We are grateful to all reviewers for their encouraging comments and extensive constructive criticism. Our answers to all individual points are given below. In the marked up manuscript, changes related to the comments of reviewer #1 are colored in magenta, those related to reviewer #2 are blue. The issue with the code availability raised by executive editor David Ham has been addressed as well.

Reviewer #1

Major Revisions:

It would be nice to see the mean spectra and histogram of the dominant scales for the case study shown in Figure 3 (please add a panel): I expect a bimodal histogram and spectra, since both small scale features and a large front are present in the case study. This bi-modality (i.e. presence of of both small and large scale features) is not represented in the stochastic rain fields produced in Section 2 (the synthetic fields considered in the article have by construction uni-modal spectra), and in fact their resulting spectra and histograms are uni-modal (e.g. Figure 4). However spectra bimodality (i.e. presence of of both small and large scale features) is bound to happen in real verification practice, and it might be badly handled by H_cd and Sp_cd. In fact, H_cd and Sp_cd are the differences of the centre of mass (of the scale histograms and of the mean spectra), and are not suitable summary statistics to compare bimodal curves (or any other non-Gaussian curve). Hemd and Spemd, on the other hand, seem more suitable statistics (to compare Gaussian or non-Gaussian curves) since based on the whole curve comparison. The authors should consider withdrawing H_cd and Sp_cd from the newly proposed wavelet-based scores. [If the authors wish to introduce a metric which measure the direction of the error, maybe they should consider a measure based on the distances along the whole curves (or the integral between the two pdf), but with a sign which accounts for the curves relative position.]

Very good suggestion, we have added the mean spectrum and scale-histogram to the plot! As expected, the latter is indeed bi-modal. Since the two dominant scales are 5 and 6 (roughly), you cannot see two peaks in the mean spectrum - there is no notion of intermediate values between scales. We have, however, seen other real examples where both curves have multiple peaks.

We agree that Hcd and Spcd are not suitable to measure structure errors in a realistic setting. A comment to that effect has been added to the final section. We do however believe that the sign of these scores is usually a helpful quantity (noting that it typically agrees with SAL's S on the error's direction). We furthermore think it's worthwhile to actually demonstrate that the conceptually more complicated EMD is indeed needed and cannot simply be replaced by the difference in centre. Hcd and Spcd are therefore not completely 'withdrawn' from the study, but we have attempted to clarify their role as auxiliary quantities, mostly giving us a sign.

Figure 4 (Section 5) shows that both mean spectra and scale histograms are sensitive to the variation of the scale parameter b and the smoothness parameter v, and that for both parameters, the curves shift in the expected direction (this is the main result). The histogram of the dominant scales seems slightly less sensitive (it shifts less), however it exhibits a smaller spread (hence smaller uncertainty: the signal is better defined). Because of this latter property, the scale histogram should be favoured, with respect to the mean spectra. Moreover, the smaller shifts of the histograms are probably simply related / due to their smaller spread (I have the feeling that the magnitude of the shift is proportional to the spread). These aspects should be mentioned in Section 5. (Note: the sensitivity of the spread to the parameters b and v is secondary: be careful not to mix it up with the main result, aka the shift).
We agree to some extent that the shift is the main effect and the change in shape is secondary. To avoid mixing the two up, we have re-structured this part of the discussion such that the shift for both parameters and both curves is discussed first, then the change in shape is mentioned. The latter result is however relevant to the experiments presented later: If both nu and b only shifted the curves, we could not distinguish between the two kinds of errors.
The perceived “difference in spread” was actually due to the fact that the plots for histograms and spectra didn’t have the same axes: The scale-axis for the histograms started at 0 even though 1 is the smallest possible central scale. After correcting the issue, the magnitude of the shift and the spread of the curves look more or less the same for histogram and spectrum. We would furthermore argue that one cannot simply compare the “spread” of these two directly anyways since they represent different mathematical entities with potentially very different degrees of freedom.

From the previous two comments, I would propose as unique new statistics H_emd.

At the end of Section 2, then authors introduce an algorithm for producing stochastic rain fields which satisfy non-stationarity and anisotropy. Some case studies are illustrated in Figure 2, and the associated verification results are discussed in Section 7.4. In my view this analysis can be removed from the article for the following reason: a) The algorithm for producing stochastic rain fields which satisfy non-stationarity and anisotropy, despite being more sophisticated than the isotropic algorithm mainly used in the article, is still not realistic (the precipitation features of Figure 2 are still far from resembling the ones for the real case illustrated in Figure 3). b) The article will result nicely well contained in illustrating “solely” the isotropic stochastic fields (you have already quite a lot of material! Moreover, this would provide a nice “excuse” for retaining the statistics based on the centres of mass -wink!-). In this case you need to add into the final discussion Section the need to analyze real cases, in
future work ...
c) For the (future) analysis of more realistic cases, I strongly suggest to consider directly real precipitation case studies (the Spatial Verification ICP cases from Ahijevych et al 2009 are available online), rather than using synthetic fields (you might end up spending

a lot of time and implementing very complex stochastic models ... to achieve the same results ...).
After some deliberation, we have decided to follow your argumentation and **remove the non-stationary example from the paper**. We agree that the additional results from this experiment are outweighed by the advantages of having a shorter more streamlined paper, especially since the realism of our non-stationary model is indeed questionable.

Minor Revisions:

Abstract and Introduction

Page 1 line 7: replace 'spatial correlation' with 'spatial structure' (or 'scale structure').
changed it.

Page 1, line 23: please quote (also) Dorninger et al (2018): "The set-up of the Mesoscale Verification Inter-Comparison over Complex Terrain project". Bull. Amer. Meteorol. Soc., 99 (9), 1887 – 1906.
Added the reference. Now the newly discovered fifth verification class is also briefly mentioned.

Page 1, line 23: replace 'avoid' with 'deal with'.
ok

Page 1, lines 16-19: rephrase… (this is a bit weak, as first sentence of the article).
The first few sentences have been replaced by a (hopefully) less lame introduction.

Page 2, line 5: I suggest adding in this paragraph one sentence introducing the fourth class of spatial verification methods, the scale-separation techniques (with the key references). Then you start the new paragraph by stating that the technique introduced in your article belongs to this latter class. Then you describe the most recent literature on variograms etc. (as from line 8 onwards). Here you need to state that the variogram-based techniques are a sub-set of the scale-separation techniques.
Re-structured the two paragraphs accordingly.

Page 2, the paragraph ending at line 22 can be joined with the one starting at line 23.
The two have been joined.

Section 2

Page 3 line 25: write 'The threshold T determines the percentage of the field which has non-zero values'. You need to state (here) that T is the base rate.
We have added that clarification.

Page 5, line 15:
When introducing the scale auto-correlation parameter b, and when discussing Figure 1,

you need to mention explicitly that smaller b are associated with larger scales, and vice-versa larger b are associated with smaller scales (this is counter-intuitive, therefore it needs to be reminded here and there in the article).
Tried to make this more explicit.

Page 5, lines 11-13: it is not clear where this statement lead to: in the article, are you imposing v > 1? Are you using random Gaussian distributions to create / perturb you parameters? Please state.
Since the model contains second derivatives of the Matern fields, it will crash if v <=1, so we only used v > 1. This technical detail is not actually necessary to understand the rest of the paper, so we simply cut it.

Page 5, line 26: define the rotation angle.
Obsolete since we cut the nonstationary model.

Section 3
Page 7, line 14 - Page 8 line 1: this is not “loosely speaking”, please redefine (in easier words) the concept of local stationarity: does it mean that locally your auto-correlation is zero? You can also decide to remain with mathematical strict definitions ... in the rest of the paragraph, you are quite technical ... however my preference is always to accompany the mathematical explanation with a sentence which explain / vulgarize the mathematical content. You might need to summarize the findings of Eckley et al (2010), Kapp et al. (2018).
You are correct, the way we speak here is not particularly loose (that formulation has been cut). We believe that further technical details about the nature of local stationary would distract from the main point - for the paper it is sufficient to know that correlations can vary in space as long as they do so slowly. In fact, neither Eckley nor Kapp give a general definition of local stationarity, they only state the specific regularity conditions imposed on the LS2W.

Section 4
Re-title section 4 as ’Wavelet spectra spatial aggregation’.
As you wish.

