 3 (168 time-steps in total). In the plots below, we will refer to the different forecasts as

- BOLAM007_1, MOL00225_1 with leadtimes +13 h, ..., +36 h



(a) June 22, 00 UTC (b) July 20, 12 UTC (c) September 28, 00 UTC

(d) August 8, 12 UTC (e) September 18, 00 UTC (f) July 10, 00 UTC

Figure 5: Representative synoptic analyses from the KNMI archive (`https://www.knmi.nl/nederland-nu/klimatologie/daggegevens/weerkaarten`) for MesoVICT test cases 1-6 (panels a-f). Case 5 is included for completeness but not used in this study.

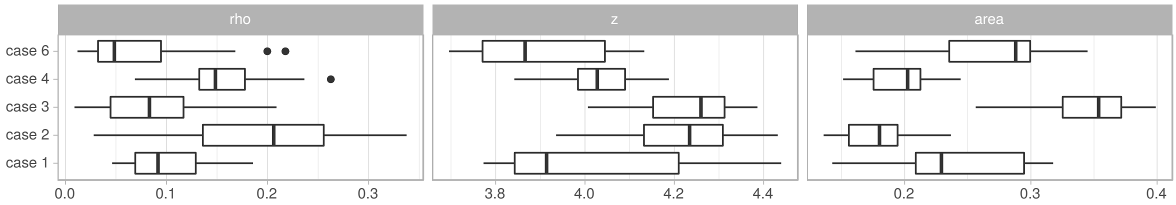


Figure 6: From left to right: Degree of anisotropy ρ, central scale z and fraction of the domain with non-zero rain, all calculated from VERA data for the five cases considered in this study.

- BOLAM007_2, MOL00225_2 with leadtimes $+37\,h, ..., +60\,h$
- BOLAM007_2, MOL00225_3 with leadtimes $+61\,h, ..., +84\,h$.

An overview of the synoptic situations in the different case studies is given in figure 5. As an objective measure of the spatial structure of the resulting rain fields analyzed in VERA, we also consider the degree of anisotropy ρ, dominant spatial scale z and total rain area in figure 6. Without discussing each case in detail, we observe that cases one and six have the weakest synoptic-scale forcing, leading to mostly isotropic structures across a relatively wide range of small and intermediate scales, covering roughly a quarter of the analysis domain – the precipitation fields are mostly convective. In contrast, cases two and four have anisotropic patterns with smaller areas. Here, the main precipitation regions are aligned along airmass boundaries and related convergence lines. Case four is smaller in scale, indicating a more prominent role of convection. Lastly, case three sees the strongest synoptic forcing from a cut-off low centered over

Germany and a related Genoa cyclone, resulting in large rain areas (up to 40 % of the domain) with a large, anisotropic pattern.

6. Verification of the MesoVICT data set

Based on the discussion above, we have a number of expectations for the outcome of the verification experiment:

- Short range forecasts are likely better than those with longer lead-times.
- The synoptically driven case three should be the more predictable over longer time-ranges than the others.
- The convective cases one and six are likely the most difficult to predict.

Whether the highly resolved MOLOCH will perform better or worse than BOLAM is unclear a priori, especially because the VERA analysis has a relatively coarse internal resolution and therefore produces smooth fields which look more similar to BOLAM. Scores which are sensitive to the spatial structure of the fields might therefore prefer the coarser forecasts. The high-resolution model could nonetheless be superior in terms of precipitation locations, especially in convectively driven situations.

In the next section, we focus on precipitation and compare the novel score to the established alternatives from section 2. The subsequent section briefly summarizes some of the results obtained for the other variables.

6.1. Precipitation

Before computing the various scores, we set all observed and predicted rainfall values below $0.1\,mm$ to zero. For the wavelet-based score d_Φ, rain intensities are replaced by their binary logarithm (setting $\log_2(0) \rightarrow \log_2(0.1)$) to reduce the impact of localized extremes and focus on the spatial distribution of rainfall as a whole (see also Buschow and Friederichs [2020]).

To give an impression of the new score's behaviour under realistic conditions, we present an example from the second MesoVICT case (figure 7). On 2007-07-20, the bulk of the observed precipitation field is linearly organized along an airmass boundary near the centre of the domain. The $23\,h$ BOLAM forecast predicts a visually similar linear feature with nearly correct placement and orientation. d_Φ registers a displacement of 16 pixels which corresponds to a single cell of the background grid in the figure. To visualize the detected errors, we have added contours around the pixels with the largest contributions to $\Delta\Phi(j)$ for $j = 3, 4, 5$. Focusing on scale 5 (shown in blue), we see that the leading edge of the simulated front lies roughly in the middle of the two squares whereas the observation align with the outer edges – a phase error of roughly half the support size at $j = 5$, i.e., 16 pixels or $128\,km$.

The previous day's BOLAM forecast (bottom part of figure 7) also simulates the linear pattern but clearly rotated and strongly displaced to the south-west. This is again easily seen by comparing the patterns within the blue box. In addition, the forecast contains a relatively intense

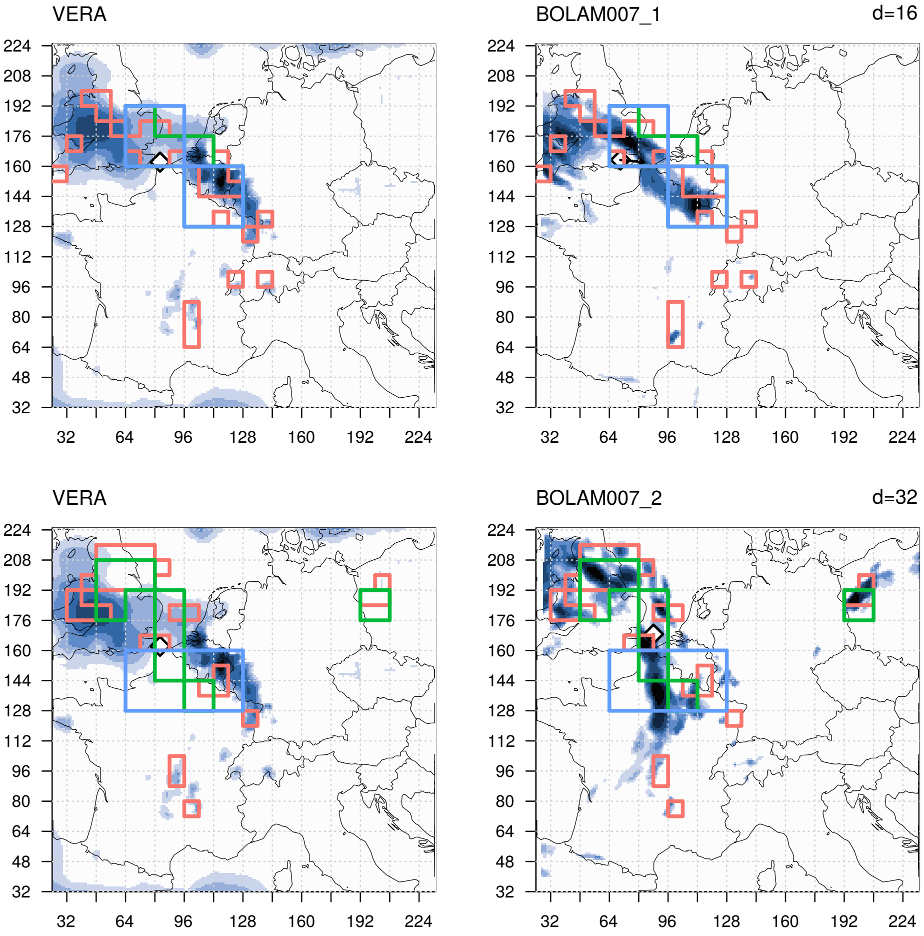


Figure 7: Phase verification for 2007-07-20 11 UTC: Contours encircle the regions with 5 % largest contributions to $\Delta\Phi$ at scales 3, 4, and 5 for the $+23\,h$ (top) and $+47\,h$ BOLAM07 forecast (bottom). The value of the maximum estimated displacement d is indicated in the top right corners. Points mark the fields' centroids, arrows point from the predicted to the observed centre.