Page 9, line 10: for the case study add the reference to Ahijevych, D., E. Gilleland, B.G. Brown, and E.E. Ebert, 2009: Application of spatial verification methods to idealized and NWP-gridded precipitation forecasts. Weather Forecast., 24 (6), 1485 – 1497.
We moved the reference from the figure caption to the body of the text.

[From the major comment: please, add a panel in Figure 3, with the mean spectra and histogram of the central scales for the shown case study. I expect a bimodal histogram and spectra, since both small scale features and a large front are present in the case study.]

See answer above.

Section 5

Re-title Section 5 as “Wavelet Spectra Sensitivity Analysis”.

Done.

Page 9, line 22: please remind here that larger (smaller) b is associated with smaller (larger) scales. Page 9, lines 26-27: eliminate the sentence “Simultaneously ... observed scales” (I do not see this in the Figure; moreover the sentence distracts from the main point). Paragraph starting at page 9, line 32 and ending at page 10, line 2 (describing the major findings of Figure 4): In this paragraph you have one main result and a secondary result. The main result is that both mean spectra and scale histogram are sensitive to the variation in the parameters b and v, and that for both parameters they shift in the expected direction. The sensitivity of the spread as you vary b or v is a secondary results (which is actually neither too visible, nor to important for your study). In the paragraph these are mixed up in the discussion, so that the latter takes away the focus from the former. Rephrase the paragraph. E.g. at page 10, line 2, I suggest writing: ’ … only affected by b: larger scales (smaller b) lead to a greater variance (panel b) whereas changes in smoothness (parameter v) do not substantially change the histogram shape’ (avoid mentioning the shift here). [From the major comment, you should also state that: 1. the scale histogram exhibits less spread, the dominant scales are better defined, and hence it is favoured wrt the mean spectra. 2. the smaller shift of the scale histogram is possibly proportional / due to its smaller spread, and not to a lack of sensitivity.]

See answer to the major comment.

Page 10, line 6: the lack of sensitivity of both the mean spectra and the scale histogram on the base rate (parameter T) is a very welcome property in a verification scoring rule (it implies that the score cannot be edged, e.g. by over-forecasting, and that the performance does not depend on the underlying climatology). This should be mentioned.

We agree that it should be mentioned, but feel that such judgemental statements should better be relegated to the discussion at the end of the paper.

Section 6

Page 11: [From the major comment: real precipitation fields might generate bi-modal spectra (whereas the synthetic fields considered in the article have by construction uni-modal spectra). H_cd and Sp_cd (page 11), are not suitable statistics for comparing bi-modal (or non-Gaussian) spectra, because they compare the centre of mass of the curves: this limitation ought to be (at least) mentioned. H_emd and Sp_emd, on the other hand, seem more suitable statistics (to compare Gaussian or non-Gaussian curves) since based on the whole curve comparison. If the authors wish to introduce a metric which measure the direction of the error (such as H_cd and Sp_cd), maybe they should consider a measure based on the distances along the whole curves (or the integral between the two pdf), but with a sign which accounts for the curves relative position.]

See answer to the major comment above.

Page 11, lines 10-13: please define EMD (either write the formula or describe how it is calculated ... “moving the dirt ... work” is visually clear, but it would be better to be more precise).
We have added a clarification of what exactly corresponds to the mass/location of the dirt piles. We furthermore mention a simple way of calculating the relevant special case of the EMD without numerical optimization (the simplification has only recently come to our attention).

Page 11, lines 13-14: by normalizing the spectra to obtain a unit sum you essentially remove the bias, and concentrate solely on the pure scale structure (how the total energy is distributed across the scales). This should be mentioned.
Good point, we now mention that.

Page 12, line 5: there is an incoherence in the naming of the Energy score, in this Section it is “Sp_e”, whereas in Figure 5 it is “SpEn”. I personally prefer the latter, or “Sp_en”, to well separate it from “Sp_emd”.
We agree, changed it to Sp_en.

Section 7
Page 14, lines 11-13 (describing the bottom panels of Figure 5, evaluating the ensembles against a RS observation): not only the RS ensemble scores best (for all scores), but also the SmS and RL exhibits the second best score and the SmL (the most dissimilar ensemble with respect to RS) exhibits always the worst score. You should mention this.
Good idea, we have added that observation.

From page 14 line 15, to page 15 line 6, need to be rephrased:
a) when comparing RL to SmS (Page 14, bottom 2 lines): the compensating error affect solely the location /mean value of the mean spectra and scale histogram, or does it affect the whole mean spectra and scale histogram? I question the phrasing ’on location of the spectra and histograms along the scale axis’ (I would eliminate this part of the sentence). In the following sentence (page 15, line 1) I question ’by their centres of mass alone’.
You are correct in that the two changed parameters affect both location and shape of the curve, we have re-ordered the relevant sentence to make it clear that we don’t mean that the shape is unaffected. The effect on the shape of the curves, however, has (to first order) the same sign: An increase in smoothness and an increase in b (decrease in scale) both lead to distributions which are less spread out. These kinds of errors *do not* compensate each other, the histogram of SmS is unambiguously too tight. We therefore stand by our statement that the centre alone is insufficient but the EMD can, in principle distinguish the two.

b) Page 15, lines 1-3: I think that the SmS and RL ensemble cannot be separated well for all scores (also Vw5), not only for Hcd (I won't attribute the lack of separation to the fact that Hcd compare centres of mass). This is possibly due to the fact that the mean spectra and scale histograms for RL and SmS are similar (From Figure 1, the top-left and bottom-right panels are more similar than the top-right versus bottom-left). Nevertheless, in the top panels of Figure 5 all scores (but V20) shows a slightly larger error for the SmS ensemble than for the RL ensemble (which is encouraging), and then even larger errors for SmL and RS (it seems to me that the scores are informative ...).
You are correct, the spectra are certainly more similar than those compared in the first half of the experiment. As discussed in the answer to the previous comment, we stand by our assertion that there is a well defined difference between the two curves which can, in principle, be measured by EMD but not cd.

c) Top panels of Figure 5: The two scores considering the sign of the error (H_cd and S) exhibit the same behaviour, not only for SmL and RS, but also for SmS (they both exhibit slightly negative values): the sentences at page 15 lines 4-5 are partially incorrect, please re-phrase them.
It has been re-phrased to note that the tendency is the same, but we stand by the claim that the signal is slightly stronger for S.

From Figure 5 and 6, it is clear that V20 is the less informative score: please add this comment (you can relate to your comment when introducing V20 in Section 6, ...).
That's right, but we feel that the unsurprisingly meagre performance of V20 is sufficiently mentioned in the final section and requires no additional discussion.

Section 7.2: The results associated to Figure 7 are very nicely discussed and very interesting! For ensembles, SpEn is the champion score followed by Vw5, whereas for deterministic Vw5 closely followed by Sp_emd are the champion scores. I am surprised of the lower performance of H_emd: why? After these results, one could be tempted to choose Vw5 as scoring rule … however its strong dependency on the base rate/climatology (Section 7.4) cannot be ignored. Maybe you can add some of this comment in the discussion?
We have attempted to clarify in the discussion that dependence on the base rate is undesirable for a pure structure score. We could however imagine verification settings where one wants to judge structure and precipitation area simultaneously. We therefore refrain from calling the wavelet-based score 'better'.

Section 7.3: please specify in the caption of Table 3 (and Table 4), or write in the text, that ExP1 = D1 = Haar, and that Exp4 = D4 is the wavelet considered in the main experiment of the article.
Good idea to remind the reader which wavelet we've been using. We have added that to the caption of Table 3 (Tab. 4's caption is just "as Tab. 3, but ...").

Section 7.4, page 18 lines 5-6: given that in the original experiment T was set to 0.2 (aka

20% of the domain was precipitation, and 80% was zero values), I imagine that with this model the precipitation area is ranging in 15-25% of the domain: can you please phrase this more clearly? (rather than using the 75%-85% range, refer to your previously fix 20% base rate ...)
That was actually just a typo, it should of course read “a uniform random variable between 15 % and 25 % of the complete domain”, not 75%-85%. Fixed it.

Discussion and conclusions

page 20, line 14: I suggest writing ’mis-representation of feature sizes (e.g. smoother representation of small-scale convective organization)’.
Changed it accordingly. The previously given example of “missing fronts” is indeed a bit misleading since that is typically a matter of displacement, not structure.