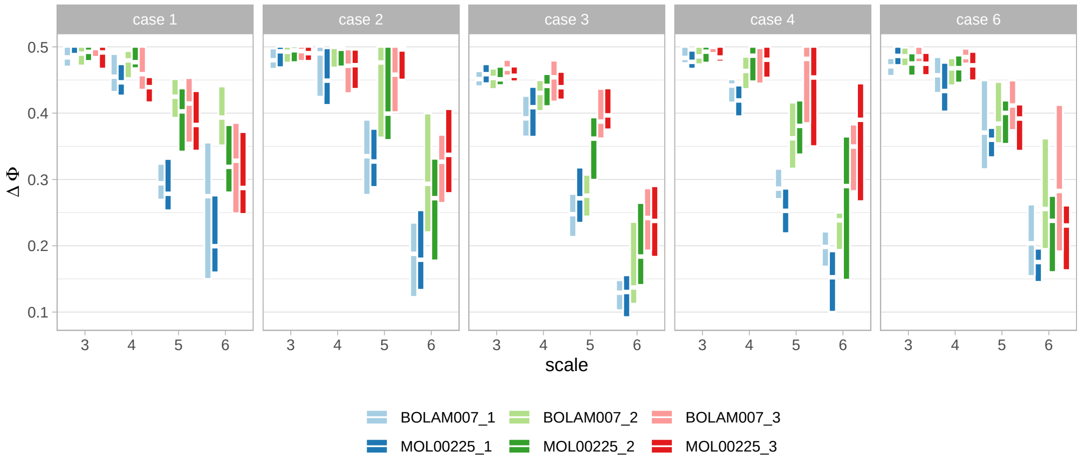


Figure 8: Inter-quartile range (colored bars) and median (white gaps) of $\Delta\Phi(j)$ for the different MesoVICT cases and forecasts.

false alarm near the north-east corner of the domain which is picked up by $\Delta\Phi(3,4)$ but not at larger scales because small-scale patterns have less impact there. Overall, we find $d_\Phi = 32$ which is the largest possible value in our case study.

In figure 7, we have marked the locations of the observed and predicted centres of mass which are the basis of SAL's location component L. We find that the centroids are almost identical in all three fields. For the 47 h forecast, the westward displacement of the front is compensated by the additional feature in eastern part of the domain, leading to a centroid displacement close to zero. This is one of two common scenarios that can lead to substantial disagreement between d_Φ and L: In a complex precipitation field with multiple objects, individual displacements (or misses and false alarms) can cancel out to create a centroid location near the centre of the domain, potentially leading to $d_\Phi >> L_1$. The opposite result can occur when forecast and observations contain precipitation regions at the same locations but with different relative intensities. In this scenario, L_1 may be large since the centroid shifts towards the most intense feature while d_Φ would likely see only small phase errors at each precipitation location.

We now begin our systematic verification of the entire data set with a look at the individual phase differences $\Delta\Phi$, separated by scale, forecast and case number in figure 8. Recalling that the worst case is $\Delta\Phi \geq 0.5$, we observe that none of the forecasts have any appreciable skill at scales smaller than $j = 4$; almost all are at least slightly skillful at the largest scale $j = 6$. As expected, the forecasts started on the previous day are almost universally superior to those with longer lead times, the advantage being most evident on scale five. The overall quality of the predictions, as well as the range from best to worst forecast, differ substantially from case to case: A clear difference between the two- and three-day forecasts is evident at all scales $j > 3$ in cases three and four. Only the most recent forecasts stand out in the first two cases.

The remarkable lack of lead-time dependence, as well as the overall mediocre performance in case six is likely due to the dominant role of convective activity (smallest observed scales in figure 6), the precise timing and location of which is hardly predictable at lead times longer than a few hours. This is also the only case where the convection permitting MOLOCH consistently out-

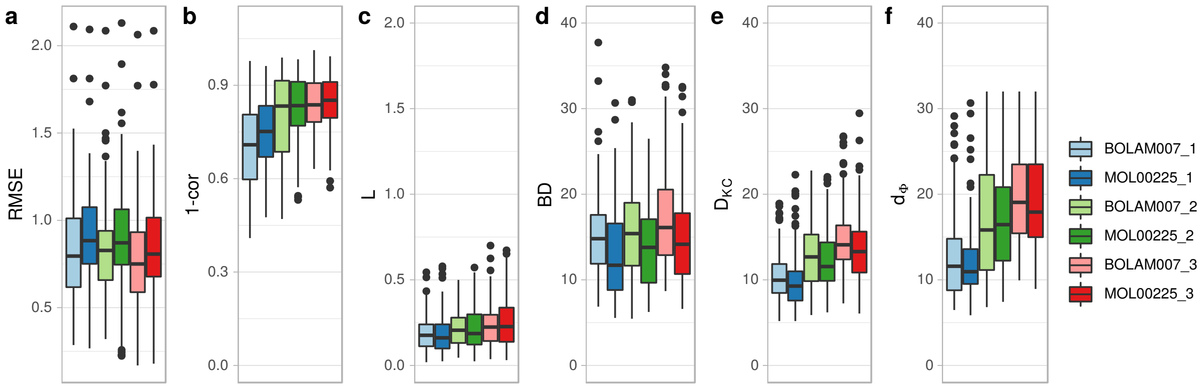


Figure 9: Distribution of RMSE (a), one minus linear correlation (b), SAL's location component (c), BD (d), D$_{KC}$ (e), and d_Φ (f) for the six competing forecasts.

performs the coarser BOLAM at all scales and lead-times. At the shortest lead times, MOLOCH furthermore has slight advantages in cases one and four which exhibit relatively small-scaled structures as well.

Conversely, case three sees the strongest synoptic forcing and was overall forecast best. The pronounced lead-time dependence indicates some remaining difficulty in predicting the precise path of the cut-off low. The difference between lead times is even stronger in case four where the formation and movement of the linearly organized precipitation patterns proved difficult to forecast more than one day in advance.

While the difference in quality between the $+12\,h$, $+36\,h$ and $+50\,h$ forecasts is thus often obvious, no clear winner emerges from the comparison between BOLAM and its finer-scaled sister model MOLOCH. This overall ranking is also reflected by the resulting displacement errors d_Φ shown in figure 9 (f). Here we see that the median displacement across all cases is around 12 pixels (roughly 100 km) for the shortest lead times, 16 for leadtimes greater than 36 h and slightly below 20 for the longest-range forecasts. The three ranges of leadtimes are clearly separated: Values that fall in the upper quartile for day one are near the median of day two and the lower quartile of day three. The worst-case value of $d_\Phi = 32$ is rare even on the third forecast day.

Interestingly, each of the other five scores shown in figure 9 (a-e) paints a different picture. As expected, RMSE (panel a) is a textbook example of the double penalty issue with hardly any difference between leadtimes but a strong preference for the coarser resolved BOLAM model. The linear correlation coefficient, shown in panel (b), mostly rewards overlap between forecast and observation and therefore naturally prefers BOLAM as well. In addition, most of the longer-ranged predictions hardly overlap the observed field at all, leading to near zero correlations in most instances on days two and three. In stark contrast to the overall bad performance with respect to correlations, SAL's location component (c) indicates low values for all forecasts with hardly any preference for either model and a very weak dependence on leadtime. The reason for this behaviour is explained by figure 10: Due to the frequently complex and wide-spread nature of these precipitation fields, their centroids are usually concentrated near the centre of the domain, leading to small values of L_1. The other half of L measures the scattering around

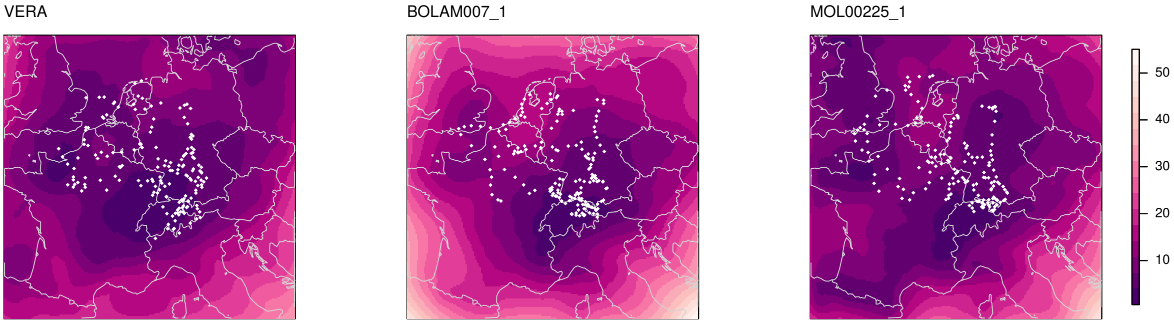


Figure 10: Average distance (in pixels) to the next rainy gridpoint for VERA and the most recent BOLAM and MOLOCH forecasts. White dots mark the position of the field's centre of mass for all cases.

the centroid and is more a structural characteristic than a measure of displacement.

The distance measure BD (panel d), shows perhaps the most surprising behaviour (cf. figure 9 e): Like RMSE, it registers almost no lead time dependence but instead of BOLAM, it clearly prefers the fine-scaled MOLOCH model! To understand this unexpected result, we must recall that distance measures are based on the distance from each rainy pixel in one field to the nearest rainy pixel in the other and therefore react sensitively to the presence or absence of small features in otherwise empty regions of the domain. Figure 10 reveals that the average distance to the nearest rain event is substantially too large in BOLAM, especially near the western and northern domain edge. While the model can decently simulate the main features of the precipitation field, it tends to neglect smaller-scaled showers throughout the rest of the domain. For the other scores, this effect is largely overshadowed by the displacement of the dominant precipitation systems.