Page 20, lines 17-25: the findings of Figure 6 and 7 are well summarized in the conclusions (page 20, lines 21-25). I would end this paragraph at line 25. The sensitivity of the Variogram score to p and w (lines 31-32) could also be added to this paragraph. Then (at page 20, line 26) I would start a new paragraph, discussing the results of the sensitivity analyses (sensitivity to T and to the wavelet choice).
The sensitivity-experiments are now in a new paragraph. The variogram score’s sensitivity was not moved because the paragraph previously ending at line 25 now also contains a remark regarding the possibility of multi-modal spectra (see answer above) and would have been too full.

Sensitivity to T: I suggest to phrase differently lines 25-30 (page 20): you need to remind that the ’perturbation of the data’ is essentially an assessment of the sensitivity of the scores to the sample climatology. I would express more concern about the loss of discrimination of the variogram scores found in section 7.4.
The remark about the sample climatology has been added. As discussed in the answer to the comment concerning 7.2, we feel that an appropriate amount of concern has been expressed - the two approaches measure different things and the wavelet-scores isolate structural characteristics more clearly.

Sensitivity to the wavelet choice: I would rephrase lines 32-33 (page 20) as ’We have also tested the sensitivity of the newly introduced wavelet-scores to the choice of the mother wavelet. We have performed ...’.
They have been rephrased, albeit with slightly different grammar to avoid the double “We have ...”.

As the last paragraph of the conclusion suggests, this study is still exploratory: there is no single score which has emerged as the recommended best score. This should be mentioned. Moreover, the paragraph could be re-phrased to include real case studies and scores which accounts for the direction of the error while applied to bi-modal spectra (as explained in the major comments).

We have slightly re-phrased the beginning of this paragraph to make it clear that the first study which applies those scores to the real world will still be experimental in nature. Whether or not there is a single best score is not really the topic of this paper. Early results from our next study, however, suggest that the difference between the two score-families is actually pretty small in real-world situations. In that case hEMD seems preferable to us, but that is a story for another time ...

Reviewer # 2

SPECIFIC COMMENTS:

P6 section 3: I have found the introduction to the wavelet theory pretty limited. I understand that you don't want to provide the mathematical details, but I think it is a difficult start for a reader whose knowledges about wavelets is short. I would therefore recommend that you to start this section with a few sentences presenting (in words) what are wavelets, why are they popular for analyzing signals (or fields in the 2D case), and to which references the reader could refer for a more detailed (and mathematical) introduction.

Fair point, we have added a few more words and literature recommendations.

P7 l5-6: I have two comments here. First of all, you may clarify that the squared weights quantify the energy spectra. Indeed, at this point of the paper you have used several times the term "spectra" but have not defined it, and you haven't used yet the term "energy".

Good idea, the term "wavelet spectrum" is now introduced at this point.

Second, I think it would be nice here to briefly mention the origin of this bias (the redundancy of the wavelet transforms?), and also which form does it takes (the energy increasing over and over with increasing scales?). In my opinion, how does this bias really affect the energy spectrum is something that has been poorly explained in the previous LS2W papers that you cite, and it would be nice to let the reader know what should he expect in case he doesn't apply the correction.

Yes, that would be nice. We have added a few words to that effect.

P7 l14-32: To my knowledge, your manuscript is the first (among the others having used the LS2W spectra for verifying precipitation fields) that investigates the choice of the mother wavelet. However, this paragraph is hard to grasp for a non-familiar reader, especially the differences between the wavelets. As a suggestion for improvement, I recommend that you add a figure of the plot of the different wavelets, so that it is easier to see the differences in terms of smoothness and support. You could also take the opportunity to refer to this figure at the very beginning of section 3, when introducing the mother wavelet function for the very first time. However, it is possible that the reader doesn't understand how one can apply a 1D function (the wavelet) to a 2D field, so it might be necessary to explain the process in few words (apply on the rows, then on the columns, etc.).

We have added a small figure, showing the 2D-version of the first four Daubechies wavelets. This hopefully gives the reader a better idea of the different basis functions and doesn't confuse them with the problem of applying the 1D mother to a 2D field: In principle, one could also calculate each coefficient by individually multiplying the field with the 2D daughter wavelet. This would be terribly inefficient, but for the purposes of this paper we don't really need to worry about the technical implementation of the RDWT.

P8 l13-15: In my opinion, the fact that the amount of negative energy averages out if we choose a wavelet smoother than D1 should not be introduced as “Preliminary experiments have shown that . . .”, but deserves a more detail paragraph and eventually supporting figures. Indeed, in my opinion, allowing negative energy is one of the biggest issue of the RDWT, so if you show that this problem vanishes by using other wavelets than the Haar wavelet, this is an important result, which should be discussed in more details.
We agree that this observation is not altogether unimportant. A figure showing the actual ratios between negative and total energy, as well as a few explanatory words, have been added to the appendix. Since the asymptotic theory of locally stationary processes is not, however, the focus of this study, we feel that more in-depth discussion of negative energy would distract readers from the core points of the paper. As you said in previous comments, the material is already not trivial if one has never worked with wavelets before. For our purposes it suffices to know that smooth wavelets mostly eliminate the problem.

P13 l15-19: It might it is necessary to give a little more explanation about the S component of the SAL (and perhaps a figure), so that your paper is self-sufficient. Moreover, you should say a few words about the ensemble version of SAL as well.
You’re probably right, it is better to have a short explanation of how S measures structure so that readers don’t need to look at a second paper just to understand this part of the experiment. A few words have been added.

MINOR COMMENTS AND TECHNICAL CORRECTIONS:

p1 l19: “a given rain field is forecast perfectly, but slightly displaced”: If there is a displacement error, then the field is not perfectly forecast. You may replace “field” by “object” or “feature”.
Done.

p1 l23: After "four main strategies", the reader expects a descriptive list of each of these strategies. This is actually what you do in the paragraph that follows, but we have to wait until p2 l6 ("the last") to be sure that you are indeed referring to these four strategies. I recommend that you make the description more explicit.
The section has been re-formulated slightly to make it clear that we have begun counting to four.

p2 l17: remove the coma
It is gone.

p2 l22: You may briefly mention here the notion of "local stationarity".
Good idea.

p2 l25: As you write "corrected RDWT", you may later in the sentence say: "to obtain an

unbiased estimate of the local wavelet spectra" (otherwise we don't know why you need to correct).
Added that remark.

p2 l33: It is not clear why does considering both the ensemble and the deterministic case "avoid the need for further data reduction".
It does not, those are just two separate things we do differently from Kapp 2018. The sentence has been clarified.

P7 l7: Isn't it "phi_{j,l,u}" instead of "phi_{j,j,u}"?
Fixed the typo.

P7 l11: You say here that the smoothing is the final step of the spectra estimation procedure. However, in the package LS2W the smoothing takes place before the bias correction by the matrix A-1. Please clarify.
Yes, Eckley does the smoothing prior to bias correction because the distribution of the uncorrected spectra should be chi-squared and we know how to smooth chi-squared variables with wavelet shrinkage. The formulation has been changed.

P7 l25-26: I don't understand what do you mean by a "sparse representation". Please clarify.
It just means that most coefficients are nearly zero. We have clarified that.

P7 l28-30: You have defined at l23 the labels "ExP" and "LeA", but these are not used until P17. Maybe you could use them here (instead of "this version of D4").
Good idea.

P8 l9: It may be nice to remind why the invariance under shift is necessary.
We have added a comment to that effect.

P9 l6-7: Maybe it would be better to replace "(i,j)" (and elsewhere where you refer to the coordinates) by other indices such as "(x,y)", to avoiding confusion with J referring to scales.
Good point, changed it to (x,y). Changed the indices in the section about the vg-score as well.

P9 l11: Please indicate the rationale behind the logarithm transformation.
Whether or not it is a good idea to log-transform real rain fields prior to the wavelet-transform is an interesting question which we plan to discuss in some detail in the next publication. Since it is really only necessary for this one example image, we feel that a full discussion is not warranted at this point. We have added a short remark stating that the log-transform leads to a more well-behaved marginal distribution and reduces the impact of extremes: If you leave the data as it is, the wavelet spectrum can be dominated by singular extreme pixels while large regions have virtually no influence. That is at odds with our intuitive idea of "structure".

P9 l25: This introduction to Fig 4a is confusing, because you say "as a function of the scale parameter", but when you look at the Figure you read "scale" for the x-axis, although the scale parameter you are referring to is in the y-axis. Please clarify.
We have attempted to clarify.

P10 Fig 4: For plots (a) and (b), you may change the style of the black dash line, as we actually don't see the dash.
We have attempted to make the dashed line a little more dashed, but the plot3D-library is a stubborn beast. A clarification has been added to the caption to make sure that everyone knows which lines we mean.

P12 Equation (6): I'm wondering if readers unfamiliar with the energy score might be confused with your definition. Indeed, the name "Energy Score" has here nothing to do with the "energy" of the spectra you are referring to, and this might be confusing with the fact that you define y and F as the observed and forecast (energy) spectrum, although in the general definition of the energy score, y and F are simply the observation and the forecast, no matter which quantity is being forecast. A more general definition of the score may reduce the risk of confusion.
Good point, we made it seem like the original definition of the energy score is somehow related to wavelets. The sentence has been clarified.

P12 l5: You never mention clearly in this paragraph that the forecast quantity at hand is a multivariate vector. Even if you bold the observation y and the realizations X and X', you should make crystal clear that it is a multivariate quantity, and give the dimension.
We have added the words "multivariate" and "vector" to bring that point across.

P12 Table 1: Actually, some of your scores (Hemd and Hcd) work for both deterministic and probabilistic forecasts, so maybe you could modify you table by either adding a column "deterministic" that you fill with "yes" or "no", or by modifying the title of your current column and fill it by "deterministic", "probabilistic" or "deterministic and probabilistic".
A further column has been added.

P13, Variogram score: It is not clear whether you apply the variogram score to quantities that represent the wavelet spectrum or the precipitation field. From p13 l5, I understand that X, y and EF refer to the vector of the spectrum, but later you refer to spatial locations, so that I figure out that your forecast quantities are fields, is that correct? Please clarify. More generally, please clarify which scores are built from the wavelet approach, and which ones from the precipitation fields directly.
That was in fact a mistake, X,y and F are not at all the same as in equation 6! We corrected that and made sure to mention at the beginning of the section that the following scores are non-wavelet alternatives. The division into wavelet and non-wavelet is also reflected in table

1.

P14 l12: I would add “for ensemble forecasts” (after “the established alternatives”), to clarify why you don’t consider the RMSE here.
It has been added.

P14 and 15, Fig 5 and 6: the energy score is here referred to as SpEn, although in the text you refer to Spe. Similarly, in Fig 7 you refer to Semd, although in the text you refer to SPemd. In addition to these corrections, I think it would be nice to use the subscripts in the Figures, so that it is fully consistent with the text.
The notation has been unified, the plots now also have the subscripts.

Executive editor comment (David Ham)

am writing as an executive editor of GMD to highlight an issue with the code availability section which needs to be remedied in the revised manuscript. Thank you for providing a reference to the full code and data used in the experiments presented in your manuscript. There are two problems with providing this data via GitHub. The first is that a reader cannot identify the exact version of the code that was used in the paper (for example, you may fix bugs or add features in the future). The second issue is that projects sometimes change the revision control system they use, or the hosting (the project might move to GitLab, for example). The solution to both of these issues is to provide a reference to a persistent archive of the exact version of the code that was used in the manuscript. This reference can, and should, be in addition to the GitHub link, so that a user can also always access the most recent version of the code.

Since your original code is hosted on GitHub, the easiest way to produce a persistent archive of a precise version is to use GitHub's Zenodo integration. For more details, see: https://guides.github.com/activities/citable-code/. Please ensure that the revised version of your manuscript contains a reference to a persistent, public archive of the exact version of the code used to produce it.

Thank you for the recommendation, we have done as you said and uploaded the version used in the paper to Zenodo: 10.5281/zenodo.3257511

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

Sebastian Buschow[1], Jakiw Pidstrigach[1], and Petra Friederichs[1]

[1]Institute of Geoscience and Meteorology, University of Bonn

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

Abstract.

The quality of precipitation forecasts is difficult to evaluate objectively because images with disjoint features surrounded by zero intensities cannot easily be compared pixel by pixel: Any displacement between observed and predicted field is punished twice, generally leading to better marks for coarser models. To answer the question 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 gridpoint-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 of Wernli et al. (2008). The 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 fifth and final 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 basic idea of the method presented in this study, which can be classified as a scale-separation technique as well, is that errors, which 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 too small- or too large-scaled. 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 is physically consistent and thus bears 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 section 2. Sections 3 and 4 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 section 5. Based on these results, we define several possible deterministic and probabilistic scores in section 6. In a second set of experiments (section 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 section 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:

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

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 $\mathbf{v} = (u, v)^T$ 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

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

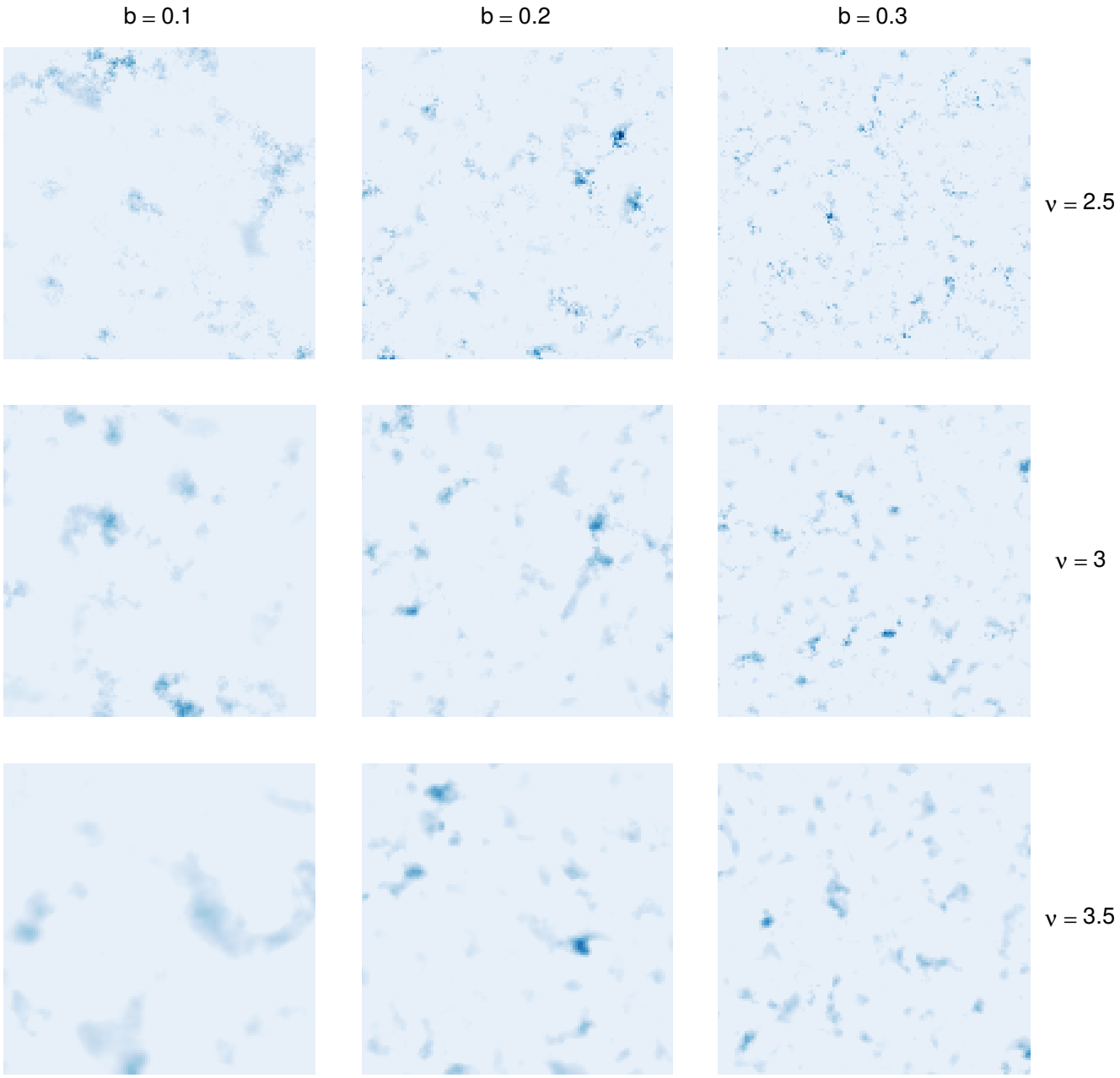

Figure 1. Example realizations of the stochastic rain-model on a 128×128 grid for various choices of scale b and smoothness ν. The threshold T was chosen such that 20 % of the field has non-zero values.