The last score included in our comparison is the field-deformation score D_{KC} (panel e) which generally prefers MOLOCH as well while noting a similar decrease in forecast quality over time as d_Φ. It is possible that this score also rewards the finer model for producing additional smaller-scaled precipitation cells in the general vicinity of the observed rain areas: The addition of hits or near misses on small scales will tend to reduce the overall mean displacement vector. Conversely d_Φ, as defined in equation 10, will focus on the most intense parts of the image due to the weighting and ignores small-scale displacements if a big displacement is present on larger scales.

6.2. Other variables

While the location scores from section 2, as well as most others in the literature, were designed specifically for precipitation or similarly intermittent fields, our approach makes no such assumptions. As a proof of concept, we now apply the exact same methodology used for precipitation to fields of absolute wind speed in 10 m height (V), 2 m potential temperature (θ) and 2 m equivalent potential temperature (θ_e).

Figure 11 summarizes the scale-dependent phase errors $\Delta\Phi$ for all four variables. The most obvious difference between precipitation and the others is a substantial improvement of $\Delta\Phi$ at small scales and long lead-times. The increased small-scale skill is particularly obvious for θ where the phase errors at scale $j = 3$ are comparable to those seen at $j = 5$ for precipitation.

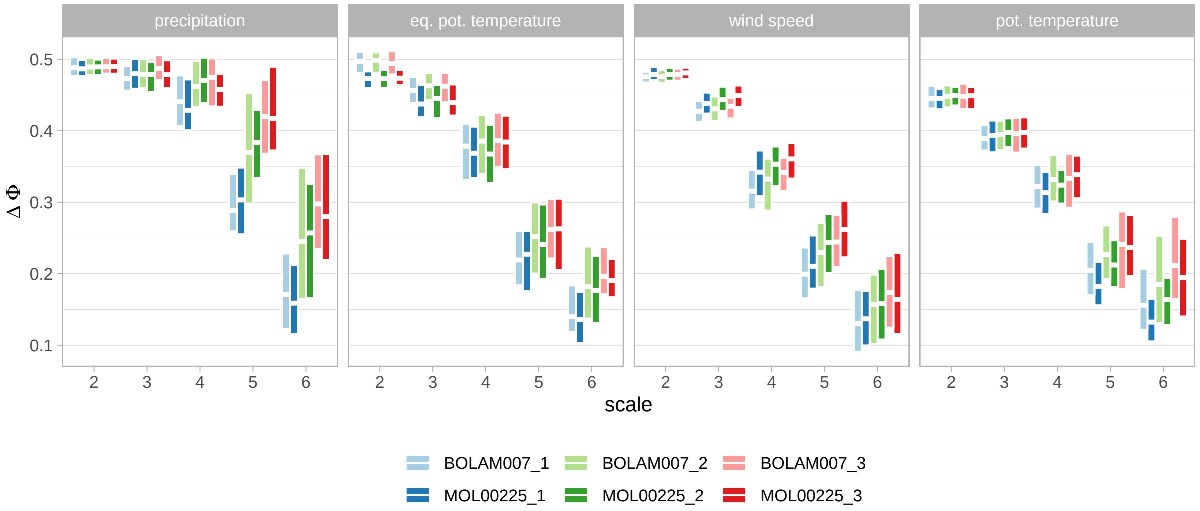


Figure 11: Inter-quartile range (colored bars) and median (white gaps) of $\Delta\Phi(j)$ for all variables and forecasts.

Wind speed and equivalent potential temperature exhibit slightly larger phase errors than θ, all three show a weak but consistent increase with lead-time. In comparing the two models, we observe that MOLOCH has some advantages for both θ_e (mostly on small scales) and θ (mostly on large scales). For wind speed, on the other hand, the coarser BOLAM model is slightly superior, especially on small scales.

An obvious explanation for the smaller displacements compared to precipitation is the presence of stationary objects like coast-lines and mountains. Depending on their representation in the model orography and land-surface fields, such features allow the forecasts to predict the location of spatial gradients in near-surface fields with high accuracy, even on small scales and multiple days in advance. To understand the qualitative difference between θ and V, we must recall that VERA enhances temperature-related variables with thermal and dynamic “fingerprints”. The interpolated station data are thereby imbued with additional fine-scaled texture far beyond the spatial resolution of the station network. This method was not applied for wind and precipitation. As a result, the finer-scaled wind features of MOLOCH appear erroneous compared to the analysis, thereby increasing the average phase errors.

An inherent advantage of the wavelet-approach is its natural capability to localize errors in space. To produce a map of average phase errors, we simply take the weighted mean over time instead of space. Figure 12 shows the results for all four variables but only one of the forecasts (images for all six forecasts look qualitatively similar). As expected, there is no coherent pattern for precipitation since the phases result form intermittent features materializing at various discrete locations across the domain. Only on the largest two scales, we see a slight tendency towards better forecast locations in the South-West and larger errors over Germany. In contrast, individual pixels with $\Delta\Phi << 0.5$ can be seen even at $j = 3$ for the other three variables. The regions of improved localization are primarily aligned along the coastlines. For θ, the Alps appear as an additional source of consistent localization which is reproduced by the model. On large scales ($j = 5, 6$), most of the pixels in the domain border on either a coast or mountain range

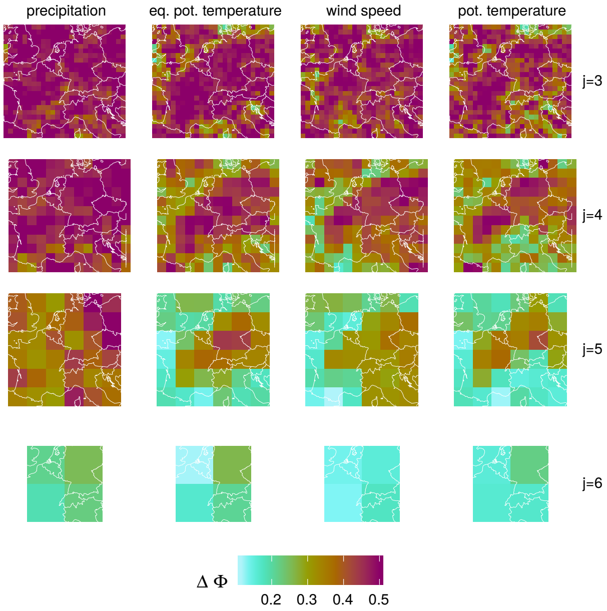


Figure 12: Weighted time-average of $\Delta\Phi(j)$ of the MOL00225_2 forecasts for all considered variables (left to right) and selected scales (top to bottom).

and consequently exhibit small $\Delta\Phi$.

7. Conclusions

In this study, we have introduced a novel location score which exploits the phase information of the dual-tree complex wavelet transform. Idealized tests with simple geometric shapes demonstrate that this score generally works as intended. The easiest of these tests consider a single, rigidly displaced feature and ask the score to reconstruct the magnitude of the shift. Like most other scores, the wavelet-approach can typically solve such problems. The localization property of the wavelets furthermore allows us to correctly assess more complex scenarios where multiple different displacements occur in different parts of the domain.

In a real-world verification setting, however, the problem is even more complicated because the existence of a well-defined location error is not guaranteed at all. This is particularly true when we consider fields like temperature and wind speed which do not naturally separate into discrete objects.

As a realistic case study, we used a subset of the MesoVICT test cases and compared one-, two- and three-day forecasts of hourly precipitation from the BOLAM-model and its higher-resolved sister-model MOLOCH to the VERA analysis. Our first experiment focused on precipitation fields and compared several scores which are sensitive to displacement errors. As expected, the point-wise RMSE uniformly prefers BOLAM. The linear correlation suffers from the same double penalty issue, but is at least capable of identifying the advantages of the one-day forecast over the others. More surprisingly, the distance measure BD exhibits the exact opposite behaviour of RMSE, preferring MOLOCH irrespective of lead-time. A look at the underlying distance maps shows that this score punishes BOLAM for neglecting small-scaled, scattered precipitation cells in parts of the domain. This result, while not uninformative, is more a property of the model climatology than the day-to-day precipitation placement. In agreement with our literature survey, SAL's L proved to have very little information on forecast performance, as indicated by miniscule differences between both models and leadtimes. It should be mentioned that this score was originally defined and optimized with a much smaller domain in mind (a single river catchment) – a setting where small numbers of features and single, well defined displacements are generally more likely than on our map of Western Europe.

With the novel wavelet-based score d_Φ, we find a clear decrease of skill with lead-time. In addition, advantages for MOLOCH are identified in the smaller-scaled, convectively driven cases 1, 4 and 6. These cases were also found to be more challenging to forecast than the synoptically dominated case 3. The most similar established score turned out to be the displacement component of DAS. The fact that this field deformation method showed a general preference for MOLOCH may be partially due to the same phenomenon as for BD. In contrast to d_ϕ neither of these scores give special weight to intense regions. We can furthermore conclude that, despite allowing divergent flow fields, DAS is not strongly sensitive to scale errors. Both it and d_Φ can thus be recommended as a complement to pure structure verification techniques and pure comparisons of the marginal distribution. When the structure is verified via SAD, d_Φ is a natural complementary score because it relies on the same wavelet transform (no additional computation needed) and utilizes exactly the information that SAD neglects.

Unlike the other scores in our inter-comparison, d_Φ requires no thresholding or object identification can therefore, in principle, be used to verify any meteorological field of interest. As a first demonstration, we have applied our method to equivalent and dry potential temperature, as well as absolute wind speed. All of these variables were considered near the surface since a relatively dense station network is needed to produce a spatial analysis. The local phases of the resulting wavelet transforms are therefore strongly influenced by mountains and coasts. Phase errors at these spatially fixed locations are likely caused by errors in the strength of local gradients, rather than a spatial displacement. Our scores represent averages over both these stationary features and more transient phenomena related to, for example, fronts, cyclones and convection. For the two temperature variables, VERA includes additional fine-scale information, resulting in slight advantages for the higher-resolved model. The localized nature of the wavelets furthermore allows us to study the distribution of consistent phase errors in space. For precipitation, this yielded little additional information due to the relatively small sample-size. The other variables, however, exhibit well-localized regions of improved location-skill, primarily along the coast lines.

While the option to apply the same location-score to a variety of atmospheric variables is doubtlessly convenient, a number of limitations must be kept in mind. Firstly, non-intermittent fields have variance in all parts of the domain and our score represents an average. It is therefore not always easy to identify the meteorological sources of the measured errors. A human might, for example, focus on the displacement of a cold front while the strongest spatial gradients, which dominate the score, are actually located on the coast. An obvious solution is to move up into the free atmosphere where surface features, as well as diurnal cycles, have less impact. This, however, exasperates the second main limitation, namely the lack of spatial observations. Interpolated data-sets like VERA can already be problematic near the surface since the density of station-networks is far coarser than the resolution of modern weather models like MOLOCH; at higher levels, spatial verification must either rely on reanalysis (which is not entirely model-independent) or novel remote-sensing data from satellites or (clear-air) radar and lidar scans.

References

Eric Gilleland, David Ahijevych, Barbara G. Brown, Barbara Casati, and Elizabeth E. Ebert. Intercomparison of Spatial Forecast Verification Methods. *Weather and Forecasting*, 24(5): 1416–1430, October 2009. ISSN 0882-8156, 1520-0434. doi: 10.1175/2009WAF2222269.1.

Manfred Dorninger, Eric Gilleland, Barbara Casati, Marion P. Mittermaier, Elizabeth E. Ebert, Barbara G. Brown, and Laurence J. Wilson. The Setup of the MesoVICT Project. *Bulletin of the American Meteorological Society*, 99(9):1887–1906, September 2018. ISSN 0003-0007, 1520-0477. doi: 10.1175/BAMS-D-17-0164.1.

Heini Wernli, Marcus Paulat, Martin Hagen, and Christoph Frei. SAL—A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts. *Monthly Weather Review*, 136(11): 4470–4487, November 2008. ISSN 0027-0644, 1520-0493. doi: 10.1175/2008MWR2415.1.

Michael Weniger and Petra Friederichs. Using the SAL Technique for Spatial Verification of Cloud Processes: A Sensitivity Analysis. *Journal of Applied Meteorology and Climatology*, 55 (9):2091–2108, September 2016. ISSN 1558-8424, 1558-8432. doi: 10.1175/JAMC-D-15-0311.1.

Florian Kapp, Petra Friederichs, Sebastian Brune, and Michael Weniger. Spatial verification of high-resolution ensemble precipitation forecasts using local wavelet spectra. *Meteorologische Zeitschrift*, 27(6):467–480, December 2018. ISSN 0941-2948. doi: 10.1127/metz/2018/0903.

Sebastian Buschow, Jakiw Pidstrigach, and Petra Friederichs. Assessment of wavelet-based spatial verification by means of a stochastic precipitation model (wv_verif v0.1.0). *Geoscientific Model Development*, 12(8):3401–3418, August 2019. ISSN 1991-9603. doi: 10.5194/gmd-12-3401-2019.

S. Buschow and P. Friederichs. Using wavelets to verify the scale structure of precipitation forecasts. *Advances in Statistical Climatology, Meteorology and Oceanography*, 6(1):13–30, 2020. doi: 10.5194/ascmo-6-13-2020.

S. Buschow and P. Friederichs. SAD: Verifying the scale, anisotropy and direction of precipitation forecasts. *Quarterly Journal of the Royal Meteorological Society*, 2021.

Heini Wernli, Christiane Hofmann, and Matthias Zimmer. Spatial Forecast Verification Methods Intercomparison Project: Application of the SAL Technique. *Weather and Forecasting*, 24(6): 1472–1484, December 2009. ISSN 0882-8156, 1520-0434. doi: 10.1175/2009WAF2222271.1.

K E Hanley, D J Kirshbaum, N M Roberts, and G Leoncini. Sensitivities of a Squall Line over Central Europe in a Convective-Scale Ensemble. *MONTHLY WEATHER REVIEW*, 141:22, 2013.

B Navascués. Long-term verification of HIRLAM and ECMWF forecasts over Southern Europe. *Atmospheric Research*, page 14, 2013.

Silvio Davolio, Francesco Silvestro, and Thomas Gastaldo. Impact of Rainfall Assimilation on High-Resolution Hydrometeorological Forecasts over Liguria, Italy. *JOURNAL OF HYDROMETEOROLOGY*, 18:22, 2017.

Barbara Früh, Jörg Bendix, Thomas Nauss, Marcus Paulat, Andreas Pfeiffer, Janus W Schipper, Boris Thies, and Heini Wernli. Verification of precipitation from regional climate simulations and remote-sensing observations with respect to ground-based observations in the upper Danube catchment. *Meteorol. Z.*, page 19, 2007.

Matthias Zimmer, Heini Wernli, Christoph Frei, and Martin Hagen. Feature-based verification of deterministic precipitation forecasts with SAL during COPS. In *Proceedings from the MAP D-PHASE Scientific Meeting in Bologna, Italy*, pages 116–121, 2008.

Petr Zacharov. Evaluation of the QPF of convective flash flood rainfalls over the Czech territory in 2009. *Atmospheric Research*, page 13, 2013.

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

David Lindstedt, Petter Lind, Erik Kjellström, and Colin Jones. A new regional climate model operating at the meso-gamma scale: performance over europe. *Tellus A: Dynamic Meteorology and Oceanography*, 67(1):24138, 2015.

A Kann, I Meirold-Mautner, F Schmid, G Kirchengast, J Fuchsberger, V Meyer, L Tüchler, and B Bica. Evaluation of high-resolution precipitation analyses using a dense station network. *Hydrol. Earth Syst. Sci.*, page 13, 2015.

Vera Maurer, Norbert Kalthoff, and Leonhard Gantner. Predictability of convective precipitation for West Africa: verification of convection-permitting and global ensemble simulations. *Meteorol. Z.*, page 18, 2017.

Eric Gilleland, Johan Lindström, and Finn Lindgren. Analyzing the Image Warp Forecast Verification Method on Precipitation Fields from the ICP. *Weather and Forecasting*, 25(4):1249–1262, August 2010. ISSN 0882-8156, 1520-0434. doi: 10.1175/2010WAF2222365.1.

Caren Marzban and Scott Sandgathe. Optical Flow for Verification. *Weather and Forecasting*, 25 (5):1479–1494, October 2010. ISSN 0882-8156, 1520-0434. doi: 10.1175/2010WAF2222351.1.

Christian Keil and George C. Craig. A Displacement and Amplitude Score Employing an Optical Flow Technique. *Weather and Forecasting*, 24(5):1297–1308, October 2009. ISSN 0882-8156, 1520-0434. doi: 10.1175/2009WAF2222247.1.

Christian Keil and George C. Craig. A Displacement-Based Error Measure Applied in a Regional Ensemble Forecasting System. *Monthly Weather Review*, 135(9):3248–3259, September 2007. ISSN 0027-0644, 1520-0493. doi: 10.1175/MWR3457.1.

A. Tafferner, C. Forster, M. Hagen, C. Keil, T. Zinner, and H. Volkert. Development and propagation of severe thunderstorms in the Upper Danube catchment area: Towards an integrated nowcasting and forecasting system using real-time data and high-resolution simulations. *Meteorology and Atmospheric Physics*, 101(3-4):211–227, October 2008. ISSN 0177-7971, 1436-5065. doi: 10.1007/s00703-008-0322-7.

George C Craig, Christian Keil, and Daniel Leuenberger. Constraints on the impact of radar rainfall data assimilation on forecasts of cumulus convection. *Quarterly Journal of the Royal Meteorological Society*, 138(663):340–352, 2012.

Heiner Lange and George C Craig. The Impact of Data Assimilation Length Scales on Analysis and Prediction of Convective Storms. *MONTHLY WEATHER REVIEW*, 142:28, 2014.