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

$$\text{Cov}\left((\Psi_\mathbf{s}, \chi_\mathbf{s}, q_\mathbf{s})^T, (\Psi_\mathbf{t}, \chi_\mathbf{t}, q_\mathbf{t})^T\right) = \Sigma_{\Psi,\chi,q} \cdot M(\,||b(\mathbf{t}-\mathbf{s})||, \nu\,). \tag{2}$$

Here, $\mathbf{t}, \mathbf{s} \in \mathbb{R}^2$ are two locations within the 2D-domain and M is the Matérn covariance function. The parameter b governs the scale of the correlations, the smoothness parameter ν determines the differentiability of the paths. The matrix $\Sigma_{\Psi,\chi,q}$ 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.,

$$\text{Cov}\left(\left[q_\mathbf{s}, \nabla \cdot q_\mathbf{s}, \nabla\chi_\mathbf{s} - \nabla \times \Psi_\mathbf{s}, \nabla^2\chi_\mathbf{s}\right]^T, \left[q_\mathbf{t}, \nabla \cdot q_\mathbf{t}, \nabla\chi_\mathbf{t} - \nabla \times \Psi_\mathbf{t}, \nabla^2\chi_\mathbf{t}\right]^T\right)$$

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 behavior. For further details, the reader is referred to Hewer et al. (2017), Hewer (2018) and references therein.

Fig. 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 consist of projecting the fields onto a series of so-called daughter wavelets $\psi_{j,l,\mathbf{u}}(\mathbf{r}) : \mathbb{R}^2 \to \mathbb{R}$, which are all obtained from a mother-wavelet $\psi(\mathbf{r})$ via scaling by $\mathbf{r} \to \mathbf{r}/j$, a shift by $\mathbf{r} \to \mathbf{r} - \mathbf{u}$ and rotation in the direction denoted by l. Such *wavelet-transforms*, which generate a series of basis functions from a single mother ψ who 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 $\{j, l, \mathbf{u}\}$ 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 $\mathbf{u}$ 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 two 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

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

be the so-called *two-dimensional locally stationary wavelet process* (henceforth LS2W). The random increments $\xi_{j,l,\mathbf{u}}$ 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 $W_{j,l,\mathbf{u}}$. For all technical details the reader is referred to Eckley et al. (2010), Kapp et al. (2018) present 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 transform's redundancy, 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 $\psi_{j,l,\mathbf{u}}$ and thus depends only on the choice of ψ and the size/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 et al., 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. Fig. 2 shows a few examples from this family. D_1 (panel a), the only family member which 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; 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). $D_{1,2,3}$ belong in both sub-families, wherever a distinction is needed, we will label the two branches of the family as *ExP*

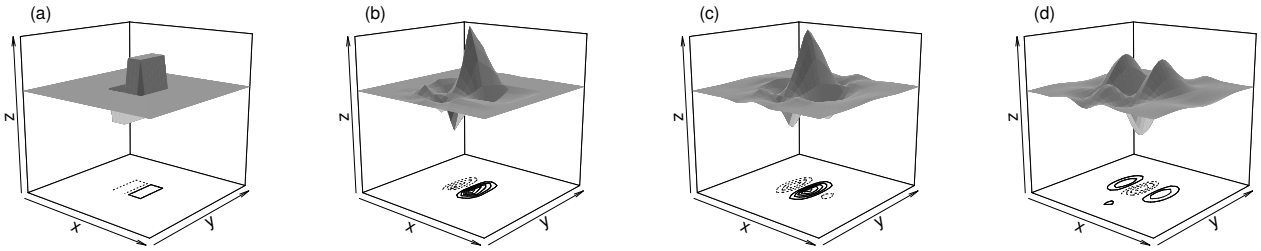

Figure 2. Two-dimensional Daubechies daughter wavelets D_1 and D_2 (a, b), as well as least asymmetric D_4 (c) and D_8 (d), vertical orientation.

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. 2 c) 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 section 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

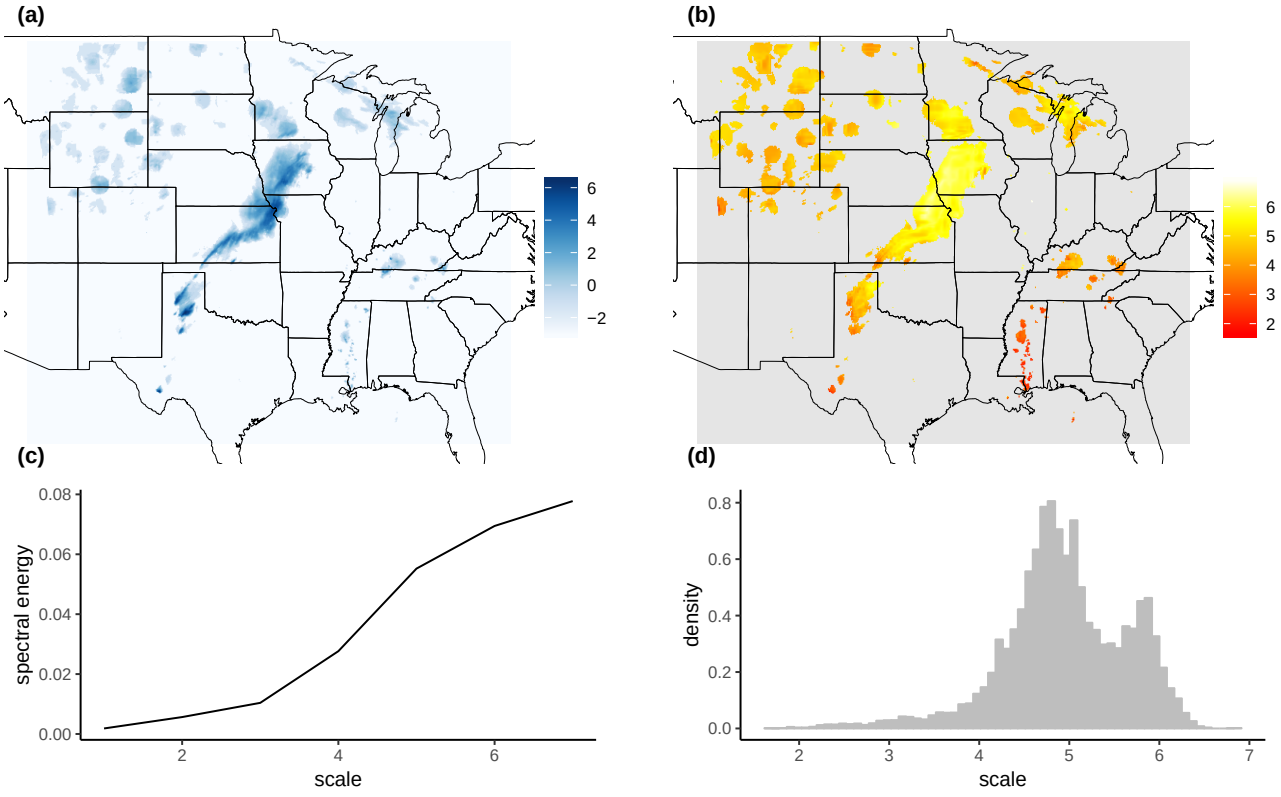

Figure 3. Logarithmized rain field (a) and corresponding map of central scales (b) from the stage II reanalysis on 13-05-2005. The field has been cut and padded with zeroes to 512×512, scales were calculated using the least asymmetric D_4 wavelet, only locations with non-zero rain are shown. Panels (c) and (d) show the corresponding mean spectrum and scale-histogram, respectively.

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, this data was originally studied by Ahijevych et al. (2009)). Here, the original rain field was logarithmized, adding 2^{-3} to all grid-points with zero rain in order to normalize the marginal distribution (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. 3 c), we can base our scores on the histogram of C over all locations pooled together (Fig. 3 d). 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.