Zhuotong Nan, Shugong Wang, Xu Liang, Thomas E. Adams, William Teng, and Yao Liang. Analysis of Spatial Similarities Between NEXRAD and NLDAS Precipitation Data Products. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 3(3): 371–385, September 2010. ISSN 1939-1404, 2151-1535. doi: 10.1109/JSTARS.2010.2048418.

Patrick S. Skinner, Louis J. Wicker, Dustan M. Wheatley, and Kent H. Knopfmeier. Application of Two Spatial Verification Methods to Ensemble Forecasts of Low-Level Rotation. *Weather and Forecasting*, 31(3):713–735, June 2016. ISSN 0882-8156, 1520-0434. doi: 10.1175/WAF-D-15-0129.1.

Fan Han and Istvan Szunyogh. A Morphing-Based Technique for the Verification of Precipitation Forecasts. *Monthly Weather Review*, 144(1):295–313, January 2016. ISSN 0027-0644, 1520-0493. doi: 10.1175/MWR-D-15-0172.1.

Simon Barthelme. *imager: Image Processing Library Based on 'CImg'*, 2018. R package version 0.41.1.

Eric Gilleland, Gregor Skok, Barbara G Brown, Barbara Casati, Manfred Dorninger, Marion P Mittermaier, Nigel Roberts, and Laurence J Wilson. A novel set of geometric verification test fields with application to distance measures. *Monthly Weather Review*, 148(4):1653–1673, 2020.

Eric Gilleland. Spatial Forecast Verification: Baddeley's Delta Metric Applied to the ICP Test Cases. *Weather and Forecasting*, 26(3):409–415, June 2011. ISSN 0882-8156, 1520-0434. doi: 10.1175/WAF-D-10-05061.1.

Stephane G Mallat. A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

B Casati, G Ross, and DB Stephenson. A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorological Applications*, 11(2):141–154, 2004.

Idris A. Eckley, Guy P. Nason, and Robert L. Treloar. Locally stationary wavelet fields with application to the modelling and analysis of image texture: Modelling and Analysis of Image Texture. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, pages no–no, June 2010. ISSN 00359254, 14679876. doi: 10.1111/j.1467-9876.2009.00721.x.

Nick Kingsbury. Image processing with complex wavelets. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 357 (1760):2543–2560, September 1999. ISSN 1364-503X, 1471-2962. doi: 10.1098/rsta.1999.0447.

I.W. Selesnick, R.G. Baraniuk, and N.C. Kingsbury. The dual-tree complex wavelet transform. *IEEE Signal Processing Magazine*, 22(6):123–151, November 2005. ISSN 1053-5888. doi: 10.1109/MSP.2005.1550194.

B. Bica, R. Steinacker, C. Lotteraner, and M. Suklitsch. A new concept for high resolution temperature analysis over complex terrain. *Theoretical and Applied Climatology*, 90:173–183, November 2007. ISSN 0177-798X, 1434-4483. doi: 10.1007/s00704-006-0280-2.

Stefano Mariani and Marco Casaioli. Effects of model domain extent and horizontal grid size on contiguous rain area (CRA) analysis: A MesoVICT study. *Meteorologische Zeitschrift*, 27(6): 481–502, December 2018. ISSN 0941-2948. doi: 10.1127/metz/2018/0897.

Marcus Paulat. *Verifikation der Niederschlagsvorhersage für Deutschland von 2001–2004*. PhD thesis, Ph. D. thesis, University of Mainz, 2007.

Matthias Zimmer. *Merkmalsbezogene Verifikation hochaufgelöster Niederschlagsvorhersagen für Deutschland*. PhD thesis, Ph. D. thesis, Johannes Gutenberg-Universität, Mainz, Germany, 168 pp . . . , 2010.

B Vincendon, V Ducrocq, O Nuissier, and B Vie. Perturbation of convection-permitting NWP forecasts for flash-flood ensemble forecasting. *Nat. Hazards Earth Syst. Sci.*, page 16, 2011.

T Haiden, A Kann, C Wittmann, G Pistotnik, B Bica, and C Gruber. The Integrated Nowcasting through Comprehensive Analysis (INCA) System and Its Validation over the Eastern Alpine Region. *WEATHER AND FORECASTING*, 26:18, 2011.

Zbynek Sokol and Petr Zacharov. Nowcasting of precipitation by an NWP model using assimilation of extrapolated radar reflectivity. *Quarterly Journal of the Royal Meteorological Society*, 138(665):1072–1082, 2012.

AF Prein, Andreas Gobiet, Martin Suklitsch, Heimo Truhetz, NK Awan, Klaus Keuler, and Goran Georgievski. Added value of convection permitting seasonal simulations. *Climate Dynamics*, 41(9-10):2655–2677, 2013.

Stefan Schneider, Yong Wang, Wolfgang Wagner, and Jean-Francois Mahfouf. Impact of ASCAT Soil Moisture Assimilation on Regional Precipitation Forecasts: A Case Study for Austria. *Monthly Weather Review*, 142(4):1525–1541, April 2014. ISSN 0027-0644, 1520-0493. doi: 10.1175/MWR-D-12-00311.1.

Jill Hardy, Jonathan J Gourley, Pierre-Emmanuel Kirstetter, Yang Hong, Fanyou Kong, and Zachary L Flamig. A Method for Probabilistic Flash Flood Forecasting. *Journal of Hydrology*, 541:480–494, 2016.

Theresa Schellander-Gorgas, Yong Wang, Florian Meier, Florian Weidle, Christoph Wittmann, and Alexander Kann. On the forecast skill of a convection-permitting ensemble. *Geosci. Model Dev.*, page 22, 2017.

F Gofa, D Boucouvala, P Louka, and HA Flocas. Spatial verification approaches as a tool to evaluate the performance of high resolution precipitation forecasts. *Atmospheric Research*, 208:78–87, 2018.

A. Use of SAL's location component in the literature

study	result of L	relevant quote
Paulat [2007]	hardly any difference between models, no lead-time dependence	
Früh et al. [2007]	not discussed	
Zimmer et al. [2008]	not discussed	
Wittmann et al. [2010]	uninformative	“The results for the location component L do not yield much information [...]”
Zimmer [2010]	no difference between models	
Vincendon et al. [2011]	hardly discussed	“The L component does not show systematic behaviour with S and A components.”
Haiden et al. [2011]	uninformative	“The location score shows little variability both for summer and winter.”
Sokol and Zacharov [2012]	hardly informative	“L is usually lower than 0.2.”
Prein et al. [2013]	not discussed	“Since we found that there are no large changes in the location (L) component between different simulations, the focus here lies on changes in the structure (S) and amplitude (A) component.”
Zacharov [2013]	not discussed	
Hanley et al. [2013]	**detailed analysis of $L_{1,2}$ useful in single cases**	
Navascués [2013]	**moderate difference between models**	
Lindstedt et al. [2015]	uninformative	“As the L component is very similar among the models, we focus on S and A.”
Schneider et al. [2014]	uninformative	“The location score L is quite similar for all regions, which is a typical behavior for this parameter when averaging it for several events”
Kann et al. [2015]	uninformative	“The location indicator [...] does not yield conclusive results”
Hardy et al. [2016]	marginal leadtime-dependence, hardly significant	
Maurer et al. [2017]	hardly informative	“The values [of L] are generally low, because the precipitation objects are large compared to the size of the evaluation domain, which causes the centers of mass to be mainly located in the middle of the domain.”
Schellander-Gorgas et al. [2017]	uninformative	“The location score [...] shows not as much variability as the other two components.”
Davolio et al. [2017]	**indicates expected improvement**	“The L component is strongly improved [...] for the NUDG simulations. Moreover, blue dots, which are those indicating largely misplaced QPF, disappear.”
Gofa et al. [2018]	no difference between models	

Table 2: Use of SAL’s L-component in the literature.

Bibliography

Ahijevych, D., Gilleland, E., Brown, B. G., & Ebert, E. E. (2009). Application of spatial verification methods to idealized and NWP-gridded precipitation forecasts. *Weather and Forecasting*, *24*(6), 1485–1497.

Baddeley, A. J. (1992). An error metric for binary images. *Robust computer vision*, *5978*.

Banghoff, J. R., Sorber, J. D., Stensrud, D. J., Young, G. S., & Kumjian, M. R. (2020). A 10-Year Warm-Season Climatology of Horizontal Convective Rolls and Cellular Convection in Central Oklahoma. *Monthly Weather Review*, *148*(1), 21–42.

Bellier, J., Scheuerer, M., & Hamill, T. M. (2020). Precipitation Downscaling with Gibbs Sampling: An Improved Method for Producing Realistic, Weather-Dependent, and Anisotropic Fields. *Journal of Hydrometeorology*, *21*(11), 2487–2505.

Benjamin, S. G., Weygandt, S. S., Brown, J. M., Hu, M., Alexander, C. R., Smirnova, T. G., Olson, J. B., James, E. P., Dowell, D. C., Grell, G. A., et al. (2016). A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Monthly Weather Review*, *144*(4), 1669–1694.

Bica, B., Steinacker, R., Lotteraner, C., & Suklitsch, M. (2007). A new concept for high resolution temperature analysis over complex terrain. *Theoretical and Applied Climatology*, *90*, 173–183.

Briggs, W. M., & Levine, R. A. (1997). Wavelets and field forecast verification. *Monthly Weather Review*, *125*(6), 1329–1341.

Brune, S., Buschow, S., & Friederichs, P. (2020). Observations and high-resolution Simulations of Convective Precipitation Organization over the Tropical Atlantic. *Quarterly Journal of the Royal Meteorological Society*, (p. qj.3751).

Brune, S., Buschow, S., & Friederichs, P. (2021). The local wavelet-based organization index – Quantification, localization and classification of convective organization from radar and satellite data. *Quarterly Journal of the Royal Meteorological Society*.

Brune, S., Kapp, F., & Friederichs, P. (2018). A wavelet-based analysis of convective organization in ICON large-eddy simulations. *Quarterly Journal of the Royal Meteorological Society*, *144*(717), 2812–2829.

Buschow, S. (2019). wv_verif software repository.

Buschow, S. (2020). *sad: Verify the Scale, Anisotropy and Direction of Weather Forecasts*. R package version 0.1.3.

Buschow, S. (2021a). Code and data for Buschow and Friederichs (2021) "Verification of Near Surface Wind Patterns in Germany using Clear Air Radar Echoes".

Buschow, S. (2021b). Measuring Displacement Errors with Complex Wavelets. In preparation.

Buschow, S., & Friederichs, P. (2018). Local dimension and recurrent circulation patterns in long-term climate simulations. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *28*(8), 083124.

Buschow, S., & Friederichs, P. (2020). Using wavelets to verify the scale structure of precipitation forecasts. *Advances in Statistical Climatology, Meteorology and Oceanography*, *6*(1), 13–30.

Buschow, S., & Friederichs, P. (2021a). SAD: Verifying the scale, anisotropy and direction of precipitation forecasts. *Quarterly Journal of the Royal Meteorological Society*.

Buschow, S., & Friederichs, P. (2021b). Verification of Near Surface Wind Patterns in Germany using Clear Air Radar Echoes. *Geoscientific Model Development Discussions*.

Buschow, S., Kingsbury, N., & Wareham, R. (2020). *dualtrees: Decimated and Undecimated 2D Complex Dual-Tree Wavelet Transform*. R package version 0.1.4.

Buschow, S., Pidstrigach, J., & Friederichs, P. (2019). Assessment of wavelet-based spatial verification by means of a stochastic precipitation model (wv_verif v0.1.0). *Geoscientific Model Development*, *12*(8), 3401–3418.

Casati, B. (2010). New developments of the intensity-scale technique within the spatial verification methods intercomparison project. *Weather and Forecasting*, *25*(1), 113–143.

Casati, B., Ross, G., & Stephenson, D. (2004). A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorological Applications*, *11*(2), 141–154.

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

Ching, J., Rotunno, R., LeMone, M., Martilli, A., Kosovic, B., Jimenez, P. A., & Dudhia, J. (2014). Convectively Induced Secondary Circulations in Fine-Grid Mesoscale Numerical Weather Prediction Models. *Monthly Weather Review*, *142*(9), 3284–3302.

Daubechies, I. (1988). Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, *41*(7), 909–996.

Daubechies, I. (1992). *Ten lectures on wavelets*, vol. 61. Siam.

Davis, C., Brown, B., & Bullock, R. (2006). Object-Based Verification of Precipitation Forecasts. Part I: Methodology and Application to Mesoscale Rain Areas. *Monthly Weather Review*, *134*(7), 1772–1784.

Davis, C. A., Brown, B. G., Bullock, R., & Halley-Gotway, J. (2009). The Method for Object-Based Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program. *Weather and Forecasting*, *24*(5), 1252–1267.

Done, J., Davis, C. A., & Weisman, M. (2004). The next generation of NWP: explicit forecasts of convection using the weather research and forecasting (WRF) model. *Atmospheric Science Letters*, *5*(6), 110–117.

Dorninger, M., Gilleland, E., Casati, B., Mittermaier, M. P., Ebert, E. E., Brown, B. G., & Wilson, L. J. (2018). The Setup of the MesoVICT Project. *Bulletin of the American Meteorological Society*, *99*(9), 1887–1906.

Ebert, E., & McBride, J. (2000). Verification of precipitation in weather systems: Determination of systematic errors. *Journal of hydrology*, *239*(1-4), 179–202.

Ebert, E. E. (2008). Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteorological Applications*, *15*(1), 51–64.

Eckley, I. A., Nason, G. P., & Treloar, R. L. (2010). Locally stationary wavelet fields with application to the modelling and analysis of image texture: Modelling and Analysis of Image Texture. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Ekström, M. (2016). Metrics to identify meaningful downscaling skill in WRF simulations of intense rainfall events. *Environmental Modelling & Software*, *79*, 267–284.

Faranda, D., Messori, G., & Yiou, P. (2017). Dynamical proxies of North Atlantic predictability and extremes. *Scientific reports*, *7*(1), 1–10.

Farchi, A., Bocquet, M., Roustan, Y., Mathieu, A., & Quérel, A. (2016). Using the Wasserstein distance to compare fields of pollutants: Application to the radionuclide atmospheric dispersion of the Fukushima-Daiichi accident. *Tellus B: Chemical and Physical Meteorology*, *68*(1), 31682.

Gabor, D. (1946). Theory of communication. Part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, *93*(26), 429–441.

Gilleland, E. (2011). Spatial forecast verification: Baddeley's delta metric applied to the icp test cases. *Weather and Forecasting*, *26*(3), 409–415.

Gilleland, E. (2021). Novel measures for summarizing high-resolution forecast performance. *Advances in Statistical Climatology, Meteorology and Oceanography*, *7*(1), 13–34.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., & Ebert, E. E. (2009). Intercomparison of Spatial Forecast Verification Methods. *Weather and Forecasting*, *24*(5), 1416–1430.

Gilleland, E., Ahijevych, D. A., Brown, B. G., & Ebert, E. E. (2010a). Verifying Forecasts Spatially. *Bulletin of the American Meteorological Society*, *91*(10), 1365–1376.

Gilleland, E., Lindström, J., & Lindgren, F. (2010b). Analyzing the Image Warp Forecast Verification Method on Precipitation Fields from the ICP. *Weather and Forecasting*, *25*(4), 1249–1262.

Gilleland, E., Skok, G., Brown, B. G., Casati, B., Dorninger, M., Mittermaier, M. P., Roberts, N., & Wilson, L. J. (2020). A Novel Set of Geometric Verification Test Fields with Application to Distance Measures. *Monthly Weather Review*, *148*(4), 1653–1673.

Goupillaud, P., Grossmann, A., & Morlet, J. (1984). Cycle-octave and related transforms in seismic signal analysis. *Geoexploration*, *23*(1), 85–102.

Haar, A. (1910). Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen*, *69*(3), 331–371.

Han, F., & Szunyogh, I. (2016). A Morphing-Based Technique for the Verification of Precipitation Forecasts. *Monthly Weather Review*, *144*(1), 295–313.

Han, F., & Szunyogh, I. (2018). A Technique for the Verification of Precipitation Forecasts and Its Application to a Problem of Predictability. *Monthly Weather Review*, *146*(5), 1303–1318.

Hewer, R., Friederichs, P., Hense, A., & Schlather, M. (2017). A Matérn-Based Multivariate Gaussian Random Process for a Consistent Model of the Horizontal Wind Components and Related Variables. *Journal of the Atmospheric Sciences*, *74*(11), 3833–3845.

Holschneider, M., Kronland-Martinet, R., Morlet, J., & Tchamitchian, P. (1990). A real-time algorithm for signal analysis with the help of the wavelet transform. In *Wavelets*, (pp. 286–297). Springer.

Horn, B. K., & Schunck, B. G. (1981). Determining optical flow. In *Techniques and Applications of Image Understanding*, vol. 281, (pp. 319–331). International Society for Optics and Photonics.

Hou, J., & Wang, P. (2019). A region-tree-based approach for the verification of precipitation forecasts. *Monthly Weather Review*, *147*(4), 1257–1275.

Jacobson, J., Kleiber, W., Scheuerer, M., & Bellier, J. (2020). Beyond univariate calibration: verifying spatial structure in ensembles of forecast fields. *Nonlinear Processes in Geophysics*, *27*(3), 411–427.

Jansen, M., Nason, G. P., & Silverman, B. W. (2009). Multiscale methods for data on graphs and irregular multidimensional situations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(1), 97–125.

Kaiser, G. (2010). *A friendly guide to wavelets*. Springer Science & Business Media.

Kapp, F., Friederichs, P., Brune, S., & Weniger, M. (2018). Spatial verification of high-resolution ensemble precipitation forecasts using local wavelet spectra. *Meteorologische Zeitschrift*, *27*(6), 467–480.