Fig. 4 (a) 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 monotonously 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. 4 (b) 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 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 characteristics are less clear: Are fields with a larger fraction of precipitating area perceived to be larger- or smaller-scaled? To investigate this, we set $b = 0.1$ and $\nu = 2.5$ and vary the rain-area between 10 % and 100 %. Fig. 4 (e) and (f) show that the

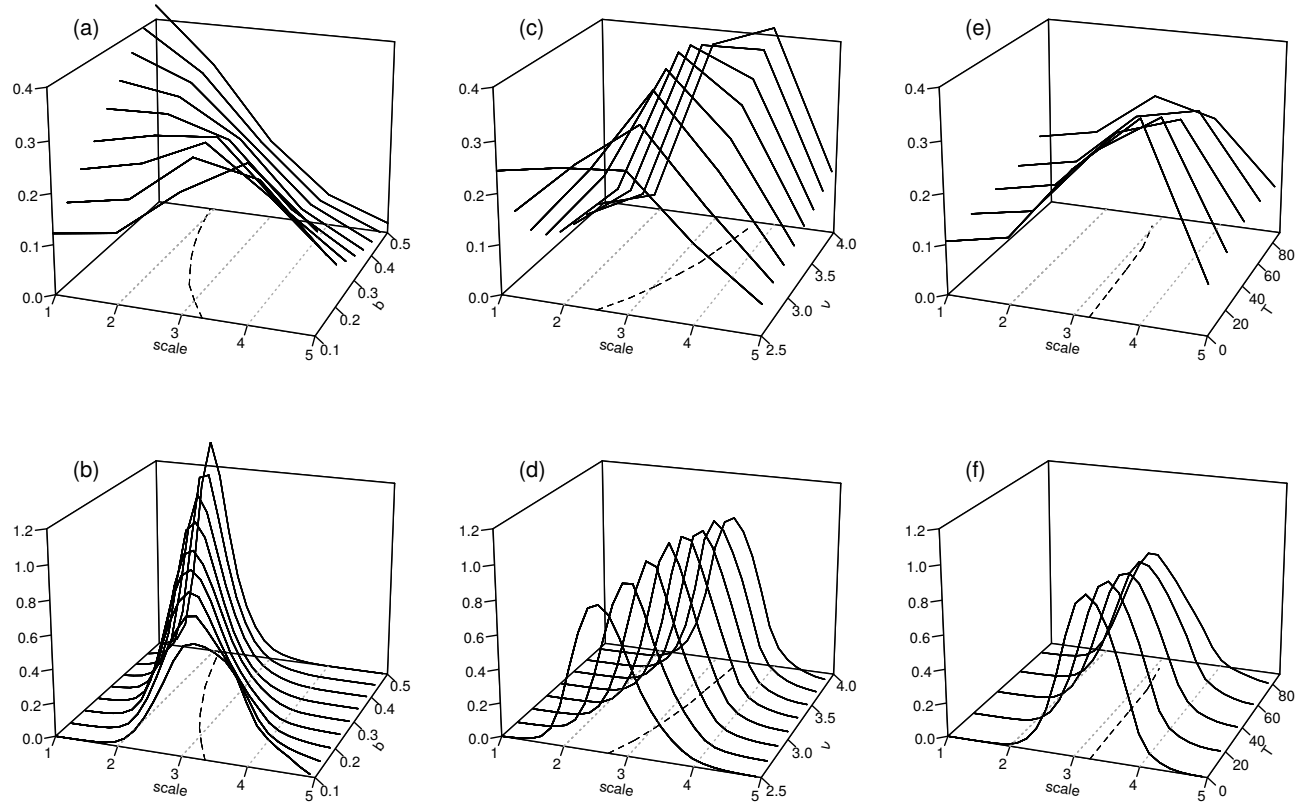

Figure 4. Mean spectra (top row) and histograms of central scale (bottom), as functions of the scale parameter b at $\nu = 2.5$ (a,b), smoothness parameter ν at $b = 0.25$ (c,d) and threshold T at $b = 0.1$, $\nu = 2.5$ (e,f). Dashed lines in the x-y-plane indicate the the respective curves' centres of mass, dotted gray lines (parallel to the y-axis) were added for orientation.

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 Wavlet-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), 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 Sp_{emd}), respectively.

Being a metric, the EMD is positive semi-definite and therefore yields no information on the direction of the error. We can obtain such a judgment 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 Sp_{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

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

where $\mathbf{y}$ is the observed vector, E_F denotes the expectation value under the multivariate distribution of the forecast F, and $\mathbf{X}$ and $\mathbf{X}'$ 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 Sp_{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 section 7.

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

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

abbreviation	description	probabilistic	deterministic
Sp_{emd}	EMD of the mean spectra	no	yes
Sp_{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
Sp_{en}	energy score of the predicted mean spectra	yes	no
RMSE	root mean square error between rain fields	no	yes
$\mathrm{V}_{w,5}$	variogram score, $w_{a,b} = \lvert\mathbf{r}_a - \mathbf{r}_b\rvert^{-1}$, $p = 0.5$	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

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

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

where $\mathbf{y}$, 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 $w_{a,b} = |\mathbf{r}_a - \mathbf{r}_b|^{-1}$, $p = 0.5$ ($\mathrm{V}_{w,5}$ below), where $\mathbf{r}_a$ 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 $\mathrm{V}_{w,5}$ 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 too large-scaled and/or too flat. In this study, we employ the original object identification algorithm of 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 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

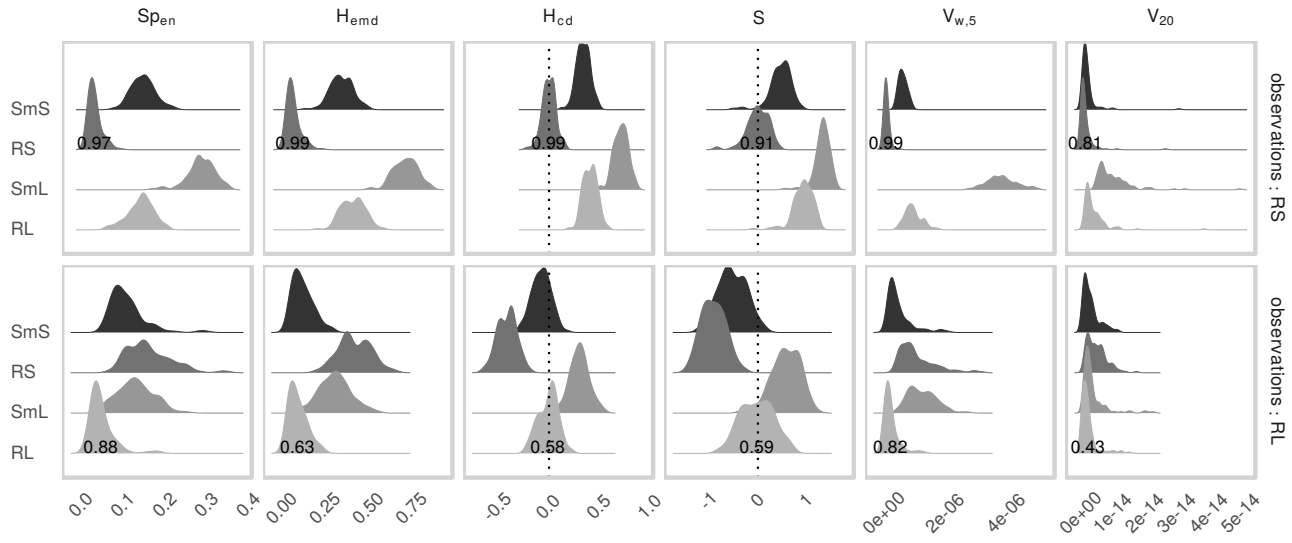

Figure 5. Distribution of all probabilistic scores for each of the forecast ensembles corresponding to the four models from table 2. Top row: Observations drawn from RS. Bottom: Observations drawn from RL. Numbers denote the fraction of cases where the forecast from the correct distribution received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

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 ten realizations from RL, SmL, RS and SmS, respectively. Only one of these ten-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 Sp_{en} as well as the established alternatives for ensemble forecasts, i.e., S, V_{20} and $\mathrm{V}_{w,5}$.