Keil, C., & Craig, G. C. (2007). A Displacement-Based Error Measure Applied in a Regional Ensemble Forecasting System. *Monthly Weather Review*, *135*(9), 3248–3259.

Keil, C., & Craig, G. C. (2009). A displacement and amplitude score employing an optical flow technique. *Weather and Forecasting*, *24*(5), 1297–1308.

Keller, J. D., Delle Monache, L., & Alessandrini, S. (2017). Statistical downscaling of a high-resolution precipitation reanalysis using the analog ensemble method. *Journal of Applied Meteorology and Climatology*, *56*(7), 2081–2095.

Kikuchi, K., & Wang, B. (2010). Spatiotemporal Wavelet Transform and the Multiscale Behavior of the Madden–Julian Oscillation. *Journal of Climate*, *23*(14), 3814–3834.

Kingsbury, N. (1999). Image processing with complex wavelets. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *357*(1760), 2543–2560.

Kingsbury, N. (2006). Rotation-invariant local feature matching with complex wavelets. In *2006 14th European Signal Processing Conference*, (pp. 1–5). IEEE.

Kumar, P., & Foufoula-Georgiou, E. (1993). A multicomponent decomposition of spatial rainfall fields: 1. Segregation of large-and small-scale features using wavelet transforms. *Water Resources Research*, *29*(8), 2515–2532.

Lack, S. A., Limpert, G. L., & Fox, N. I. (2010). An object-oriented multiscale verification scheme. *Weather and forecasting*, *25*(1), 79–92.

Leistedt, B., McEwen, J. D., Vandergheynst, P., & Wiaux, Y. (2013). S2LET: A code to perform fast wavelet analysis on the sphere. *Astron. and Astrophys.*, *558*(A128), 1–9.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., & Gneiting, T. (2017). Forecaster's dilemma: extreme events and forecast evaluation. *Statistical Science*, (pp. 106–127).

Lobligeois, F., Andréassian, V., Perrin, C., Tabary, P., & Loumagne, C. (2014). When does higher spatial resolution rainfall information improve streamflow simulation? An evaluation using 3620 flood events. *Hydrology and Earth System Sciences*, *18*(2), 575–594.

Loritz, R., Hrachowitz, M., Neuper, M., & Zehe, E. (2021). The role and value of distributed precipitation data in hydrological models. *Hydrology and Earth System Sciences*, *25*(1), 147–167.

Lucas, B. D., & Kanade, T. (1981). An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, (p. 674–679). Morgan Kaufmann Publishers Inc.

Mallat, S. (1999). *A wavelet tour of signal processing*. Elsevier.

Mallat, S. G. (1989). A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE transactions on pattern analysis and machine intelligence*, *11*(7), 674–693.

Mariani, S., & Casaioli, M. (2018). Effects of model domain extent and horizontal grid size on contiguous rain area (CRA) analysis: A MesoVICT study. *Meteorologische Zeitschrift*, *27*(6), 481–502.

Marzban, C., & Sandgathe, S. (2009). Verification with Variograms. *Weather and Forecasting*, *24*(4), 1102–1120.

Marzban, C., & Sandgathe, S. (2010). Optical Flow for Verification. *Weather and Forecasting*, *25*(5), 1479–1494.

Mass, C. F., Ovens, D., Westrick, K., & Colle, B. A. (2002). Does increasing horizontal resolution produce more skillful forecasts? The results of two years of real-time numerical weather prediction over the Pacific Northwest. *Bulletin of the American Meteorological Society*, *83*(3), 407–430.

Messori, G., & Faranda, D. (2021). Technical note: Characterising and comparing different palaeoclimates with dynamical systems theory. *Climate of the Past*, *17*(1), 545–563.

Murphy, A. H., & Winkler, R. L. (1987). A general framework for forecast verification. *Monthly weather review*, *115*(7), 1330–1338.

Nason, G. P., Von Sachs, R., & Kroisandt, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(2), 271–292.

Nelson, J. D. B., Gibberd, A. J., Nafornita, C., & Kingsbury, N. (2018). The locally stationary dual-tree complex wavelet model. *Statistics and Computing*, *28*(6), 1139–1154.

Nerini, D., Besic, N., Sideris, I., Germann, U., & Foresti, L. (2017). A non-stationary stochastic ensemble generator for radar rainfall fields based on the short-space Fourier transform. *Hydrology and Earth System Sciences*, *21*(6), 2777–2797.

Poll, S., Shrestha, P., & Simmer, C. (2017). Modelling convectively induced secondary circulations in the terra incognita with TerrSysMP. *Quarterly Journal of the Royal Meteorological Society*, *143*(707), 2352–2361.

Radanovics, S., Vidal, J.-P., & Sauquet, E. (2018). Spatial verification of ensemble precipitation: an ensemble version of SAL. *Weather and Forecasting*, *33*(4), 1001–1020.

Roberts, N. M., & Lean, H. W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*(1), 78–97.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, *40*(2), 99–121.

Scheuerer, M., & Hamill, T. M. (2015). Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities. *Monthly Weather Review*, *143*(4), 1321–1334.

Schlager, C., Kirchengast, G., Fuchsberger, J., Kann, A., & Truhetz, H. (2019). A spatial evaluation of high-resolution wind fields from empirical and dynamical modeling in hilly and mountainous terrain. *Geoscientific Model Development*, *12*(7), 2855–2873.

Scovell, R. W. (2020). Applications of Directional Wavelets, Universal Multifractals and Anisotropic Scaling in Ensemble Nowcasting; A Review of Methods with Case Studies. *Quarterly Journal of the Royal Meteorological Society*, (p. qj.3780).

Selesnick, I., Baraniuk, R., & Kingsbury, N. (2005). The dual-tree complex wavelet transform. *IEEE Signal Processing Magazine*, *22*(6), 123–151.

Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A., & Vandergheynst, P. (2013). The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, *30*(3), 83–98.

Skinner, P. S., Wicker, L. J., Wheatley, D. M., & Knopfmeier, K. H. (2016). Application of Two Spatial Verification Methods to Ensemble Forecasts of Low-Level Rotation. *Weather and Forecasting*, *31*(3), 713–735.

Skok, G. (2015). Analysis of fraction skill score properties for a displaced rainband in a rectangular domain. *Meteorological Applications*, *22*(3), 477–484.

Skok, G. (2016). Analysis of fraction skill score properties for a displaced rainy grid point in a rectangular domain. *Atmospheric Research*, *169*, 556–565.

Skok, G., & Hladnik, V. (2018). Verification of Gridded Wind Forecasts in Complex Alpine Terrain: A New Wind Verification Methodology Based on the Neighborhood Approach. *Monthly Weather Review*, *146*(1), 63–75.

Skok, G., & Roberts, N. (2018). Estimating the displacement in precipitation forecasts using the fractions skill score. *Quarterly Journal of the Royal Meteorological Society*, *144*(711), 414–425.

Stucki, P., Froidevaux, P., Zamuriano, M., Isotta, F. A., Messmer, M., & Martynov, A. (2020). Simulations of the 2005, 1910, and 1876 Vb cyclones over the Alps – sensitivity to model physics and cyclonic moisture flux. *Natural Hazards and Earth System Sciences*, *20*(1), 35–57.

Sweldens, W. (1995). Lifting scheme: a new philosophy in biorthogonal wavelet constructions. In *Wavelet applications in signal and image processing III*, vol. 2569, (pp. 68–79). International Society for Optics and Photonics.

Theis, S. (2005). *Deriving probabilistic short-range forecasts from a deterministic high-resolution model*. Ph.D. thesis, University of Bonn.

Thurston, W., Fawcett, R. J. B., Tory, K. J., & Kepert, J. D. (2016). Simulating Boundary-Layer Rolls with a NWP model. *Quarterly Journal of the Royal Meteorological Society*, *142*(694), 211–223.

Torrence, C., & Compo, G. P. (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological society*, *79*(1), 61–78.

Valens, C. (1999). A really friendly guide to wavelets. *ed. Clemens Valens*.

Vidakovic, B., & Mueller, P. (1994). Wavelets for kids. *Instituto de Estadística, Universidad de Duke*.

Weniger, M., & Friederichs, P. (2016). Using the SAL Technique for Spatial Verification of Cloud Processes: A Sensitivity Analysis. *Journal of Applied Meteorology and Climatology*, *55*(9), 2091–2108.

Weniger, M., Kapp, F., & Friederichs, P. (2017). Spatial verification using wavelet transforms: A review. *Qarterly Journal of the Royal Meteorological Society*, *143*(702), 120–136.

Wernli, H., Paulat, M., Hagen, M., & Frei, C. (2008). SAL—A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts. *Monthly Weather Review*, *136*(11), 4470–4487.

Willeit, M., Amorati, R., Montani, A., Pavan, V., & Tesini, M. S. (2015). Comparison of spectral characteristics of precipitation from radar estimates and COSMO-model predicted fields. *Meteorology and Atmospheric Physics*, *127*(2), 191–203.