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

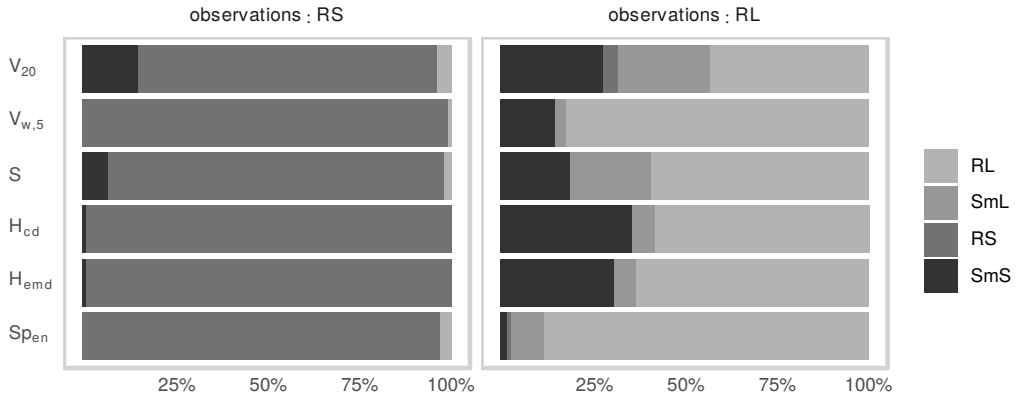

Figure 6. Percentage of cases where each of the four ensembles corresponding to the models in table 2 was deemed the best forecast, separated by score and the model of the observation.

Fig. 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 too large-scaled. 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 smaller-scaled, 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}.

Fig. 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, Sp_{en} hardly ever erroneously prefers SmS over RL. Instead, SmL is wrongly selected most frequently, leading 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

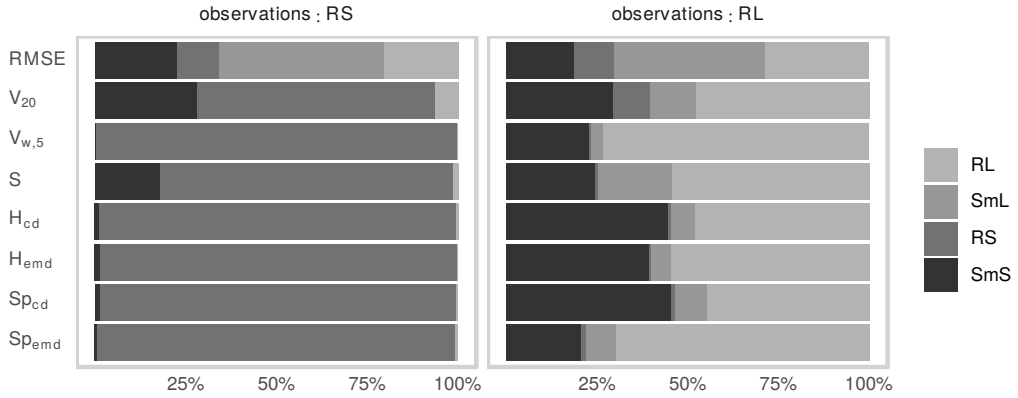

Figure 7. As Fig. 6, but for the deterministic verification experiment.

the size of the forecast ensembles is reduced from ten to one. 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 (Sp_{emd}, Sp_{cd}, H_{emd} and H_{cd}), we again calculate $\mathrm{V}_{w,5}$ 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.

Fig. 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. $\mathrm{V}_{w,5}$ is similarly successful as the best wavelet-based score, faring marginally better than Sp_{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 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 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 process for several Daubechies wavelets. Recalling the results of our objective wavelet selection (section 3 and appendix A), we expect no dramatic effects.

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*		
	Sp_{emd}	Sp_{cd}	H_{emd}	H_{cd}	Sp_{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*		
	Sp_{emd}	Sp_{cd}	H_{emd}	H_{cd}	Sp_{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

Table 3, listing the overall success rates for each tested wavelet, mostly confirms this expectation: In the deterministic case, Sp_{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 (Sp_{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.

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.

		Sp_{en}	Sp_{emd}	Sp_{cd}	H_{emd}	H_{cd}	$\mathrm{V}_{w,5}$	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	-

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 section 5 indicate that the wavelet-based scores should be relatively robust to small changes in the threshold T (cf. Fig. 4 e 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 of 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 judgment, $\mathrm{V}_{w,5}$ 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 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 is 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 mis-representation of feature sizes (e.g. smoother representation of small-scale convective organization) while errors in smoothness correspond to forecast models with too-coarse resolution, 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 of their centres (H_{cd} and Sp_{cd}), their earth mover's distance (H_{emd} and Sp_{emd}), and the energy score (Sp_{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 ($\mathrm{V}_{w,5}$). The less robust V_{20} as well as the SAL's structure component S and the simplistic RMSE were clearly out-performed. H_{cd} and Sp_{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 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. 3 d), 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 however undesirable.

The variogram-score's two free parameters, 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 sub-optimal choice for our purposes.

Now that the merits of wavelet-based structure scores have been demonstrated in a controlled environment, further test 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. The specific version used in this paper was also archived at https://doi.org/10.5281/zenodo.3257510.

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 $\mathbf{y} = (y_1, \dots, y_n)^T$ be a vector with non-negative entries satisfying $\sum_i y_i = 1$. For our purposes, its entropy is defined as

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

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 $\int x^q \psi(x) dx = 0$ for $q = 0, \dots, N-1$. This implies that poly-