Wong, M., & Skamarock, W. C. (2016). Spectral Characteristics of Convective-Scale Precipitation Observations and Forecasts. *Monthly Weather Review*, *144*(11), 4183–4196.

Yano, J.-I., & Jakubiak, B. (2016). Wavelet-based verification of the quantitative precipitation forecast. *Dynamics of Atmospheres and Oceans*, *74*, 14–29.

Yu, B., Zhu, K., Xue, M., & Zhou, B. (2020). Using new neighborhood-based intensity-scale verification metrics to evaluate WRF precipitation forecasts at 4 and 12 km grid spacings. *Atmospheric Research*, *246*, 105117.

Zhou, B., Simon, J. S., & Chow, F. K. (2014). The Convective Boundary Layer in the Terra Incognita. *Journal of the Atmospheric Sciences*, *71*(7), 2545–2563.

Zschenderlein, P., Pardowitz, T., & Ulbrich, U. (2019). Application of an object-based verification method to ensemble forecasts of 10-m wind gusts during winter storms. *Meteorologische Zeitschrift*, (p. 90341).

Acronyms

BD Baddeley's Delta. 41, 43, 44, 58

CRA Continuous Rain Area verification method. 38, 44, 45, 58

DAS Displacement and Amplitude Score. 38, 39, 41–44

DTCWT Dual-Tree Complex Wavelet Transform. 27–32, 54, 56, 62, 65

DWT Discrete Wavelet Transform. 13, 20, 26–28, 30, 65

EMD Earth Mover's Distance. 52, 53, 55

FSS Fractions Skill Score. 34–36, 41–45, 47, 61

ICP Spatial Forecast Verification Methods Inter-Comparison Project. 3, 33, 46, 47

ISS Intensity Scale Skill-Score. 35, 36, 42–45, 48, 62

LSW Locally Stationary Wavelet Process. 21, 23, 31, 49

MesoVICT Mesoscale Verification Intercomparison over Complex Terrain. 3, 46, 47, 50, 56, 59

MODE Method for Object-Based Diagnostic Evaluation. 38, 42, 43, 45, 58

MRA Multi-Resolution Analysis. 12, 15, 16, 18, 20, 21, 27, 35, 36, 42, 44, 48, 49, 65

MSE Mean Square Error. 34, 35, 62

NWP Numerical Weather Prediction. 63

RDWT Redundant Discrete Wavelet Transform. 21, 23, 25, 26, 30, 51, 54, 62

RMSE Root Mean Square Error. 45, 59

SAD Scale, Anisotropy and Direction verification. 58, 61, 63

SAL Structure, Amplitude and Location score. 37, 41–45, 47, 49, 58, 59, 62, 63

SALAD Scale, Anisotropy, Location, Amplitude and Direction verification method (hypothetical). 194

Acknowledgments

I am immensely grateful to the many people who have helped me, professionally and personally, during my work on this project. Weather forecasts can be ranked by their quality, your contributions can not. I therefore thank, in alphabetical order, Joseph Bellier, Armin Blanke, Sebastian Brune, Barbara and Reiner Buschow, Barbara Casati, Anne Debus, Shashi Dixit, Petra Friederichs, Alex Gibberd, Eric Gilleland, Lucia Hallas, Andreas Hense, Florian Kapp, Felix Kleinert, Berthold Laue, Stefano Mariani, Franka Nawrath, Timon Netzel, Velibor Pejcic, Jakiw Pidstrigach, Ingeborg Rassow, Tobias Scharbach, Julian Steinheuer, Clemens Ullrich, Rich Wareham, Michael Weniger. I am furthermore thankful for the help from five anonymous reviews, the editors and typesetters at GMD, ASCMO and QJRMS, all organizers of the ICP and MesoVICT project and the members of the R-language core team. My work on this project was funded by the German Research Foundation DFG under grant no. FR 2976/2-1.

BONNER METEOROLOGISCHE ABHANDLUNGEN

Herausgegeben vom Institut für Geowissenschaften der Universität Bonn, Abteilung Meteorologie, durch Prof. Dr. H. FLOHN (Hefte 1-25), Prof. Dr. M. HANTEL (Hefte 26-35), Prof. Dr. H.-D. SCHILLING (Hefte 36-39), Prof. Dr. H. KRAUS (Hefte 40-49), ab Heft 50 durch Prof. Dr. A. HENSE.

Heft 1-69: siehe `https://www.ifgeo.uni-bonn.de/abteilungen/meteorologie/bibliothek/bonner-meteorologische-abhandlungen-bma`

70-92: open access, verfügbar unter `https://bonndoc.ulb.uni-bonn.de/xmlui/handle/20.500.11811/1627`

Heft 70: ***A S M Mostaquimur Rahman***: Influence of subsurface hydrodynamics on the lower atmosphere at the catchment scale, 2015, 98 S. + XVI.

Heft 71: ***Sabrina Wahl***: Uncertainty in mesoscale numerical weather prediction: probabilistic forecasting of precipitation, 2015, 108 S.

Heft 72: ***Markus Übel***: Simulation of mesoscale patterns and diurnal variations of atmospheric CO_2 mixing ratios with the model system TerrSysMP-CO_2, 2015, [erschienen] 2016, 158 S. + II

Heft 73: ***Christian Bernardus Maria Weijenborg***: Characteristics of Potential Vorticity anomalies associated with mesoscale extremes in the extratropical troposphere, 2015, [erschienen] 2016, 151 S. + XI

Heft 74: ***Muhammad Kaleem***: A sensitivity study of decadal climate prediction to aerosol variability using ECHAM6-HAM (GCM), 2016, 98 S. + XII

Heft 75: ***Theresa Bick***: 3D Radar reflectivity assimilation with an ensemble Kalman filter on the convective scale, 2016, [erschienen] 2017, 96 S. + IX

Heft 76: ***Zied Ben Bouallegue***: Verification and post-processing of ensemble weather forecasts for renewable energy applications, 2017, 119 S.

Heft 77: ***Julia Lutz***: Improvements and application of the STatistical Analogue Resampling Scheme STARS, 2016, [erschienen] 2017, 103 S.

Heft 78: ***Benno Michael Thoma***: Palaeoclimate Reconstruction in the Levant and on the Balkans, 2016, [erschienen] 2017, XVI, 266 S.

Heft 79: ***Ieda Pscheidt***: Generating high resolution precipitation conditional on rainfall observations and satellite data, 2017, V, 173 S.

Heft 80: ***Tanja Zerenner***: Atmospheric downscaling using multi-objective genetic programming, 2016, [erschienen] 2017, X, 191 S.

Heft 81: ***Sophie Stolzenberger***: On the probabilistic evaluation of decadal and paleoclimate model predictions, 2017, IV, 122 S.

Heft 82: ***Insa Thiele-Eich***: Flooding in Dhaka, Bangladesh, and the challenge of climate change, 2017, V, 158 S.

Heft 83: ***Liselotte Bach***: Towards a probabilistic regional reanalysis for Europe, 2017 [erschienen] 2018, VI, 114 S.

Heft 84: ***Yen-Sen Lu***: Propagation of land surface model uncertainties in terrestrial system states, 2017, [erschienen] 2018, X, 120 S.

Heft 85: ***Rüdiger Hewer***: Stochastic physical models for wind fields and precipitation extremes, 2018, 99 S.

Heft 86: ***Sebastian Knist***: Land-atmosphere interactions in multiscale regional climate change simulations over Europe, 2018, VIII, 147 S.

Heft 87: ***Jessica Keune***: Integrated terrestrial simulations at the continental scale: Impact of groundwater dynamics and human water use on groundwater-to-atmosphere feedbacks during the European heatwave in 2003, 2019, IX, 172 S.

Heft 88: ***Christoph Beekmans***: 3-D Cloud Morphology and Evolution Derived from Hemispheric Stereo Cameras, 2019, [erschienen] 2020, VIII, 118 S.

Heft 89: ***Nils Weitzel***: Climate field reconstructions from pollen and macrofossil syntheses using Bayesian hierarchical models, 2019, [erschienen] 2020, XII, 153 S.

Heft 90: ***Alexander Kelbch***: Investigations to quantify individual exposure to solar ultraviolet erythemal radiation including cloud meteorological impact, 2020, III, 107 S.

Heft 91: ***Mari L. Schmidt***: Improvement of hail detection and nowcasting by synergistic combination of information from polarimetric radar, model predictions, and in-situ observations, 2020, VI, 136 S.

Heft 92: ***Sebastian Brune***: Der Wavelet-basierte Organisationsindex als Maß der konvektiven Organisation über Deutschland und dem tropischen Atlantik, 2021, IV, 121 S.

Heft 93: ***Sebastian Buschow***: Spatial Verification with Wavelets, 2022, V, 195 S.

Institut für Geowissenschaften
Abteilung Meteorologie
Mathematisch Naturwissenschaftliche Fakultät
Universität Bonn