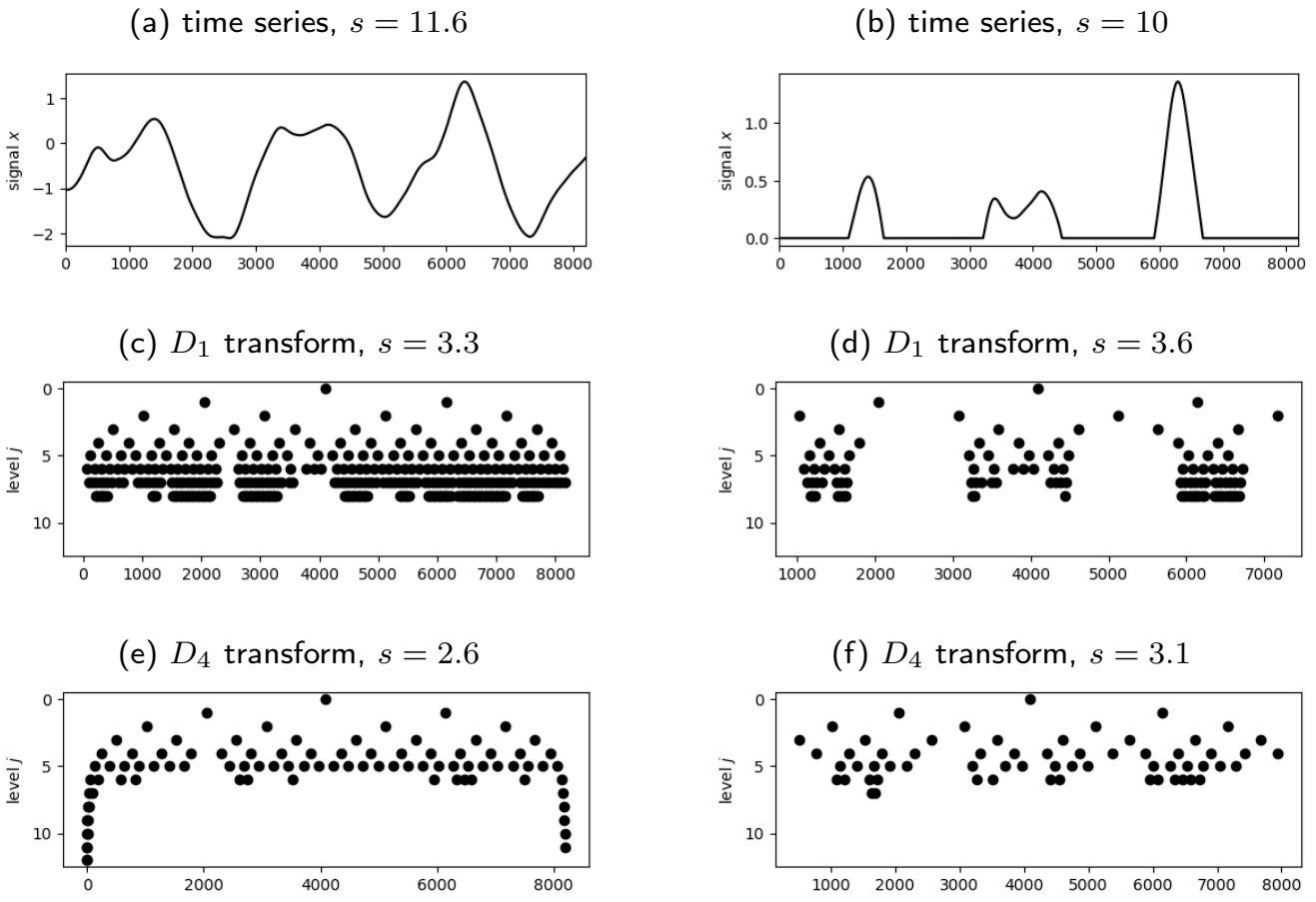

Figure A1. Realization of a one-dimensional Gaussian random vector with covariance $M(\nu = 3.5, b = 2)$ (a) and the corresponding values of the D_1- and least asymmetric D_4-transform (c and e) which are greater than 0.1. (b), (d) and (f) are the corresponding plots for the case where the vector is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

nomials 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 N-th 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 contains 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

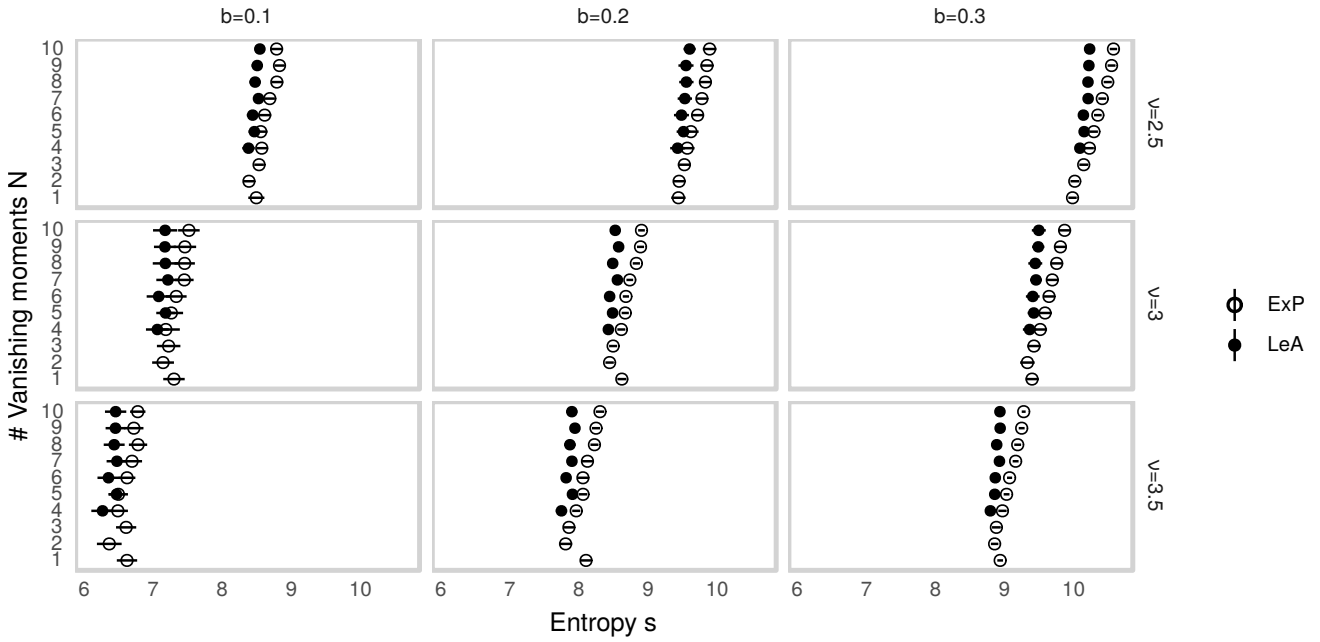

Figure A2. Entropy of the wavelet-transformed synthetic rain fields from Fig. 1 as a function of the wavelet's order N. Empty and filled dots correspond to the extremal phase- and least asymmetric versions of D_N. Lines indicate one standard deviation, estimated from ten realizations.

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, ν as in Eq. (2), but only one variable and one spatial dimension). Fig. 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.

Fig. 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

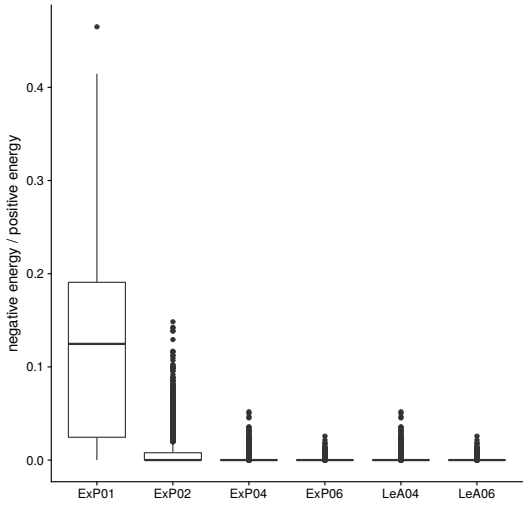

Figure A3. Ratio of negative to positive energy in the mean spectra (data-set from section 7.3, all models from table 2, six selected mother wavelets as in table 3).

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 behavior: 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 equation 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 to see 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 section 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.

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.

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, Wea. Forecasting, 24, 1485–1497, 2009.

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

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

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.

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

Daubechies, I.: Orthonormal bases of compactly supported wavelets, Communications on pure and applied mathematics, 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. Wea. 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, Bulletin of the American Meteorological Society, 99, 1887–1906, https://doi.org/10.1175/BAMS-D-17-0164.1, 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, Meteor. Appl., 15, 51–64, 2008.

Eckley, I.A., Nason, and G.P.: LS2W: Implementing the Locally Stationary 2D Wavelet Process Approach in R, Journal of Statistical Software, 43, 1–23, http://www.jstatsoft.org/v43/i03/, 2011.

Eckley, I. A., Nason, G. P., and Treloar, R. L.: Locally stationary wavelet fields with application to the modelling and analysis of image texture, Journal of the Royal Statistical Society: Series C (Applied Statistics), 59, 595–616, 2010.

Ekström, M.: Metrics to identify meaningful downscaling skill in WRF simulations of intense rainfall events, Environmental Modelling & Software, 79, 267–284, 2016.

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

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

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

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, Journal of the American Statistical Association, 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. Wea. Rev., 146, 1303–1318, 2018.

Hewer, R.: Stochastisch-physikalische Modelle für Windfelder und Niederschlagsextreme, Ph.D. thesis, University of Bonn, http://hss.ulb.uni-bonn.de/2018/5122/5122.htm, 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, Journal of the Atmospheric Sciences, 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, Meteor. Z., pp. 467–480, 2018.

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

Mallat, S.: A wavelet tour of signal processing, Elsevier, 1999.

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

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

Nason, G.: wavethresh: Wavelets Statistics and Transforms, https://CRAN.R-project.org/package=wavethresh, 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, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62, 271–292, 2000.

Radanovics, S., Vidal, J.-P., and Sauquet, E.: Spatial verification of ensemble precipitation: an ensemble version of SAL, Wea. Forecasting, 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. Wea. Rev., 136, 78–97, 2008.

Rubner, Y., Tomasi, C., and Guibas, L. J.: The earth mover's distance as a metric for image retrieval, International journal of computer vision, 40, 99–121, 2000.

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

Schlather, M., Menck, P., Singleton, R., Pfaff, B., and team, R. C.: RandomFields: Simulation and Analysis of Random Fields, https://CRAN.R-project.org/package=RandomFields, r package version 2.0.66, 2013.

Theis, S., Hense, A., and Damrath, U.: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach, Meteor. 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, 58, American Mathematical Soc., 2003.

Weniger, M., Kapp, F., and Friederichs, P.: Spatial verification using wavelet transforms: A review, Quart. 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. Wea. Rev., 136, 4470–4487, 2008.